

# **Ontology Based News Generation Framework Using Neural Models**

*PhD Thesis submitted to  
Cochin University of Science and Technology  
in partial fulfillment of the requirements  
for the award of the degree of  
Doctor of Philosophy  
under the Faculty of Technology*

by  
**Shine K. George**  
(Reg.No.4528)

*Under the guidance of*  
**Prof. (Dr.) Jagathy Raj V. P.**  
**(Supervising Guide)**

and

**Prof. (Dr.) K. V. Pramod**  
**(Co-Guide)**



**Department of Computer Applications  
Cochin University of Science and Technology  
Kochi - 682 022, Kerala, India**

December, 2018

# Ontology Based News Generation Framework Using Neural Models

*Ph.D. Thesis*

*Author*

**Shine K. George**

Department of Computer Applications  
Cochin University of Science and Technology  
Kochi - 682 022, Kerala, India  
Email: shineucc@gmail.com

*Supervising Guide:*

**Prof. (Dr.) Jagathy Raj V. P.**

Professor  
School of Management Studies  
Cochin University of Science and Technology  
Kochi - 682 022, Kerala, India.  
Email: jagathy@cusat.ac.in

*Co-Guide:*

**Prof. (Dr.) K. V. Pramod**

Emeritus Professor  
Department of Computer Applications  
Cochin University of Science and Technology  
Kochi - 682 022, Kerala, India.  
Email: pramodkv4@gmail.com.



Department of Computer Applications  
Cochin University of Science and Technology  
Kochi - 682 022, Kerala, India.

December, 2018



**Department of Computer Applications**  
**Cochin University of Science and Technology**  
Kochi - 682 022, Kerala, India.

## **Certificate**

This is to certify that the thesis entitled "**Ontology Based News Generation Framework Using Neural Models**" is a bona fide record of the research work carried out by Mr. Shine K. George, under our supervision and guidance in the Department of Computer Applications, Cochin University of Science and Technology. The work presented in this thesis or part thereof has not been included in any other thesis submitted previously for the award of any degree.

**Prof. (Dr.) Jagathy Raj V. P.**  
*(Supervising Guide)*

**Prof. (Dr.) K. V. Pramod**  
*(Co-Guide)*

Kochi- 682 022  
December 07, 2018





**Department of Computer Applications**  
**Cochin University of Science and Technology**  
Kochi - 682 022, Kerala, India.

## **Certificate**

This is to certify that the thesis entitled "**Ontology Based News Generation Framework Using Neural Models**" submitted to Cochin University of Science and Technology by Mr. Shine K. George for the award of degree of Doctor of Philosophy under the Faculty of Technology contains all the relevant corrections and modifications suggested by the audience during the pre-synopsis seminar and recommended by the Doctoral Committee.

**Prof. (Dr.) Jagathy Raj V. P.**  
*(Supervising Guide)*

**Prof. (Dr.) K. V. Pramod**  
*(Co-Guide)*

Kochi- 682 022  
December 07, 2018



## *Declaration*

I hereby declare that the work presented in this thesis entitled **“Ontology Based News Generation Framework Using Neural Models”** is based on the original research work carried out by me under the supervision and guidance of Dr. Jagathy Raj V. P., Professor, School of Management Studies, Cochin University of Science and Technology and Dr. K. V. Pramod, Emeritus Professor, Department of Computer Applications, Cochin University of Science and Technology. The work presented in this thesis or part thereof has not been included in any other thesis submitted previously for the award of any degree.

Kochi- 682 022  
December 07, 2018

**Shine K. George**  
Register No: 4528





---

## *Acknowledgements*

First and foremost, I would like to thank God Almighty for giving me the grace, wisdom, and health to complete this project successfully.

I take his opportunity to convey my profound and sincere gratitude to my teacher and research supervisor, Dr. Jagathy Raj V. P., Professor, School of Management Studies, Cochin University of Science and Technology, for guiding me through my study, research and career as a teacher over the years.

It's my honor and privilege to express my heartfelt gratitude to my Co-Guide Dr. K.V. Pramod, Emeritus Professor, and former Head, Department of Computer Applications, Cochin University of Science and Technology, for his inspiring guidance and constant encouragement. I am thankful to Dr. B. Kannan, Head, Department of Computer Applications, Cochin University of Science and Technology for his help in the completion of this thesis.

I would like to thank my teacher Dr. Santhosh Kumar Gopalan, Head, Department of Computer Science, Cochin University of Science and Technology who helped me in building a career as a teacher. His optimistic approach to everything made my period of research work fruitful.

I extend my gratitude to all teaching and non-teaching staff of my department for their cordial relation, sincere cooperation, and valuable help.

I wish to place a record of gratitude to Tinto James, Maria Alice Issac, Jwala Jose and Dr. P. V. Mathai for their significant association.

I thank all the research scholars of my department, especially for their timely support and suggestions.

I am blessed to have the unconditional love of my parents K. K George, Sheela George, and mother-in-law Roothamma Chacko whose encouragement and support always kept me going through hard times. I dedicate this accomplishment to them.

Finally, but most importantly, I would like to thank my wife Shinu Chacko, who spent many years managing our children Ishaan Eldho Shine and Isha Mary Shine single-handedly in my absence. Without their sacrifice, love, and support, this thesis would not have been possible.

*Shine K. George*

## Abstract

As technology is progressing, news channels have been shifting to a new trend of instant news which gets updated each second. The growing network of news channels have resulted in massive production of news which gives journalists the hectic task of creating news stories from a huge amount of stored data within short time periods. Thereby, there is a need for an automated method that enables faster production of news. News story production basically involves the shortening of the topic's history and inclusion of the latest updates about the topic. The present work aims at proposing a framework which uses neural net models for generating news from keywords and to study the relation between the number of keywords mined by ontology and the quality of news generated. Two neural net models, one based on Char-RNN and another based on Sequence to Sequence model are tested with two datasets namely BBC news dataset and a Cricket Commentary dataset. The results are analyzed with both automatic evaluation measures and human evaluation.

**Keywords:** Ontology, Information Extraction, Deep Learning, Recurrent Neural Networks, Long Short Term Memory.



# Contents

Abstract .....	i
List of Tables .....	ix
List of Figures .....	xi
List of Abbreviations .....	xii

## **Chapter 1**

### **Introduction ..... 01 - 16**

1.1 Introduction.....	01
1.2 Background.....	02
1.3 Problem Formulation.....	07
1.4 Research Objectives.....	07
1.5 Research Methodology.....	08
1.5.1 Problem Identification and Derivation of Objectives.....	10
1.5.2 Literature Review.....	10
1.5.3 Design and Development of Research Framework.....	11
1.5.4 Construction of Proposed News Generation Framework.....	11
1.5.5 Design of Evaluation Approach and Application Development Environment.....	12
1.5.6 Testing and Evaluation .....	12
1.5.7 Analysis and Interpretation .....	13
1.6 Expected Research Outcome.....	13
1.7 Outline of the Thesis.....	14
1.8 Conclusion .....	16

## **Chapter 2**

### **Literature Review ..... 17 - 54**

2.1 Introduction.....	17
2.2 Semantic Web.....	18
2.2.1 Architecture of the Semantic Web.....	19
2.3 Ontology .....	21
2.3.1 Ontology Components .....	22
2.3.2 Ontology – Degrees of formalization .....	23
2.3.3 Ontology Levels.....	24

2.3.4	Ontology Applications.....	24
2.4	Deep Learning Techniques .....	25
2.4.1	Recurrent Neural Network (RNN).....	29
2.4.2	Long Short Term Memory (LSTM).....	29
2.4.3	Sequence to Sequence Networks (seq2seq).....	31
2.4.4	Applications of Recurrent Neural Networks .....	32
2.5	Ontology Based News Archiving and Extraction Systems.....	34
2.6	Text Generation Based on Neural Networks .....	41
2.7	Gist of Observations .....	51
2.8	Research gaps .....	52
2.9	Motivation.....	53
2.10	Conclusion .....	54

### ***Chapter 3***

#### **Development of Research Framework.....55 - 60**

3.1	Introduction.....	55
3.2	Research Strategy .....	55
3.3	Research Framework.....	56
3.3.1	Creating Experimental Framework .....	57
3.3.2	Evaluation .....	58
3.4	Conclusion .....	60

### ***Chapter 4***

#### **Keyword Generation Using Ontology ..... 61 - 80**

4.1	Introduction.....	61
4.1.1	Ontology Completeness .....	62
4.2	Ontology Based News Extraction and Archiving.....	64
4.3	Open Calais Ontology.....	67
4.4	YouTube – 8M Dataset.....	69
4.5	Evaluation Results of the Open Calais Based News Extraction System.....	70
4.5.1	Limitations of Open Calais Ontology .....	73
4.5.2	Evaluation Results of the News Extraction System Based on Open Calais and Local Ontology .....	76
4.6	Conclusion .....	78

## *Chapter 5*

### **Proposed News Generation Framework ..... 81 - 112**

5.1	Introduction.....	81
5.2	Choosing Recurrent Neural Networks for News Generation.....	82
5.3	Natural Text Generation Using Neural Network .....	83
5.3.1	Pre-processing.....	85
5.3.2	Tokenize Text .....	85
5.3.3	Removal of Infrequent/Stop words.....	86
5.3.4	Vocabulary and Inverse vocabulary formation.....	86
5.3.5	Building batches.....	86
5.3.6	Model Building.....	87
5.3.7	Initialization.....	87
5.3.8	Creating Computational graph .....	87
5.3.9	Forward Pass .....	88
5.3.10	Loss Calculation.....	88
5.3.11	Running Training Session .....	88
5.3.12	Checkpoints.....	88
5.3.13	Sampling.....	89
5.3.14	Output Generation.....	89
5.4	Mathematics of Neural Network Based Language Model.....	90
5.4.1	Neural language models.....	91
5.4.1.1	Artificial Neurons.....	91
5.4.1.2	Activation functions .....	92
5.4.1.3	Feed-Forward Neural Networks.....	96
5.4.1.4	Recurrent Neural Networks .....	97
5.4.1.5	LSTM .....	100
5.4.1.6	Encoder-Decoder model .....	101
5.5	Proposed News Generation Framework .....	102
5.5.1	Vanilla Char-RNN-LSTM Model.....	105
5.5.2	Sequence to Sequence (Seq2Seq) Model.....	109
5.6	Uniqueness of Proposed News Generation Framework .....	112
5.7	Conclusion .....	112

## **Chapter 6**

### **Design of Evaluation Approach and Application Development Environment ..... 113 - 131**

6.1	Introduction.....	113
6.2	Evaluation Approach .....	113
6.2.1	Automatic Evaluation .....	114
6.2.1.1	Bilingual Evaluation Understudy (BLEU).....	114
6.2.1.2	Recall-Oriented Understudy for Gisting Evaluation (ROUGE).....	115
6.2.1.3	Limitations of Automatic Evaluation.....	115
6.2.2	Human Evaluation.....	116
6.2.2.1	Team Selection.....	117
6.2.2.2	Evaluation Criteria.....	119
6.2.2.3	Evaluation Method.....	120
6.2.2.4	Limitations of Human Evaluation Study.....	121
6.2.3	Evaluation Procedure Followed in This Work .....	122
6.3	Datasets Used .....	123
6.4	Application Development Environment.....	124
6.4.1	Aws p2. xlarge Instance .....	124
6.4.2	TensorFlow.....	126
6.4.3	Python.....	127
6.4.4	CUDA.....	129
6.4.5	AWS Deep Learning AMIs for Machine Learning Practitioners .....	130
6.5	Conclusion .....	131

## **Chapter 7**

### **Testing and Evaluation ..... 133 - 171**

7.1	Introduction.....	133
7.2	Model Training.....	136
7.2.1	Char-RNN model.....	136
7.2.2	Seq2Seq model .....	139
7.3	Test Results.....	142
7.3.1	Automatic Evaluation .....	142



7.3.1.1	BLEU Metric (Char-RNN model).....	142
7.3.1.2	BLEU Metric (Seq2Seq model).....	146
7.3.1.3	ROUGE Metric (Char-RNN model) .....	150
7.3.1.4	ROUGE Metric (Seq2Seq model) .....	151
7.3.1.5	Comparing Two Neural Net Models Using BLEU-2 and ROUGE-L .....	154
7.3.2	Human Evaluation Results.....	155
7.3.2.1	Based on Evaluation Criteria .....	156
7.3.2.2	Based on Evaluators Traits .....	159
7.3.2.3	Summary of Human Evaluation .....	163
7.4	Sample Output .....	164
7.5	Correlating Automatic and Human Evaluation.....	165
7.6	Discussion and Inferences .....	167
7.6.1	Char-RNN-LSTM Model.....	167
7.6.2	Sequence to Sequence Model.....	169
7.6.3	Comparing Char-RNN and Sequence to Sequence Model.....	170
7.7	Conclusion .....	171

**Chapter 8**

**Summary and Conclusion .....173 - 184**

8.1	Introduction.....	173
8.2	Summary.....	173
8.3	Implications of the Study.....	177
8.4	Contributions.....	179
8.5	Limitations .....	180
8.6	Suggestions for Future Research .....	182
8.7	Conclusion .....	183

**References .....185 - 204**

**Appendices ..... 205 - 222**

**List of Publications ..... 223 - 224**

**Index ..... 225 - 226**

**About the Author.....227**



## ||||| **List of Tables** |||||

Table 1.1	Archiving systems of news channels in Kerala .....	06
Table 2.1	Ontology based news extraction systems .....	35
Table 2.2	Sequence to Sequence model application areas in text generation .....	43
Table 4.1	Search Query Code Representation.....	71
Table 4.2	Semantic Similarity (Open Calais Ontology).....	72
Table 4.3	Semantic similarity (Open Calais and Local Ontology).....	77
Table 6.1	P2 instance details .....	126
Table 7.1	BLEU score of Char-RNN model (BBC news dataset).....	143
Table 7.2	BLEU score of Char-RNN model (Cricket Commentary dataset) .....	143
Table 7.3	BLEU score of Seq2Seq model (BBC news dataset) .....	146
Table 7.4	BLEU score of Seq2Seq model (Cricket Commentary dataset) .....	147
Table 7.5	ROUGE score of the Char-RNN model.....	150
Table 7.6	ROUGE score of the Seq2Seq model .....	152
Table 7.7	Human ratings (Fluency).....	156
Table 7.8	Human ratings (Adequacy) .....	157
Table 7.9	Human ratings (Total Quality) .....	158
Table 7.10	Human ratings (language proficiency).....	161
Table 7.11	Sample Output.....	164



## ||| List of Figures |||

Figure 1.1	Illustrates typical use case .....	04
Figure 1.2	Illustrates order of visual scenes as per sample news script.....	05
Figure 1.3	Research Methodology .....	09
Figure 2.1	Semantic Web Stack .....	19
Figure 2.2	Structure of a single node.....	26
Figure 2.3	LSTM cell state .....	30
Figure 2.4	Seq2Seq model.....	31
Figure 2.5	Language model.....	33
Figure 2.6	Trade-off between rule based and neural based systems .....	41
Figure 2.7	Text generation using neural network.....	42
Figure 3.1	Research Framework .....	56
Figure 4.1	Ontology Based News Archiving and Extraction .....	65
Figure 4.2	News Extraction Process .....	66
Figure 4.3	LSA algorithm steps.....	70
Figure 4.4	Evaluation Result.....	73
Figure 4.5	Ontology Based News Extraction System incorporating Local Ontology.....	75
Figure 4.6	Evaluation Result (Using local ontology along with Open Calais).....	78
Figure 5.1	Text generation process.....	84
Figure 5.2	Sigmoid function .....	93
Figure 5.3	tanh function.....	94
Figure 5.4	Softmax function .....	95
Figure 5.5	RNN and Feed-Forward neural network.....	97
Figure 5.6	Recurrent Neural Network Language model .....	98
Figure 5.7	Sequence to Sequence model.....	101
Figure 5.8	News Generation Framework based on Ontology .....	103
Figure 5.9	Char-RNN based model.....	106
Figure 5.10	Char-RNN.....	107

Figure 5.11	Seq2Seq based model.....	109
Figure 5.12	Encoder-Decoder network .....	110
Figure 6.1	Evaluation Approach.....	122
Figure 7.1	Testing and Evaluation Framework .....	134
Figure 7.2	Average training loss of the Char-RNN model (BBC news dataset) .....	137
Figure 7.3	Average training loss of the Char-RNN model (Cricket Commentary dataset) .....	138
Figure 7.4	Average training loss of the Seq2Seq model (BBC news dataset) .....	140
Figure 7.5	Average training loss of the Seq2Seq model (Cricket Commentary dataset) .....	141
Figure 7.6	BLEU scores of Char-RNN model (BBC news dataset) .....	144
Figure 7.7	BLEU scores of Char-RNN model (Cricket Commentary dataset) .....	145
Figure 7.8	BLEU scores of Seq2Seq model (BBC news dataset).....	148
Figure 7.9	BLEU scores of Seq2Seq model (Cricket Commentary dataset) .....	149
Figure 7.10	ROUGE-L score of Char-RNN model (BBC news dataset and Cricket Commentary dataset) .....	151
Figure 7.11	ROUGE-L score of Seq2Seq model (BBC news dataset and Cricket Commentary dataset) .....	153
Figure 7.12	Comparison of models based on ROUGE-L .....	154
Figure 7.13	Comparison of models based on BLEU-2 .....	155
Figure 7.14	Human ratings based on prescribed qualities (Fluency) .....	156
Figure 7.15	Human ratings based on prescribed qualities (Adequacy) .....	157
Figure 7.16	Human ratings based on Total Quality .....	159
Figure 7.17	Human evaluation based on language proficiency .....	162

## List of Abbreviations

ANN	Artificial Neural Networks
BBC	British Broadcasting Corporation
BLEU	Bilingual Evaluation Understudy
CNN	Convolutional Neural Network
CSV	Comma-Separated Values
CUDA	Compute Unified Device Architecture
EC2	Elastic Compute Cloud
GAN	Generative Adversarial Network
GANN	Gated Attention Neural Network model
GI	Geographical Indication
GPU	Graphics Processing Unit
GRU	Gated Recurrent Units
IoT	The Internet of Things
IPTC	International Press Telecommunications Council
IT	Information Technology
LCS	Longest Common Subsequence
LSA	Latent Semantic Analysis
LSTM	Long Short Term Memory Network
MCC	Marylebone Cricket Club
NLG	Natural Language Generation
NLP	Natural Language Processing
NN	Neural Network
OpenMP	Open Multi-Processing
OWL	Web Ontology Language
POS	Parts of Speech
RDF	Resource Description Framework

RDFS	RDF Schema
RIF	Interchange Format
RNN	Recurrent Neural Network
ROUGE	Recall Oriented Understudy for Gisting Evaluation
Seq2Seq	Sequence to Sequence Networks
SQL	Structured Query Language
SVD	Singular Value Decomposition
TF-IDF	Term Frequency–Inverse Document Frequency
UI	User Interface
URI	Unique Resource Identifier
XML	Extended Marked-up Language

.....



**INTRODUCTION**

- 1.1 Introduction
- 1.2 Background
- 1.3 Problem Formulation
- 1.4 Research Objectives
- 1.5 Research Methodology
- 1.6 Expected Research Outcome
- 1.7 Outline of the Thesis
- 1.8 Conclusion

**1.1 Introduction**

As technology is developing, there is a progressive and significant reduction in the amount of work that humans have to do. This phenomenon also has an impact on news generating agencies including print media (like newspapers) and visual media (like news broadcasting services). When newly available technologies were incorporated into the conventional processes of news production, it gave a considerable scope for a customization with which the users could set their preferences for the information accessed. Customization enhances user experience by allowing them to control their interaction, and therefore, plays an important role in facilitating the user in making their choices while considering the tight competition within this service sector.

With the numerous sources that provide information for generating news, there arises the necessity for a need- specific customization which enables them to provide a strong user experience both online and offline. Another reason for growth of its importance is the availability of the huge amount of data. This chapter discusses the background in the first part. Later, the problem statement, objectives, research method and expected outcome are explained. Finally, the chapter presents an overview of the chapters to follow.

## **1.2 Background**

Public life has made news channels a part and parcel of their existence. A number of basic story forms are being used in the news agencies like voice overs, voice over with sound bites and news reading out stories. A short clip of an interview, speech or music taken out from a full-length audio or video is called a sound bite. A ‘reader’s story’ which usually omits visuals may upgrade to a voice over story according to its news value. They usually lack production value. A report or a visual news story is a sequence of semantically related visual scenes edited with relevant sound bite and a voice over.

With development in Information Technology, and with media emerging as a primary source of information, news content generation has undergone exponential growth. With the large number of news medias and organizations that report news on an instant pace, news production has reached an intensely competitive level. News channels

have established an instant news habit that gets updated virtually every few seconds. Archiving becomes a challenging process due to the amount of news produced. The journalists access news archives to get details about the past events related to the currently trending news updates. Finding the apt content from a whole archive is time consuming for the journalist. They are asked to make stories at an instant pace as the refreshing process takes place rapidly.

News producers, editors and news desk journalists who are responsible in making a news script and compiling stories face difficulty to meet the demand for faster news delivery. Preparing a news script is the initial step in the process of news generation. A thorough search in the news library is required to gather the necessary information. The news correspondent would get a number of news articles for a single query. The journalist needs to search the archive using many keywords to get the required details. The right information may not be available in the first search itself. It is a hectic task for the journalist to check on each search results and to screen out unnecessary information and select the information needed for the story

On the contrary, if the news reporter gets a relevant and comprehensive machine generated news from the archive with regard to a current event, based on his/her search query, that would be an experience which is more customized. This will help to frame a detailed news script by considering all the available information about a news topic thus helping to reduce the time required for news generation.

For example, a news story was framed in the context of the 2018 football world cup which presented Kylian Mbappe as an emerging talent from the side of France. His skills were revealed in the pre quarter match where Argentina played against France. The usage scenario is depicted by Figure 1.1

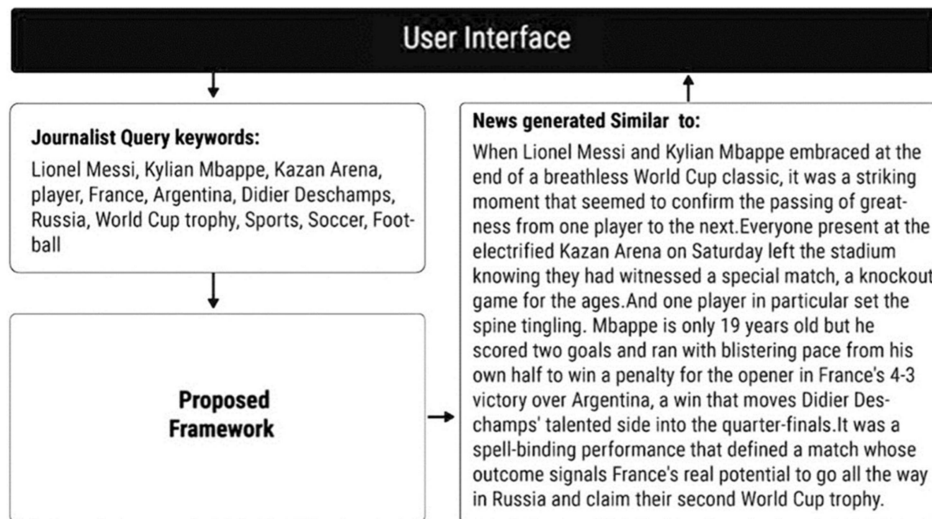


Figure 1.1: Illustrates typical use case

The figure above represents the generation of news by the proposed framework. Here the generation of news is implemented in three steps. The journalist gives the keywords as the input, which then gets processed by the framework and the news is generated. With the machine generated news, a news desk journalist can easily collect the suitable visual scenes from the library. The correspondent edits and

integrates the visuals using a visual editor. Figure 1.2 gives more information about the visual scenes which are to be extracted for the news event mentioned above.

REQUIRED VISUAL LIST & ITS ORDER	
Reuters latest news feed	Kylian Mbappe practising visual latest
00:01:30 - 00:02:00 Taken from Tape 2345 or computer From archive	Lionel Messi and Kylian Mbappe embracing
00:22:30 - 00:23:00 Taken from Tape 2346 or computer	Gallery visuals of France Vs Argentina world cup 2018 match and Mbappe visual
01:01:10 - 00:01:01:25 from Tape 2346 or computer	Mbappe attacking + Mbappe's goal + Penalty taking visual
02:05:15 - 02:05:25 from Tape 2347 or computer visual	France team visual + Didier Deschamp's












**Figure 1.2:** Illustrates order of visual scenes as per sample news script

The above figure represents the order with which these scenes are to be connected. The length of each visual is determined by the news story's length. The purpose of the example described here is to showcase a possible innovation in news generation process which reduces the complexity of the work involved.

There are more than 75 news channels working in India where the medium of communication is English, Hindi or other regional languages. This study was based on news channels in the Kerala state, in India. In Kerala, both national and regional news channels are available. A major finding of the study was that all the channels are following keyword based search in the news archive. Some of them are still using

video tape for storage. Only a few news channels are having advanced server based storage systems. Table 1.1 provides details of the study conducted about the archiving systems followed by the news channels in Kerala, India.

**Table 1.1: Archiving systems of news channels in Kerala**

News channel	Application type	Search method	Storage
	Stand-alone	Keyword based	Tape/Computer
	Server based ERP	Keyword based	Server based
	Stand-alone	Keyword based	Tape
	Stand-alone	Keyword based	Computer
	Stand-alone	Keyword based	Computer
	Manual	Keyword based	Tape
	Stand-alone	Keyword based	Computer
	Stand-alone	Keyword based	Computer
	( Dalet –total media solution)	Keyword based	Server-based
	( Diva –total media solution)	Keyword based	Server based
	Stand-alone	Keyword based	Computer/Tape

Out of the 11 channels considered, seven channels use stand-alone type of application for news generation as per details shown in Table 1.1. Three of the channels use server based storage, six of them use stand-alone computer based storage and two of them use tape/-server based storage.

### **1.3 Problem Formulation**

While considering the huge production of news and insufficient automated tools, news archiving and extraction have turned out to be demanding tasks. The initial phase of present work will be dealing with how customization in digital archiving can be achieved using ontology and the significance of the completeness of ontology in this domain. The main aim of this work is to propose a framework which uses neural net models for generating news from keywords and to study the relation between keywords which are extracted by the ontology and the quality of generated news.

### **1.4 Research Objectives**

The main objective of this work is to develop a framework for generating news from keywords using neural net models. More specifically, the objectives of the present study are:

- To design and develop a research framework to conduct the present study.

- To scrutinise the impact of using ontology in news extraction and the significance of ontology with completeness property in generating more keywords.
- To develop a framework for news generation from keywords extracted by ontology using neural net models.
- To design an evaluation approach suitable for the present study.
- To compare the evaluation results of neural net models used in the news generation framework with different datasets and conclude on the best performing news generation framework.
- To study the relation between the quality of news generated and the number of keywords.

## **1.5 Research Methodology**

Selecting a research strategy is necessary for solving problems in a scientific and systematic way. Different problems require diverse research strategies. Various types of research methods are used in computer related research because of the diverse nature of computer technologies. The research work discussed here follows the experimental approach. The research methodology followed in this work is summarized in Figure 1.3.



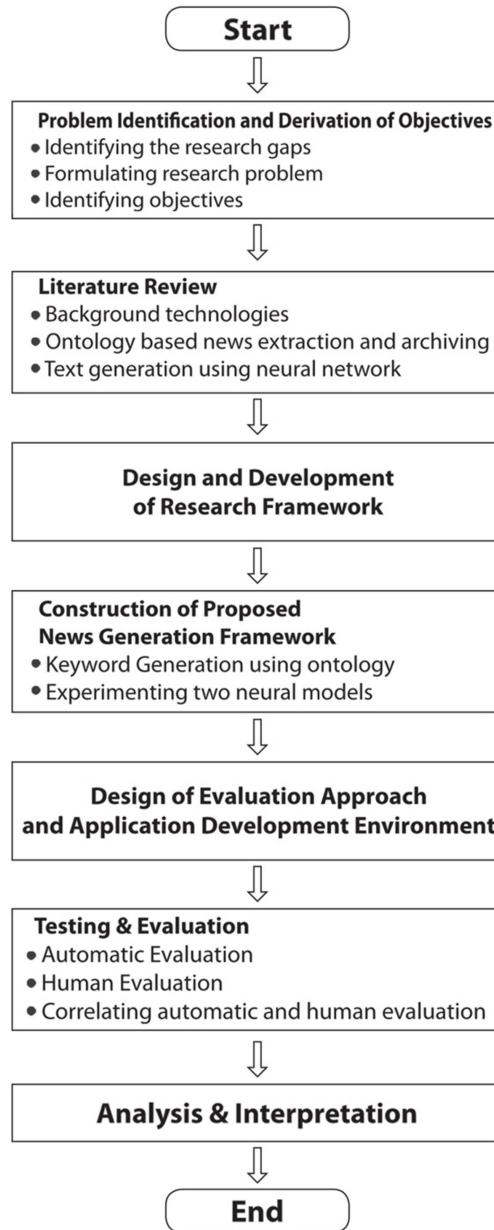


Figure 1.3: Research Methodology

Sometimes, there is a requirement of a specific environment to analyse the reason of the phenomena under study. The requirement of this experiment arises when the variable under the study has no proven relation. The domain under study may be complex to formulate a mathematical explanation or relation. Here the only alternative to analyse the problem is experimental approach. This approach has three components namely data sample, dependent variables and independent variables. Values of dependent variables may change as a result of change in the independent variables. The changes are measured using statistical or other evaluation methods which are suitable for the domain under consideration.

### **1.5.1 Problem Identification and Derivation of Objectives**

Problem identification and derivation of objectives is done in this phase. It starts with examining the research gaps explored during the initial study. Based on this, the problem is formulated and research objectives of this work are carefully established.

### **1.5.2 Literature Review**

Selecting the tools and techniques which aids to achieve the objectives is significant for this particular study. The statistical or rule based models used in text generation have inherent drawbacks. These models function with limited number of handmade rules and a small vocabulary, but fails with huge datasets [136,140]. Another major drawback is its inability to remember long-term dependencies which is highly required for the problem discussed in this work. This is clearly

mentioned in the research gaps. RNN-LSTM network overcomes these limitations in an effective manner. Hence the LSTM network, a variant of the Recurrent Neural Network is selected for this work. The head words or keywords which are supposed to be used for news generation are not developed manually in this problem. It is extracted from news text using ontology which is one of the powerful semantic web technologies.

Various ontology based news extraction and classification systems have to be studied to understand the power of ontology to grasp knowledge concepts. Neural net models which are developed not only for generating news but also for text generation is required to be studied here to get a better understanding about the works already done in this field. Some of the statistical or rule based text generation models which uses keywords have to be considered since a few neural text generation systems using keywords as input have been developed so far.

### **1.5.3 Design and Development of Research Framework**

The conceptual framework is designed and developed in this phase which acts like a blueprint. It provides an outline of the plan to conduct the present research.

### **1.5.4 Construction of Proposed News Generation Framework**

Here the aim is to make the technical environment to run the experiments to understand the phenomena and to get the objectives achieved.

The first step involved in this phase is the extraction of keywords from the news using ontology. These keywords are fed to the neural net models for the purpose of training and testing purposes. Two neural net models are experimented in this work. The selection or creation of the neural model algorithms to be used in this work has to be decided in this phase. How to prepare data for neural model training is also an issue to be solved in this phase.

### **1.5.5 Design of Evaluation Approach and Application Development Environment**

The purpose of the evaluation here is to test the proposed news generation framework to find out whether it achieves the research objectives. The major part in the design of evaluation approach in this work involves selecting the suitable technique to test the machine-generated news. The evaluation approach followed in this work consists of automatic and human evaluation. Technical environment required for the development of the proposed framework is also decided in this step.

### **1.5.6 Testing and Evaluation**

The results are tested based on the evaluation approach followed in this work. Automatic evaluation metrics are very effective in machine translation scenarios. However, they are not popular in text generation. One of the reasons behind this is that the machine generated text would not be totally similar to the referral text since the machine is going

through a learning process using selected neural network algorithms and news is generated based on its learning data.

So here, the human evaluation approach is also considered as part of the evaluation strategy. After all, the end user is a human and definitely the feedback has to be taken from human evaluators. But one cannot neglect the limitation of human evaluation either. The evaluation approach followed in the work is to derive a conclusion by correlating the automatic evaluation scores with human evaluation scores.

### **1.5.7 Analysis and Interpretation**

Finally, the analysis and interpretation is made based on the experimental results.

## **1.6 Expected Research Outcome**

Most importantly, the results will provide a number of benefits to news agencies in news creation process as well as archiving. The result will trigger advanced study on keyword based text generation which are applicable to various information extraction scenarios. One of the objectives of this research work is to study the relation between the number of keywords and the quality of news generated. This particular objective of this work creates the scope for future work in creating ontology in different domains with high degree of completeness.

## **1.7 Outline of the Thesis**

The rest of the thesis is organised as follows:

**Chapter 2** explains the background technology concepts necessary for the objectives to be fulfilled. Here, an introduction to the semantic web and the concept of ontology is analysed in detail. An outline of deep learning is also explained in this chapter. Thorough discussion about recurrent neural network and how LSTM overcomes RNN's limitation to remember the long term dependency is also done. This chapter reviews news extraction and archiving systems which use ontology. Applications of neural networks in text generation genre are also reviewed in chapter 2. Finally, research gaps are derived and motivation of this research is formed based on the literature review.

**Chapter 3** introduces the research framework applied in this study. It also explains the procedure to be followed to accomplish the objectives, based on literature survey and background technologies.

**Chapter 4** discusses the ontology's effectiveness in understanding knowledge in a domain and the role played by ontology in the field of digital library archives. The details of application developed to analyse the benefits of ontology based news extraction system and the limitation of ontology used in that

application is also provided here. This chapter emphasises the necessity for ontology completeness in generating more number of keywords.

**Chapter 5** proposes a news generation framework which uses ontology and neural networks. At the chapter's beginning, the steps involved in the generation of text are detailed. How the keywords extracted by the ontology aids in generating news using a neural model is explained in the chapter. Different deep learning algorithms are attempted to obtain the best results and details are explained in this chapter.

**Chapter 6** focuses on the evaluation approach followed in this work. This chapter provides a description of overall evaluation procedures. Manual and automatic evaluation are employed here since it's a text generation process. The work uses two datasets of different nature and in depth description of these datasets are given in the chapter. The final section mentions the technical environment used for developing the proposed framework.

**Chapter 7** deals with results analysis and discussion. The automatic evaluation and human evaluation results are detailed here. The neural net models with different datasets are compared on the basis of their results. The last section is a discussion based on automatic and manual evaluations.

**Chapter 8** recapitulates the thesis and mentions possible future research directions.

## **1.8 Conclusion**

This chapter deals with the background of the complete work. Further, the research problem was formulated and objectives were set. The experimental research approach selected for this study and its expected outcome was also noted down. Finally, the outline of the thesis is presented in this chapter.

*.....❧.....*



**LITERATURE REVIEW**

2.1	Introduction
2.2	Semantic Web
2.3	Ontology
2.4	Deep Learning Techniques
2.5	Ontology Based News Archiving and Extraction Systems
2.6	Text generation Based on Neural Networks
2.7	Gist of Observations
2.8	Research Gaps
2.9	Motivation
2.10	Conclusion

**2.1 Introduction**

Ontologies are extensively used in information archival and retrieval scenarios since it can represent knowledge in a domain and its complex relations. Several applications are developed in various fields which uses the capabilities of ontology.

Research about making a computer function as the human brain had started in the previous century. Neural network concepts derived its inspiration from the human brain. Neural networks are the foundation stones of deep learning. As the power of the machine substantially went up, a number of applications were created based on deep learning techniques. This work derives its motivation from the results of such studies. Deep learning algorithms and ontology provide a solution for the specified problem in this study.

There are numerous ontology based news classification and archiving systems developed to overcome the limitation of traditional keyword based systems. The use of neural networks in news generation, summarization, headline generation etc. has increased in recent times largely due to the accessibility of higher computational power.

This chapter's first section presents a general understanding of the semantic web and ontology. The next section provides more details about deep neural networks. In this chapter, ontology-based news archiving systems as well as neural based text generation frameworks, and in particular news-related text are reviewed to get an idea about the recent developments in the problem domain discussed in this work. Finally, research gaps are explored based on the literature review and the motivation for this work is presented.

## **2.2 Semantic Web**

The content of the World Wide Web is designed exclusively for humans to read. Due to its semi-structured feature, it is a complex task for computer programs to manipulate this information meaningfully. One of the challenges faced by information technology is to deliver apt information to the right person at the desired time. Achievement of this goal requires seamless association with people, software agents and various IT systems. Vibrant communities need such a correspondence to facilitate their elevation and utilize the data to the maximum. As per Tim Berners-Lee's vision, this IT challenge can be fulfilled using

semantic web which is an extension of World Wide Web through which web content can be shown in a form that can be grasped by software agents for effective gathering and integration of data [1,2,3].

### 2.2.1 Architecture of the Semantic Web

Tim Berners-Lee, known as the inventor of World Wide Web, portrays the structure of the Semantic Web in the model of a Semantic Web stack. The stack visualizes the hierarchy of languages, wherein the capabilities of layers lying below are utilized by those on top. It portrays how technology that is standardized is organized scientifically to make the semantic web possible [4]. Figure 2.1 illustrates the components of the Semantic Web stack.

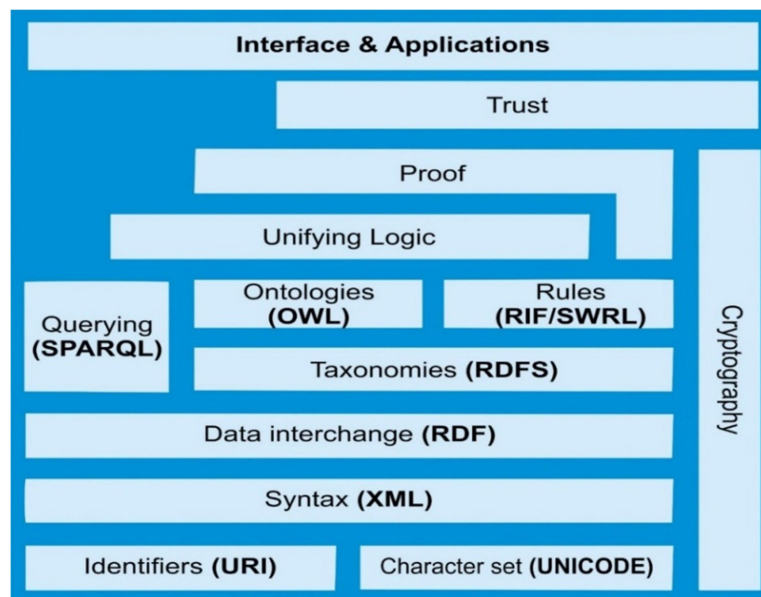


Figure 2.1: Semantic Web Stack (Tim-Berners-Lee, 2015)

Semantic Web stack comprises of three main layers. The Lower layer includes hypertext web technologies i.e. Unique Resource Identifier (URI) and Character Set (Unicode). The main function of URI is to distinguish physical or abstract resources and the text is manipulated in different languages with the Unicode. The XML (Extended Marked-up Language) placed on top of that, and is represented as a language which encodes documents in a structured and readable format for machines.

The Middle layer consists of Semantic Web technologies [5]; both RDF (Resource Description Framework) and RDFS (RDF Schema) which are formed on XML syntax. RDF adds semantics to data which is structured by XML. RDF is a model used for data interchange on the web which creates statements about resources in the form of triples (Subject, Predicate, Object).

RDFS is the basic schema language, which provides terminological knowledge for RDF in classes, property hierarchies, semantic interdependencies etc.

OWL (Web Ontology Language) at the top level of the stack uses the RDFS syntax to portray more complex knowledge even the one which is only implicit in the domain of interest. OWL, is a fully structured knowledge model which includes relations of various kinds of concepts. SPARQL is a Semantic Web standard for querying RDF-based information for presenting the results [6].

Apart from OWL, RIF (Rule Interchange Format) gives complex sharp relations that cannot be directly traced using the description logic used in OWL. The top layers of the stack (Logic, Proof, and Trust), deal with the logical and semantic validation of ontologies that are still potential ideas that can be implemented to realize the scope of the Semantic Web. Moreover, cryptography layer covers most layers from bottom to the top of the stack which ensures and verifies the reliability of the statements in Semantic Web. “User Interface” and “Applications” constitute the last layer that helps humans in using the Semantic Web applications [4].

### **2.3 Ontology**

In the last couple of years, ontology has been a hot keyword in Information Technology. Originally ontology is a branch of philosophy, which studies the essence of existence in objective things. In computer science, ontology is an important element in content theories in the domain of Artificial Intelligence, which studies object classification, object attribute and relations between objects. It attempts to give a terminology for domain knowledge description. An inevitable role is played by ontology in fields like information exchanging, system integration, knowledge-based software development etc.

There were many definitions put forward about ontology [7,8,9,10,11] but the best known (in computer science) can be attributed to Gruber [12,13]:

**“An ontology is an explicit specification of a conceptualization”**

(Citation)

In the above context, conceptualization means an abstract model of some aspects of the world, taking the form of a definition of the properties of important concepts and relationships. An explicit specification enables information to be processed by the machine.

From this broad definition, Borst [14] and Fensel [15] stress the fact that “there must be an agreement on specified conceptualization”. The capability to reuse an ontology will almost be nullified when the specified conceptualization is not accepted

### **2.3.1 Ontology Components**

Normally, an Ontology describes following aspects [16].

- (1) Instances: These are the basic/ fundamental components in ontology that are actual or abstract objects like text and numbers. Instances are not vital to ontology.
- (2) Classes: Object is an instance of a class that can be an individual, or another class.
- (3) Attributes: objects in ontology are explained by assigning attribute values to them. An attribute carries a minimum of one name and one value, which stores specific information of an object.

- (4) Relations: An important contribution of attributes is to describe relations between two objects. Usually a relation is an attribute which has another object in the ontology as its value. In general, the set of relations describe the whole semantic scenario in a domain.
- (5) Events: events are objects about time, or instantiated object resources.

The main aim of ontology is to capture domain knowledge, provide shared understanding to domain knowledge, and guarantee that only one set of vocabulary is used and to clearly define terms and assure a connection with all the layers. Generally, the creation of Ontology makes knowledge reusable and sharable to an extent

### **2.3.2 Ontology – Degrees of formalization**

In the applications that use ontologies, various degrees of formalization are considered. Navigli [17] introduces six levels for the degree of formalization in ontologies from the least to the most formalized knowledge resources:

- Unstructured text: Just a text string with no structure
- Terminology: A group of terms that express concepts for a domain
- Glossary: A terminology with a textual definition for every concept
- Thesaurus: Provides information about the relationship between words like synonyms and antonyms.

- Taxonomy: A hierarchical classification of concepts
- Ontology: A fully structured knowledge model, which includes things, their properties, and their relationship to other things.

### 2.3.3 Ontology Levels

Guarino [9,18] suggests the opportunity to develop different kinds of ontology with the level of generality.

- 1) Top-level ontologies are generic ontologies which are independent of a domain.
- 2) Domain ontologies and task ontologies are formed based on a domain by specializing the terms introduced in the top-level ontology.
- 3) An “application ontology” is developed for a specific use or application that cannot be shared or used by another community. Application ontologies depend both on domains and the specified task of interest.

### 2.3.4 Ontology Applications

Several systems like the semantic question and answering systems absorb the advantages of ontology to its best [19]. Including ontological knowledge like in information retrieval processes can provide a solution to many problems currently faced by these types of systems. A mentionable role is played by ontology in query expansion, semantic formalization, natural language understanding and information abstraction.



Ontologies are widely used in online science activities or e-science, especially when there is a necessity of management and data integration of workflows. Though an extensive growth of the digital library domain could be seen in the last two decades, its interdisciplinary feature paves way for a huge number of concepts to be captured, classified, created and structured. Thereby a potential role is to be undertaken by ontologies for its common vocabulary which shares information in a domain. Ontology can be used for digital library collaboration, interoperation, research, education and modelling. A number of ontologies have been evolved with different approaches in this domain [20,21].

## **2.4 Deep Learning Techniques**

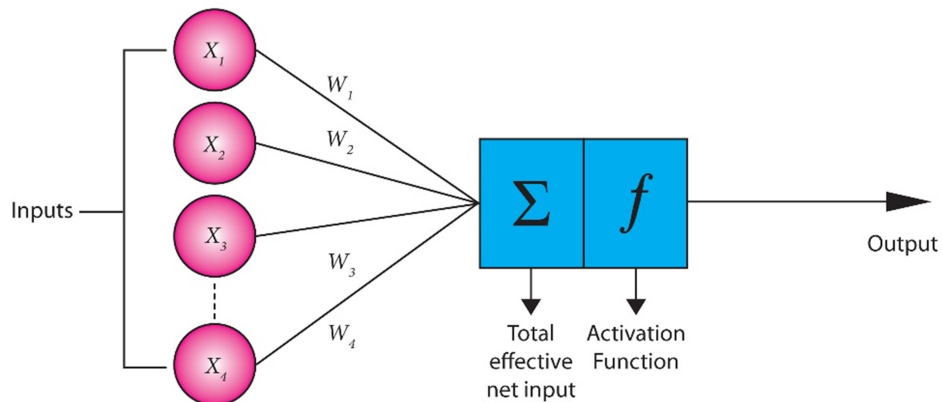
In the last few years, with the advancement in different aspects of computer technologies, deep learning has helped in providing better accuracy in complex applications. Availability of bigger size training data and enhancement in computer infrastructure like multi-core CPU/GPUs accelerated the growth of deep learning.

Deep learning is a technique in artificial intelligence which helps computers to study and correct the errors from its past and conceive the world around them as a hierarchy of concepts. To elaborate, deep learning is a machine learning approach that simplifies working with data and provides flexibility and excellent power. However, unlike with basic machine learning techniques, humans needn't describe all the

required information while performing a task. The programs can grasp all the needed knowledge from its experiences from the past.

The inventor of one of the first neurocomputers, Dr. Robert Hecht-Nielsen, defines a neural network [22] as “...a computing system made up of a number of simple, highly interconnected processing elements, which process information by their dynamic state response to external inputs.” (Citation)

Artificial neural networks in general when represented graphically [23], where nodes are the neurons and edges are the synapses. Figure 2.2 illustrates the structure of a node.



**Figure 2.2: Structure of a single node**

Each node contains three basic components, which are shown in Figure 2.2:

- 1) Weighted inputs
- 2) Transfer (summation) function, which calculates the sum of signals multiplied by concrete weights
- 3) Activation function, which maps the result of the net input to the output of the neuron

If the signal is too weak (weaker than the set threshold), the propagation would be hindered thereby stopping in the neuron.

In general, the network may or may not have more layers. If there are many layers, the first one would be the input layer and the last one, the output layer. The input layer is that part of the network, where signals enter through an external input. Neurons of this layer serve to process the external input and transfer it further. After the input layer, there might be more than one hidden layer, whose neuron acts exactly as stated in the Figure 2.2 provided above. The neurons of the output layer act like those of the layer that is hidden, but their output is propagated into the final output of the whole network. If we allow the feedback edges, edges that go within the same layer or to one of the previous layers, it would mean there can appear cycles. Neural networks with cycles are called recurrent.

Artificial neural networks can learn in different ways. Supervised, unsupervised or reinforcement learning are some examples. Consider the newly created artificial neural network as the brain of a new-born child. The new-born child is not capable of classifying items on their colour and thereby has to learn it first. He learns the way in which he is taught to assign an item to the suitable category, for example blue. After that his mother assures if he either assigned it correctly or it should have belonged to another colour, for example green. In the next trial for the similar coloured item, there is a greater probability of assigning the item correctly. And this is almost the similar way as how the supervised learning works.

The most widely used supervised learning algorithm is the Backpropagation algorithm. For every input in the learning set, a model output is given. The Backpropagation algorithm has two major phases:

- 1) Forward phase – Computation of outputs of all the neurons in the network, the end error is computed based on the model.
- 2) Backward phase – There is propagation of error back through the network; the weights are being changed for the error rates to be minimized.

Backpropagation leads to the error being minimized, functioning to minimum, which isn't necessarily global. The speed of the training can be changed to prevent overlearning with each single incorrectly recognized input [24,25].

### **2.4.1 Recurrent Neural Network (RNN)**

Recurrent Neural Networks (RNN) are feed-forward networks supplemented by additional feedback edges, which provide back the context. The context cannot be stored using normal neural networks. In general, this context might be considered as a state of the network in the previous time step. RNN remembers the context in memory.

Though the networks had achieved success in learning short-range dependencies, they haven't been showing any worth mentioning achievement in learning mid-range or long range dependencies mainly because of the problems of vanishing and exploding gradients [26]. While back propagating the error across many time steps, using standard learning algorithms, both vanishing and/or exploding gradient problems can occur. Both are caused by the incapability of RNN to learn the mid-range or long range dependencies [27]. The exploding gradient problem appears when the long-term components grow exponentially than the short term ones, which leads to their explosion.

The vanishing gradient problem is the opposite behaviour, which appears when long-term components grow exponentially from fast to zero. It is impossible for the RNN model to find any links between distant events.

### **2.4.2 Long Short Term Memory (LSTM)**

Normally RNNs have the ability to use context information i.e. predicting the next word if the previous word is given, but this becomes

harder as the distance between the dependencies within the sequences grow. For instance, given the sequence “My pet name” even the standard RNN model would predict that the next word is “is”. Consider the sequence “Last summer I went to Kashmir for a vacation. It was ... I have decided that next year I will return to” ... The model is expected to predict “Kashmir”, but it is difficult for the RNN model to remember the dependency. Long short term memory network (LSTM) is a variant of RNN developed to solve this issue [28,29]. The key concept to LSTM's is the cell state as shown in Figure 2.3, a memory that keeps the important information that the cell has seen. This state can for instance contain more information on the sentence, predict if it is a thing or a person and based on that use the correct pronoun. In each interaction with a LSTM cell, the internal workings decide what should be done with the state, whether to add or remove information to it.

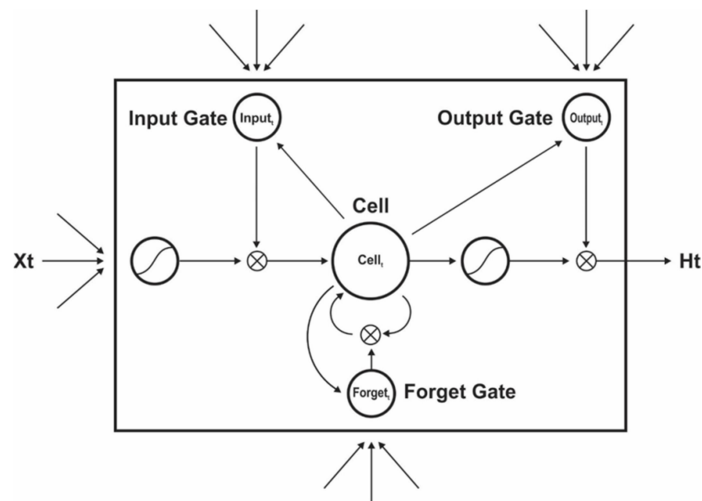


Figure 2.3: LSTM cell state [30]

There are three gates for a LSTM cell state. The input and output gates control the input and output into the memory cells and inner state of the memory cell is reset by the forget gate once its context is out of date. The newest and simplified version of LSTM network [31] is the GRU (Gated Recurrent Units) which is capable of handling long term dependencies. Different applications use GRU to negotiate the vanishing gradient problem.

### 2.4.3 Sequence to Sequence Networks (seq2seq)

It is composed of an encoder and a decoder RNN that generates the output sequence. Figure 2.4 describes Sequence to Sequence model structure.

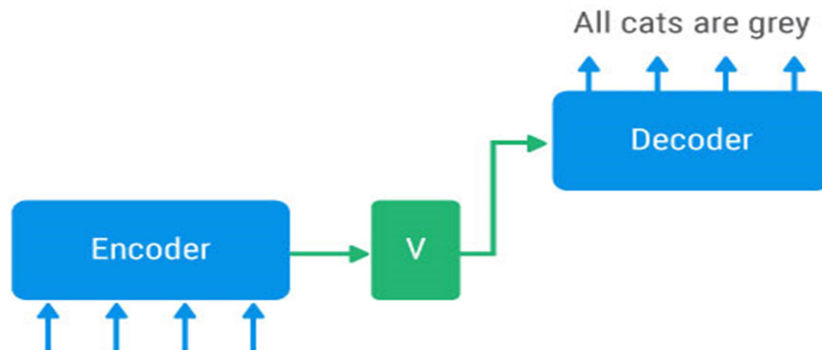


Figure 2.4: Seq2Seq model

RNN, in its hidden state of the encoder is used to compute a generally fixed-size context variable  $V$  as in the above figure which represents a semantic summary of the input sequence which is given as

input to the decoder. RNN can be with sequences which might not be necessarily of the same length.

The basic RNN architecture was first proposed by Cho et al. [32] and shortly after by Sutskever et al. [33] also. They were the first two people to obtain state-of-the-art translation using this approach and to be termed as the encoder-decoder or sequence-to-sequence architecture. The idea is simple: (1) an encoder / reader or input RNN processes the input sequence. The context  $C$  is emitted by the encoder, usually as a simple function of its final hidden state. (2) a decoder / writer or output RNN is conditioned on that fixed-length vector to generate the output sequence.

#### 2.4.4 Applications of Recurrent Neural Networks

RNNs are flexible to be trained in a genre of sequential data. Some of its applications can be seen in many natural language processing tasks. Some examples of it are given below

##### 1) Machine Translation

In the source language, there would be sequence of words to be given as input in the source language (be it German or Malayalam). The requirement is to convert input text to a target language like English.

There is a significant difference between language modelling and machine translation. The output starts in machine translation only after the input has been seen as there might be a requirement of information



captured from the input sequence in the first word of the translated sentence [35]

## 2) Language Modelling and Text Generation

These are the tasks in which a few words would be given in a sequence and the probability of each word would be predicted by the network using previous words of a sentence. The RNN language model example is shown in Figure 2.5 below [34].

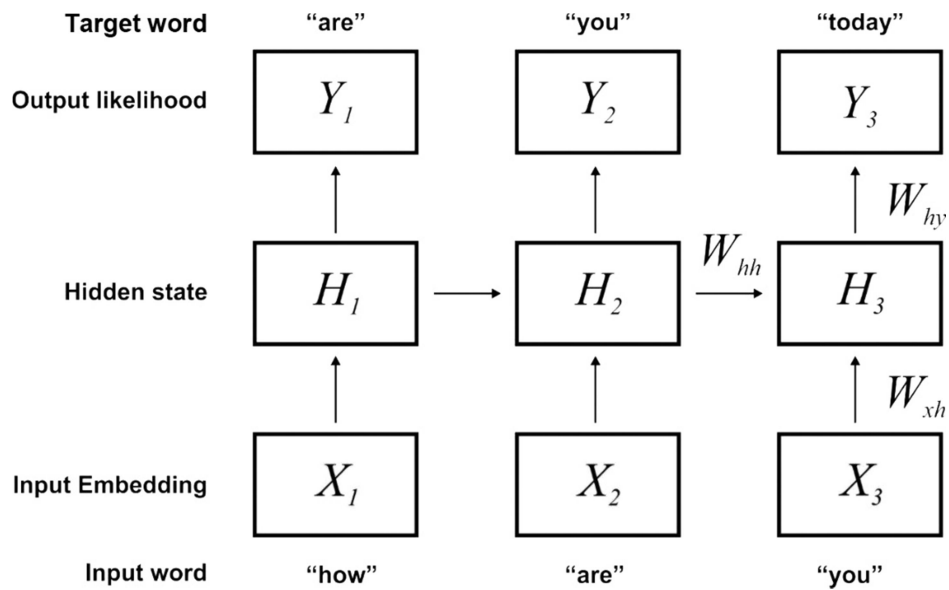


Figure 2.5: Language model

Language Models give probability distribution over sequence of words. A generative model can be obtained using this in which new text can be made by sampling words, considering the output probabilities. In

language modelling, the input is generally a word sequence and output is simply a predicted word sequence.

### 3) **Speech Recognition**

The last few years have seen the effective collaboration of hidden Markov models and neural networks for speech recognition. The process of speech recognition is achieved by conversion of sound sequence to phoneme sequence using a classifier [36,37]. Multiple hidden layers are made to use by the RNN to identify latent dependencies in executing speech recognition.

## 2.5 **Ontology Based News Archiving and Extraction Systems**

The difficulty to apprehend semantic knowledge of the application domain (which includes axioms, concepts and inherent properties) is a snag in the traditional news extraction systems. When dealing with customized news systems, knowledge about the domain through which news would be accessed is inevitable. The comprehensive results via generic browsing would stay as a hindrance to precise searches. The numerous discrete domain models would highlight the need-specific customization effect.

Domain models with advanced semantic structures are more effective in providing content specific news than those having relationships with alludes. Issues like substandard quality of information can get a solution through ontologies which provide a semantically rich structured approach in modeling a domain. Recent years have seen the

mushrooming of a number of ontology driven extraction frameworks. Customized extraction and being able to peek into the semantics of the news content are add-on benefits of the ontology driven systems. Table 2.1 summarizes the reviewed ontology based extraction systems.

**Table 2.1: Ontology based news extraction systems**

Approach	Techniques	Reviewed Systems	Year
Semantic Based	Ontology Driven	SmartPush [38]	2001
		SeAN [39]	2001
		aceMedia personalization system [40]	2005
		myPlanet [41]	2001
		SenSee [42]	2007
		Valet et al. [43]	2006
		Hermes framework [44]	2007
		Athena [45]	2010
		ePaper [46]	2009
		News@hand [47]	2008
		infoSlim [48]	2009
		Personalized Financial News Recommendation [143]	2015
		Ontology based recommender system of economic articles [144]	2013
		Ontology-based Recommender System for Online Forums [145]	2011
		Ontology-based Top-N Recommendations on New Items [146]	2014
Ontology-Based Recommendation of Editorial Products [147]	2018		

As shown in the above table, most of the reviewed systems are in the news. Some of them are concentrating on customization in the field of news recommendations.

The deficiency of commercial systems like the lack of customizable ontologies can be eliminated through Smart Push [38], which uses structured content using domain ontologies. The semantically structured news articles are juxtaposed using a user profile to obtain relevant news articles and a user feedback system is employed to incorporate the dynamic interests of the users.

SeAN [39] is a system that works along with user models, which use an ontology which splits it up into various dimensions (sections and sub sections) to give a better view of a user profile rather than other systems that illustrates conceptual domains. The dimensions include cognitive characteristics, lifestyle, interests and expertise. An interesting aspect about the system is its behavior tracking feature. SeAN follows a same parallel structure to newspaper editorial systems. Using ontologies to describe user models is an interesting approach to customized news.

AceMedia [40] system, a part of the huge framework of AceMedia project, makes use of the semantic insights of AceMedia framework and aids in developing a layer of device adaptive capabilities for the semantic aware user. The system is an intended framework that aids customization facilities in multimedia content management. The framework is based on an ontology based illustration of the domain through which user

preferences are collaborated with content semantics. Automatic learning capabilities are developed using ontologies to update user profiles. The resultant interests of the users are analyzed along with the available metadata to facilitate browsing, retrieval of data and browsing of content. An open platform is used that facilitates adaptive capability extensions.

MyPlanet [41], is a news service which is an extension of the PlanetOnto news publishing system. An ontology driven interest – profiling tool enables the users to go for their preferences. It also supplements ontology driven heuristics to search for news items related to the user’s interests.

Customized access to TV content is provided by the SenSee system [42] in a cross media environment. The focus of the system is in finding programs that support the interests of individual and group TV viewers. A hybridized view of data from heterogeneous and web sources is obtained through the system. The category of customization achieved here faces issues related to data integration and context modeling. IFancy is a TV guide application which uses Sensee Framework. The system integrates various data sources which are connected and mapped to external vocabularies using ontologies.

Valet et al. [43] proposed a system that aims at contextualization within customization. Within the structure, a novel contextual knowledge modeling scheme is proposed for the contextual activation of semantic user preference to make a good synchronization with user preferences

and activities. For example, in an interactive retrieval process, use of semantic concepts for representation of meanings and usage of ontology based information forms the main essence of the system. It also helps in compiling implicit textual messages, with a much general representation of user preference. The twofold benefits include increased accuracy and reliability by avoiding the risk of irrelevant news preferences getting in the way of retrieval activity.

Hermes framework [44], is a framework that makes customized service using input, output and an internal processing. The input comprises of concepts opted by the user and predefined RSS feeds of news items. The grouping of these items using concept and knowledge base can be termed as internal processing. The customized news that is produced is known as the output. The system offers a semantic based approach to retrieve news items from a domain ontology.

Athena [45], this system can be considered as an extension of the Hermes Framework. It uses many methods to find the user interests like user profiles, news items and many other similarity measures. The ontology used on Hermes network is the center of Athena which provides the relationship between concepts and the domain concepts

The ePaper [46] project is a prototype of a system that also gives an electronic newspaper reading facility for its users. The system gives the feel and look of a real newspaper while updating customized news aggregated from many news providers on a medium formatted mobile

device. The system works as a client server application. The system's functions include collection and grouping of news, predicting the relevance of the news items based on content and redistributing the processed, customized news. The system was developed at Deutsche Telekom libraries and sponsored by Deutsche Telecom Co.

News@hand [47] works by making news suggestions with collaborating information and content features. Profiles and news items are portrayed in terms of domain ontologies. Semantic relations along with these concepts are used to mend the above representations and inserted with the recommendation process. The system unlike others uses controlled and structured vocabulary to elaborate the preferences of the users and news texts. Since this is ontology based, the system is less ambiguous, its portable, the proposal's nature is multisource and domain is independent from subsequent recommendation algorithms.

InfoSlim [48] system employs semantic techniques to explain user preferences and news items to enhance metadata information into the key word vector. This allows in making a measure of similarities between user profile and item profile not only in lexical- level cosine- based method, but also in semantic level ontology based level. Thus this method gives precision to recommendations, reflect users interest and mobile resources.

Financial news recommendation system works on an algorithm [143] which provides a way to find articles based on user interest. Here, unstructured text data is represented as concept terms and stored in a

domain ontology. User profile is created and updated to predict the actual interest of users automatically and by the experiment of this algorithm in a dataset proved that the algorithm discussed here performs better than traditional systems.

Online forums provide a platform to discuss various topics of interest. One of the problems involved in this domain is information overloading due to the huge volume of discussion data. Ontology based information retrieval system [145] is proposed to solve this issue by considering the semantics. The ontology based system can suggest discussion topics based on user interest and can avoid duplicate posts. Evaluation proved that the ontology based system outperforms the traditional information retrieval applications.

Major publishers do analyze the catalogue of their products to decide which items are to be marketed in a particular venue. Earlier it was done by publishing editors manually. It was a hectic task considering the increasing number of products. An ontology based Smart Book Recommender system [147] was developed for Springer Nature to solve this issue and the experiments show that the application is effective.

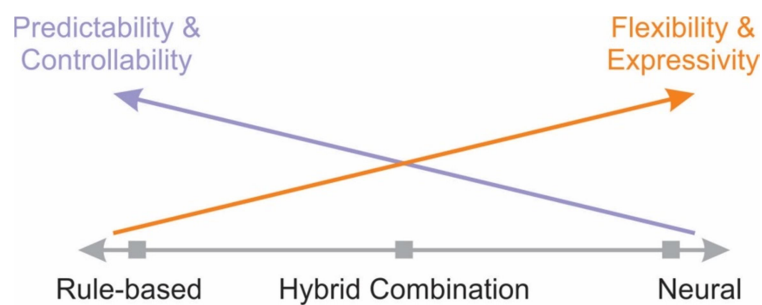
Collaborative filtering technique is widely used in recommendation systems and data sparsity is one of the drawbacks of these systems. Data sparsity is due to low score in user rating matrix. Ontology based recommendation using matrix factorization is proposed [146] to handle the missing values in the user rating matrix.



Personalized information is required for decision makers to formulate their decisions especially in the economic sector. To get a personalized view, the information has to be structured using concept terms and the extraction process has to use this structured information. Ontology based recommendation systems for economic articles [144] is used in this scenario and user profile as well as semantic description of articles is represented using ontology.

## 2.6 Text Generation Based on Neural Networks

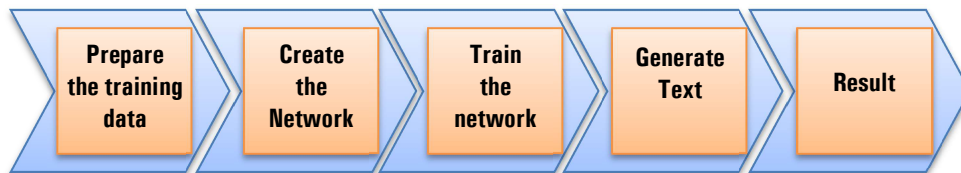
Neural networks have become capable in many natural language processing (NLP) tasks like machine translation and sentiment analysis. The main function done by NLP is text generation and the conditioning of it. Figure 2.6 illustrate the trade-offs of two types of systems in text processing field. In conventional times, the techniques used were rule or template based or models like n-gram or long linear models [49,50]. One drawback these systems faces are the requirement of hand engineering to scale in template based models [51].



**Figure 2.6: Trade-off between rule based and neural based systems**

Rule based systems are not up to the mark in terms of flexibility and expressivity as it is evident from the Figure 2.6. Neural net models are yet to prove its power in natural language processing field in term of predictability and controllability. Neural Networks, despite their benefits are underrated and not much put into application.

Machine generated texts have become a trend and numerous methods are available today to assist production of texts in different types and qualities. The complexity of models capturing the languages has undergone a significant growth with the growth in computational power. The Figure 2.7 summarizes the text generation procedure using neural model.



**Figure 2.7: Text generation using neural network**

The text generation process starts with preparing of training data. The main step involved here is to select or create the suitable neural network model. After creating or selecting the model, the network is trained using the pre-processed data. Finally, text is generated using appropriate input.

The Table 2.2 summarizes some of the application areas of sequence to sequence model in text generation.

**Table 2.2: Sequence to Sequence model application areas in text generation**

Sl.No	Application	Problem Description	Reference	Year
1	Machine Translation	Converting a text from a reference language to a target language.	[86], [87], [88], [89], [91],[92],[33]	2014, 2015, 2017, 2018
2	Text Summarization	Summarize a large text document	[84], [85], [86], [90], [93], [94],[95], [96],[97],[98], [99]	2015, 2016, 2017, 2018
3	Headline Generation	Headline generation from a news text	[55]	2015
4	Question and Answer Generation	Generating questions from a text, Answer generation from a text and a question	[100],[101], [102],[103], [104],[105],[106]	2015, 2016, 2017
5	Natural Language Inference	Finding a natural language hypothesis X is inferred from a natural language statement Y	[124]	2001
6	Semantic Parsing	Creating SQL query from manual written description	[111],[112]	2017, 2018
7	Image Captioning	Creating a caption based on image content	[113], [114], [115], [116],[82],[83],[117]	2014, 2015
8	Video Captioning	Creating a caption based on video content	[118],[119],[120], [121]	2015, 2016
9	Speech Recognition	Convert a speech to text and vice versa.	[80],[81],[122],[123]	2014, 2015, 2016
10	Dialogue generation	It used to generate a dialogue between two agent's e.g., between a robot and human	[107],[108],[109], [110]	2016, 2017

The Table 2.2 details the application areas of Seq2Seq model in the text generation field. A lot of studies are taking place in this domain and new text generation applications are developed on daily basis.

Storyline generation framework [52], unlike most systems that are taught to extract news from different time periods, is based on neural network and focuses on squeezing out the crisp news with categories. While conventional systems work on framing coherent stories from relevant news, Storyline generation framework also uses probabilistic graphic models. The model was assessed on mainly three news corpora.

The main issues of natural language generation (NLG) were resolved with the RNN networks. Yet it fails to study the production of machine generated news comments which is an innovative approach. The process requires considering opinions of a larger sample. Gated Attention Neural Network model (GANN) [53] uses gated attention technique to compensate for the contextual relevance issue. Methods like random sampling and relevance controls are used for its effectiveness.

Session-based Recurrent Neural Network system [54], is used in recommending catchy and informative news articles. It is an important function of news sites which can be made possible through Neural Network which encapsulates the preferences of the users and adjusts referral results accordingly.

The Recurrent Neural Networks were employed by Lopyrev [55] to coin headlines from news article. The sentence which is provided as the input is transfigured into a distributed representation of one word at a time by the encoder and for the hidden layers to be fed. Each word in a headline is generated using the output of a hidden layer by the decoder using attention mechanism. The algorithm uses attention mechanism to calculate importance of each word fed to the decoder.

Grangier et al. [56] introduced a model based on neural networks for text generation that uses Wikipedia's biographies as dataset. Biographical sentences are made using the fact tables on the dataset employing conditional neural language models for it. The model was polished with global and local conditioning. The model thus generated describes people's biographies in the manner of structured data.

A single convolutional neural network architecture had been developed by Collobert and Weston [57] that masters features with the limited yet sufficient knowledge. A neural network was trained as to find some outputs like POS tags, semantic roles, semantically similar words and chunks used for language modelling etc. The network is aided for these tasks using weight sharing. The process can aid many NLP tasks like parts of speech (POS), semantic role labelling and learning a language model.

The possibilities of RNN for generation task has been showcased by Roemmele et al. [58] and how this system can be made use for

analysing generation systems. The data driven approach is used in developing creative help application. The particular system offered suggestions which modelled the context of the originating story than user's story which limits the compatibility. For quality evaluation, modification to suggestions has been tracked by creative help. Language generation can be analysed through user interaction, where the evaluations are to be conducted separately.

Sutskever et al. [59] proposed that text generation can be made possible using RNN with HF optimizer providing solutions for challenging sequence problems. An RNN variant was introduced in which multiple connections were used to make input character which determines transition mix from one hidden state vector to another. The relevance of RNN was illustrated in this work by using them for language modelling tasks.

A method has been proposed by Uchimoto et al. [60] for generating sentences using keywords or headwords. The system consisted of generation-rule acquisition, candidate-text sentence construction and evaluation. Each headword for the generation-rule is acquired automatically during the generation rule acquisition phase. The text is generated in the form of dependent trees by the construction part and complimentary information was used to substitute for the missing information. The evaluation part has a model that generated a proper text, when keywords are given. The model considered word n-gram information as well as words information dependency.

Ayana et al. [61] made an extensive study about existing and recent developments in crafting a neural headline using Recurrent Neural Networks, a learning method capable of mapping documents to the headlines.

Machine translation uses Sequence to Sequence model which interprets text from a single language, (for example, English) into another, (for example, Malayalam). Input sequence consists of words in one language and sequence of words will be the output in another language. It accomplishes more spare time and lessens translation costs [86,87,88,89,91,92]. The Sequence to Sequence model by Sutskever et al. [33] saw English to French translation being performed with Long Short Term Memory (LSTM) networks, using an encoder and a decoder. Since the encoder is mapped to a vector of fixed size, the length of the input sequence is flexible. The result was fruitful primarily due to the implementation of LSTM cells.

Text summarization is the process of making a short, precise, and familiar synopsis of a more extended text. By using encoder-decoder model, a long document is summarized to a short one in a very effective way. Modified text summarization methodologies would address the method of creating proportion of content data to find vital information. There are two distinct methods of content summarization: indicative and informative. The text length of the indicative summarization is less than 10 percentage of original content. The other type of summarization framework gives squeezed data and the length of synopsis is in between

20 to 30 percent of the main content. There are three courses for reducing records. These are identifying topic, interpretation and generation of summary. Head line generation is also a variant of text summarization. Generally long news articles contain vast measure of data. Numerous times because of absence of time, individuals can't pursue the entire news article. Consequently, headline is required with the goal to get the final thought of the news without perusing the entire news article. For accomplishing the point, LSTM units with encoder-decoder Recurrent Neural Network were utilized [84,85,86,90,93,94,95,96,97,98,99].

Automated Question Answering and creating questions are two dynamic zones of Natural Language Processing with the first one overwhelming the previous decade and the last one on the way to rule the next decade. Because of the huge measures of data accessible electronically in the Internet-era, automated Question Answering is expected to satisfy data needs in a fruitful and powerful way. Automated Question Answering is the assignment of giving answers consequent to queries asked in a natural language. Typically, the answers are retrieved from a large collections of documents. Generating questions from a text document or an image is achieved by LSTM. The input is a piece of text and the output will be a set of questions related to the text or image. Given a text document or an image and a question, finding the answer to the question is also possible by making a generative machine understanding neural net model that adapts together to answer questions dependent on records. [100,101,102,103,104,105,106]



Generating automatic SQL queries from a given manual narrative is performed using Recurrent Neural Networks. The feed data to the neural net model is text description and the model produces a SQL query based on that. The SQL command equivalent to that description will be delivered with help of semantic parsing. Semantic parsing is the way to map a characteristic language sentence into a formal portrayal of its significance [111,112].

Inference has been a focal theme in man-made brainpower from the beginning, yet while programmed strategies for formal derivation have progressed massively, not much advancement has been made on natural language inference, that is, deciding if a natural language speculation P can legitimately be surmised from a natural language preface H. The difficulties of natural language inference are different from those experienced in formal derivation. Regardless of its extraordinary straightforwardness, the model referred here accomplishes surprisingly great outcomes on a standard NLI assessment.

The Stanford RTE framework, utilizes composed reliance trees as a mediator in the semantic structure, and looks for a minimal effort arrangement between trees for P and H, utilizing a cost display which consolidates both lexical and auxiliary coordinating expenses. This framework is a run of the mill way among a class of ways to deal with NLI dependent on inexact chart coordinating [124].

Image captioning is the generation of an apt caption that explains the content of the provided image. Input is an image (sequence of layers) to the caption (sequence of words) describing that image. Consequently, producing a natural language portrayal of an image is an assignment near the essence of image understanding. The RNN-LSTM based neural systems learn how to portray the substance of images. These models comprise of two sub-models namely a question recognition and restriction display, which separate the data and their spatial relationship in images individually. Each expression of the description will be naturally adjusted to various objects of the info image on its creation. Many image captioning applications are developed based on the encoder-decoder model [113,114,115,116,82,83,117].

Generating a caption that explains the content of the video would be a difficult task since the caption should aptly portray the essence of the entire video clip. Transcripts are the full, exact and final content of a video. Some transcripts incorporate just talked discourse; while others incorporate depictions of non-exchange sound and melody verses. Subtitles obtained from either a transcript or screenplay of the discourse or analysis in movies, television programs, video recreations, and such, usually displayed at the bottom of the screen. They can be a type of composed interpretation of a discourse in a specific language. The input fed to the sequence model is a sequence of images or video and the outputs generated are the video captions [118,119,120,121].

Speech recognition is the capacity of a machine or program to distinguish words and expressions in talked language and make them to a machine-clear arrangement. Simple speech recognition programming has a constrained vocabulary of words and expressions, and it might just recognize these on the off chance that they are talked plainly. More complex programming can acknowledge regular speech. The Sequence to Sequence model takes a segment of speech as input and convert it to text and vice versa [80,81,122,123].

Dialogue generation is used to make a dialogue between two agents e.g., between a robot and human. It is a computer framework proposed to chat with a human with a coherent structure. The primary information sources are text, speech, graphics, gestures, and different modes for correspondence on both the input and output frameworks [107,108,109,110].

## **2.7 Gist of Observations**

As a result of the literature review, the following observations are drawn.

- Ontology is a powerful tool for knowledge representation. Its use in information extraction and archiving field is significant. Ontology enables the machine to understand the underlying semantics of information content. Thus its usage in information extraction scenario would provide a customized experience based on user's need.

- Neural Networks immense power is evident in different application areas. LSTM network is widely used in text processing applications due to its power to remember long term dependencies and is far more capable than statistical or rule based models in this aspect.
- The applications of neural network and especially Seq2Seq neural network model in different text processing scenarios are reviewed.

## 2.8 Research gaps

Based on the study conducted in this particular problem domain, the following knowledge gaps that are yet to be researched have been identified.

- The already used text generation techniques like rule based, statistical and data driven techniques have their limitations. They have numerous predefined rules and are domain specific. These models work well for small vocabulary but are ineffective in domains like news.
- Text generation from head words or key words are done by some statistical models. But they fail when considering the case of long term dependencies.
- Deep learning techniques like RNN-LSTM network is used for text generation but much research has not been done on generating text from keywords.

- There are only a few studies done in finding the relation between the quality of text generated using neural net models and the number of keywords.

## **2.9 Motivation**

With the recent developments in the media sector, news production has peaked to an incessant level with an enormous rise in the rate of growth of news stories. Log systems in a visual news broadcasting centre is not particular about following a standard. A study was conducted to get information about the archiving systems followed in news channels and it shows that some news channels are still in the premature stage of archiving. The complex nature and size of the content, as well as the limitations in time for describing, cataloguing and sorting the received information, makes the management of archives a difficult task. Thereby it can be seen that such news systems share many problems and characteristics with the World Wide Web.

Semantic Web technologies [3] are applicable in situations like these and hence a process like ontology based news extraction and archiving can be effective [62]. For automatic generation of news from an archive, the semantic knowledge of a journalist's query as well as the related news events stored in the archive as per user query is to be grasped by the machine. An important role can be done by ontology in this aspect. The news concepts which are the inputs for news generation can be extracted from the archive and user query using ontology.

Neural networks have attained significant results in text generation problems [63]. It is far more expressive than traditional statistical or rule based models [136]. RNN especially LSTM which is one version of RNN can be employed for generation of news if all details connected with the news events have been fed in the archive as per the reporter's requirements. The framework described here would give a much easier experience that is customized for the journalist's requirement.

## **2.10 Conclusion**

The first section of this chapter discussed about the ontology tool and its applications, followed by deep learning techniques and its applications.

Subsequently in this chapter, ontology based news extraction systems are reviewed to get a better view about the various usage scenarios. After that, Neural Networks based natural language text generation applications are also discussed here to get a wider perspective of the problem under study. Finally, the research gaps are identified and formulated and the motivation of this research work is mentioned.

*.....✍.....*

## DEVELOPMENT OF RESEARCH FRAMEWORK

- 3.1 Introduction
- 3.2 Research Strategy
- 3.3 Research Framework
- 3.4 Conclusion

### 3.1 Introduction

The problem that is taken up in this research aims to propose a news generation framework. This chapter gives an idea about the research strategy and major stages of the work. The research method followed in this study is based on observable experiments or empirical evidence.

### 3.2 Research Strategy

The overall context of this work is to use experimental approach. Experimental research provides the finest approach to find the reason of a specific situation [135]. One of the objectives of this work is to study the relation between the number of keywords and the quality of news generated. Here in this study, the dependent variable is the quality of news generated and independent variable is keywords derived by ontology.

### 3.3 Research Framework

The framework provides an overall outline of research process from beginning to the end of investigation to get the solution to the objectives. The detailed structure is provided in Figure 3.1.

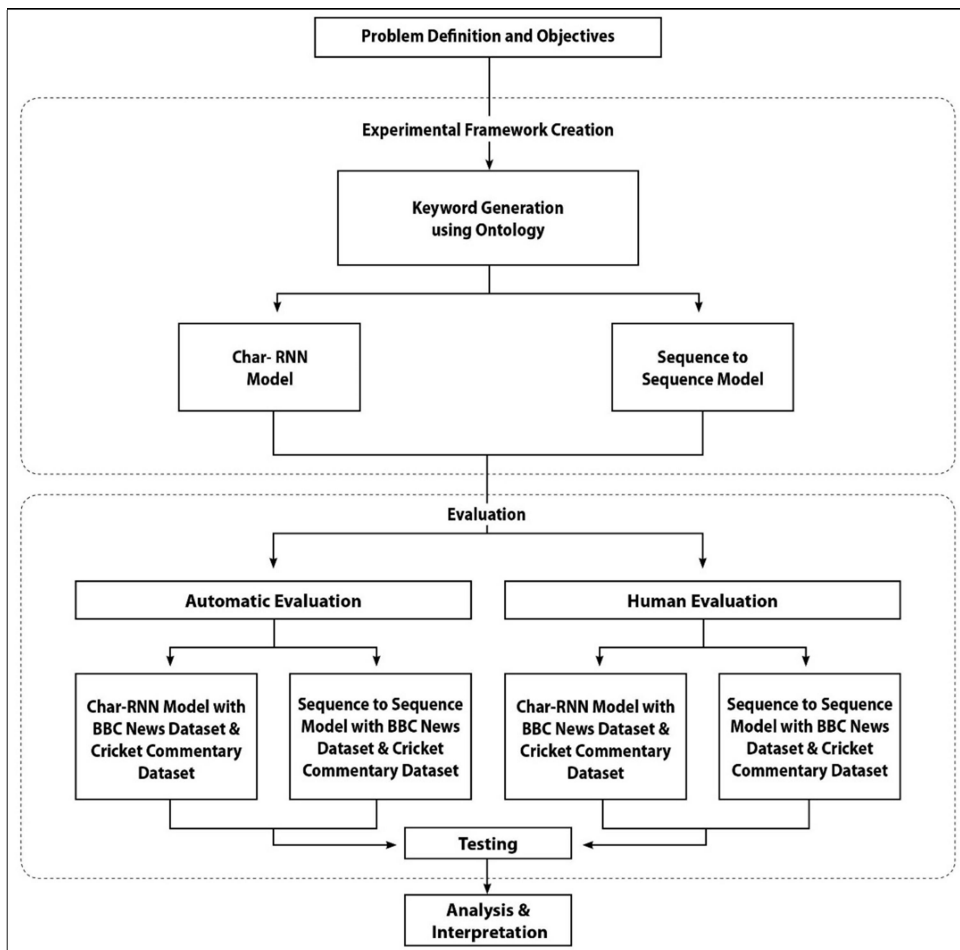


Figure 3.1: Research Framework



The research framework described here is formulated based on the literature review and the selected background technologies that help to achieve the objectives.

Generating keywords by using ontology from news text is the first step in the proposed news generation framework. The procedure is explained below in detail.

### **3.3.1 Creating Experimental Framework**

Here, a technical environment is created to study the problem. This includes creating keywords using ontology and generating text using suitable neural network models.

Ontology is an efficient mechanism that symbolizes knowledge in a domain. At first, ontology based news extraction and archiving application was developed to study the effectiveness of ontology in digital archiving. Open Calais ontology [68] is used for this purpose. The understanding of high-level and low-level concepts in a domain is a measure of ontology completeness. Putting it another way, if the ontology can grasp the domain knowledge, it definitely improves the understanding of any event related to that domain. The limitation to understand low level concepts by the Open Calais ontology used in the above mentioned application was explored by evolving a new system which uses both Open Calais and a custom made local ontology. The point derived from the above mentioned experiment is that with

ontology completeness of a high degree, more keywords will be spotted by the ontology. The detailed discussion is done in Chapter 4.

Two models are framed to study the research problem. One neural net model is based on Char-RNN and another is based on Sequence to Sequence model which uses Word RNN. Char-RNN model is chosen to explore the unreasonable effectiveness of Recurrent Neural Networks. Word RNN is employed in the Sequence to Sequence model by considering the inputs to the news generation model which are keywords extracted by ontology. So Word RNN may perform in a better way when compared to other models under consideration. The details are elaborated in Chapter 5.

### **3.3.2 Evaluation**

The automatic evaluation technique like BLEU [74,75] and ROUGE [76,77] are used here. Basically BLEU and ROUGE are document similarity measures. These methods are very effective in machine translation scenario. Yet they are not popular in text generation. One of the reasons behind it is that the machine generated text would not be totally similar to the referral text. So here, human evaluation approach is also considered as part of the evaluation strategy. A human evaluation approach was designed based on certain criteria, however one cannot neglect the limitation of human evaluation also. The evaluation strategy followed in this work is a combination of these

two approaches and it formulates a conclusion by correlating the automatic evaluation and human evaluation scores.

Training and testing of the framework is done by two datasets. One dataset is BBC news dataset which is available in the internet. The second dataset is developed especially for this research work. The notable point in these two datasets is that the BBC dataset comprises of lesser number of keywords for each news extracted by Open Calais ontology than Cricket Commentary dataset. In the case of Cricket Commentary dataset, the Open Calais along with an ontology developed specifically to conduct this study were used to generate keywords. The main purpose for doing this is to analyse whether the neural net models created in this research performs better with a dataset having more keywords to prove the necessity for a high degree of ontology completeness.

The front-end language used for framework development is Python and backend is Tensor Flow. The application is developed in Keras framework and the technical environment used in this work is Amazon AWS Deep Learning AMI. The details are discussed in Chapter 6.

The results analysis is performed and discussed on the outputs based on human and automatic evaluation. Char-RNN model with BBC dataset, Char-RNN model with Cricket Commentary dataset, Sequence to Sequence model with BBC news dataset and Sequence to Sequence

model with Cricket Commentary dataset are the four cases to discuss. The experimental results and discussion based on it is provided in Chapter 7.

### **3.4 Conclusion**

The research framework followed in this thesis is explained through this chapter. Here, experimental approach is selected to conduct research work. This chapter gives an idea about how can the objectives be achieved and the evaluation approach followed in this work. Each step in the research framework is elaborated in the subsequent chapters.

*.....❧.....*

## KEYWORD GENERATION USING ONTOLOGY

- 4.1 Introduction
- 4.2 Ontology Based News Extraction and Archiving
- 4.3 Open Calais Ontology
- 4.4 YouTube – 8M Dataset
- 4.5 Evaluation Results of the Open Calais Based News Extraction System
- 4.6 Conclusion

### 4.1 Introduction

One of the main objectives of this research work is to study the relation between number of keywords derived by ontology and the quality of generated news. Ontology completeness is a property which contributes to the knowledge level identified by the ontology. It decides the number of keywords extracted by the ontology in the context of this study.

The chapter starts by introducing the concept of ontology completeness and discusses its role in the archiving scenario. Later, an ontology based news archiving and extraction system is developed and

discussed to understand the effectiveness of ontologies in a classification and archiving scenario. Then the limitation of ontology used in this application in terms of completeness is discussed. A use case is developed to address the knowledge gaps in the Open Calais ontology which is used for application development. Sample ontology is developed based on the use case and used along with Open Calais in the news extraction application to grasp the necessity of ontology completeness for a better outcome in the research scenario under consideration.

#### **4.1.1 Ontology Completeness**

The proficiency with which ontology replicates the real world is termed as the completeness of the ontology. An incomplete ontology omits specifications and sub concepts. The property of describing all the ontology parts in its wholeness is called completeness. A set of completely presented and described ontology parts can be equated with granular characteristic, i.e. how microscopically the information has been efficiently subdivided for easy understanding. The standard of detailing depends on the type of ontology and prescribed specification. Thereby ontologies can be divided into generic and domain (specific) ontologies.

The two mutually opposing features of ontology are given below:

- a) Generality (Universality): when ontology is to be applied for general purposes of the domain.
- b) Specificity: When ontology is applied to a narrower domain in specifying the minute details of the domain. Here the emphasis is on the depth of the ontology.

Consider the digital library scenario. The present world has seen numerous libraries emerging from long-term personal digital libraries and specialised digital libraries. The available information about the growth rate of data has made it possible to assign a considerable significance to the techniques that aid in organising information and various structures like glossaries, taxonomies, ontologies, thesaurus, semantic works etc. They are used to organise information, its processing, for aiding in its selection, organisation and dissemination. The information collections are heterogeneous too [20,64].

However, the complex nature associated with optimising the management processes of information resources and the difficulty in developing and managing digital libraries can be attributed to the multiplicity and elusiveness of related information and technologies in the digital environment. There is dire need for a knowledge discovery approach based on bottom-up automated knowledge extraction and top-down knowledge creation when the amount of huge data available is considered [20,65,66].

Many studies have proposed ontologies as an effective substitution for the organisation of information, which would meet the information needs, help in the transformation of traditional services to digital ones, whereby the next generation libraries would be more active in offering customized information according to the needs of each individual. Ontologies in digital libraries can be used for easy manipulation and processing of information in the digital format.

In the case of digital libraries discussed above, the ontology completeness is an essential property for effective information organization and extraction scenarios.

## **4.2 Ontology Based News Extraction and Archiving**

As per the study conducted in news channels in Kerala, India, all archival systems in news channel libraries are keyword based. The proposed system focuses on extracting and archiving news in a news channel library based on ontology. It serves two major functions. The first function is the generation of keywords or conceptual terms from news, and the storage of this news using ontology. The second function deals with retrieval of the desired news using ontology. The following Figure 4.1 represents the ontology based news extraction and archiving system.



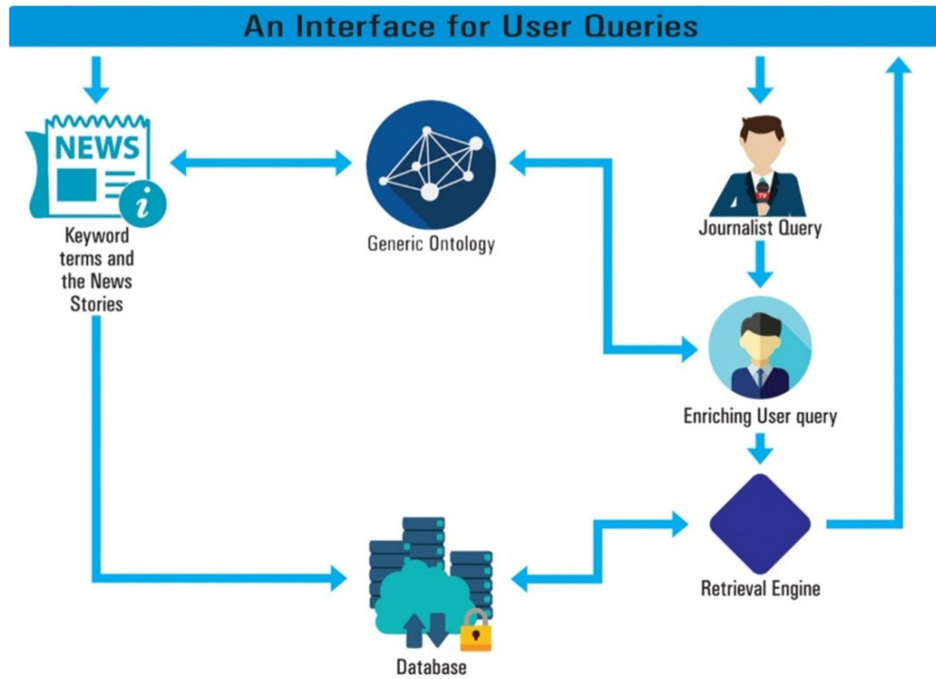


Figure 4.1: Ontology Based News Archiving and Extraction

The initiation of the process shown in the above figure is marked with the stockpiling of news content aided with conceptual terms opted by ontology. The following steps are used in archiving.

- Ontology tags are fished out from the news content in the process of archiving news.
- The database is stored with corresponding news and keywords.

The news extraction steps are portrayed in Figure 4.2.

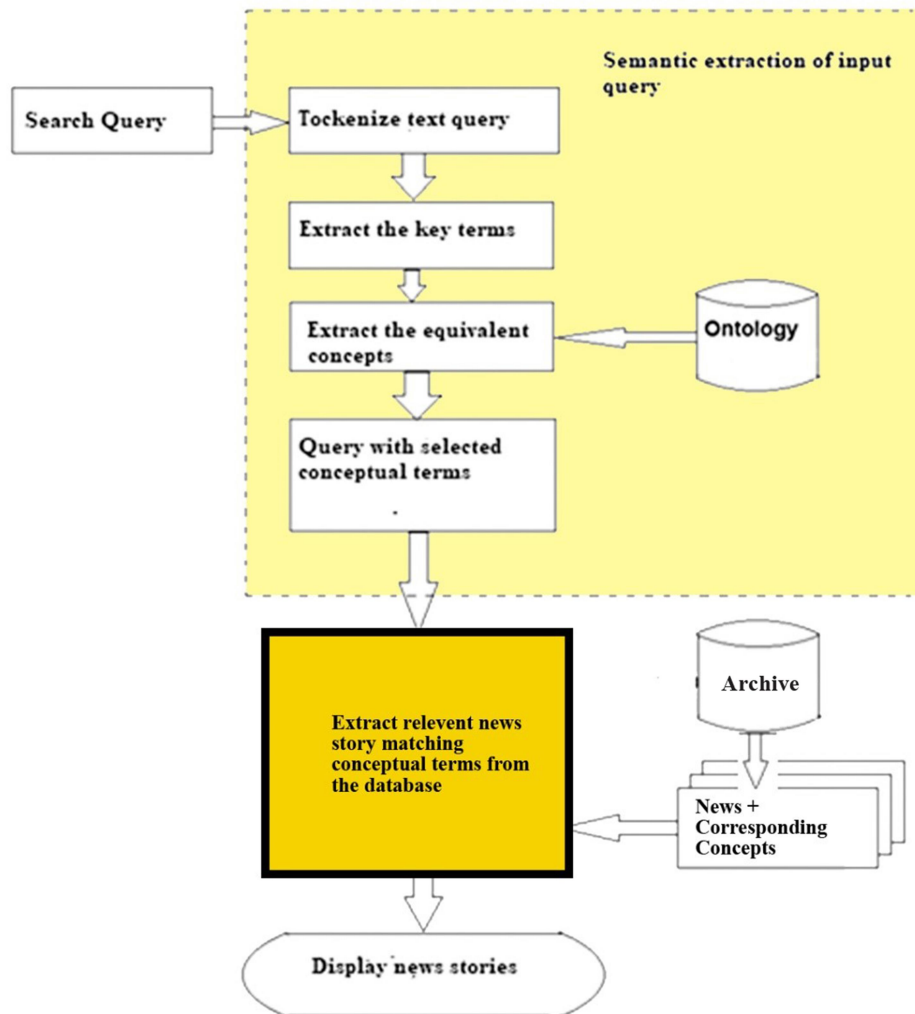


Figure 4.2: News Extraction Process

The news extraction procedure in Figure 4.2 consists of the following steps

- 1) Queries are inserted by the journalists or the news reporters via User Interface (UI)
- 2) The subsequent search query is tokenized, i.e. the words in the text are split.
- 3) Insignificant stop words like ‘and’, ‘about’, ‘the’ etc. are pulled out from the query.
- 4) Extraction of equivalent terms from the ontology and running the query to the database.
- 5) The keywords extracted from the query and concept terms in the stored news content in the database are matched.
- 6) Display of the news content based on concept terms which is similar to the search query.

The application to showcase the Proof of Concept was built in Drupal, an effective content management system by considering the availability of easily pluggable contributed modules. The ontology employed in this implementation is Open Calais and the dataset used to analyse the application is YouTube – 8M dataset.

### **4.3 Open Calais Ontology**

Open Calais [68] is an ontology that follows the news classification standardisation formed by the International Press Telecommunications Council (IPTC), developed by Thomson Reuters.

IPTC is formed by prominent news agencies in the world to protect telecommunication standards [67]. For the last couple of years, IPTC has focused their activities in developing and publishing industry standards to encourage the exchange of news data in common media types.

The IPTC has more than 50 companies, associations and organisations as its members all over the world. The members are mainly thought leaders and technology experts from news agencies and media professionals in news production and delivering. IPTC standards have a huge role to play in efficient news exchange between media organisations and world news developments. The main aim of the organisation is making data distribution an easier task. Technical standards are improvised to enhance information transfer and management between consumers, intermediates and content providers. IPTC is flexible to open standards, which increases its accessibility. The International Press Telecommunication Council taxonomy has three layers - the subject, the subject matter and the subject detail. The taxonomy has the ability to classify news articles based on content.

The Open Calais ontology is capable of analysing and highlighting the most suitable key terms from the content to deliver the results in the form of social tags with respect to events, topics and relations. Open Calais follows 17 top level topics of IPTC taxonomy. It includes politics, sports, weather, health etc.

Open Calais explains the unstructured text in four ways. IPTC topic is the first one. The second one can be termed as social tags which are derived from Wikipedia taxonomy. The broad taxonomy is dynamic in nature and similar to how Wikipedia topics are updated. This user generated topics are always updated and incorporates all major events. The third one is high quality taxonomy, available as a part of intelligent tagging called the Reuters Classification Schema. The fourth one termed as the Industry Codes Taxonomy, drawn from Thomson Reuters' analysis on how to associate relevant industries to an unstructured data.

Various taxonomies are used to extract key words using Open Calais from high level (IPTC) to Reuters Classification Schema to a more general taxonomy taken from Wikipedia.

#### **4.4 YouTube – 8M Dataset**

Many large labelled datasets are available today which have made many breakthroughs in machine perception and learning possible. The YouTube-8M dataset [69] was announced by Google in 2016, including numerous videos labelled in numerous classes, with a hope of finding a similar innovation and development in understanding videos. A cross section of the society is portrayed through YouTube-8M with millions of video IDs. About 20 domains of video content like sports, entertainment, news, hobbies, commerce, jobs, health and education are included.

News video sources like News Broadcasting, CBS news, ABS-CBN news, Newscaster, etc. were selected for the study mentioned here and only news with English annotations were extracted from the above mentioned sources.

#### 4.5 Evaluation Results of the Open Calais Based News Extraction System

Measuring semantic similarity is a generally acceptable tool in assessing the accuracy of ontology based information retrieval applications. In order to compute the semantic similarity between the extracted news stories and the submitted search query, Latent Semantic Analysis (LSA) algorithm is used [71,72]. LSA algorithm processing steps are described in Figure 4.3.

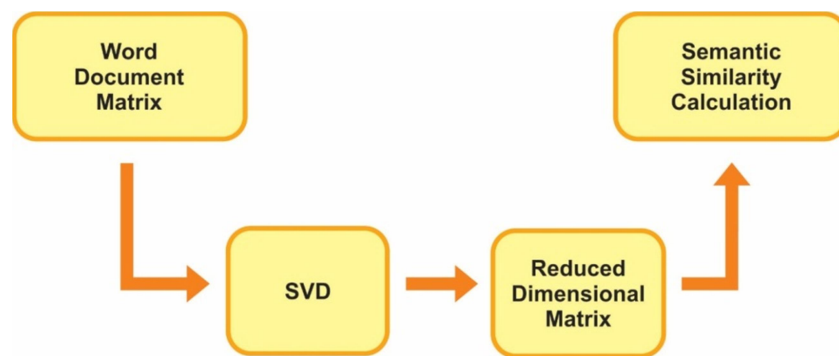


Figure 4.3: LSA algorithm steps

At first as shown in the above figure, word document matrix is made by assigning weight, using tf-idf (term frequency–inverse document frequency). The constructed matrix is decomposed using

SVD (singular value decomposition). Let vector space matrix be X. SVD of X is

$$X=U \sum V^T \dots\dots\dots (4.1)$$

Where U AND V are orthogonal matrix.  $\sum$  is a diagonal matrix

Using SVD, singular values are taken which are non-zero entries of the diagonal matrix  $\sum$ . These non-zero entries are employed to reduce the order of the matrix and a reduced dimensional matrix is formed. Word document similarity is calculated using dot product of word vector and document vector. This measurement is used for finding semantic similarity of news retrieved and corresponding search query. Table 4.1 provides the details of search query code representation.

**Table 4.1: Search query code representation**

<i>Query Code</i>	<i>Query</i>	<i>Ontology Terms</i>
Q1	Obama and Iran	Obama ,Iran
Q2	Obama about Israel nation	Obama,Israel,nation
Q3	Obama and Iraq	Obama,Iraq
Q4	Syria war and Obama	Syria,Obama,war
Q5	Bush and Obama	Bush,Obama
Q6	Obama and Cheney	Obama,Cheney
Q7	Romney and Obama	Romney,Obama
Q8	President election and Obama	President,Obama,election
Q9	Senator Barack Obama	Senator,Obama,Barack Obama

A sample of 9 queries for evaluation was prepared for demonstration purpose as displayed in Table 4.1. For each query, a query code was given for representation purpose and key terms were generated using Open Calais ontology. Semantic similarity is measured using LSA between search query and the search results. The details are given in Table 4.2.

**Table 4.2: Semantic Similarity**

<i>Average Semantic Similarity (In Percentage)</i>		
<i>Query Code</i>	<i>Normal Keyword Search</i>	<i>Ontology Based Search</i>
Q1	57.11	89.73
Q2	59.43	83.53
Q3	39.13	87.12
Q4	41.23	88.56
Q5	67.68	84.32
Q6	53.12	93.28
Q7	49.71	91.77
Q8	47.71	90.21
Q9	36.33	94.19

In Table 4.2, an average semantic similarity of the extricated news content is presented in terms of two extraction techniques with the search query. A trial was done with the Normal Keyword Search extraction and Ontology based search for each query code. Clearly



Ontology based news extraction gave a better performance. Figure 4.4 is a graphical representation of evaluation result.

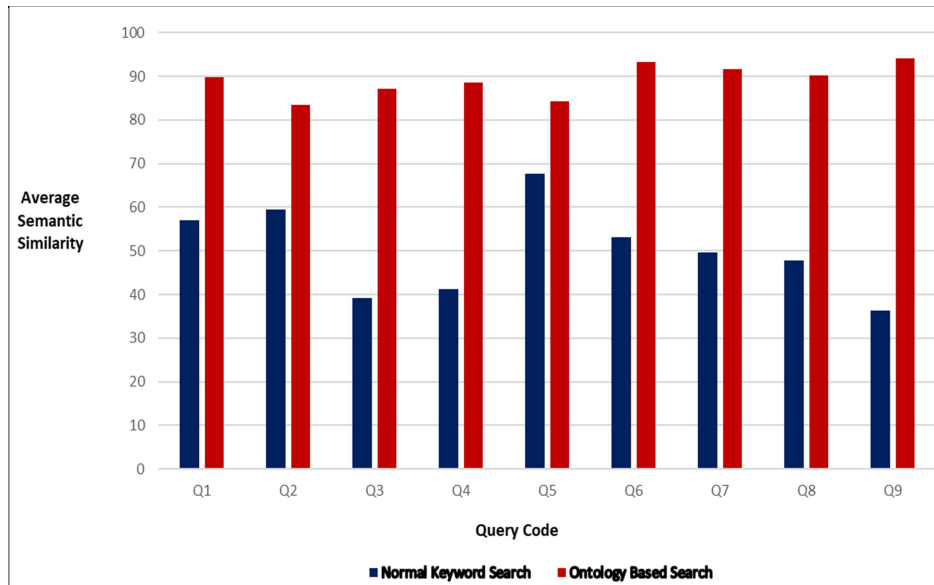


Figure 4.4: Evaluation Result

In Figure 4.4, the bar chart highlights the differences between the two approaches. Thereby it is found that ontology based archival system is efficient than the normal keyword based system.

#### 4.5.1 Limitations of Open Calais Ontology

Domain ontology plays a vital role in sharing and reusing of information. As a tool of knowledge representation, there is still a need for the acceptance of domain ontology as a well-engineered product. One main hindrance is the designing and maintenance of high quality ontologies.

Open Calais, as a system can scrutinize and extricate suitable key terms from the content and present the results in terms of social tags, establishments in relation with the topic, relations and events. Many of the news topics in YouTube-8M dataset were not extracted by Open Calais. That means the number of keywords identified by Open Calais is less. If Open Calais has more knowledge, the extraction and archiving will be more efficient. This is pointing to the low degree of completeness of Open Calais.

Ontology completeness depends on the anticipated purpose of usage. Open Calais is developed for a broad purpose rather than a specialized intention. One of the drawbacks of Open Calais is that it follows only 17 top level concepts designed by IPTC. That is, it can extract the general concepts in the news domain but not the specialized low level concepts. The limitation of Open Calais is evident from different studies already conducted [149]. In a study for introducing a methodology [148] for extracting multi-scale topic clusters from a text corpus, the dataset used was Vox media news articles. Open Calais could not extract the knowledge terms from a considerable fraction of Vox media dataset used in the above mentioned work.

One of the ways to check the completeness of an ontology is by finding the knowledge gaps. Use cases can be developed to address these knowledge gaps and finding the missing one. Such missing information can be added to the ontology to make it more complete.

To prove this point, a new ontology is developed for testing purposes in a specific domain and the news archiving system already developed was extended with the newly built ontology. One of the areas that remain partially uncovered by Open Calais is local knowledge. The newly formed ontology is called local ontology which consists of traditional knowledge in terms of geographical tags [70]. Figure 4.5 represents the application which uses both local and generic ontology (Open Calais).

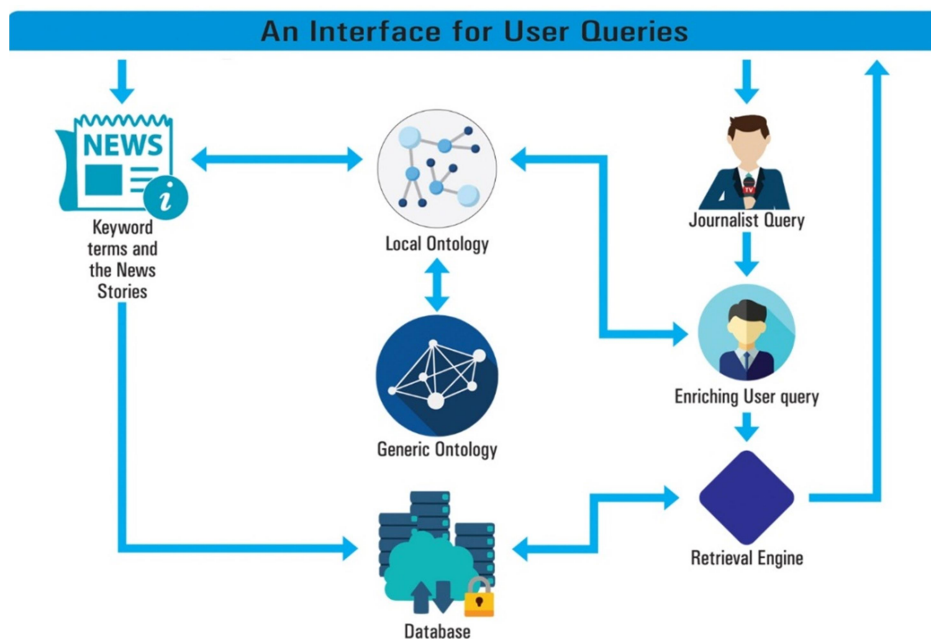


Figure 4.5: Ontology Based News Extraction System incorporating Local Ontology

Local ontology is incorporated with Open Calais (generic ontology) to analyze how the generic ontology fails in extracting traditional knowledge. Local ontology which when combined with generic ontology in the application mentioned in Figure 4.5 was proven to give fruitful results in extracting keywords / concepts related to traditional knowledge.

#### **4.5.2 Evaluation Results of the News Extraction System Based on Open Calais and Local Ontology**

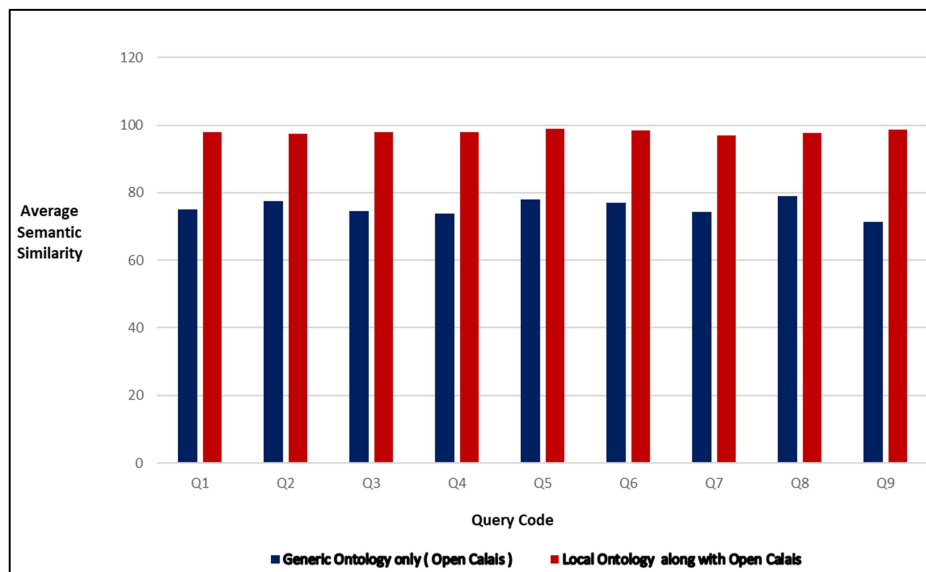
The application developed here is examined using GI. A product with a specific geographical origin that possesses a reputation for its origin would be marked with a peculiar representation called the geographical indication (GI). An inevitable connection is shared between indigenous knowledge and geographical indications in protecting local knowledge of a particular geographical area. Some of the sample GI tags included in the local ontology are used as search query for the testing purpose. The details are given in the Table 4.3.

**Table 4.3: Semantic similarity (Using local ontology along with Open Calais)**

Query code	GI tag	Category	State	Country	Avg. Semantic Similarity in percentage	
					Generic Ontology Only	Using Local Ontology along with Open Calais
Q1	Kachai Lemon	Agricultural	Manipur	India	75.1	98
Q2	Purandar Fig	Agricultural	Maharashtra	India	77.6	97.5
Q3	Malabar pepper	Agricultural	Kerala	India	74.5	98
Q4	Pokkali Rice	Agricultural	Kerala	India	73.8	97.8
Q5	Vazhakulam Pineapple	Agricultural	Kerala	India	78.1	98.9
Q6	Aranmula Kannadi	Handicraft	Kerala	India	77	98.5
Q7	Alleppey Coir	Handicraft	Kerala	India	74.2	97
Q8	Bhagalpur Silks	Handicraft	Bihar	India	79	97.6
Q9	Dharwad Pedha	Food Stuff	Karnataka	India	71.4	98.7
Q10	Banglar Rasogolla	Food Stuff	West Bengal	India	78.7	98

Table 4.3 provides insight about the advantage of using a local ontology with generic ontology by measuring the semantic similarity of query and search results. For example, for query about ‘Vazhakulam Pineapples’ (Vazhakulam is a place in Kerala famous for its pineapples), instead of listing results for ‘Vazhakulam’ and ‘Pineapples’, the user gets crisp information about the searched query when the local ontology is

incorporated with Open Calais ontology. Figure 4.6 is a graphical representation of the results shown in the above table.



**Figure 4.6: Evaluation Result (Using local ontology along with Open Calais)**

It is evident from the above figure that Open Calais fails to extract geographical tags used in the search query in testing the system. It points to the limitation of Open Calais to extract knowledge terms from a traditional knowledge domain.

## 4.6 Conclusion

The chapter emphasizes the ontology completeness property to identify high and low level concepts. The number of keywords extracted by ontology having high degree of completeness will be high. To prove

this concept, an application was developed which aids the study of the advantages of ontology in retrieval and news archiving in which a data set like YouTube-8M was applied to test the system. The main ontology used was Open Calais which is significantly black marked for its constricted digital voraciousness as an ontology with completeness property. The ontology's limitation is shown with the implementation of a domain specific local ontology. The showcased drawback highlights the requirement of ontology completeness which proves to be a major point when the huge diversity of news topics is considered. Adding new definitions, properties and missing entities are indeed important when reading optimization of ontology is considered. Topping the defined sources with much relevant data can give the knowledge-base a comprehensive source.

The news ontology should continuously be evolving since anything in the universe can be a news story. A dynamic evolution of news ontology is inevitable considering the present scenario where news stories have to be developed very frequently. The expansion of modeled concepts is necessary in news ontology. A constant reshaping of the ontology is demanded after the inclusion of new concepts in order to meet the integrity and consistency of the knowledge base which seldom affects the philosophical aspects in ontological decisions.

The crafting, refining and evolution of operational domain ontology requires numerous resources, time and strenuous expertise of professionals

for a couple of years. This is especially true in the case of news ontology where a numerous topics have to be dealt with. Though time consuming and strenuous, the production of high quality ontologies is inevitable for the effective development of domain applications

The next chapter introduces news generation framework which uses keywords derived by ontology to generate news. Ontology keywords / tags play an integral role in the news generation process. The completeness property of ontology discussed in detail in this chapter is a prominent factor which affects the number of keywords extracted by the ontology.

....❧....



## PROPOSED NEWS GENERATION FRAMEWORK

- 5.1 Introduction
- 5.2 Choosing Recurrent Neural Networks for news generation
- 5.3 Natural Text Generation Using Neural Network
- 5.4 Mathematics of Neural Network Based Language Model
- 5.5 Proposed News Generation Framework
- 5.6 Uniqueness of Proposed News Generation Framework
- 5.7 Conclusion

### 5.1 Introduction

The chapter explains the evolution of an experimental framework to generate news from keywords. The last chapter discussed the capability of ontology in extracting knowledge and the effectiveness of ontology in the field of information extraction. The study conducted in the last chapter is significant, considering that the inputs of the proposed news generation framework are keywords mined by ontology. The steps involved in text generation using Recurrent Neural Networks and mathematical foundation of RNN language modelling are described in the first part of this chapter. Then the details about the architecture used in the proposed framework and the procedure used in generating news are explained. The two neural net models used in the proposed framework are also mentioned along with the algorithms in this chapter. Finally, a conclusion is presented.

## 5.2 Choosing Recurrent Neural Networks for News Generation

A lot of researchers prefer Recurrent Neural Networks over Convolutional Neural Networks for language generation due to several reasons. The working of RNN is based on the processing of sequential information. That is, the output is based on previous results. So it is naturally suited for language generation due to this sequential nature. Words in a language are semantically inter-related and follow the same sequential behaviour.

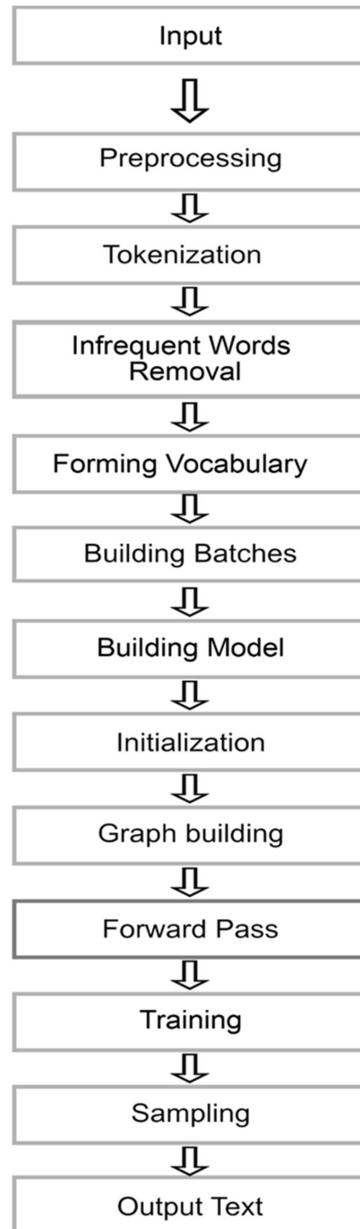
RNN can handle long sequences of variable length and the computational steps are flexible than in CNN [150]. CNN is used successfully in classification problems whereas RNN has proved its efficiency in language modelling. However, it is not proper to make a conclusion based on just the reasons mentioned above that RNN is better than CNN due to the fact that there are some good studies where CNN performs well in language modelling. Basically the success depends on how much global semantics is required for each problem.

According to Goodfellow [151], representation of text as a continuous space poses some issues in language generation using GAN. GAN is proven to be the state-of-the-art technology for image generation whose ability in language modelling is yet to be proved. Some of the studies about GAN are also emphasizing the need for improvement in language modelling [152].

RNN-LSTM is the current state-of-the-art-technology used for language generation. It is the most successful language model application developed so far that relies on the power of Recurrent Neural Networks. Many of the literature available in this domain refer to RNN as the proven tool capable for language modelling. Here, in this work, the neural net models suggested are based on RNN-LSTM considering the facts discussed in this section.

### **5.3 Natural Text Generation Using Neural Network**

Various learning algorithms can generate natural languages, yet the layers that are hidden make neural networks outperform these algorithms. The model gives crisp predictions and is likely to learn better with more information about the features and its complexity. But neural networks can't comprehend the sequence in which the current state is affected by its previous states. A clear solution was provided for this issue with the Recurrent Neural Networks. Each hidden layer of the Recurrent Neural Network is used for computing the corresponding input at that time step, the previous time step and output. Therefore, the Recurrent Neural Networks have an ability to remember data with hidden layers of different time steps. But vanishing gradient problem persisted which demanded an improvement. The Long Short Term Memory is a potential successor. LSTM has the capability to decide if to discard a previous idea or to keep it. It also spots the requirement to update the current state using the previous state. The text generation process is explained below in detail and illustrated through Figure 5.1.



**Figure 5.1: Text generation process**

As shown in the Figure 5.1, the text generation process starts with the pre-processing of data. Here TensorFlow is used for creating computational graph. Training is done using the input data and the model is generated. The step wise process is explained in detail below.

### **5.3.1 Pre-processing**

Pre-processing is the first and most important stage in text generation. The input text is made appropriate by pre-processing of the input data to be fed into RNN, which includes vocabulary and inverse vocabulary formation, eliminating infrequent words and tokenisation of text. Normally, source data for training will be available as raw text. It needs to be transformed based on the needs of the neural net model which is going to be used in the process. Removing commas, punctuations, non- alphabetic characters and normalizing all words to lower case are some of the cleaning task done in the pre-processing stage.

### **5.3.2 Tokenize Text**

A sequence of characters grouped together to make a useful semantic unit for processing is termed as a token. Since making predictions on each word is a necessity, requirement for the tokenisation of the text still persists. The text is divided into sentences and the sentences into words. There are methods available in the front-end language, Python for this purpose.

### 5.3.3 Removal of Infrequent / Stop words

Words which appear only once or twice have to be removed. Since the training depends on the vocabulary size, the removal of infrequent words would prevent the large size of vocabulary. The model wouldn't be able to learn from infrequent words since there are not much examples of those words, thereby they are removed.

### 5.3.4 Vocabulary and Inverse Vocabulary Formation

The deep learning models cannot understand natural languages and hence require mapping of words for model training purposes. Each word in the vocabulary has been mapped to the index based on the sentences. An Inverse Vocabulary is maintained in which an index to word mapping is stored.

### 5.3.5 Building batches

Batches are formed to process the number of parallel sentences in the network. The input in any feed-forward network is a matrix of shape  $[x*y]$  where  $x$  is batch size and  $y$  is feature Size. Batches are formed from the vocabulary which has been divided into sentences.

The first word in each of the sentence of the batch is processed in parallel, then second word of sentences of each batch and so on. In a batch, all sentences are handled in parallel yet the network views only one word of sentence at a time and computes accordingly.

### **5.3.6 Model Building**

The RNN is trained to build the language model after the input text has been pre-processed. The output text is produced by the language generation model after the input text has undergone the deep learning process. The parameters of the network are set up and the variables for training are initialised. Then the word probabilities are predicted by implementing the forward propagation, after which the loss is calculated. The text is generated finally after training RNN with stochastic gradient and back propagation through time.

### **5.3.7 Initialization**

Initialising the variables and setting network parameters is the first step for training RNN. The size of the hidden layers, number of hidden layers, batch size, number of epochs, learning rate etc. are some of the variables which have to be initialized in this phase.

### **5.3.8 Creating Computational graph**

The computational graph specifies the operations to be done and is created by TensorFlow. TensorFlow makes use of this graph to depict the connections between individual operations. These graphs are powerful tools but have a complex nature. The input and output of the graph are multidimensional arrays. Every time RNN runs, batch data is fed into placeholders. These placeholders are start nodes of the graph. The RNN state is also fed into the placeholder. This state is the output of the previous run.

### **5.3.9 Forward Pass**

The building part of the graph which does RNN computation is included in the forward pass. Basically forward pass deals with calculation process. One iteration includes both forward and backward passes. The popular term used for one iteration is an epoch. RNN is treated as a deep neural network with recurring weights in every layer.

### **5.3.10 Loss Calculation**

The model's progress after each optimization / iteration is known by calculating the loss. Minimizing the loss function with respect to parameters of the model using different optimization techniques like back-propagation is an objective in learning the model.

### **5.3.11 Running Training Session**

The dataflow graph runs in session using TensorFlow. Data is generated on each epoch. The time to execute an epoch depends on the complexity of the model. In this work, two types of models are experimented. Sequence to Sequence model is complex than Char-RNN model and takes more time to train.

### **5.3.12 Checkpoints**

If the model is stopped, there is a probability of losing all the work since it may take a couple of weeks for the deep learning model to acquire the required knowledge. The requirement of optimisation slows down the process of training the model. This issue is avoided with



model checkpointing in which the state of the system is maintained if a failure occurs. Each time an improvement in loss is observed at the end of the epoch, it records all of the network weights to a file. Numerous iterations control the frequency with which these checkpoints are written. While the model is trained, it will periodically write checkpoint files to the specified location. If anything happens wrong during the training period, the process can be restarted from the last checkpoint saved to the disk.

### **5.3.13 Sampling**

The RNN language generation model is developed after training. Sampling is done after completing the model. The first word is provided in the initial phase and the model is made to learn the dependencies between conditional probabilities of the words in the sequence of the text and the provided words so that the next word would be generated after the third word using the first two words. A specific number of words will be generated by it.

### **5.3.14 Output Generation**

The news text that is generated using the words in the given input news, after the sampling process is stored in output file. Two models and two datasets are used in the research work mentioned here. After obtaining the results, a comparison is done to find which RNN model in a specific dataset gives more realistic news.

## 5.4 Mathematics of Neural Network Based Language Model

Most of the details discussed here are referred from Ian Good fellow's book [137]. Statistical language model is applied in different natural language problems. Suppose there is a need to search a sequence of words  $w_1, \dots, w_m$  from a given data consisting of several news annotations. Searching is done recursively using probability measures. The probability  $P$  of  $w_1, \dots, w_m$  which computed recursively by the equation

$$P(w_1, \dots, w_m) = \prod_{i=1}^m P(w_i | w_1, \dots, w_{i-1}) \dots\dots\dots (5.1)$$

Where  $w_0$  stands for empty word.

When there are a number of words, it becomes difficult to compute  $P(w_i | w_1, \dots, w_{i-1})$ . In this situation, the probability of a word is limited on a set of  $n$  words ( $n$ -gram model). In this case, the approximate probability equation is

$$P(w_1, \dots, w_m) \approx \prod_{i=1}^m P(w_i | w_{i-n}, \dots, w_{i-1}) \dots\dots\dots (5.2)$$

$P(w_i | w_{i-n+1}, \dots, w_{i-1})$  is computed by counting the word sequence ( $w_{i-n+1}, \dots, w_{i-1}, w_i$ ) and the sequence of word ( $w_{i-n+1}, \dots, w_{i-1}$ ) in the given data. In other words,

$$P(w_i | w_{i-N+1}, \dots, w_{i-1}) = \frac{\text{count}(w_{i-1}, \dots, w_{i-1}, w_i)}{\text{count}(w_{i-N+1}, \dots, w_{i-1})} \dots\dots\dots (5.3)$$

This approach has the defect of producing zero probability for a sensible combination of words. For example, consider a 3-gram model “three boys sitting”.

$$P(\text{sitting} | \text{three boys}) = \frac{\text{count}(\text{sitting})}{\text{count}(\text{three boys})} \dots\dots\dots (5.4)$$

Naturally the counting ‘sitting three boys’ will be zero even though a collection of data is there. Therefore, the ratio also will be zero. This is one of the limitations of statistical language modelling [140].

#### **5.4.1 Neural language models**

The language model using neural networks was introduced after a research of many years and it was first proposed by Bengio et al. [141]. It was popularized after the introduction of recurrent neural network by Mikolov et al. [142] and replaced the standard n-gram technique. The evolution of Recurrent Neural Network Language Model is detailed in the subsequent sections.

##### **5.4.1.1 Artificial Neurons**

Artificial neurons are made of several components. If n input signals are given to a neuron, it gives a vector in  $\mathbf{x} \in \mathbb{R}^n$ .

Each input  $x_j$  to the  $k^{\text{th}}$  neuron is multiplied by a weight  $w^{(kj)}$ . They are called synaptic weights and represent connection between different

neurons. Bias  $b^{(k)}$  can be assigned to the  $k^{\text{th}}$  neuron. Using the weight and bias, the activation of neuron is expressed by the equation

$$s^{(k)} = b^{(k)} + \sum_{j=1}^n w^{(kj)} x^{(j)} \dots\dots\dots (5.5)$$

This can be compactly written in the form

$$\begin{aligned} s^{(k)} &= \sum_{j=0}^n w^{(kj)} x^{(j)} \dots\dots\dots (5.6) \\ &= \mathbf{w}^T \mathbf{x} \end{aligned}$$

Where  $x^{(0)} = 1$  and  $w^{(k0)} = b^{(k)}$

The output  $\hat{y}^{(k)}$  of  $k^{\text{th}}$  neuron is computed by an activation function  $\sigma$  acting on the activation  $s^{(k)}$ . The equation derived from the above data is

$$\hat{y}^{(k)} = \sigma(s^{(k)}) \dots\dots\dots (5.7)$$

#### 5.4.1.2 Activation functions

Activation functions are nonlinear functions which are used to assign probability measures for variables. The commonly used nonlinear functions in language models are sigmoid, tanh and softmax functions. Softmax and sigmoid functions are naturally arising functions in the analysis of exponential distributions [138].

The sigmoid function provides a real valued output. The sum of the probability need not be 1. The sigmoid score is calculated using the function given in Eq. (5.8).

$$f(x) = \text{sigmoid}(x) = \frac{1}{1+e^{-x}} \dots\dots\dots (5.8)$$

Figure 5.2 represents sigmoid function and the first derivative of sigmoid is non-negative and non-positive. It is often used in artificial neural networks to introduce the nonlinearity in the model. In the graph given below, the X axis represents the input  $x$  which are fed to the sigmoid function whereas Y axis represents the sigmoid function  $f(x)$ .

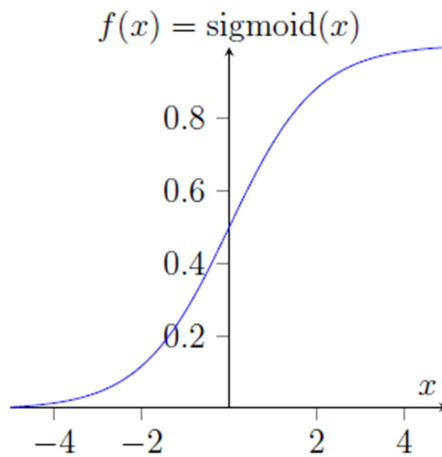


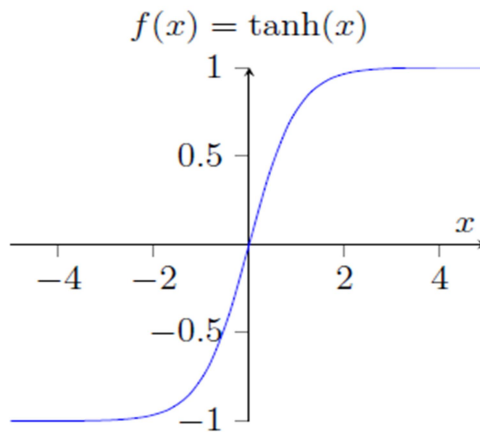
Figure 5.2: Sigmoid function

The graph takes a S-shape and sigmoid score increases with input value as shown in the Figure 5.2. The maximum value is 1. The sigmoid score ranges from 0.9 to 1 at the top of the graph. It can be used as activation function for building neural networks.

The tanh function is often referred as hyperbolic tangent function. Tanh function is calculated using the equation Eq. (5.9).

$$f(x) = \tanh(x) = \frac{1-e^{-2x}}{1+e^{-2x}} \dots\dots\dots (5.9)$$

Figure 5.3 represents tanh function and it is another version of logical sigmoid. The function is differentiable. The tanh function is mainly used in classification between two classes. Feed-forward nets use tanh as an activation function. The X axis of the graph is the input  $x$  and Y axis is the corresponding tanh function value.



**Figure 5.3: tanh function**

As in the Figure 5.3, the output ranges from -1 to +1 and it suits more for neural network in certain situations. The graph takes an S-shape as in the case of sigmoid function.

The softmax score is derived from the function given in Eq. (5.10).

$$\text{softmax}(x_i) = \frac{\exp(x_i)}{\sum_{j=1}^n \exp(x_j)} \dots\dots\dots (5.10)$$

Softmax function is portrayed in Figure 5.4 and is used in different layers in neural network building. The calculated probability will be in between 0 and 1 and sum of the probabilities is 1.

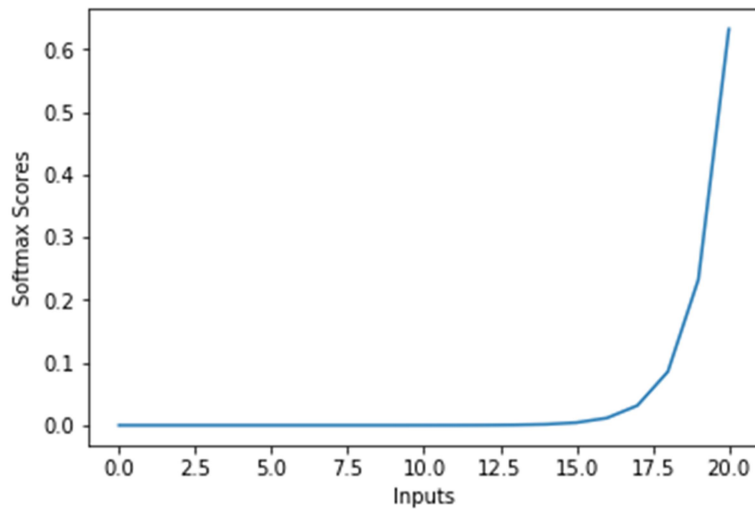


Figure 5.4: Softmax function [160]

The Figure 5.4 shows the fundamental property of softmax function. The X axis of the graph denotes the input while the Y axis provides corresponding softmax score. For high input value, the probability will be high.

### 5.4.1.3 Feed-Forward Neural Networks

Feed-forward neural networks have three components.

- 1) Input layer
- 2) A group of hidden layers
- 3) An output layer

If the feed forward neural network has one hidden layer, activation has to be done for the input layer which is represented by

$$f^{(1)}(\mathbf{x}) = \sigma^{(1)}(\mathbf{W}_{(1)}^{(T)} \mathbf{x}) \dots\dots\dots (5.11)$$

Now this activated output is the input for the hidden layer. The activation of the new input to the hidden layer is represented by

$$f^{(2)}(f^{(1)}(\mathbf{x})) = \sigma^{(2)}(\mathbf{W}_{(2)}^{(T)}(f^{(1)}(\mathbf{x}))) \dots\dots\dots (5.11)$$

Taking this as the input for the output layer, the output function can be calculated as below

$$f(\mathbf{x}) = f^{(3)}(f^{(2)}(f^{(1)}(\mathbf{x}))) = \sigma^{(3)}(\mathbf{W}_{(3)}^{(T)}(\sigma^{(2)}(\mathbf{W}_{(2)}^{(T)}(f^{(1)}(\mathbf{x})))) \dots (5.12)$$



#### 5.4.1.4 Recurrent Neural Networks

Feed-forward neural networks are modified into recurrent neural networks for better output. Figure 5.5 provides the graphical representation of Feed-forward networks and Recurrent Neural Networks. In Feed forward network, the flow of information from the input vector  $\mathbf{x}$  to the output vector  $\mathbf{y}$  goes in only one direction. That is, no cycles exist. In RNN, there are cyclic connections also.

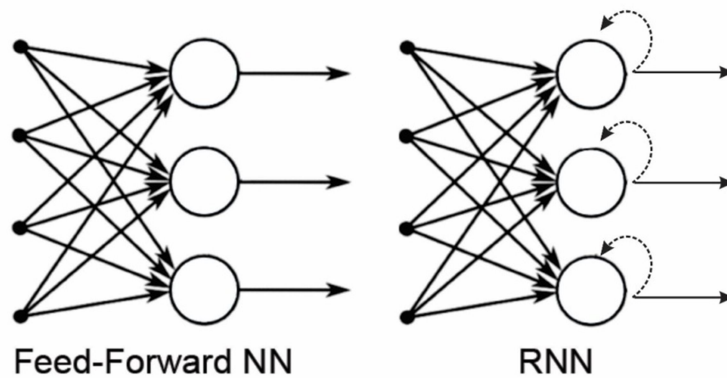


Figure 5.5: RNN and Feed-Forward neural network

As represented in Figure 5.5, RNN has cycles whereas Feed-forward network does not have cycles. The cycles make RNN to remember context.

In this case, the computation of the input of  $t^{\text{th}}$  hidden layer is done by using the output of the  $t-1$  hidden layer, the input vector  $\mathbf{x}^{(t)}$  and some parameter  $\theta$ .

This is described by the equation

$$\mathbf{h}^{(t)} = g(\mathbf{h}^{(t-1)}, \mathbf{x}^{(t)}, \theta) \dots\dots\dots (5.13)$$

An example of a Recurrent Neural Network Language model is given in the Figure 5.6. RNN maps an input sequence  $\mathbf{x}$  to an output sequence of  $\mathbf{o}$ . The loss  $L$  is measured with the training target.

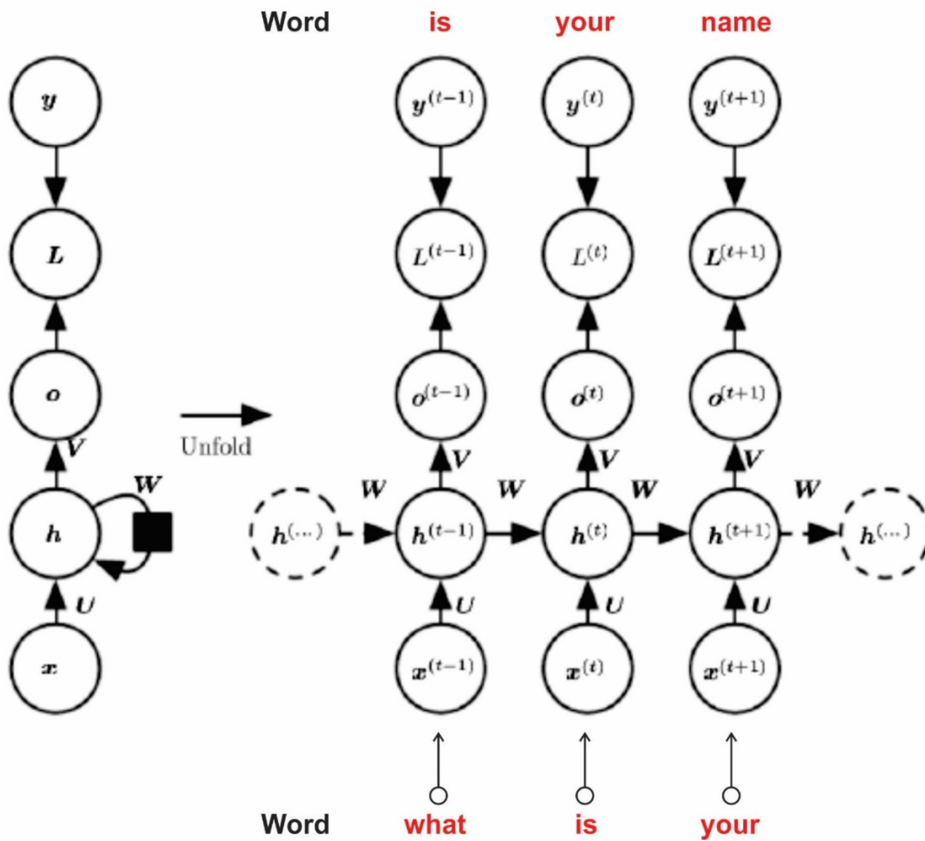


Figure 5.6: Recurrent Neural Network Language Model [137]

In the example mentioned above in the Figure 5.6, the target sequence is provided by the input sequence shifted one-time step to the left. The input to the hidden layer is parametrized by a weight matrix  $\mathbf{U}$ , the connection between hidden layers is parametrized by weight matrix  $\mathbf{W}$  and the connection between hidden layer to output layer is parametrized by a weight matrix  $\mathbf{V}$ .

Then, for an input vector sequence  $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(\tau)}$ , the activations of the hidden units from 1 to  $\tau$  are measured by the formula

$$\mathbf{h}^{(t)} = \sigma_h(\mathbf{b} + \mathbf{W}\mathbf{h}^{(t-1)} + \mathbf{U}\mathbf{x}^{(t)}) \dots\dots\dots (5.14)$$

Where  $\mathbf{b}$  denotes a bias and  $\sigma_h$  denotes the element wise activation function in the hidden layer.  $\mathbf{h}^{(0)}$  denotes initial hidden state. Now the output is computed using the formula

$$\hat{\mathbf{y}}^{(t)} = \sigma_y(\mathbf{c} + \mathbf{V}\mathbf{h}^{(t)}) \dots\dots\dots (5.15)$$

Where  $\mathbf{c}$  denotes bias and  $\sigma_y$  represents the activation function of the output layer

The loss is calculated as negative log likely wood of  $\mathbf{y}^{(t)}$  given the input sequence  $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(\tau)}$

If  $L^{(t)} = -\log P_{\text{model}}(\mathbf{y}^{(t)} | \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(\tau)})$ ,

The loss function can be written as

$$L(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(\tau)}, \mathbf{y}^{(1)}, \dots, \mathbf{y}^{(\tau)}) = \sum_1^{\tau} L^{(t)} \dots\dots\dots (5.16)$$

### 5.4.1.5 LSTM

RNN's are further modified by introducing input, forget and output gates together with a special type of cell called memory cell to eliminate the long term dependency problem.

The input gate controls flow of information into the cell and decides whether it is worth to preserve or not. The forget gate decides how useful is the information in the memory cell before feeding it back to the self-feedback loop. That is, it forgets and resets the information in the memory cell.

The output gate decides which part of the memory state is to be carried out to the hidden state and hence control the output flow of the network.

LSTM model is defined by the following equations.

$$g_i^{(t)} = \sigma(b_i^g + \sum_j u_{i,j}^g x_j^{(t)} + \sum_j W_{i,j}^g h_j^{(t-1)}) \dots\dots\dots (5.17)$$

$$f_i^{(t)} = \sigma(b_i^f + \sum_j u_{i,j}^f x_j^{(t)} + \sum_j W_{i,j}^f h_j^{(t-1)}) \dots\dots\dots (5.18)$$

$$s_i^{(t)} = f_i^{(t)} s_i^{(t-1)} + g_i^{(t)} \sigma(b_i + \sum_j u_{i,j} x_j^{(t)} + \sum_j W_{i,j} h_j^{(t-1)}) \dots (5.19)$$

$$q_i^{(t)} = \sigma(b_i^o + \sum_j u_{i,j}^o x_j^{(t)} + \sum_j W_{i,j}^o h_j^{(t-1)}) \dots\dots\dots (5.20)$$

$$h_i^{(t)} = \tanh(s_i^{(t)}) q_i^{(t)} \dots\dots\dots (5.21)$$

Where  $g_i^{(t)}$ ,  $f_i^{(t)}$ ,  $s_i^{(t)}$ ,  $q_i^{(t)}$  are input gate, forget gate, internal state and the output gate. The input vector is given by  $x^{(t)}$  and the current hidden vector by  $h^{(t)}$ .  $b$ ,  $U$ ,  $W$  are biases, input weights and recurrent weights respectively for different gates.  $\sigma$  denotes the logistic sigmoid function [153].

#### 5.4.1.6 Encoder-Decoder model

Figure 5.7 represents Sequence to Sequence model. Sequence to Sequence model is effective when the input and output sequences are of variable length. This model is also referred as Encoder-Decoder network.

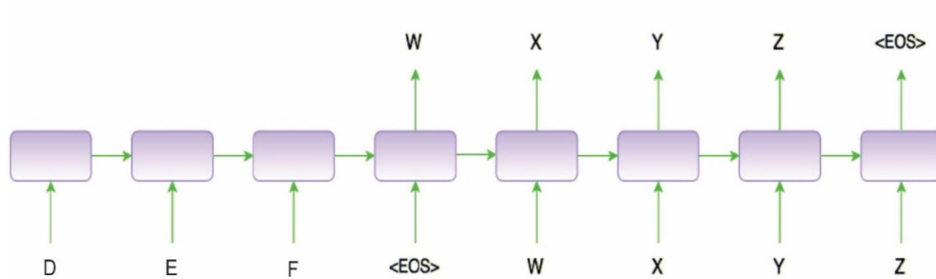


Figure 5.7: Sequence to Sequence model

The Sequence to Sequence model described in Figure 5.7 uses two RNN. The input and output sequences are represented using two different RNN. This model is trained to maximize the probability of an output according to the given problem. Probabilistic representation of this problem can be stated as

$$\max_y p(\mathbf{y}|\mathbf{x}) \dots\dots\dots (5.22)$$

$\mathbf{y}$  is the output sequence whereas  $\mathbf{x}$  is the input sequence.

The encoder RNN reads the input sequence  $\mathbf{x}$  and it reaches various hidden states. The last hidden state is denoted by  $\mathbf{c}$ . This vector is called summary vector. This summary vector is the input to the decoder RNN. The decoder RNN produces output using previous hidden states, output and also using the summary vector  $\mathbf{c}$ . The hidden states of decoder RNN can be represented by the equation

$$s_t = f(s_{t-1}, y_{t-1}, c_t) \dots\dots\dots (5.23)$$

The encoder- decoder  $p(\mathbf{y}|\mathbf{x})$  is given by

$$p(\mathbf{y}|\mathbf{x}) = \prod_{t=1}^T p(y_t|y_{t-1}, \dots, y_1, \mathbf{x}) \dots\dots\dots (5.24)$$

The conditional probability of the next output in the encoder-decoder model is given by

$$p(y_t|y_{t-1}, \dots, y_1, \mathbf{x}) = g(s_t, y_{t-1}, c_t) \dots\dots\dots (5.25)$$

Where  $g$  is the softmax function [138].

### 5.5 Proposed News Generation Framework

The main objective of this research work is to develop a framework which generates news from keywords. The proposed news generation framework is represented in Figure 5.8. The framework discussed here follows a unique process. Ontology is used for generating

keywords whereas RNN-LSTM based neural net models are used for generating news from keywords.

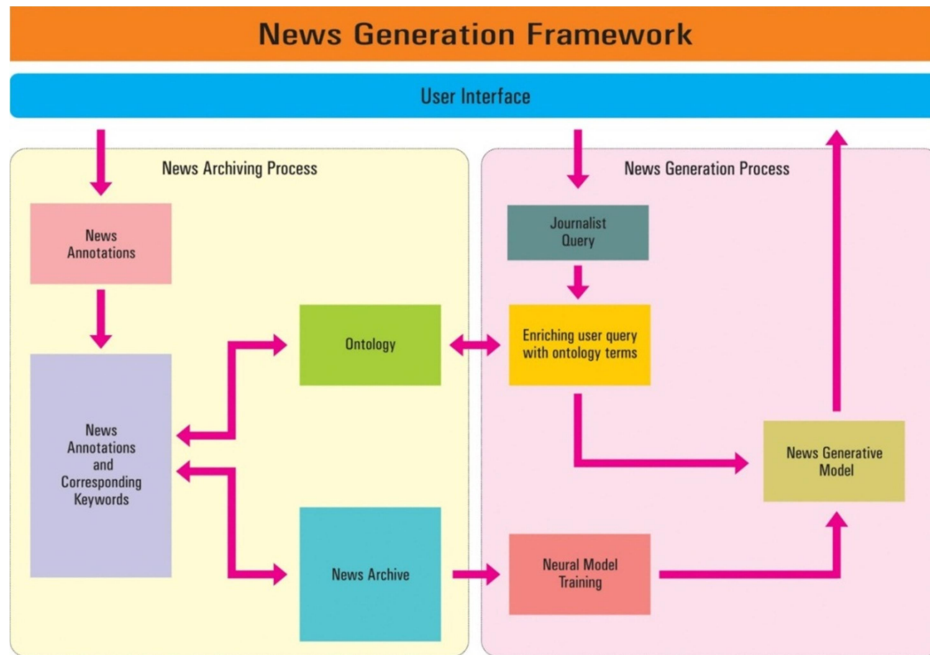


Figure 5.8: News Generation Framework based on Ontology

As discussed in the literature review, ontology is a proven mechanism in representing concepts and its complex relationships. So it is widely used in information extraction scenario. RNN is the most popular neural networks in text generation. So the proposed framework incorporates the power of ontology and Recurrent Neural Networks in news generation using keywords. Main processes in the proposed news generation framework as shown in Figure 5.8 are explained below.

- **Archiving Steps**

- 1) News annotations and corresponding keywords mined by ontology are stored in an archive

- **Neural Net Model Training Steps**

- 1) The news archive is searched and the text annotations of the news videos are extracted.
- 2) Concept terms / tags / keywords of each news annotation are extracted using ontology.
- 3) Keywords and corresponding news annotations are provided as input for neural network model training.

- **News Generation Steps**

- 1) Inputs are keywords. Keywords are mapped with ontology to get concept terms before entering into the trained model as input.
- 2) Using the trained model, News is generated.

Char-RNN and Sequence to Sequence model are the two neural net models experimented in this work.

The Sequence to Sequence language model mentioned in the literature [33,156] is a source of inspiration for many applications in language translation field [155]. The Sequence to Sequence model experimented in this research is adopted from above mentioned works



with some changes. The purpose of using Sequence to Sequence model in this study is to generate news from keywords. The referred works are basically translation models. So here in training phase, instead of having text in one natural language as input sequence in the translation model, the input sequence is keywords derived by ontology. The output sequence consists of news annotations. For testing purpose, keywords are used to generate news text. Hyper parameters are tuned according to the needs of news generation framework and the dataset used.

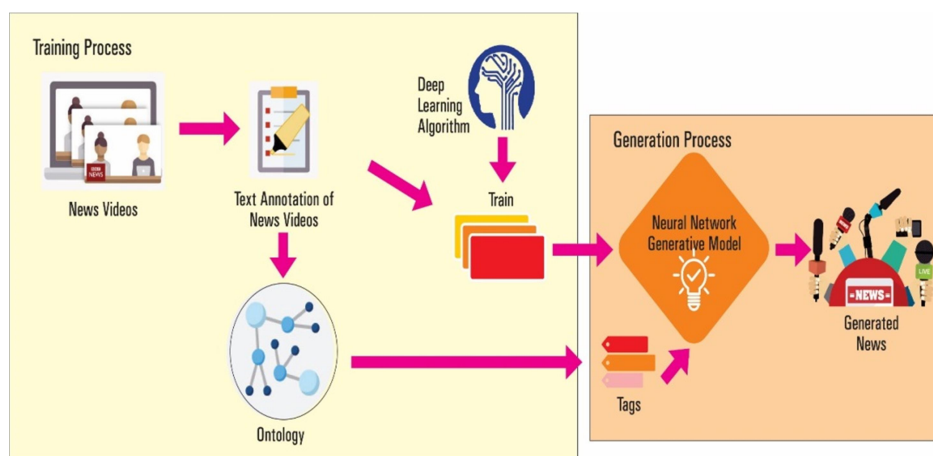
There is plenty of literature [55,139] available for generating text using Recurrent Neural Networks. Many text generation applications are developed based on it including the implementation of Andrej Karpathy [157,158]. The Char-RNN model used in this work follows the same procedure except in the testing phase. For news text generation, keywords are given as input. Hyper parameter tuning is done based on framework requirements.

The detailed explanation of each neural net model is given in the following sections.

### **5.5.1 Vanilla Char-RNN-LSTM Model**

A language model is a type of a machine learning algorithm which learns the statistical structure of a language and reproduces parts of language by predicting the next character based on previous characters.

The architecture diagram of Char-RNN model is illustrated in Figure 5.9. Only news annotations are provided for model training and keywords are the input for neural news generation.



**Figure 5.9: Char-RNN based model**

As shown in the Figure 5.9, two main processes are there namely training and text generation. The trained model generates new text one character at a time. The detailed working of Char-RNN network is shown in Figure 5.10 below. Here LSTM is used for character-level language model training.

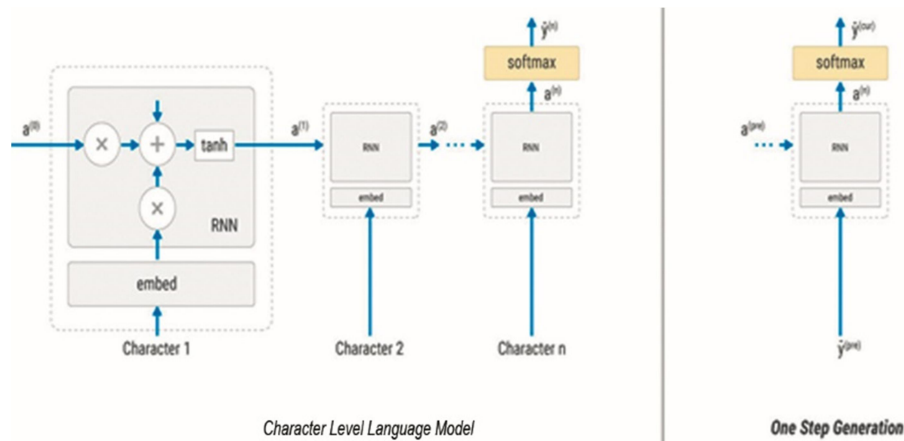


Figure 5.10: Char-RNN

Char-RNN learns to predict the next character in a sequence. A huge text data is given as input here and tries to model the probability distribution of the next character in the sequence given a sequence of previous characters. The detailed working procedure of Char-RNN shown in Figure 5.10 is given below

- **Procedure**

- 1) Unique characters are extracted from the news annotations.
- 2) These characters are then mapped to an arbitrary index value / numeric value thereby providing a better understanding for the machine.
- 3) The conversion from characters to numeric and from numeric to original characters are done using two dictionaries
- 4) Optimization is done for each character and the gradient loss is computed.

- 5) Using the computed gradient, training of the model is done by updating the parameters.
- 6) Ontology concept terms are the input sequence for testing. Each character from every inputted keywords is taken and N characters of news is generated

Algorithm 1 provides training and testing procedure of Char-RNN model.

***Algorithm 1 Vanilla RNN-LSTM model***

*Training Steps:*

*Input: Input news annotations  $\mathbf{x}$*

*Output: Trained Vanilla Char-RNN-LSTM Model.*

***Start Procedure***

***For several epochs of training do***

***For each character  $c_i$  in  $\mathbf{x}$  do***

*Run encoding on  $c_i$*

*Run one step of NN optimization and Compute gradients of the loss*

*Update the parameters according to this gradient*

***End for***

***End for***

***End Procedure***

*Testing Steps:*

*Input: Input Ontology Tags  $\mathbf{x}$*

*Output: News generated*

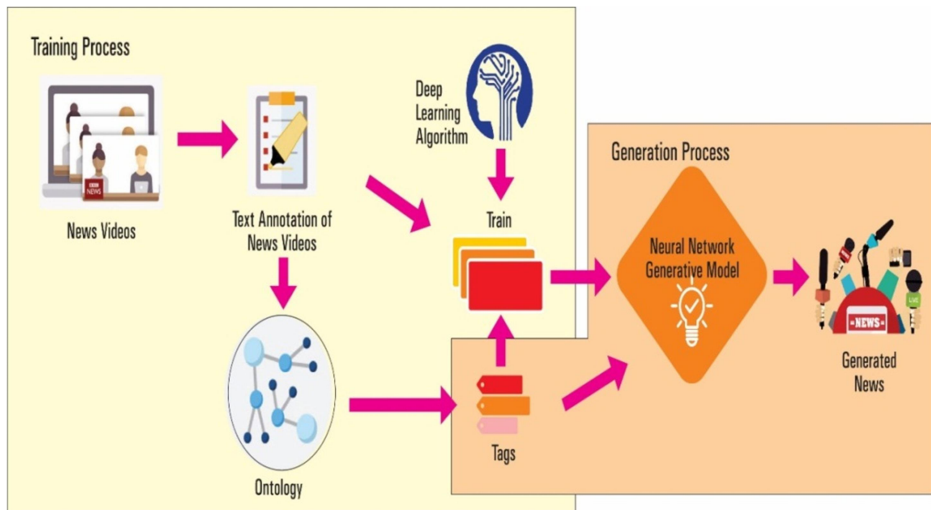
***For each tag  $ta_i$  in  $\mathbf{x}$  do***

*Generate the next  $n$  characters of news using the trained model*

***End for***

### 5.5.2 Sequence to Sequence (Seq2Seq) Model

In a high-level view, a seq2seq model has an encoder, decoder and intermediate step as its main components. The Seq2seq model solves a specific limitation of a deep neural network. This particular model allows input and output sequence of variable length. Figure 5.11 represents Seq2Seq model for news generation. Both news annotations and ontology keywords are employed in model training. Ontology tags are the input for news generation.



**Figure 5.11: Seq2Seq based model**

The training and generation processes are the two main processes in Seq2Seq model as shown in the Figure 5.11. Figure 5.12 describes the encoder-decoder network. Ontology tags are used as input sequence and the news annotations are output sequences.

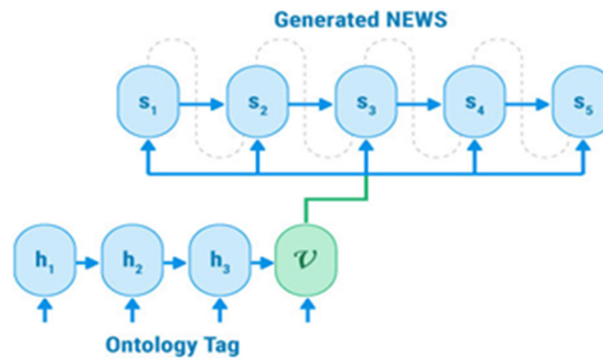


Figure 5.12: Encoder-Decoder network

The encoder reads the input sequence, word by word and emits a context which provides the essence of the input sequence. Based on this context, the decoder generates the output sequence, one word at a time while looking at the context and the previous word during each time step. The detailed working of Encoder-Decoder network shown in Figure 5.12 is given below.

▪ **Procedure**

- 1) Data preprocessing is executed by mapping input tags into arbitrary numeric values. 'UNK' signifies the unknown words in the news annotation which are excluded in input vocabulary. Zero padding is done to make the sequences of same length and unique tags are included in the vocabulary.
- 2) The LSTM layer receives the fixed size vector from the input sequence generated by the encoder.

- 3) The parameters are updated with the gradient loss which is calculated.
- 4) The decoder generates the output sequence from vector outputted by encoder.
- 5) The input is tested using ontology tags. The news is generated word by word using the trained model.

The algorithm 2 given below explains the training and testing steps.

**Algorithm 2 seq2seq model**

*Training Steps:*

*Input: Input ontology terms  $\mathbf{x}$ , and news annotations  $\mathbf{y}$*

*Output: Trained seq2seq model.*

**Start Procedure**

**For** batch of input and output sequences  $\mathbf{x}$  and  $\mathbf{y}$  **do**

*Run encoding on  $\mathbf{x}$  and get the last encoder state  $\mathbf{a}_n$*

*Run decoding by feeding  $\mathbf{a}_n$  to decoder and get the sampled output sequence*

*Calculate the loss and update the parameters*

**end for**

**End Procedure**

*Testing Steps:*

*Input: Input Ontology Tags  $\mathbf{x}$*

*Output: News generated*

**For** batch of input sequence  $\mathbf{x}$  **do**

*Using the trained model, the decoder generates news word by word for the given input sequence.*

**end for**

## 5.6 Uniqueness of Proposed News Generation Framework

The proposed framework is distinct in terms of process steps followed in generating news. Ontology is used for generating keywords from news while RNN-LSTM is used for generating news from keywords. There is no direct relation between ontology and neural networks. Both these technology is used in the framework developed as part of the present study. The reason for using ontology in keyword generation is that it is best suited to grasp knowledge in a specific domain. Ontology tags / keywords are used for both training and testing the neural net model.

## 5.7 Conclusion

Ontology based news generation framework using neural networks was proposed in this chapter. The basic concept behind the framework was the generation of news from keywords. At the beginning of the chapter text generation process in general was detailed. Then the mathematical background of Recurrent Neural Network Language Model was discussed. The two neural net models experimented here includes Sequence to Sequence model and Vanilla Char-RNN model. Sequence to Sequence model used ontology tags / keywords for training and generating news. Char-RNN model used keywords for news generation purpose only. The evaluation approach to test the proposed framework discussed in this chapter is described in the next chapter.

.....✂.....



## DESIGN OF EVALUATION APPROACH AND APPLICATION DEVELOPMENT ENVIRONMENT

- 6.1 Introduction
- 6.2 Evaluation approach
- 6.3 Datasets Used
- 6.4 Application Development Environment
- 6.5 Conclusion

### 6.1 Introduction

The chapter explains the evaluation approach followed in this research work. The proposed framework which is discussed in the previous chapter is analysed through the evaluation procedure presented here. The details of the datasets used for training and testing the application are given in this chapter. It also mentions the detailed information about the technical environment where the application was developed.

### 6.2 Evaluation Approach

A language model can most efficiently be tested by applying it on a designated environment to evaluate how the model improves the overall experience. The practical side of it is a bit strenuous when the cost and dependency on the application is taken into account. There is a

requirement of a metric which is independent of the environment. The evaluation methods used in this work are automatic evaluation and human evaluation.

### **6.2.1 Automatic Evaluation**

The desirable properties of automatic evaluation metrics include affordable cost, fastness and repeatable property. It plays a big role in areas of natural language processing. Two commonly used metrics are the Bilingual Evaluation Understudy (BLEU) score [74,75] and the Recall Oriented Understudy for Gisting Evaluation (ROUGE) [76,77] score. Both differ in terms of precision and recall. While precision based BLEU measures the quantity of generated text that appears in the reference text, recall based ROUGE measures how much of the reference text appears in the generated text. BLEU and ROUGE as well as the limitations of automatic evaluations are discussed in the following sections.

#### **6.2.1.1 Bilingual Evaluation Understudy (BLEU)**

The main objective of BLEU was to decrease the time and human effort in the evaluation of translations. It was first used for evaluating machine translation. It is inexpensive, quick, language independent and has a high correlation with human judgments. Though BLEU calculates the N-gram overlap between the reference sentence and hypothesis, and consider penalty for each short sentence, it does not take into account the paraphrasing or synonymy. The output of BLEU varies from 0 to 1,

with the similarity of candidate text with the reference text increasing with value closer to 1. The value 1 indicates a similar candidate text to the reference text, which is rare in human translations. Zero corresponds to a non-existing overlap.

#### **6.2.1.2 Recall-Oriented Understudy for Gisting Evaluation (ROUGE)**

ROUGE compares the generated text or its translation with the references based on the N gram. There are different variations of this metric ROUGE-N, ROUGE-L, ROUGE-W, and ROUGE-S. ROUGE-N computes the overlap of N-grams. For instance, ROUGE-2 calculates the overlap of bigrams between the hypothesis and the references. ROUGE-L – measures longest matching sequence of words using LCS. LCS does not require consecutive matches (which is an advantage), but in-sequence matches that reflect sentence level word order. As for the BLEU metric, a ROUGE score of 1 corresponds to a perfect match with the reference sequence, whereas a score of zero means no overlap at all.

#### **6.2.1.3 Limitations of Automatic Evaluation**

The task of computationally evaluating generative models is challenging unlike classification and information retrieval systems. These generation models create the possibility of numerous answers being accepted. There is a chance that the generated text can differ from the reference text. Evaluation metrics like recall or accuracy or precision cannot be easily applied in generative models. The complex nature of

the natural language makes automatic evaluation techniques incapable of checking the quality of the text. Automatic evaluation has the tendency to replicate the text without taking into consideration certain human centric factors like grammatical rules, lexical quality, spelling, grammar etc.

Automatic evaluation metrics are not able to analyse synonyms since both measures overlap the tokens between the generated sequence and reference. The semantic meaning of the source text can be generated in the machine text with different words, but it will be judged with a poor score of BLEU / ROUGE. Another major problem is with distinguishing outputs of good and medium quality text generated.

### **6.2.2 Human Evaluation**

Since the end user of the system is a human, there is a necessity for some human interaction in both providing inputs and in evaluating the results generated by the application. In the human evaluation process, human interaction is used in two ways, direct interaction with the application and indirect interaction with the application.

In the evaluation process, two datasets were taken namely a Cricket Commentary Dataset and BBC News Dataset. Both these datasets were provided to the two models subjected for evaluation which are, Char-RNN model and Sequence to Sequence model. Thereby there were four cases i.e. Char-RNN model with Cricket Commentary dataset, Char-RNN model with BBC news dataset, sequence to

sequence model with Cricket Commentary dataset and sequence to sequence model with BBC news dataset. Many evaluators were employed as a part of human evaluation to cross check the results. Of the numerous features, the main features assessed in this evaluation process would be fluency, adequacy and total quality. Selection of human evaluators, evaluation criteria, methods and limitations of human evaluation is discussed below.

#### **6.2.2.1 Team Selection**

The human evaluation team is selected based on: -

- a) **Age:** Age is a decisive factor that affects the evaluation process. Evaluators were selected from different age groups to receive a better sample. There is a chance of a specific age group being influenced by similar ideologies or thoughts which can make a discrepancy in the evaluation results. Thereby it was necessary to select people of different ages. 20 evaluators were from the age group of 20 to 30, 20 from 30 to 40 age group and 5 from age group 50 to 60.
- b) **Gender:** Male and Female psychology are marked by significant differences. There is a chance for the results to be narrow if a specified gender is selected. Thereby 25 male evaluators and 20 female evaluators were selected.
- c) **Computer proficiency:** This is a prime factor while dealing with the evaluation process. An evaluator who is not an expert in computer might come on to his / her own conclusions. There is a

possibility that such evaluations can be marred by personal judgments and the scores might be less than the actual deserved result, whereas a computer expert would know the real ability of the networks and would correctly analyse and give apt scores. Moreover, a computer expert would be comfortable in direct interaction with the application than a non-expert. However, to obtain a broader range of evaluation, people from both categories were selected.

- d) Language proficiency (basic, expert): There are two categories while dealing with language proficiency. Evaluators with a basic level of English might see the evaluation scores efficient when compared to their skills whereas an English language expert might take stringent measures of scrutiny and critically evaluate the generated text which would give a serious variation in their scores when compared with basic level evaluators. Anyhow, people from both categories were selected and the average score of their evaluation results was considered while forming the conclusion.

Assessment consists of evaluating two news text generation application models by using two different datasets.

- 1) Char-RNN Model
  - a) Using Cricket Commentary dataset
  - b) Using BBC news dataset

- 2) Sequence to Sequence Model
  - a) Using Cricket Commentary dataset
  - b) Using BBC news dataset

#### **6.2.2.2 Evaluation Criteria**

When evaluation of the two models is taken, the main features that were analysed were adequacy, fluency and total quality.

Adequacy is the judgment on how well the translation of a phrase transmits the meaning of the original, and specifically it checks if the content of the source sentence has been preserved and the sense has not been distorted.

Fluency is a judgment on how well-formed and fluent the translation is; it does not scrutinize the information contained in a phrase, rather its overall grammaticalness and correctness.

Total Quality is the extent to which the document maintains the expected quality.

The evaluators were provided with a questionnaire with a score level from 0 to 10 except in the judgment of the source of news (human or computer). Yes, or no question was included in the questionnaire to judge news source and these documents were used to determine the final result. Appendix 2 provides the format of the questionnaire.

### 6.2.2.3 Evaluation Method

Two evaluation methods are chosen in this human evaluation. One is direct interaction with the application framework and another is the indirect interaction. The details are given below.

- Direct interaction with application

As the initial step to test in real environment, the application is provided with keywords as inputs and the output is generated. The output is then evaluated using the referral text. The application would be operated by the evaluators. Their main work is to evaluate the quality of the result that is provided by the application. In the method of direct interaction, the computer experts may find it easy to adapt to the system.

- Indirect interaction with the application

In the method of indirect interaction, the referral news is compared with automatically generated text and the human evaluators come to a conclusion. The referral and the generated text for evaluation would be provided to the evaluator in an excel sheet for convenience. An interaction with the machine is not compulsory in indirect interaction. In this case a non-expert who is not familiar with the technicalities of computers would find the method comfortable.



#### **6.2.2.4 Limitations of Human Evaluation Study**

One prominent feature of human evaluation is that it was highly subjective. Human nature is essentially dynamic therefore there was a chance that the judgment made by the same person at different time periods could be different. Though people from various age groups and genders were selected, it was not possible to strictly take into account all those traits. Environment plays an important role in shaping perceptions of the people and thereby the whole evaluation process got affected by it. A person working in a thought provoking and peaceful environment can bring out better results than a person working in a noisy environment. The health condition and the psychological moods of the person would also affect the judgments. Different users from different backgrounds would have different judgments. Different people might be familiar with different usages of expressions / slangs and pronunciation which can affect the quality of the analysis.

When a language is considered, proficiency is a huge factor. Basic level speakers might be using the language as a mere language of secondary communication whereas language experts have a tendency to scrutinize it critically with their level of knowledge. The experts may give lower scores whereas basic level speakers may give a comparatively higher level of scores. Another limitation is with variation in the results with computer expertise. An issue that come up in human evaluation is the use of punctuation marks. While dealing with language proficiency many evaluators consider punctuations as a decisive point to evaluate

the results. As humans, each person would have his / her own version of the evaluation criteria which can affect the end result. These are the important limitations in human evaluation.

### 6.2.3 Evaluation Procedure Followed in This Work

While automatic evaluation techniques like BLEU and ROGUE have proved its efficiency in checking the similarity between texts (like translation), there are significant drawbacks while considering text generation. But while considering the unreliability of the method and the features that the specified end user needs, there arises the necessity for a human centric approach. But human evaluation is also highly subjective. It might be appropriate to analyse the linguistic quality of the generated texts using existing automatic metrics, especially if manual evaluations are supplemented by metric evaluations. Figure 6.1 represents the evaluation approach followed in this study.

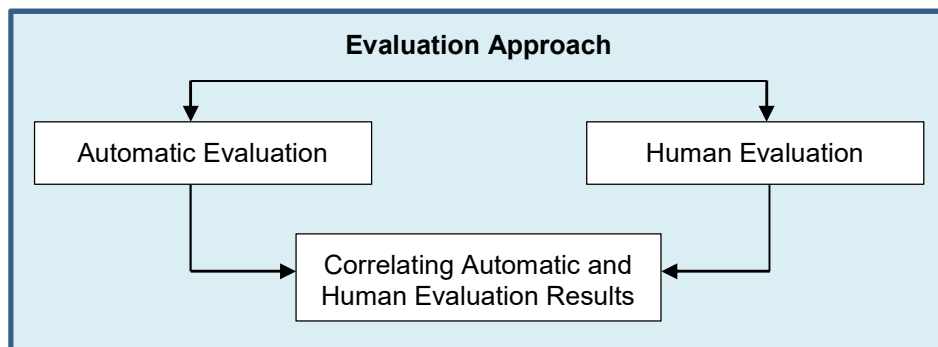


Figure 6.1: Evaluation Approach

As shown in Figure 6.1, a combined approach is used for evaluation here. The evaluation of results by correlation of automatic metrics with human ratings is the approach followed in this work. Both automatic and human evaluation results would be collected and analysed to conclude which system is more efficient in generating news.

### **6.3 Datasets Used**

Two datasets are used to evaluate the proposed framework.

BBC news dataset was used to test and train the model which consisted of 2225 raw news. The news belonged to different categories like politics, sports, entertainment, business and technology which were collected in the year 2004-2005 [78]. Two thirds of this dataset was used to train models whereas the rest was used for testing purpose. The extraction of keywords from each news text in the BBC news dataset is done by Open Calais ontology. The news annotations along with corresponding concepts / tags are stored as CSV file for training and testing the neural net model for news story generation [161].

Cricket Commentary dataset was created by scraping commentaries of all the matches in IPL (Indian Premier League) 2016 from Cricbuzz website [79]. Since Open Calais Ontology was not able to extract all the key terms in the cricket domain, a cricket taxonomy was made by referring different sources including MCC cricket laws [159]. Custom made taxonomy along with Open Calais ontology was employed for mining key terms from IPL 2016 Cricket commentaries. The

commentaries were split into small texts based on overs and text length. Pre-processing of Cricket Commentary dataset was a time consuming job. Commentators often digressed from Cricket. They move to discuss on environment, history, venue details, food habits, tourism spots etc. Sometimes, pep talks are also a part of the commentary. These were irrelevant details and sometimes it could mislead the machine learning process. Avoiding or erasing it from the dataset was a strenuous job and it took several days of manual work to remove these details and clean the dataset. The split commentary text along with the corresponding ontology tags was stored in a CSV file for model training and testing. The dataset was uploaded to kaggle for reference [162].

## **6.4 Application Development Environment**

The application was developed using Amazon EC2 P2 instance. AWS provides a virtual, ready-to-use deep learning environment. AWS Deep Learning AMI (Ubuntu 16.04) is used for application development in this work. The system was developed in Keras framework using Python as front-end language and TensorFlow as backend. The technical environment used for application development will be discussed in the following sections.

### **6.4.1 Aws p2. xlarge Instance**

Amazon EC2 P2 Instances are great, versatile examples that give GPU-based parallel register abilities [131]. P2 is a broadly useful GPU application which utilizing CUDA and OpenCL, is a perfect world suited

for machine adapting, elite databases, computational liquid elements, computational finance, seismic examination, atomic model, genomics, rendering, and other server-side outstanding burdens requiring gigantic, parallel, floating point processing power.

Amazon Linux AMI with pre-introduced Deep Learning systems, for example, Caffe, Mxnet, NVIDIA AMI with GPU driver and CUDA toolbox were used for quick on boarding.

- **P2 Features**

P2 instances deliver up to 16 NVIDIA K80 GPUs, 64 vCPUs and 732 GB of host memory, with a joined 192 GB of GPU memory, 40 thousand parallel processing cores, 70 teraflops of single accuracy floating point execution, and more than 23 teraflops of floating point performance. Group P2 instances in a scale-out form with Amazon EC2 ENA-based Enhanced Networking can run top performance, low-latency computing grid. P2 is appropriate for Deep Learning systems, for example, MXNet, that scale out with close perfect proficiency [130].

- **P2 Instance Details**

P2 uses Intel Broadwell Xeon's CPU and is capable for machine learning applications which require huge computational power. The P2's support NVIDIA GK210 GPU's and its latest advancement P3 run on Tesla V100 which can handle very huge machine learning operations. Table 7.1 summarizes the P2 instance capability details.

**Table 6.1: P2 instance details**

<b>Instance Name</b>	<b>GPU's</b>	<b>vCPUs</b>	<b>Ram (GB)</b>	<b>Bandwidth</b>
P2.xlarge	1	4	61	High
P2.8xlarge	8	32	488	10Gbps
P2.16xlarge	16	64	732	20Gbps

The above table provides P2 instance details and the instances used for application development in the present study are P2.xlarge and P2.8xlarge.

#### **6.4.2 TensorFlow**

Machine learning is a complex discipline. In any case, executing machine learning models is far less crushing and troublesome than it used to be, on account of machine learning systems, for example, Google's TensorFlow - that facilitates obtaining data, preparing models, serving forecasts, and refining future outcomes [128,132].

Made by the Google Brain group, TensorFlow is an open source library for numerical calculation and huge scale machine learning. TensorFlow brings together a huge number of machine learning and deep learning models, making it very valuable.

- **How TensorFlow functions**

TensorFlow enables to make dataflow graphs - structures that depict how data travels through a diagram, or a progression of handling hubs. Every hub in the chart speaks to a numerical task,

and every association or edge between hubs is a multidimensional Data exhibit, or tensor.

Python can be learnt easily and worked with, which gives advantageous approaches to express how abnormal state deliberations can be coupled together. TensorFlow uses both Python and CPP languages. The real math activities are not performed in Python. The libraries of changes that are accessible through TensorFlow are composed as elite CPP pairs.

- **TensorFlow advantages**

The greatest advantage TensorFlow accommodates is its support for machine learning, deep learning and reinforcement learning. It offers TensorBoard for visualization. It can be used on multiple CPU's and GPU's. Google made this software as open source considering its importance in the machine learning domain.

### **6.4.3 Python**

Python language is modeled as simple to learn and run anywhere. It is valuable for various applications, including machine learning applications. Popular IT organizations depend on Python broadly, including Instagram and Google.

It is a dynamic programming language like Java, and used as a general purpose language as well as used in several application domains. It gives solid support to integrate new programming innovations. It is suited for substantial and complex undertakings with evolving necessities.

Python is one of the quickest developing open source programming languages. It additionally frames the base for different top applications and is utilized crosswise over various ventures [127].

- **Easy to use**

The major highlight of Python programming language is it's easy to use format and readability. Due to this ability, the code can be easily maintained. Both the experts and beginners would find this language comfortable.

- **Direct and rapid**

The Python community offers quick and effective help to clients, and a huge number of engineers endeavor to discover and solve bugs and to give upgrades to the programming language. The community offers quick input from multiple points and Python provides easy adaptation of code. Python additionally allows quick adjustment of code. This programming language can be named as ready-to-run, requiring simply direct code to be executed. Testing and playing around the code turns out to be considerably easier with Python.

- **Ease of use with IoT**

The Internet of Things or IoT has opened up colossal opportunities, and Python can assume a key job in using these opportunities. The programming language is shaping into a prominent factor for IoT, with new stages.



- **Asynchronous coding**

Python has been a compelling choice for composing Asynchronous code, which uses a single loop for doing work in little units. This is due to the reason that it is simpler to write and maintain with no difficulty in handling deadlock or different issues.

- **A less restricted programming approach**

When contrasted with Java, Python utilizes a considerably less constrained multi-paradigm programming approach. For example, it is not necessary to make a different class for printing 'Hi World' in Python, While we need to do so in Java. Python is multi-paradigm and supports almost all programming styles. In Python, anything can be an object.

#### **6.4.4 CUDA**

Numerous algorithms for image handling and pattern recognition have been actualized on GPU for quicker computational power. The Implementation utilizing GPU experiences two issues. One is the requirement of knowing computer graphics and related languages and then there is the need for cooperation between CPU and GPU. CUDA (compute unified device architecture) is a parallel programming platform with which these issues can be solved and can be easily programmed which is developed by NVIDIA. Besides, OpenMP (Open Multi-Processing) is utilized to simultaneously process various inputs with

single guidance on multi-core CPU, which results in effectively using the memories of GPU. [126]

#### **6.4.5 AWS Deep Learning AMIs for Machine Learning Practitioners**

Deep learning innovation is developing at a quick pace – with everything from structures and algorithms to new strategies and speculations from the scholarly world and industry, emerging frequently. This often causes unpredictability for designers who require apparatuses for rapidly and safely testing algorithms, streamlining for particular adaptations of structures, running tests and benchmarks, or teaming up on undertakings beginnings with a clear canvas. Virtual conditions give the opportunity and adaptability to do this.

The Base AMI comes pre-introduced with the fundamental building block for deep learning. This incorporates NVIDIA CUDA libraries, GPU drivers, and framework libraries to accelerate and scale machine learning on Amazon Elastic Compute Cloud (EC2) instances. Base AMI is just like an open platform and user can create the environment by selecting their choice of technology [133].

- **Conda-based Deep Learning AMI**

The Conda-based AMI uses Anaconda virtual environment and it is easy to switch between the frameworks. It is a completely prepared virtual condition to run deep learning applications. It is very user friendly and often updated with the latest version of

frameworks. The Conda-based Deep Learning AMI comes bundled with the most recent versions of deep learning frameworks [129,134].

- **AWS Deep Learning AMI (Ubuntu 16.04)**

AWS deep learning AMI (Ubuntu 16.04) is used for developing the news generation framework in this research study. Deep Learning AMI accompanies famous Deep learning frameworks upgraded for superior performances on Amazon EC2 instances. It Incorporates Apache MXNet, TensorFlow, PyTorch, Caffe, Caffe2, Keras, Chainer, CNTK and Theano. The Deep learning systems are introduced in Conda situations to give a solid and separate condition for machine learning professionals [125]. Deep Learning frameworks are pre-configured with most recent updations of NVIDIA CUDA, cuDNN and Intel speeding up libraries for better performance along with Amazon EC2 instances

## **6.5 Conclusion**

The evaluation approach followed for the fulfilment of this work is discussed in detail through this chapter. Since it is a text generation application both automatic and manual evaluation is done. The dataset details are provided in this chapter. In the latter part of the section application implementation environment is explained in detail. The experimental results and discussion based on results is done in the next chapter.





**TESTING AND EVALUATION****Contents**

- 7.1 Introduction
- 7.2 Model Training
- 7.3 Testing Results
- 7.4 Sample Output
- 7.5 Correlating Automatic and Human Evaluation
- 7.6 Discussion and Inferences
- 7.7 Conclusion

**7.1 Introduction**

The evaluation approach followed in this work consists of both human and automatic evaluation methods. Testing and evaluation framework are portrayed in Figure 7.1. Three parameters are used for testing the neural net models under human evaluation. They are fluency, adequacy and total quality.

As shown in the Figure 7.1, automatic evaluation uses BLEU metric and ROUGE metric. Automatic and human evaluation results are correlated and finally, a discussion is done based on the test results.

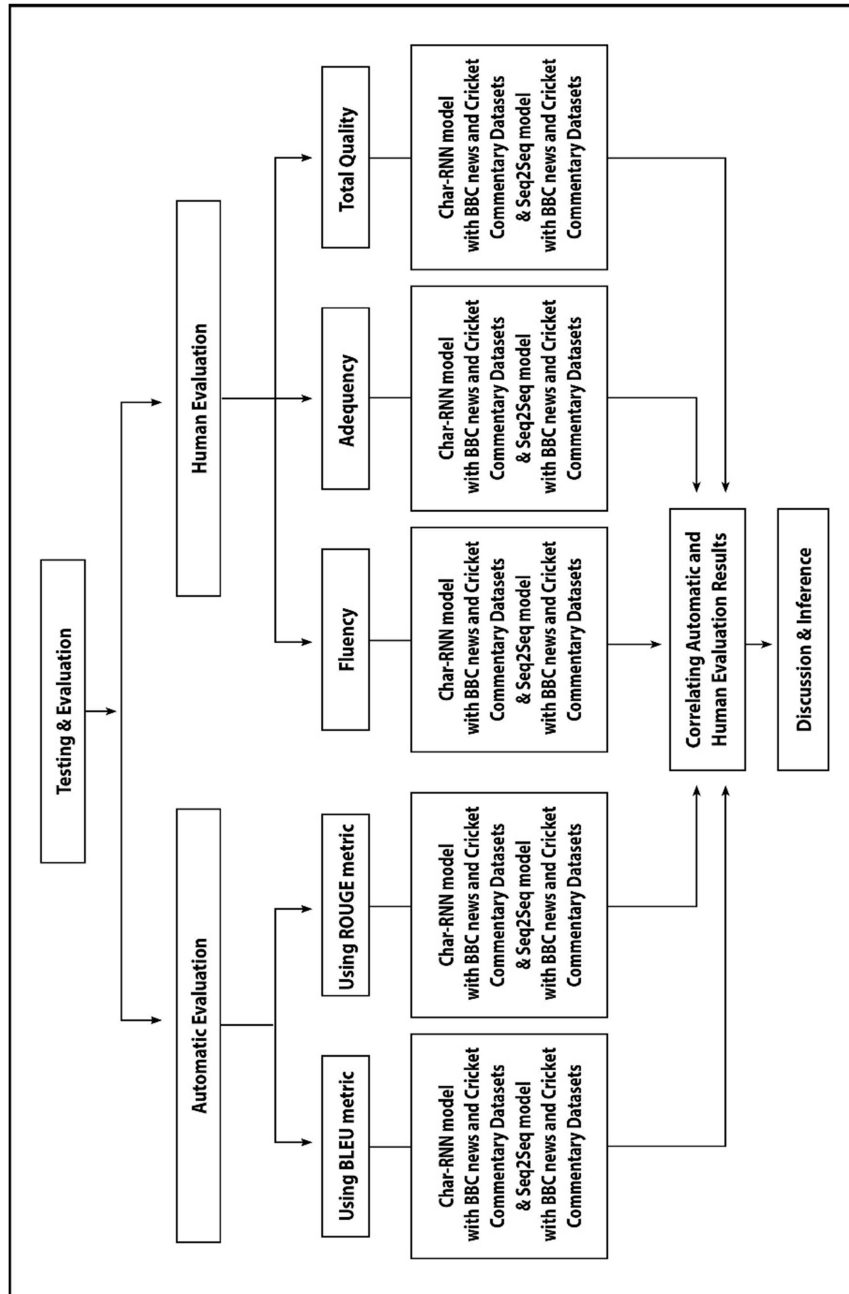


Figure 7.1 Testing and Evaluation Framework

The first section of this chapter provides details about model training. Automatic evaluation results are described using tables and diagrams in the second section. The third section of this chapter elaborates results based on human evaluation. Finally, a discussion is done by correlating automatic and human evaluation results. The experiment uses a BBC news dataset as well as a Cricket Commentary dataset. The screenshots of dataset CSV files are given in Appendix 1.

One of the generic assumptions made during the implementation of the framework is to ensure that the size of the generated news text is of a minimum of 1000 characters. This provides a better evaluation process using BLEU and ROUGE. The datasets used in this work were prepared by including news of almost the same length. This is required to reduce the zero-padding to a minimum especially in the case of Sequence to Sequence neural net model. News corresponds to the same topic were included in the dataset for better learning by the neural net models. The BBC news dataset was chosen carefully by including only five categories like politics, sports, entertainment, business and technology which were collected in the year 2004-2005. Cricket Commentary dataset was created from commentaries in IPL (Indian Premier League) 2016 from Cricbuzz website.

## 7.2 Model Training

Two neural net models were trained using the implementation set up provided by Amazon. Sequence to Sequence model took more time for training than Char-RNN model. Screen shots of training of models are provided in Appendix 3. The training details are discussed in the following sections.

### 7.2.1 Char-RNN model

Char-RNN model was trained for 10000 epochs. In Char RNN model, there are three layers. Each layer consists of 256 nodes. The learning rate of Vanilla Char RNN model is 0.003 and loss is calculated using `softmax_cross_entropy_with_logits`. The optimizer used was RMSProp. Figure 7.2 provides a graph which represents average training loss of the Char-RNN model based on the number of epochs. Checkpoints were made for every 100 epochs.

Activation functions are used to determine the output of neural network. The softmax function is a more generalized logistic activation function which is used for multiclass classification task. The sigmoid function is normally used for binary classification task. The tanh function is very similar to the sigmoid function. It is actually just a scaled version of the sigmoid function. Sigmoid and tanh functions are sometimes avoided due to the vanishing gradient problem. In the neural net model used in this work, softmax is used as an activation function



since the model tries to predict next character or word and it is similar to a multiclass classification task.

The loss value calculation is an important part in neural net model training. Its value decides how good the model is. The value of loss should be less for a good model. The loss values can also be in lower range when the model is over-fitted. It is possible to understand the model's performance on training and validation set by analysing loss score. Normally, the loss is not represented as a percentage. It is the total sum of the errors created for each example in training or validation sets.

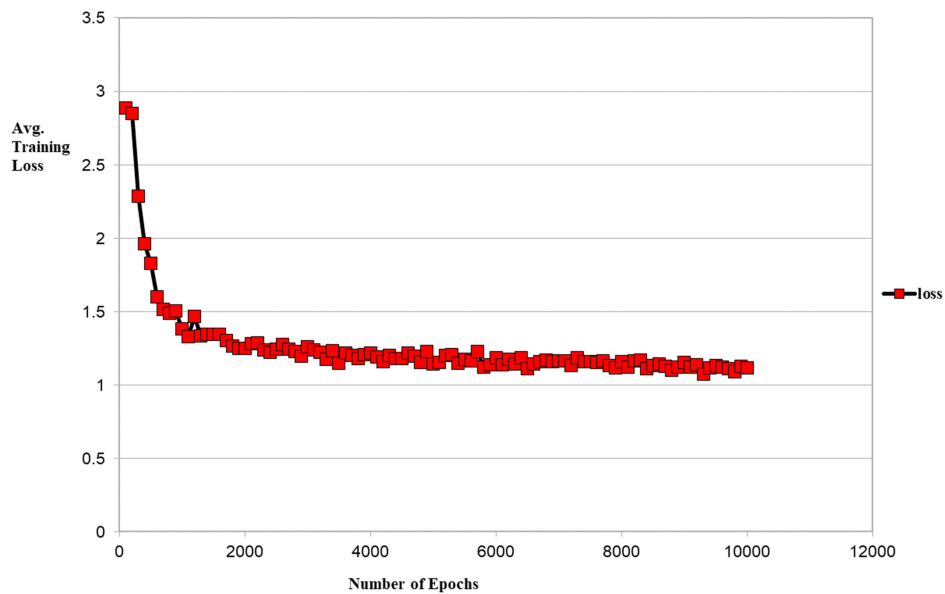
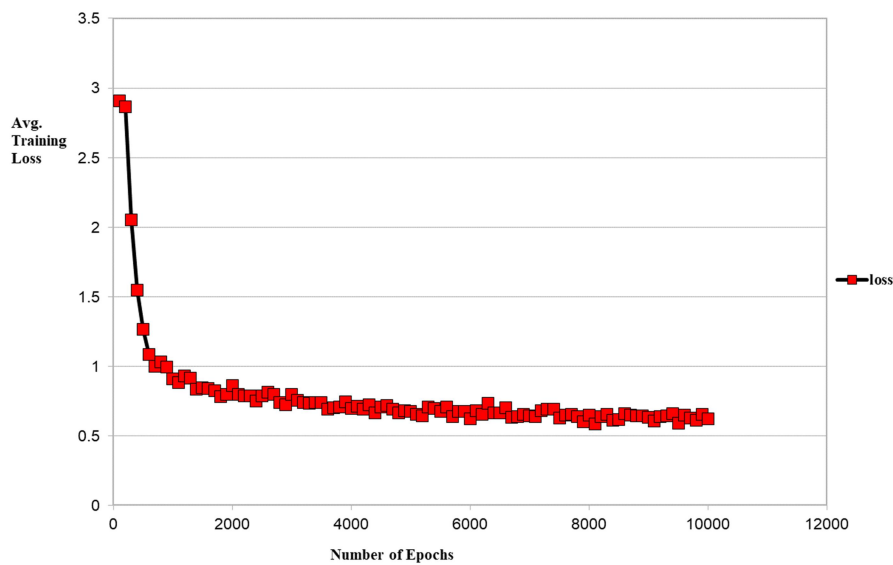


Figure 7.2: Average training loss of the Char-RNN model (BBC news dataset)

In Figure 7.2, The X axis represents the number of epochs and the Y axis represents the average training loss. After 10000 epochs, loss value reaches close to 0.5. The training of Char-RNN model took almost 96 hours to complete 10000 epochs. Figure 7.3 shows the average training loss of the Char-RNN model with the Cricket Commentary dataset. The main aim in a learning phase of the neural net model is to reduce the loss function score with respect to the parameters set for the model by changing the weight vector values using various optimization algorithms.



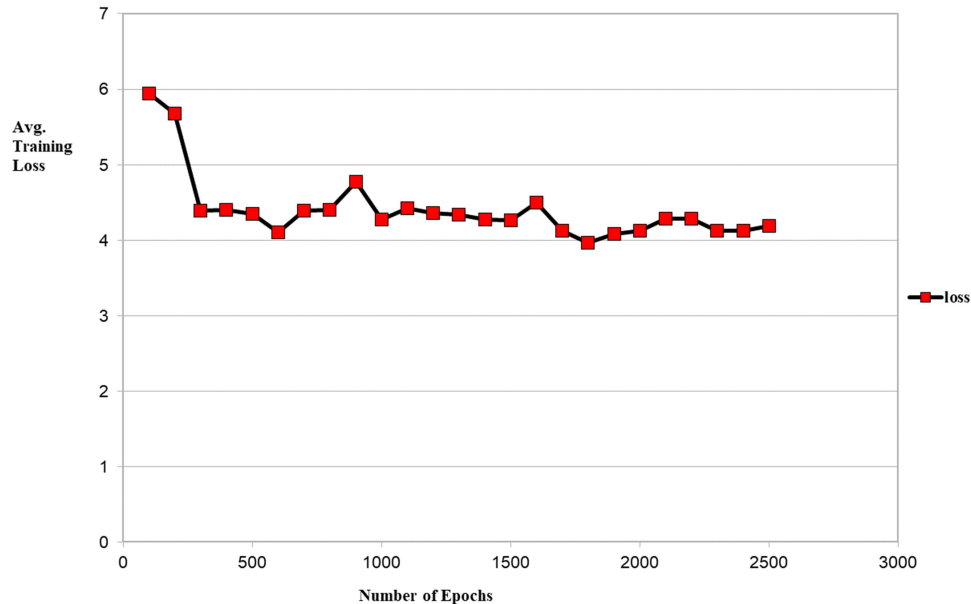
**Figure 7.3: Average training loss of the Char-RNN model (Cricket Commentary dataset)**

The graph shown in Figure 7.3 is plotted using the number of training epochs and the average training loss. During the initial stage of training, the loss value was close to 3 and decreased as the number of training epochs increased. A loss is calculated after each iteration. Ideally, the loss is expected to be reduced after a sufficient number of iterations.

### **7.2.2 Seq2Seq model**

Sequence to Sequence model was trained for 2500 epochs. In Sequence to Sequence neural net model, encoder and decoder has 3 layers each. The number of nodes in each layer is 500. The learning rate of Sequence to Sequence neural net model is 0.001. The loss was calculated using `categorical_crossentropy` and optimizer is `RMSProp`.

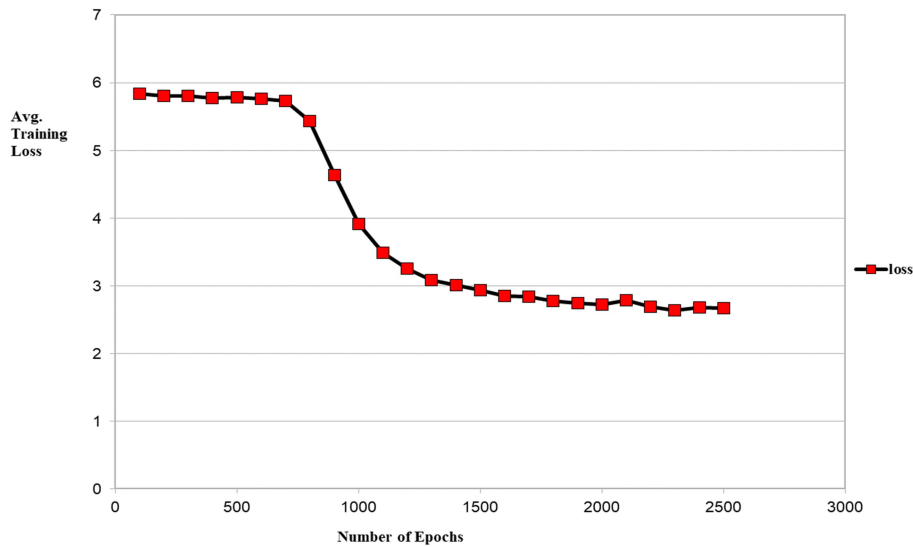
The graphical representation of training loss of Sequence to Sequence model with BBC news dataset is shown in Figure 7.4. The training of Seq2Seq model with BBC news dataset took longer time than the same model with Cricket Commentary dataset. One of the reasons was the high vocabulary size required for the BBC news dataset compared to the Cricket Commentary dataset. Loss score is often utilized to fine-tune the hyperparameter values for the neural net models. It is a measure of the inconsistency between the predicted and actual value. Ideally, the loss is a non-negative value and is an average of the losses at each time step.



**Figure 7.4: Average training loss of the Seq2Seq model (BBC news dataset)**

In Figure 7.4, X-axis represents the number of training epochs and Y-axis denotes the average training loss. During the initial stage of training, loss value reached 6. Training loss of Seq2Seq model with the BBC news dataset does not decrease considerably compared with the other cases. Figure 7.5 shows the average training loss of the model when trained with the Cricket Commentary dataset. A good model is a model where the predicted output is close to the training output. The loss function sometimes called a cost function. It calculates the loss by comparing the predicted character to the target character. The loss is the

measure of the mistakes done by a neural net model in predicting the output.



**Figure 7.5: Average training loss of the Seq2Seq model (Cricket Commentary dataset)**

Sequence to Sequence model with the Cricket Commentary dataset shows decrease in loss after 800 training epochs. Initial training loss was above 5 but after the completion of 2500 epochs, the loss was in between 2 and 3 as represented in the Figure 7.5.

An epoch is a hyperparameter which is to be fixed before starting the model training. One epoch is completed when the dataset is passed both forward and backward through the neural network only once. One epoch is not sufficient for proper training of the model. Sometimes, it requires several epochs as in the cases discussed in the present study.

Sometimes, one epoch is too big to feed to the computer at once. So, it is divided into several smaller batches. A batch is the total number of training examples present in a single batch and an iteration is the number of batches needed to complete one epoch. In this experiment, the batch size used for the Char-RNN model is 64 whereas the Sequence to Sequence model batch size is 50.

### 7.3 Test Results

The evaluation approach followed in the present study is described in Chapter 6. Automatic and human evaluation results are given in this section. Automatic evaluation results are presented first followed by the human evaluation ratings. BLEU and ROUGE scores of generated news are calculated as a part of the automatic evaluation.

#### 7.3.1 Automatic Evaluation

The ROUGE scores were calculated using the Python package `pyrouge` using the original Perl script `ROUGE-1.5.5.pl`. The BLEU scores were computed with the built in BLEU method `nlk.translate.bleu_score` from the Python package NLTK 3.2.5 [154].

BLEU metric scores are represented using tables and diagrams in the following section followed by ROUGE results.

##### 7.3.1.1 BLEU Metric (Char-RNN model)

All n-gram matches between reference news and machine generated news were obtained and BLEU scores were measured in this work where

$n = 1$  to 4. Table 7.1 and 7.2 summarizes BLEU values of Char-RNN model. The BLEU score provides an overall assessment of model quality.

**Table 7.1: BLEU score of Char-RNN model (BBC news dataset)**

Epoch	BLEU-1	BLEU-2	BLEU-3	BLEU-4
100	0.07919291	0.03600362	0.02111043	0.01492017
200	0.08016491	0.0370115	0.02115638	0.01418193
500	0.2577802	0.08885826	0.03697561	0.02143184
1000	0.30320515	0.10988054	0.043327	0.02466201
2000	0.31102969	0.11991875	0.04701409	0.02580221
5000	0.31503532	0.11723873	0.0459208	0.02448898
10000	0.31141876	0.11613108	0.04820172	0.02638924

The table above is a tabulated result of the values obtained with the BLEU metric in the case of BBC news dataset. As shown in the table, there was no considerable increase in the BLEU value after 5000 epochs.

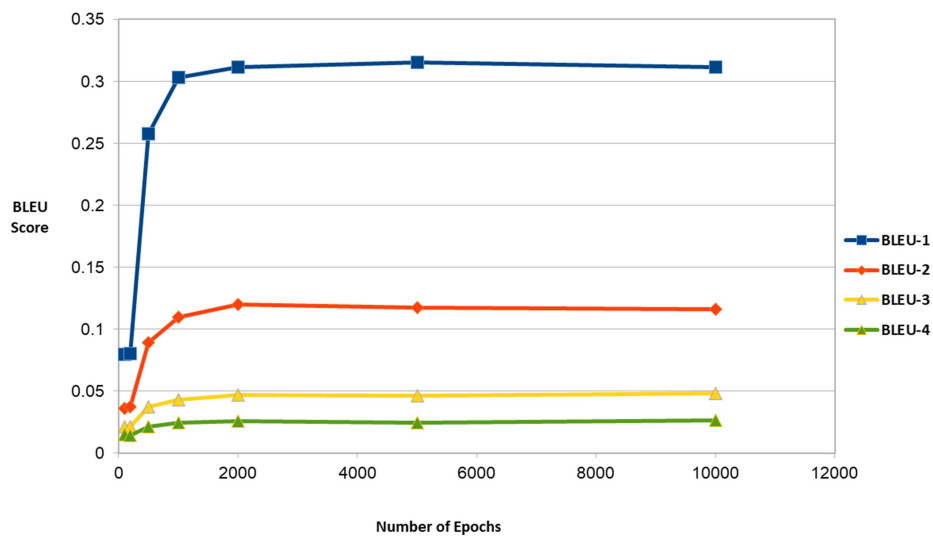
**Table 7.2: BLEU score of Char-RNN model (Cricket Commentary dataset)**

Epoch	BLEU-1	BLEU-2	BLEU-3	BLEU-4
100	0.12707224	0.1000	0.07575722	0.0666749
200	0.11904762	0.06532553	0.05252713	0.05066197
500	0.34057971	0.13399752	0.07202726	0.04965434
1000	0.35652174	0.16414127	0.08084437	0.05056503
2000	0.3364486	0.15757673	0.09030465	0.06972041
5000	0.35211268	0.1647197	0.08577557	0.06429451
10000	0.3537415	0.16086022	0.11746087	0.09487732

From Table 7.2, it is evident that there was no significant improvement in the BLEU score as in the case of Char-RNN model with BBC news dataset after 5000 epochs. The Char-RNN model with Cricket Commentary dataset provides high BLEU value than the BBC news dataset.

Figure 7.6 is the graphical representation of the BLEU score values of the Char-RNN model with the BBC news dataset. BLEU score, which indicates how similar the generated news is to the reference news, with values closer to one representing more similar news.

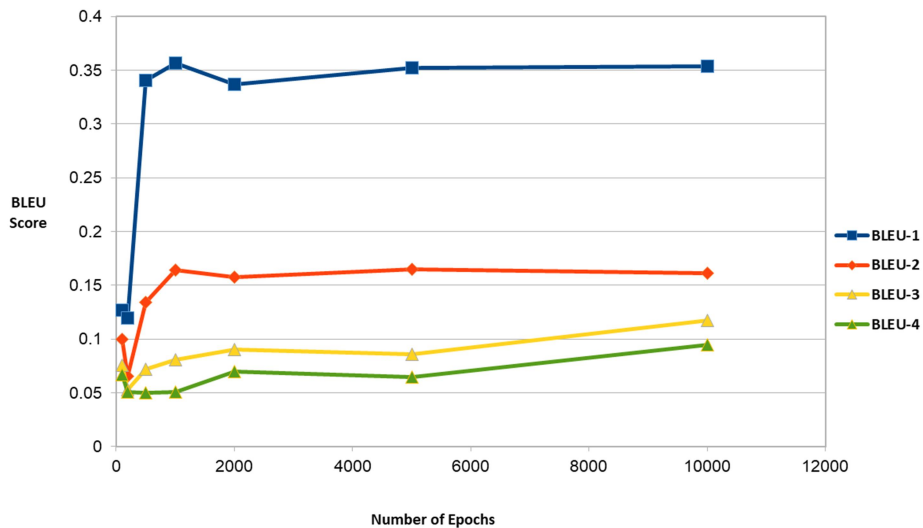
BLEU score value of 0 means that the machine-generated news has no overlap with the reference news while a value of 1 means there is perfect overlap with the reference news.



**Figure 7.6: BLEU scores of Char-RNN model (BBC news dataset)**



In figure 7.6, The X axis represents the number of epochs and the Y axis represents the BLEU score. BLEU score stabilizes after 5000 epochs. The BLEU scores of Char-RNN model with the Cricket Commentary dataset is graphically represented in the figure below. The X axis represents the number of epochs whereas Y axis denotes the BLEU score.



**Figure 7.7: BLEU scores of Char-RNN model (Cricket Commentary dataset)**

From Figure 7.7, it is clear that maximum value was achieved by BLEU when the number of epochs is in between 5000 and 10000. The Char-RNN model with the Cricket Commentary dataset has better BLEU value than the same model with the BBC news dataset.

### 7.3.1.2 BLEU Metric (Seq2Seq model)

Table 7.3 and 7.4 provides BLEU scores of Sequence to Sequence model with the BBC news dataset and the Cricket Commentary dataset respectively.

**Table 7.3: BLEU score of Seq2Seq model (BBC news dataset)**

Epoch	BLEU-1	BLEU-2	BLEU-3	BLEU-4
100	0	0	0	0
200	0.14849741	0.08573502	0.04949914	0.02857834
300	0.24622588	0.14215858	0.08207529	0.04738619
400	0.30858461	0.17816141	0.10286154	0.05938714
500	0.29816577	0.17214609	0.09938859	0.05738203
600	0.3454022	0.19941805	0.11513407	0.06647268
700	0.31843699	0.18384968	0.10614566	0.06128323
800	0.30592122	0.1766237	0.10197374	0.05887457
900	0.32209715	0.18596288	0.10736572	0.06198763
1000	0.34644385	0.18500019	0.11548128	0.06667315
1100	0.35595514	0.20551079	0.11865171	0.0685036
1200	0.3520304	0.20324485	0.11734347	0.06774828
1300	0.34857143	0.20124781	0.11619048	0.0670826
1400	0.34	0.19629909	0.11333333	0.06543303
1500	0.31931714	0.18435784	0.10643905	0.06145261
1600	0.31979427	0.18463331	0.10659809	0.06154444
1700	0.33428571	0.19299995	0.11142857	0.06433332
1800	0.30784981	0.17878383	0.10342312	0.05976983
1900	0.27019035	0.15599447	0.09006345	0.05199816
2000	0.32018075	0.18485644	0.10672692	0.06161881
2100	0.32719466	0.18890593	0.10906489	0.06296864
2200	0.31616669	0.18253893	0.1053889	0.06084631
2300	0.31706489	0.1830575	0.1056883	0.06101917
2400	0.31022997	0.17911136	0.10340999	0.05970379
2500	0.31930302	0.18434969	0.10643434	0.0614499

The above table represents the values of BLEU-1, BLEU-2, BLEU-3 and BLEU-4 in the case of BBC news dataset. After 1000 training epochs itself, this model provides better result than the Char-RNN model.

**Table 7.4: BLEU score of Seq2Seq model (Cricket Commentary dataset)**

Epoch	BLEU-1	BLEU-2	BLEU-3	BLEU-4
100	0.0000	0.0000	0.0000	0.0000
200	0.0000	0.0000	0.0000	0.0000
300	0.0000	0.0000	0.0000	0.0000
400	0.0000	0.0000	0.0000	0.0000
500	0.0000	0.0000	0.0000	0.0000
600	0.0000	0.0000	0.0000	0.0000
700	0.17762286	0.10255061	0.05920762	0.03418354
800	0.3372549	0.19471421	0.1124183	0.06490474
900	0.34117647	0.19697833	0.11372549	0.06565944
1000	0.34246682	0.19772331	0.11415561	0.06590777
1100	0.3541449	0.20446566	0.1180483	0.06815522
1200	0.37647059	0.2173554	0.1254902	0.0724518
1300	0.36470588	0.21056304	0.12156863	0.07018768
1400	0.37254902	0.21509128	0.12418301	0.07169709
1500	0.38431373	0.22188363	0.12810458	0.07396121
1600	0.37254902	0.21509128	0.12418301	0.07169709
1700	0.36470588	0.21056304	0.12156863	0.07018768
1800	0.34509804	0.19924245	0.11503268	0.06641415
1900	0.36078431	0.20829892	0.12026144	0.06943297
2000	0.36078431	0.20829892	0.12026144	0.06943297
2100	0.38823529	0.22414775	0.12941177	0.07471592
2200	0.35294118	0.20377068	0.11764706	0.06792356
2300	0.36470588	0.21056304	0.12156863	0.07018768
2400	0.34901961	0.20150657	0.11633987	0.06716886
2500	0.34509804	0.19924245	0.11503268	0.06641415

The values shown in the above table indicates that the Sequence to Sequence model with the Cricket Commentary dataset has achieved higher BLEU score than the Sequence to Sequence model with the BBC news dataset.

Figure 7.8 is a graphical representation of the BLEU score of Seq2Seq model with Cricket Commentary dataset. BLEU score interpretation in language generation is a difficult task. BLEU score below 0.10 is almost useless and in between 0.10 and 0.20, it is very hard to make the gist. BLEU score above 0.20 and below 0.30 makes sense in text generation scenario. But there is every possibility that there may be significant grammatical errors. BLEU score between 0.30 and 0.40 indicates a good score. BLEU score above 0.40 can be taken as high-quality text generation

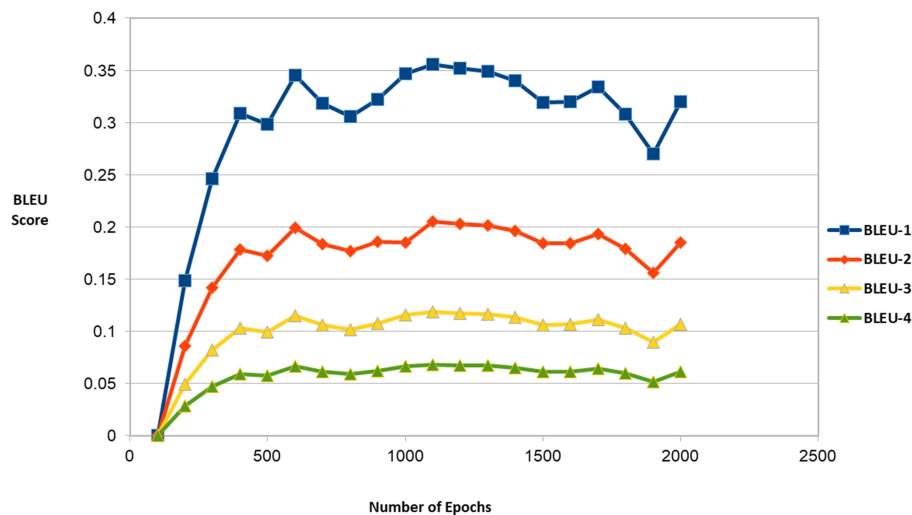
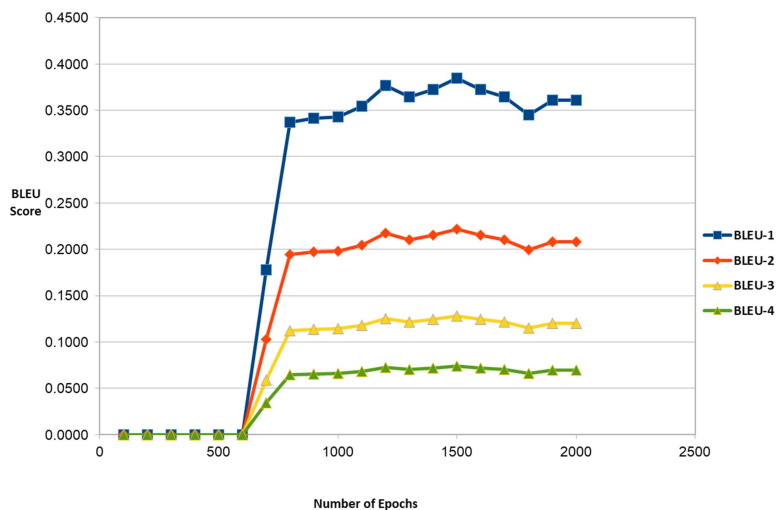


Figure 7.8: BLEU scores of Seq2Seq model (BBC news dataset)

The figure above represents the BLEU score of Seq2Seq model with the BBC news dataset. The X axis represents the number of epochs and the Y axis represents the BLEU score. The BLEU score is better here, compared to that of the Char-RNN model.

Figure 7.9 is a graphical representation of BLEU score of Seq2Seq model with the Cricket Commentary dataset. Seq2Seq model with the Cricket Commentary dataset gives better BLEU values than BBC news dataset.



**Figure 7.9: BLEU scores of Seq2Seq model (Cricket Commentary dataset)**

In the initial stages, the model with the BBC news dataset gave better results than the model with the Cricket Commentary dataset. But after 800 epochs, Seq2Seq model with the Cricket Commentary dataset showed much improvement in terms of BLEU score as shown in Figure

7.9. BLEU metric has several limitations. Tokenizing the reference and candidate text before computing BLEU score affects the final BLEU score. It performs badly with individual sentences. BLEU is basically a corpus-based metric. BLEU metric is not good in capturing the meaning and grammatical aspect of the content.

### 7.3.1.3 ROUGE Metric (Char-RNN model)

ROUGE is used to measure recall value and it is best suited to evaluate summary. It has several variants. ROUGE concentrates on the adequacy of generated news rather than fluency. The ROUGE value ranges from 0 to 1. The value closer to 1 obviously points to high quality generated news. Table 7.5 provides ROUGE score of the Char-RNN model. Figure 7.10 graphically represents the ROUGE-L score of Char-RNN model with the BBC news dataset and the Cricket Commentary dataset.

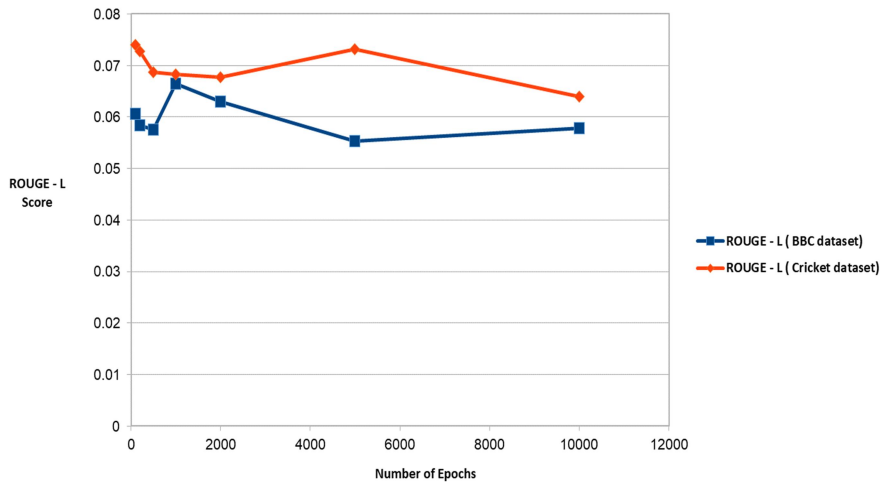
**Table 7.5: ROUGE score of the Char-RNN model**

Epoch	ROUGE - L (BBC news dataset)	ROUGE - L ( Cricket Commentary dataset)
100	0.06057	0.07407
200	0.0584	0.07273
500	0.05758	0.06864
1000	0.06644	0.06825
2000	0.06297	0.06766
5000	0.05537	0.07321
10000	0.0578	0.064

The ROUGE values in the above table show that there was no considerable change in the ROUGE score after 10000 epochs. The ROUGE score of the Char-RNN model with the Cricket Commentary

dataset had a slightly better value than the same model with the BBC news dataset.

The figure below is a pictorial representation of Rouge-L scores of char-RNN model with the number of epochs. The X axis represents the number of epochs and the Y axis represents the ROUGE -L scores.



**Figure 7.10: ROUGE-L score of Char-RNN model (BBC news dataset and Cricket Commentary dataset)**

From the above figure, it is clear that the Char-RNN model with the Cricket Commentary dataset performed better in terms of ROUGE score. Significant changes were not seen in the case of ROUGE score by training the model for 10000 epochs.

#### 7.3.1.4 ROUGE Metric (Seq2Seq model)

The type of ROUGE used in this experiment is ROUGE-L. It measures the longest matching sequence of words Longest Common

Subsequence (LCS). LCS does not require consecutive matches but in-sequence matches that reflect sentence level word order.

In general, ROUGE score above 0.20 indicates that the generated news makes sense in terms of adequacy. Table 7.6 gives the ROUGE score of the Sequence to Sequence model.

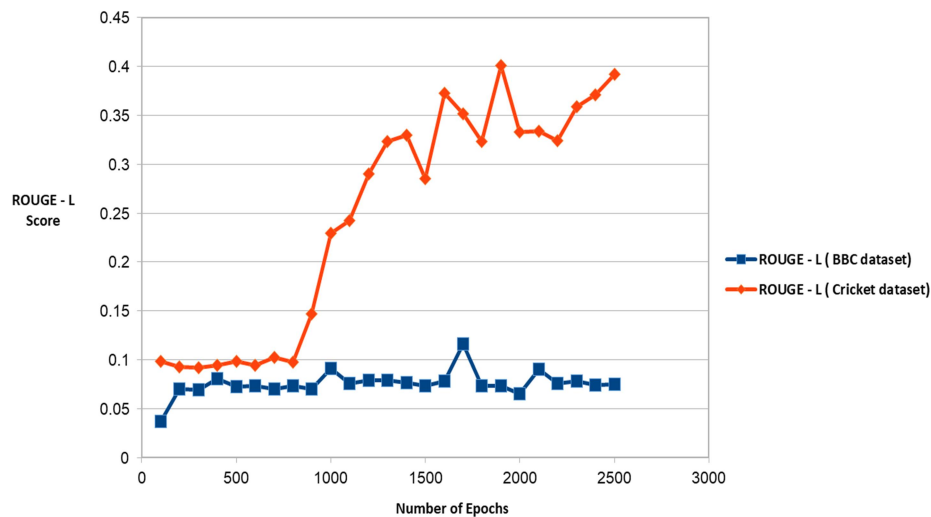
**Table 7.6: ROUGE score of the Seq2Seq model**

Epoch	ROUGE – L ( BBC news dataset)	ROUGE - L ( Cricket Commentary dataset)
100	0.0368	0.09808
200	0.07043	0.09249
300	0.06941	0.09231
400	0.08061	0.09405
500	0.0726	0.09827
600	0.07326	0.09423
700	0.06995	0.10232
800	0.07336	0.09747
900	0.07038	0.14686
1000	0.09079	0.22965
1100	0.07559	0.24263
1200	0.07931	0.29017
1300	0.0792	0.32336
1400	0.07695	0.33013
1500	0.07338	0.28571
1600	0.07849	0.37255
1700	0.11622	0.35196
1800	0.07352	0.32365
1900	0.07305	0.4009
2000	0.06557	0.33333
2100	0.09034	0.33386
2200	0.07559	0.32398
2300	0.07824	0.35874
2400	0.07415	0.37122
2500	0.07508	0.39204



The values in the above table prove that the Seq2Seq model with the Cricket Commentary dataset outperforms the model with the BBC news dataset in terms of ROUGE score.

Figure 7.11 graphically represents the ROUGE-L score of Sequence to Sequence model with the BBC news dataset and the Cricket Commentary dataset respectively.



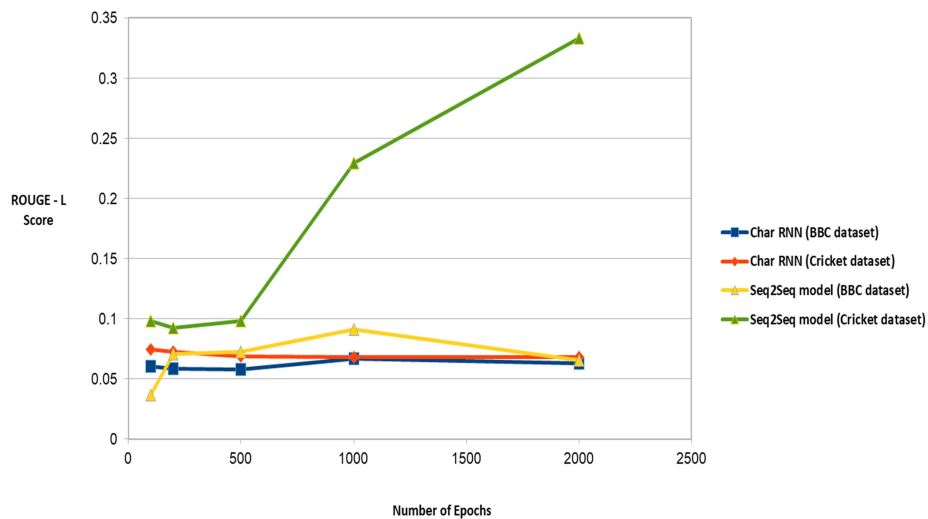
**Figure 7.11: ROUGE-L score of Seq2Seq model (BBC news dataset and Cricket Commentary dataset)**

As seen in the figure above, the Seq2Seq model with the BBC news dataset maintains almost a steady ROUGE value till 2500 epochs. But the Seq2Seq model with the Cricket Commentary dataset shows a sudden jump in the ROUGE score after 800 epochs. This graph is a

clear indication of the better performance of the model with the Cricket Commentary dataset.

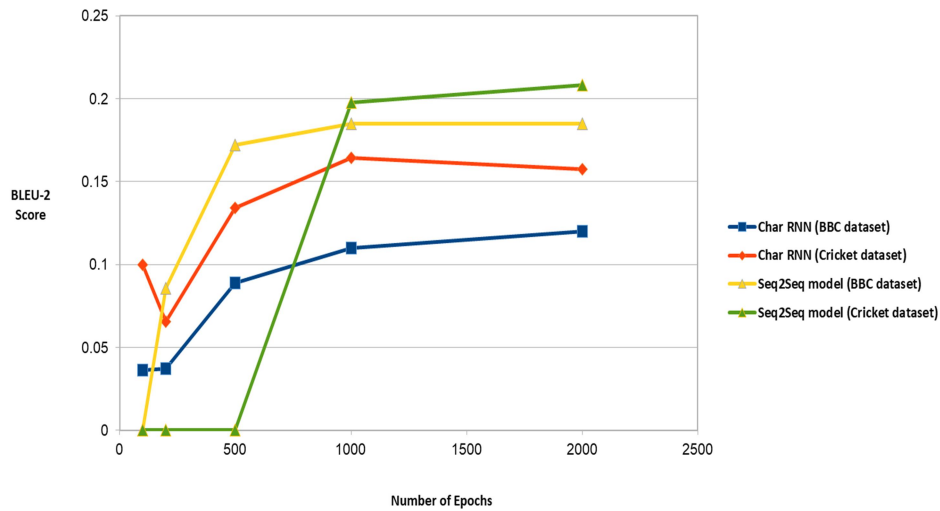
### 7.3.1.5 Comparing Two Neural Net Models Using BLEU-2 and ROUGE-L

The comparison based on ROUGE-L and BLEU-2 scores of neural models with the two datasets are illustrated in Figure 7.12 and Figure 7.13. In the graph given below, The X axis represents the number of epochs while Y axis denotes ROUGE-L scores.



**Figure 7.12: Comparison of models based on ROUGE-L**

Each of the test cases is shown with differently coloured lines in Figure 7.12. The Sequence to Sequence model with the Cricket Commentary dataset is the best performing case among the four cases in terms of ROUGE score. In the graph given below, X axis represents the number of training epochs whereas Y axis represents the BLEU-2 scores.



**Figure 7.13: Comparison of models based on BLEU-2**

From Figure 7.13, it is seen that the Sequence to Sequence model with the Cricket Commentary dataset performs better than the other 3 cases in terms of BLEU-2. The second best model is Seq2Seq model with the BBC news dataset. The third is the Char-RNN model with the Cricket Commentary dataset and the fourth one is the Char-RNN model with the BBC news dataset.

### 7.3.2 Human Evaluation Results

Human evaluation approach is detailed in chapter 6. The main evaluation parameters are Fluency, Adequacy and Total Quality. The results are presented in the following section.

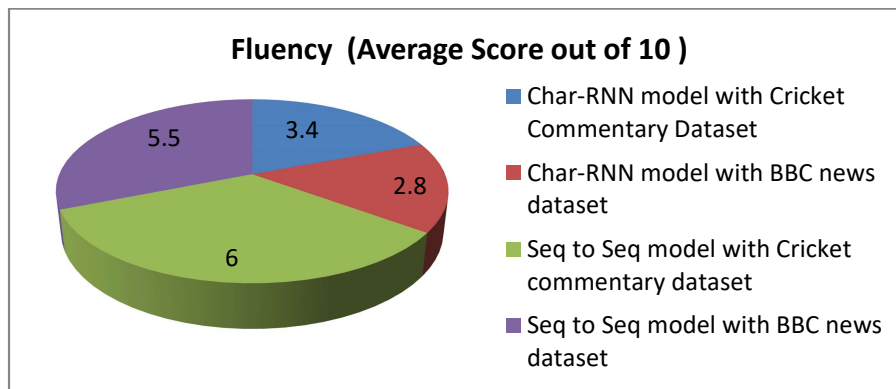
### 7.3.2.1 Based on Evaluation Criteria

The first feature analysed was Fluency. The individual results from human evaluators were collected, its average was taken out of 10 and tabulated as given in Table 7.7.

**Table 7.7: Human ratings (Fluency)**

Model	Fluency (out of 10)
Char-RNN model with Cricket Commentary Dataset	3.4
Char-RNN model with BBC news dataset	2.8
Seq2Seq model with Cricket Commentary dataset	6
Seq2Seq model with BBC news dataset	5.5

As in the Table 7.7, the Sequence to Sequence model with the Cricket Commentary dataset shows the highest value for fluency followed by Sequence to Sequence model with the BBC news dataset. Char-RNN model for Cricket Commentary dataset ranks third, followed by the Char-RNN model with the BBC news dataset. The Fluency scores are graphically represented in Figure 7.14.



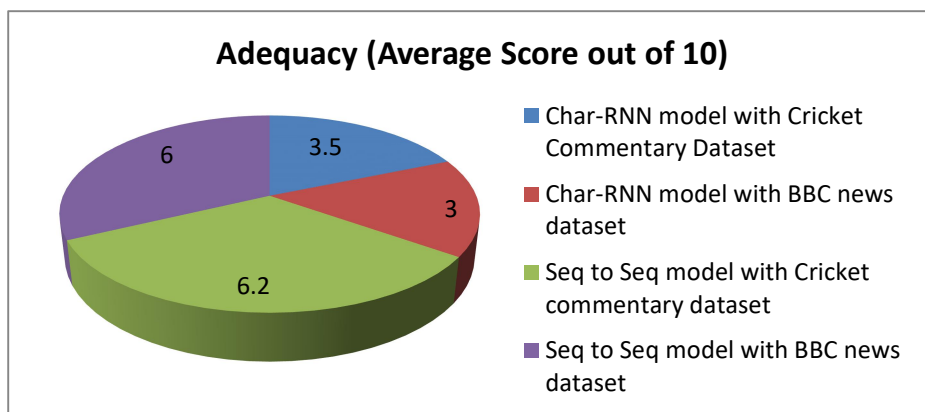
**Figure 7.14: Human ratings based on prescribed qualities (Fluency)**

The figure above represents the evaluation of the quality of fluency (on a scale of 10) in the four cases. The results are the average of all the individual evaluation results. Table 7.8 is a tabulated result of human evaluation in terms of the quality of Adequacy.

**Table 7.8: Human ratings (Adequacy)**

Model	Adequacy ( mark out of 10)
Char-RNN model with Cricket Commentary Dataset	3.5
Char-RNN model with BBC news dataset	3
Seq2Seq model with Cricket Commentary dataset	6.2
Seq2Seq model with BBC news dataset	6

From Table 7.8, the maximum value for Adequacy was obtained by the Seq2Seq model with the Cricket Commentary dataset and the lowest was obtained by the Char- RNN model with the BBC news dataset. Adequacy scores are graphically represented in Figure 7.15.



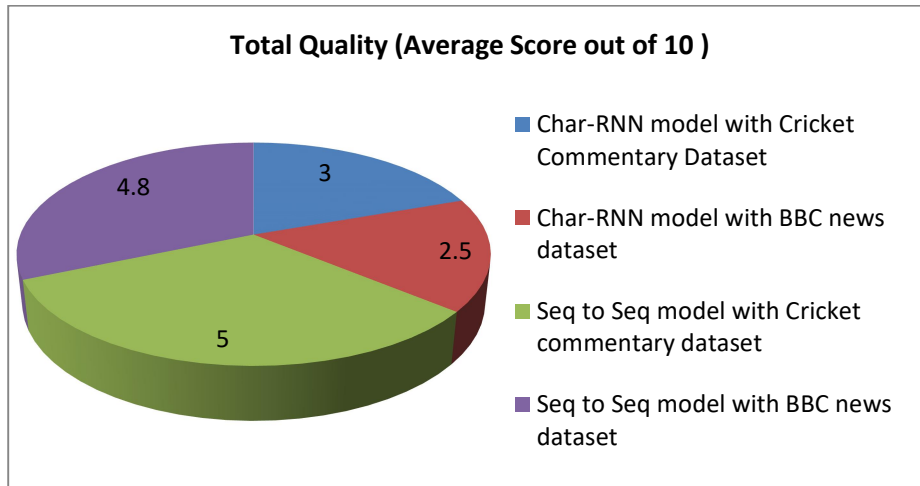
**Figure 7.15: Human ratings based on prescribed qualities (Adequacy)**

As shown in Figure 7.15, when the quality of Adequacy is considered, the maximum value was attributed to the Sequence to Sequence model with the Cricket Commentary dataset with a value of 6.2 out of 10 followed by the Sequence to Sequence model with the BBC news dataset with a numerical value of 6. Then comes Char-RNN model with Cricket Commentary dataset and the fourth one is Char-RNN model with the BBC news dataset. Table 7.9 is a tabular representation of the values obtained when human evaluation was conducted for the feature of Total Quality.

**Table 7.9: Human ratings (Total Quality)**

<b>Model</b>	<b>Total Quality (mark out of 10)</b>
Char-RNN model with Cricket Commentary Dataset	3
Char-RNN model with BBC news dataset	2.5
Seq2Seq model with Cricket Commentary dataset	5
Seq2Seq model with BBC news dataset	4.8

The table above shows that the maximum score for Total Quality was obtained by the Sequence to Sequence model with the Cricket Commentary dataset. The figure below is a pictorial representation of the values obtained when human evaluation was conducted with the four cases for the feature of Total Quality.



**Figure 7.16: Human ratings based on Total Quality**

Total Quality represents the quality of the evaluated content and judges if the content is worth presenting and acceptable. In the ranking list, the first was the Sequence to Sequence model with the Cricket Commentary dataset with an average score of 5 out of 10. The second was the Sequence to Sequence model with BBC news dataset with a score of 4.8 followed by the Char-RNN model with the Cricket Commentary dataset with a score of 3 and finally the Char-RNN model with the BBC news dataset with score of 2.5, as shown in Figure 7.16.

### 7.3.2.2 Based on Evaluators Traits

The main evaluation criteria were Fluency, Adequacy and Total Quality. To understand in detail whether the computer expertise and language proficiency of the human evaluators affect the human ratings, evaluation is conducted to test those behaviours too.

**a) Computer Expert's v/s Non Expert's**

Two categories of evaluators, experts and non-experts were grouped from the selected team for the evaluation. It was challenging for the non-experts to have direct interaction with the application whereas computer experts found it comfortable to have direct interaction with the application. But it did not affect the final human ratings.

**b) English Language Proficiency (Basic v/s Expert)**

In this experiment, there were two groups of human evaluators - basic level English speakers and expert level English speakers. This would be beneficial in collecting results from different strata of English language proficiencies. A more critical approach was expected from English language experts. The following features were analysed to measure language proficiency of the system under test.

- *Appropriateness of Words*: the efficiency of the system in using apt words at appropriate situations. This feature is important as it contributes to the meaning of the sentence as per the context.
- *Spelling*: if the generated system is dependable with the spellings of words. A properly spell checked document is needed to make it impressive to the user.



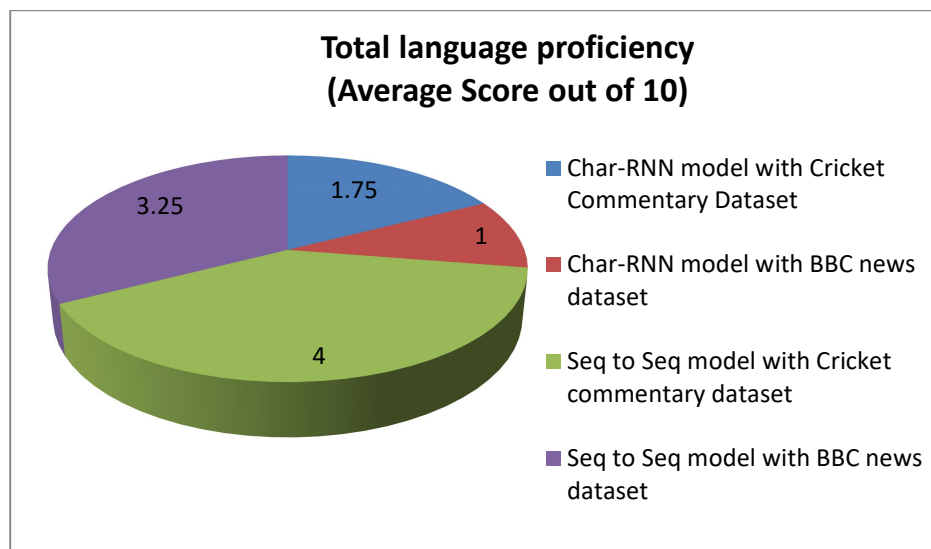
- *Grammar*: If the system follows grammatical rules, tenses and correct forms of words. Using the correct forms of words and following grammatical rules are important while considering a document.
- *Comprehensibility*: how far the result conveys the right meaning to the users. If the proper sense is not conveyed, the total process would be in vain. Therefore, comprehensibility is important.

Table 7.10 provides language proficiency score. The scores are average scores out of 10. It is evident from the scores in the below table that the language proficiency results justify the human scores of Fluency, Adequacy and Total Quality.

**Table 7.10: Human ratings (language proficiency)**

Model	Appropriateness of words	Spelling	Grammar	Comprehensibility	Language proficiency (Average score out of 10)
Char-RNN model with Cricket Commentary Dataset	2	1	2	2	1.75
Char-RNN model with BBC news dataset	1	1	1	1	1
Seq2Seq model with Cricket Commentary dataset	4	4	4	4	4
Seq2Seq model with BBC news dataset	3	4	3	3	3.25

Individual evaluations were collected and the average score was calculated for the above features. Language proficiency was calculated as the sum of the average values of the above mentioned features. Details are shown in Table 7.10. Figure 7.17 provides a graphical representation of evaluation results based on language proficiency.



**Figure 7.17: Human evaluation based on language proficiency**

The Sequence to Sequence model with the Cricket Commentary dataset showed the highest value for appropriateness of words, spelling, grammar and comprehensibility as represented in Figure 7.17.

### **7.3.2.3 Summary of Human Evaluation**

The human evaluation process was carried out successfully. On a scale of 10, more number of people concluded that the Sequence to Sequence model with the Cricket Commentary dataset gave out the best results. The second best result was given by the Sequence to Sequence model with the BBC news dataset. Third was the Char-RNN model with the Cricket Commentary dataset and the fourth was the Char-RNN model with the BBC news dataset. Clearly the Sequence to Sequence model proved to be efficient in news generation than the Char-RNN model. About 50% out of the 45 people predicted that the Sequence to Sequence model with the Cricket Commentary dataset was machine generated; about 60 % of the evaluators predicted that the second ranked Sequence to Sequence model with the BBC news dataset was machine generated; about 90% evaluators thought that the third ranked Char-RNN model with Cricket Commentary dataset was machine generated; and about 95% of the evaluators believed that the Char-RNN model with the BBC news dataset was machine generated.

## 7.4 Sample Output

Table 7.11 provides a sample output of the Sequence to Sequence model with the Cricket Commentary dataset.

**Table 7.11: Sample Output**

---

Original text	faulkner to rahane one run rahane is hurried by this back of a length delivery from round the wicket he swipes it awkwardly to midwicket faulkner to smith one run smith stays legside on this fuller length delivery and then drives it to longoff easy single dwayne bravo to smith one run slightly misses the line but smith cannot make full use of it tickles it off his pads to backward square leg for a singledwayne bravo to rahane one run bravo is back to what he does best slow ball on the stumps rahane is forced to work to midwicket
Ontology Keywords	sports, cricket, fielding, ball, length delivery, delivery, off, leg, square leg, pads, drives, misses, legside, round, swipes, back drives, length backward, full, midwicket, smith, Faulkner, rahane, Dwayne, bravo
Machine Generated	faulkner to rahane one run rahane is hurried by this back of a length delivery from round the wicket he swipes it awkwardly to midwicket faulkner to smith one run smith stays legside on this fuller length delivery and then drives it to longoff easy single dwayne bravo smith smith one slightly one run to to to to to to to to to to to to to to to to to to

---

The referral text is given in the first row of the above table. Keywords extracted by ontology are given in the second row and the final row of the Table 7.11 provides the machine generated text. More sample outputs are given in the Appendix 4.

### **7.5 Correlating Automatic and Human Evaluation**

BLEU focuses on precision and it is used for measuring the fluency of generated texts. ROUGE is often used to measure recall value and provides information about completeness or the level of satisfaction the generated text provides when compared with the reference text.

The generated news may not be exactly similar to reference news since the neural net model is learning from the entire dataset. So comparison with the reference news by automatic evaluation is not an apt procedure for news generation. BLEU and ROUGE metrics are only document similarity measures and its score depends on the reference text. Due to these limitations of automatic evaluation metrics while dealing with computer generated text, the process of testing machine news becomes difficult. There is a necessity for automatic evaluation results to be cross verified. The quality of the news being generated can be checked if it is actually useful to the user of the system. Manual evaluation is therefore important for language generation model.

Four cases are tested in this evaluation process.

- Char-RNN model with BBC news dataset
- Char-RNN model with Cricket Commentary dataset
- Seq2Seq model with BBC news dataset
- Seq2Seq model with Cricket Commentary dataset

When BLEU and ROUGE score of the above four cases are examined, they clearly indicate that the Seq2Seq model with the Cricket Commentary dataset is the best performing case among the above mentioned cases. Another observation that can be derived from automatic evaluation results is that the Seq2Seq model is better suited for news generation framework than the Char-RNN model. Human Evaluators also rated the text generated by the Seq2Seq model with the Cricket Commentary dataset as the best case in terms of Fluency, Adequacy and Total Quality. Human evaluators assessed the Sequence to Sequence model as the apt model for the proposed news generation framework.

The approach chosen in this research is to evaluate the two models with automatic and human evaluation. Then the results are correlated to draw a conclusion. Here automatic and human evaluation results correlate with each other and provide the same results. The tested cases are ranked in the following order given below as per the evaluation approach followed in this study.

- 1) Seq2Seq model with Cricket Commentary dataset

- 2) Seq2Seq model with BBC news dataset
- 3) Char-RNN model with Cricket Commentary dataset
- 4) Char-RNN model with BBC news dataset

## **7.6 Discussion and Inferences**

As per the evaluation results, the Sequence to Sequence model outperforms the Char-RNN model. Due to the limitations of Open Calais ontology, an ontology in Cricket domain is developed and used along with the Open Calais to build the Cricket Commentary dataset. The objective was to make more keywords from the commentary and to study how the quality of the generated text is affected by the number of keywords. This approach was not used with the BBC news dataset due to the practical difficulty in creating custom ontology in different topics which are included in the BBC news dataset. A detailed discussion is done in the following sections based on the test results.

### **7.6.1 Char-RNN-LSTM Model**

One of the advantages of the Char-RNN model is its smaller discrete working space. The number of characters in English language including all punctuation marks is only less than 100. Character based models may generate uncommon words with small probability. Some of them are meaningful words whereas some words can be utter nonsense too. Since the generation is based on characters, there may be plenty of words with incorrect spelling and generated text may not be meaningful. Since the working space is small (97 English characters), the required

computational power is reasonable. It is the simplest model in the domain of Recurrent Neural Network and so it is pretty simple to learn. Char-RNN does not have the skill to scrutinise spelling in addition to syntax, semantics etc., and it is one of the major limitation of Char-RNN to be used in the language generation domain.

The character based models may perform well provided that the dataset should be big enough and require more hidden layers for training. It should be trained for a longer period of time too. But the above mentioned changes will require more computational power.

- Char-RNN-LSTM model with BBC news dataset and Cricket Commentary dataset

The Char-RNN model was trained for 10000 epochs and after 5000 epochs, there was no considerable improvement in the performance. The test results clearly show that the model is working better with the Cricket Commentary dataset. Automatic evaluation metric scores are correlated with the human evaluation. One of the reasons is that the Cricket Commentary dataset has lesser number of unique words compared to the BBC news dataset. Cricket Commentary dataset only deals with the Cricket domain and the details are about twenty-twenty matches in one Indian Premier League season. So prediction space is low. Whereas in BBC news dataset, it consists of several topic areas like politics, technology, sports etc.



Another important reason is the increase in the number of keywords in the Cricket Commentary dataset compared to the BBC news dataset. Since keywords are used for generating commentary, the number of keywords is a prominent factor influencing the text generation.

### **7.6.2 Sequence to Sequence Model**

Word RNN is used in the Sequence to Sequence model used in this work. When compared to Char-RNN, here discrete working space is large since vocabulary consists of thousands of words. In Char-RNN, only less than 100 characters needs to be addressed. Since there are fewer characters than words, it results in smaller input space. Word-RNN requires huge memory compared to Char-RNN because of the large size of vocabulary.

Word RNN has an important edge over character-level models. The Word RNN takes input sequence as words whereas Char-RNN takes it as characters. Character based models have to consider more dependencies over more time steps than word based models. So the learning task is difficult for character model than the word model. Obviously, word language models result in lower number of spelling errors than character models.

If you want to generate a text based on words, then Word RNN may be the right option. Misspelled words or other words not appearing in the vocabulary are treated as special "unknown" tokens in Word

RNN. So it is evident that word RNN models are not flexible. Choosing the right neural net model depends on certain points like dataset size, the languages and the pre-processing required. The Word RNN models take more time for the training process and generate more consistent texts compared to Char-RNN model. Choosing Word RNN or Char-RNN is more of a trial and error decision as well as a trade-off between data and computational power available.

- Sequence to Sequence Model with BBC news dataset and Cricket Commentary dataset

The model was trained for 2500 epochs and the testing results clearly indicate that the model works better with the Cricket Commentary dataset. One reason is that the vocabulary of the Cricket Commentary dataset is smaller compared to the BBC news dataset. Another reason is the number of keywords for generation and training. The Cricket Commentary dataset has more keywords than the BBC news dataset. Data on a single topic is covered in Cricket Commentary dataset whereas in BBC news dataset, there are news on different topics. It is clearly understandable from the output that the inefficiency in extracting more number of keywords by the ontology from the original news hindered the skill of the neural net model to understand word relations.

### **7.6.3 Comparing Char-RNN and Sequence to Sequence Model**

The inputs given in the Seq2Seq approach for generating news are keywords derived by the ontology from the original news. The model

was expected to perform better as ontology keywords are inputs and Word RNN is used in Sequence to Sequence model instead of Char-RNN. Seq2Seq model overtakes the Char-RNN model when its performance is considered. After 1000 epochs itself, the Sequence to Sequence model shows better result than the Char-RNN model.

## **7.7 Conclusion**

This Chapter sums up the experimental results. The chapter provided a discussion based on test results towards the end. An automatic evaluation was done using BLEU and ROUGE metrics. Then the results were compared with human evaluation results. BBC news dataset and Cricket Commentary dataset were used in training and testing the framework. Char-RNN model was trained for 10000 epochs while Seq2Seq model was trained for 2500 epochs. Seq2Seq model took more time for training compared to Char-RNN model due to the complexity of the Sequence to Sequence model. BLEU and ROUGE scores showed that both models performed better with the Cricket Commentary dataset than the BBC news dataset. Human ratings also provided the same results. The experiment raises the necessity for sufficient number of ontology tags / keywords for the proposed news generation framework to perform better. As per the test results, Seq2Seq model was the suitable neural net model for the proposed news generation framework. The next Chapter summarizes the present study and its outcome.

.....❧.....



## SUMMARY AND CONCLUSION

- 8.1 Introduction
- 8.2 Summary
- 8.3 Implications of the Study
- 8.4 Contributions
- 8.5 Limitations
- 8.6 Suggestions for Future Research
- 8.7 Conclusion

### 8.1 Introduction

This Chapter summarises and concludes the thesis. Section 8.2 provides an overview of the thesis, summarising the research motivation and objectives. Section 8.3 discusses the implications of the research outcomes reported in this work. Section 8.4 includes the academic, community and industrial contributions of this study. Limitations of the present research is mentioned in Section 8.5 and future directions to enhance the work are detailed in Section 8.6. Finally, a conclusion is presented in section 8.7.

### 8.2 Summary

This section recapitulates the essence of this work. The main objective was to evolve a news generation framework using keywords

mined by ontology. A secondary objective of this study was overcoming the limitations of statistical models, traditionally used in natural language generation tasks. N-gram models were used mostly in statistical modelling. One defect of N-gram models was the performance degradation with the increasing amount of data training. Obviously, it required a large memory. Another limitation of these models was the limited power of handling long histories of input. Long-term dependencies are significant for an efficient text generation model since the connections between words often span across several sentences. Recurrent neural network language modelling provides a far better performance in the aspects mentioned above than statistical models. The model selected for this research was Recurrent Neural Network language model to avoid these drawbacks.

Two neural net models were studied out of which sequence to sequence model provided a better result in the technical environment specified in this work. One of the hurdles faced in achieving this research objective in the implementation stage was computational power and the issue of getting a suitable dataset. The model may get a better chance to learn semantic relations if a huge number of interrelated news stories are incorporated into the dataset. The evaluation stage posed another problem. The evaluation task was hectic since the models were generative models. Though it was possible to identify how close the automatically generated news was to the human generated news based on BLEU and

ROGUE scores, a human evaluation was still required due to the limitation of automatic evaluation techniques.

So the evaluation approach used metrics like BLEU, ROGUE and human evaluation. Then the results were correlated to derive a conclusion.

Another major objective of this research work was to make a study with regard to the quality of the generated news and the number of keywords. Thereby, the framework was tested with two datasets and a number of keywords. Open Calais ontology extracted only minimum number of keywords from the BBC news dataset. So, a new Cricket term ontology was developed and used along with Open Calais ontology to extract more concept terms from a cricket commentary dataset. This process was inapplicable in the case of the BBC news dataset since it contained news of different topics and building comprehensive ontology for each domain was a massive task. The experiment showed that both the models studied, viz. the Char-RNN model and the Sequence to Sequence model performed well with the Cricket Commentary dataset than the BBC news dataset. The rate of success in the process of making semantically enriched machine news by using Cricket Commentary dataset showed that the standard of machine generated news depended on the number of keywords extracted by the ontology used.

There are plenty of applications developed for generating text using Neural Network and some of the significant works are discussed in the literature review chapter [55,56,57,58,59]. The approach followed in the above-mentioned works are different from the study discussed in this thesis and the major difference is in the input data. All these applications use text as input for training the neural net model. The framework proposed in this work follows a unique process in generating news. Here keywords are used as input for generating news and that makes this framework highly applicable and practical in any information extraction scenario. The power of ontology in representing domain knowledge is used here for the preparation of training data as well as for the news generation task too. Here, the keywords or tags extracted by the ontology from the training news data is used for generating news using the framework. Thus the framework proposed in this work is unique and performance comparison with existing text generation applications may not be appropriate.

One of the uniqueness of this work is the use of ontology in preparing training data for the neural net models used in the proposed framework. Ontology is an efficient knowledge representation tool and provides a structure for the training data. It is less ambiguous and scalable too when comparing with other techniques. So here, ontology is used for generating keywords and utilized as background knowledge for deep learning tasks. The news is generated by the architecture using the keywords extracted by the ontology from the news data.



Another novel fact that this research work emphasizes is the need for completeness property of the ontology which is used in the news generation framework. The Ontology used in this work is Open Calais and the limitation of this ontology in representing all concepts in the news domain was discussed in Chapter 4. The results show that both the neural net models discussed in this work provide better results with Cricket Commentary dataset than BBC news dataset. The reason for the better performance is that each news in the Cricket Commentary dataset has many keywords associated with it. The quality of the output depends on the number of keywords. The number of concept keywords extracted by the ontology increases when the ontology is comprehensively representing the knowledge in the specific domain. That is, a high degree of ontology completeness is significant in deciding the quality of the news generated using the proposed framework.

### **8.3 Implications of the Study**

The findings presented in this thesis have broader implications in terms of enhancing the current news creation as well as digital archiving process. News story creation is a time consuming task considering the availability of a huge volume of news stories and insufficient archiving tools. The journalist may not get the required news events related to the latest update from a first search of the archive. So he / she is forced to do several searches and time-consuming screening of the search results to get the apt materials. The process suggested in this research helps to improve this current scenario and would provide the user a customized

experience. By using the suggested framework, news story generation will become faster and easier to manage.

Providing useful information for the public good, in a timely fashion is a challenge in the development of responsive public services. The study conducted here had wide-ranging implications in this scenario. Information extraction from public archives using the framework developed in this work would provide potentially valuable information with a rich content to the public.

The research done here has implication on any digital archives like cultural heritage, traditional knowledge, indigenous medical knowledge etc. These domains have a significant role in the social life of humans, by preserving social characteristics and knowledge for the next generation. There is scope for a wide range of possible innovations using traditional knowledge that can contribute to the development of various domains. The framework developed as a part of this study can be effectively used in information extraction from digital archives.

Another implication of this study is that it emphasises the advantages of using ontology in every information extraction domain. The usage of ontology provides customization based on needs. The study proved that quality of news generated had increased with the number of keywords which emphasised the need for developing comprehensive ontology in any domain where there is a need for information extraction systems.

## **8.4 Contributions**

The significance and implications of this study in academic, societal and industrial arenas are pointed out here.

### **a) Contributions to the Academia:**

- Through this research, it was proven that carefully chosen keywords derived from ontology can be used to generate quality text. Further work can focus on improving the quality of the generated text by using different deep learning algorithms.
- The study emphasises on the necessity of ontology completeness in respective domains to become the knowledge effectively readable by machines.

### **b) Contributions to the Society:**

- The application developed in this research can be used in digital archives in public libraries. It will help the public to get the relevant information for their requirements
- The application developed in this study can be used in information kiosks to get updated details based on user requirements within a short time period.

### **c) Contributions to the Industry:**

- The results obtained are directly applicable in news media libraries and the journalist can get a customized experience when they search for news content.

- Considering the huge amount of news generated and the increasing number of news media houses, those news organizations which use the application developed in this research work will get a competitive platform over other news organizations.

## 8.5 Limitations

The results of this work were subject to some constraints which if solved, can result in fruitful developments. Though a few constraints are specific to the study, others hold hope for the possibility of future works if they are effectively solved.

One main clampdown specific to this research was the non-availability of suitable datasets. This challenge is common in deep learning research. Both dataset used in this work has less than 3000 news text. The BBC news dataset consisted of news with different topics too. As the number of inter-related news in the BBC news dataset are lesser, the learning process may suffer in comprehending the semantic relations between the news items. The availability of apt datasets may not be an issue in future since there are a lot of ongoing research work in deep learning. The contributions of these research works will improve the availability of suitable datasets.

The training details of Seq2Seq model with BBC news dataset showed that there is no considerable decrease in training loss after 500

epochs. That means the model with BBC news dataset is not converging as expected. It can happen due to several reasons including the selected dataset, algorithm issues, hyperparameter values etc. This can be overcome by hyperparameter tuning, selecting suitable algorithms and datasets etc. This issue was not concentrated in the present study since the main purpose was to construct a framework that can generate news from keywords derived by ontology.

The effectiveness of BLEU and ROUGE metric in measuring the quality of machine generated news is yet to be proved. It is evident from some of the experimental results of this work. For example, BLEU score of Sequence to Sequence model with Cricket Commentary dataset is very poor up to 600 epochs while ROUGE shows better value. It points to the need for a more suitable automated evaluation technique to judge language generation models.

Another challenge common in research which involves deep learning is the need for high computational power which would find a solution in the future with the fast pace of growth in the computer hardware field.

Though Open Calais ontology used for the evaluation can extract higher level concepts, it was less efficient in extracting lower level concepts which reduces the number of keywords and affects the news generation process. The unavailability of a comprehensive ontology especially in specific news domain demands the development of advanced ontological systems. The rapid development of technology

holds the promise for better innovations in ontologies and thereby provides a scope for better news generation systems.

## 8.6 Suggestions for Future Research

Presented here are a few ideas that can be used in the long run to enrich or extend the work mentioned in this thesis. The proposed framework effectively carries out its ability of news generation. The necessity of a suitable dataset is evident from the results obtained. The results are indeed promising, encouraging further research in this area. Yet with more training and hyperparameter tuning with suitable datasets, these systems would provide encouraging results.

With the advancements happening in the deep learning field as well as with rapid technological growth, there arises a need to research the possibility about making a hybrid model using different deep learning algorithms to upgrade the framework discussed in this research.

The requirement for an optimum automatic evaluation technique to measure the quality of machine generated text is evident from the present study. BLEU and ROUGE are state-of-the-art techniques in the machine translation field, but fails to prove their capability in assessing the quality of generated text. There is a scope for further research in this domain.

The adequacy of the generated news and its fluency can be improved by increasing the number of input ontology tags. The ontology that the

framework employs was not fully effective in extracting low level key words / concept terms. It is indeed hectic challenge to create a comprehensive ontology capable of doing this, particularly in the news domain. Different information generation domains can also utilise this research work to make their effort fruitful. So, developing comprehensive ontologies that are domain specific can be considered as one of this work's prime extensions.

## **8.7 Conclusion**

This final Chapter of the thesis summarizes the entire research work. It begins by providing an introduction in Section 8.1 regarding what all topics are broadly covered in this Chapter. In the next section 8.2, a summary was provided by revisiting the main objectives of this research work. The main objective was to evolve a news generation framework using keywords extracted by the ontology. Then this section mentioned secondary objectives including finding alternatives to the traditionally used statistical models. It was then mentioned and elaborated that neural network models were found to be a much more practical and effective alternative. The two neural net models studied in this work viz. Char-RNN model and Sequence-to-Sequence model were briefly explained. Finally, in this section, another key objective of this study viz. finding any possible correlation between the number of keywords and the quality of news generated, was mentioned. In Section 8.3, the implications of this study were mentioned, not only in news generation but in searching public archives in various domains.

Especially, the effectiveness of using an ontology, as found in this research, implies that ontology may be put to effective use in all domains where automated archiving and retrieval of knowledge are required. In Section 8.4, contributions possible from this study in three broad domains of Academia, Society, and Industry were listed out in detail. While academia can use this study to experiment further with better deep learning algorithms and better ontology completeness, society can use this framework in applications like digital archives and information kiosks. Industry use of this research will be mainly in media organizations, where implementing this framework would provide a competitive edge in digital archiving and automated news generation. In Section 8.5, the limitations of this study were elaborated upon, including the non-availability of suitable datasets and the need for better automated evaluation tools than BLEU and ROUGE. The fact that a greater completeness of the ontology can improve the results was stressed upon here. Some of the limitations would progressively get solved as greater computational power is developed and as newer technologies enable better ontology completeness. Finally, in Section 8.6, a few suggestions on future research directions based on this study were mentioned, with the prime suggestion being the development of detailed domain specific ontologies for greater success in similar applications.

*.....✎.....*



## References

- [1] M. Uschold and M. Gruninger, “Ontologies and semantics for seamless connectivity,” *ACM SIGMOD Record*, vol. 33, no. 4, p. 58, Jan. 2004.
- [2] Lobna Karoui, Marie-Aude Afaure, and Nacéra Bennacer Seghouani, “Ontology Discovery from Web Pages: Application to Tourism,” *Workshop W6 on Knowledge Discovery and Ontologies*, pp. 115–120, Sep. 2004. ECML/PKDD 2004.
- [3] T. Berners-Lee, J. Hendler, and O. Lassila, “The Semantic Web,” *Scientific American*, vol. 284, no. 5, pp. 34–43, 2001.
- [4] Semantic Web stack. [Online]. Available: [https://en.wikipedia.org/wiki/Semantic\\_Web\\_Stack](https://en.wikipedia.org/wiki/Semantic_Web_Stack). [Accessed: 7-Sep-2018].
- [5] P. Hitzler, “Foundations of Semantic Web Technologies,” Jun. 2009.
- [6] Search RDF data with SPARQL,” IBM - United States, 10-May-2005. [Online]. Available: <https://www.ibm.com/developerworks/xml/library/j-sparql/>. [Accessed: 5-Sep-2018].
- [7] Neches R, Fikes RE, Finin T, Gruber TR, Senator T, Swartout WR, “Enabling technology for knowledge sharing,” vol. 12(3): 1991.
- [8] A. Gomez-Perez, and V. R. Benjamins, “Overview of Knowledge Sharing and Reuse Components: Ontologies and Problem-Solving Methods,” *Workshop on Ontologies and Problem-Solving Methods: Lessons Learned and Future Trends*, OAD:1999.
- [9] N. Guarino, “Formal Ontology and Information Systems,” 1st International Conference on Formal Ontology in Information Systems (FOIS’98). Trento, Italy. IOS Press, Amsterdam, pp. 3–15, 1998.

## *References*

---

- [10] S. Staab, and R. Studer, and Springer Verlag, Handbook on Ontologies. International Handbooks on Information Systems. 2004.
- [11] R. Studer, V. Benjamins, and D. Fensel, “Knowledge engineering: Principles and methods,” *Data & Knowledge Engineering*, vol. 25, no. 1-2, pp. 161–197, 1998.
- [12] T. Gruber, Ling Liu and M. Tamer Özsu, and Springer-Verlag, “Encyclopedia of Database Systems”, 2008.
- [13] T. R. Gruber, “A translation approach to portable ontology specifications,” *Knowledge Acquisition*, vol. 5, no. 2, pp. 199–220, 1993.
- [14] W. Borst, and University of Twente, “Construction of Engineering Ontologies for Knowledge Sharing and Reuse,” Ph.D.Dissertation, 1997.
- [15] D. Fensel, “Ontologies,” *Dynamic Networks of Formally Represented Meaning*, 2001.
- [16] “Ontology Components.” [Online]. Available: [https://en.wikipedia.org/wiki/Ontology\\_components](https://en.wikipedia.org/wiki/Ontology_components). [Accessed: 15-Sep-2018].
- [17] P. Velardi, R. Navigli, A. Cucchiarelli, and F. Dantonio, “From Glossaries to Ontologies,” *Extracting Semantic Structure from Textual Definitions*, 2008.
- [18] N. Guarino, van Heijst, Schreiber, and Wielinga, “Understanding, Building, and Using Ontologies,” *Using Explicit Ontologies in KBS Development*, vol. (46): pp. 293-310.
- [19] A. Bouziane, D. Bouchiha, N. Doumi, and M. Malki, “Question Answering Systems: Survey and Trends,” *Procedia Computer Science*, vol. 73, pp. 366–375, 2015.

- [20] N. Ferran, E. Mor, and J. Minguillón, “Towards personalization in digital libraries through ontologies,” *Library Management*, vol. 26, no. 4/5, pp. 206–217, 2005.
- [21] H. Nguyen, “Trends in digital library research,” a knowledge mapping and ontology engineering approach. [Online]. Available: repository.vnu.edu.vn/bitstream/VNU\_123/325/1/Nguyen Hoang Son.pdf . [Accessed: 18-Sep-2018].
- [22] M. Caudill, “Neural Network Primer,” Part I. *AI Expert*, vol. 2, no. 12, Feb. 1989.
- [23] Artificial Neuron Model, 12-Feb-2018. [Online]. Available: <http://andrewjamesturner.co.uk/images/ArtificialNeuronModel.png> [Accessed: 21-Aug-2018].
- [24] N. Patel, “Artificial Neural Networks.” [Online]. Available: <https://ocw.mit.edu/courses/sloan-school-of-management/15-062-data-mining-spring-2003/lecture-notes/NeuralNet2002.pdf>. [Accessed: 06-Oct-2018].
- [25] J. Burger, “A Basic Introduction To Neural Networks.” [Online]. Available: <http://pages.cs.wisc.edu/~bolo/shipyard/neural/local.html>. [Accessed: 02-Aug-2018].
- [26] Z. C. Lipton, “A Critical Review of Recurrent Neural Networks for Sequence Learning,” 2015. [Online]. Available: <http://arxiv.org/abs/1506.00019>. [Accessed: 15-Aug-2018]. CoRR, volume abs/1506.00019.
- [27] Y. Bengio, P. Simard, and P. Frasconi, “Learning long-term dependencies with gradient descent is difficult,” *IEEE Transactions on Neural Networks*, vol. 5, no. 2, Mar. 1994. pp. 157–166, ISSN 1045-9227, doi:10.1109/72.279181.

## References

---

- [28] S. Hochreiter, J. Schmidhuber, “Long Short-Term Memory,” *Neural Computation*. [Online]. Available: <http://dx.doi.org/10.1162/neco.1997.9.8.1735>. [Accessed: 18-Aug-2018]. volume 9, no. 8, Nov. 1997: pp. 1735–1780, ISSN 0899-7667.
- [29] Martin Sundermeyer., Ralf Schlüter and Hermann Ney, “LSTM neural networks for language modelling,” 13th annual conference of the international speech communication association, Portland, Oregon, USA, pp. 194–197, 2012.
- [30] “long short term memory,” 14-May-2015. [Online]. Available: <http://blog.otoro.net/2015/05/14/long-short-term-memory/>. [Accessed: 06-Sep-2018].
- [31] Junyoung Chung et. al, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” CoRR, preprint arXiv:1412.3555, 2014.
- [32] Cho, Kyunghyun, van Merriënboer, Bart, Gulcehre, Caglar, Bougares, Fethi, Schwenk, Holger, and Bengio, Y, “Learning Phrase Representations using RNN,” *Encoder-Decoder for Statistical Machine Translation*, 10.3115/v1/D14-1179, 2014.
- [33] Sutskever, Ilya, Vinyals, Oriol, and V. Le, Quoc, “Sequence to Sequence Learning with Neural Networks,” vol. 4, pp. 3104–3112, 2014.
- [34] “Modeling a billion words.” [Online]. Available: <http://torch.ch/blog/2016/07/25/nce.html>. [Accessed: 12-Jul-2018].
- [35] “Understanding LSTM Networks — colah’s blog | Spurensucher ...” [Online]. Available: <https://peterschuthblog.wordpress.com/2018/04/01/understanding-lstm-networks-colahs-blog/>. [Accessed:12-Oct-2018].

- [36] “Speech Recognition.” [Online]. Available: <https://gab41.lab41.org/speechrecognition-you-down-with-ctc-8d3b558943f0>. [Accessed: 13-Apr-2018].
- [37] A. Graves, A. R. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 6645–6649, 2013.
- [38] Sami Jokela, Marko Turnpeinen, Teppo Kurki, Eerika Savia, and Reijo Sulonen, “The Role of Structured Content in a Personalised News Service,” 34th Hawaii International Conference on System Sciences, pp. 1–10, 2001.
- [39] Liliana Ardissono, Luca Console, and Iliara Torre, “An Adaptive System for the Personalised Access to News, in AI Communications,” pp. 129–147, 2001.
- [40] David Vallet, Phivos Mylonas, Miguel A. Corella<sup>1</sup>, José M. Fuentes<sup>1</sup>, Pablo Castells<sup>1</sup>, and Yannis Avrithis<sup>2</sup>, “A Semantically-enhanced personalization framework for knowledge driven media services,” 2005.
- [41] Yannis Kalfoglou, John Domingue, Enrico Motta, Maria Vargas-Vera, Simon Buckingham Shum, and myPlanet, “an ontology-driven Web-based personalised news service,” 2001.
- [42] P. Bellekens, L. Aroyo, G. J. Houben, A. Kaptein, and K. V. D. Sluijs, “Semantics-Based Framework for Personalized Access to TV Content: The iFanzzy Use Case,” The Semantic Web Lecture Notes in Computer Science, pp. 887–894, 2007.
- [43] D. Vallet, P. Castells, M. Fernandez, P. Mylonas, and Y. Avrithis, “Personalized Content Retrieval in Context Using Ontological Knowledge,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 17, no. 3, pp. 336–346, 2007.

## *References*

---

- [44] Borsje, Jethro, Levering, Leonard, Embregts, Hanno and Frasinca, Flavius Frasinca, Erasmus University Rotterdam, and Hermes, “an Ontology-Based News Personalization Portal,” Jan. 2007.
- [45] W. Ijntema, F. Goossen, F. Frasinca, and F. Hogenboom, “Ontology-based news recommendation,” in Proceedings of the 1st International Workshop on Data Semantics - DataSem 10, 2010. Lausanne, Switzerland.
- [46] B. Shapira, P. Shoval, N. Tractinsky, and J. Meyer, “ePaper: A personalized mobile newspaper,” *Journal of the American Society for Information Science and Technology*, vol. 60, no. 11, pp. 2333–2346, 2009.
- [47] I. Cantador, A. Bellogín, and P. Castells, “Ontology-Based Personalised and Context-Aware Recommendations of News Items,” 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2008.
- [48] F. Gao, Y. Li, L. Han, “InfoSlim: an ontology-content based personalized mobile news recommendation system,” 5th International Conference on Wireless Communications Networking and Mobile Computing (WiCom), vol. 41, pp. 1–4, 2009.
- [49] S. F. Chen and J. Goodman, “An empirical study of smoothing techniques for language modeling,” Proceedings of the 34th annual meeting on Association for Computational Linguistics -, 1996.
- [50] P. Koehn, F. J. Och, and D. Marcu, “Statistical phrase-based translation,” Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL 03, 2003.

- [51] R. Jozefowicz, O. Vinyals, M. Schuster, N. Shazeer and Y. Wu, “Exploring the limits of language modeling,” arXiv preprint arXiv:1602.02410, 2016.
- [52] D. Zhou, L. Guo, and Y. He, “Neural Storyline Extraction Model for Storyline Generation from News Articles,” Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 1727–1736, 2018.
- [53] H. T. Zheng, W. Wang, W. Chen and A. K. Sangaiah, “Automatic Generation of News Comments Based on Gated Attention Neural Networks,” in IEEE Access, vol. 6, pp. 702–710, 2018.
- [54] K. Park, Lee, Jisoo, and Choi, Jaeho, “Deep Neural Networks for News Recommendations,” Conference ACM, pp. 2255–2258, 2017.
- [55] K. Lopyrev "Generating news headlines with recurrent neural networks.," arXiv preprint arXiv, 2015.
- [56] R. Lebret, D. Grangier, and M. Auli, “Neural Text Generation from Structured Data with Application to the Biography Domain,” Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016.
- [57] R. Collobert and J. Weston, “A unified architecture for natural language processing,” Proceedings of the 25th international conference on Machine learning - ICML 08, pp. 160–167, 2008.
- [58] M. Roemmele and A.S. Gordon, “Creative help: a story writing assistant,” International Conference on Interactive Digital Storytelling, Springer International Publishing, pp. 81–92, 2015.

## *References*

---

- [59] I. Sutskever, J. Martens and G.E. Hinton, “Generating text with recurrent neural networks,” 28th International Conference on Machine Learning, pp. 1017–1024, 2011.
- [60] K. Uchimoto, H. Isahara and S. Sekine, “Text generation from keywords,” 19th international conference on Computational linguistics, vol. 1, no. 8, pp. 1–7, 2002.
- [61] Ayana, S. Q. Shen, Y.K. Lin, C.C. Tu, Y. Zhao, Z.Y. Liu, and M.S. Sun, “Recent Advances on Neural Headline Generation,” Journal of Computer Science and Technology, vol. 32, no. 4, pp. 768–784, 2017.
- [62] C. Wimalasuriya, Daya and Dou, Dejing: Ontology based information extraction: an introduction and a survey of current approaches, Journal of Information Science, Vol. 36, No. 3. pp. 306–323, 2010.
- [63] A. Karpathy: The unreasonable effectiveness of recurrent neural networks, (2015) <http://karpathy.github.io/2015/05/21/mn-effectiveness>.
- [64] A.P. Sheth, C. Ramakrishnan, and C.J. Thomas, “Semantics for the Semantic Web,” The Implicit, the Formal and the Powerful. International Journal on Semantic Web & Information Systems, vol. 1, pp. 1–18, 2005.
- [65] M. López “Advanced Information and Knowledge Processing Ontological Engineering,” Methodologies and Methods for Building Ontologies, IJCAI-99 workshop on Ontologies and Problem-Solving Methods (KRR5) Stockholm, Sweden, pp. 107–197, Aug. 1999.
- [66] S. H. Nguyen, “Trends in digital library research: a knowledge mapping and ontology engineering approach”, PhD thesis, University of Technology, Sydney, 2013.



- [67] “Home Page,” IPTC. [Online]. Available: <https://iptc.org/>. [Accessed: 02-Oct-2018].
- [68] “Open Calais,” Open Calais ontology. [Online]. Available: <http://www.opencalais.com/>. [Accessed: 29-Sep-2018].
- [69] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan and S. Vijaya-narasimhan, “YouTube-8M: A large-scale video classification benchmark,” arXiv preprint, arXiv:1609.08675, 2016.
- [70] M. Blakeney, “Protection of Traditional Knowledge by Geographical Indications,” *International Journal of Intellectual Property Management*, vol. 3, no. 4, p. 357, 2009.
- [71] A. Rozeva and S. Zerkova, “Assessing semantic similarity of texts – Methods and algorithms,” 43rd International Conference Applications of Mathematics in Engineering and Economics AIP Conference Proceedings1910, 060012, 2017.
- [72] Boling, Chelsea and Das, Kumer, “Semantic Similarity of Documents Using Latent Semantic Analysis,” National Conference On Undergraduate Research (NCUR) University of Kentucky, Lexington, Apr. 2014.
- [73] C. M. Bishop, “Neural networks for pattern recognition,” Oxford university press, 1995.
- [74] Z. Shi, X. Chen, X. Qiu, and X. Huang, “Toward Diverse Text Generation with Inverse Reinforcement Learning,” *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pp. 4361–4367, 2018.

## References

---

- [75] K. Papineni, S. Roukos, T. Ward, and W.J. Zhu, “Bleu,” Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL 02, pp. 311–318, 2001.
- [76] C.Y. Lin and E. Hovy, “ROUGE: A Package for Automatic Evaluation of summaries,” Proceedings of the ACL Workshop on Text Summarization Branches Out, vol. 10, 2004.
- [77] J. Steinberger and K. Ježek, “Evaluation Measures for Text Summarization, Computing and Informatics,” Proceedings of the 9th ACM symposium on Document engineering - DocEng 09, vol. 28, pp. 251–275, 2009.
- [78] D. Greene, and P. Cunningham, “Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering,” in Proc. 23rd International Conference on Machine learning, ICML '06, ACM, New York, NY, USA, pp. 377–384, 2006.
- [79] “Indian Premier League, 2016 schedule, live scores and results,” Cricbuzz. [Online]. Available: <https://www.cricbuzz.com/cricket-series/2430/indian-premier-league-2016/matches>. [Accessed: 10-Oct-2018].
- [80] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, “End-to-End Attention-Based Large Vocabulary Speech Recognition,” 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4945–4949, 2016, doi:10.1109/icassp.2016.7472618.
- [81] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, and G. Chen, “Deep Speech 2: End-to-End Speech Recognition in English and Mandarin,” In International Conference on Machine Learning, pp. 173–182, 2016.

- [82] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention,” In *International Conference on Machine Learning*, pp. 2048–2057, 2015.
- [83] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and Tell: A Neural Image Caption Generator,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3156–3164, 2015.
- [84] A. M. Rush, S. Chopra, and J. Weston, “A Neural Attention Model for Abstractive Sentence Summarization,” *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 379–389, 2015.
- [85] S. Chopra, M. Auli, A. M. Rush, and S. Harvard, “Abstractive Sentence Summarization with Attentive Recurrent Neural Networks,” In *NAACL-HLT*, pp. 93–98, 2016.
- [86] A. See, P. J. Liu, and C. D. Manning, “Get to the Point: Summarization with Pointer-Generator Networks,” In *Association for Computational Linguistics*, vol. 1, pp. 1073–1083, 2017.
- [87] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” *arXiv preprint arXiv:1409.0473*, 2014.
- [88] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using rnn encoder-decoder for statistical machine translation,” *arXiv preprint arXiv:1406.1078*, 2014.
- [89] T. Luong, H. Pham, and C. D. Manning, “Effective Approaches to Attention-Based Neural Machine Translation,” In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1412–1421, 2015.

## References

---

- [90] R. Nallapati, F. Zhai, and B. Zhou, “Summarunner: A Recurrent Neural Network Based Sequence Model for Extractive Summarization of Documents,” In *AAAI*, pp. 3075–3081, 2017.
- [91] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. Rush, “Opennmt: Open-Source Toolkit for Neural Machine Translation,” *Proceedings of Association for Computational Linguistics, System Demonstrations*, pp. 67–72, 2017.
- [92] M. X. Chen, O. Firat, A. Bapna, M. Johnson, W. Macherey, G. Foster, L. Jones, N. Parmar, M. Schuster, and Z. Chen, “The Best of Both Worlds: Combining Recent Advances in Neural Machine Translation,” pp. 76–86, 2018.
- [93] J. Ling and A. Rush, “Coarse-to-fine attention models for document summarization,” in *Proceedings of the Workshop on New Frontiers in Summarization*, 2017, pp. 33–42, 2018.
- [94] Q. Zhou, N. Yang, F. Wei, and M. Zhou, “Selective Encoding for Abstractive Sentence Summarization,” In *Association for Computational Linguistics*, vol. 1, no. 9, pp. 1095–1104, 2017.
- [95] J. Tan, X. Wan, and J. Xiao, “Abstractive document summarization with a graph-based attentional neural model,” in *Association for Computational Linguistics*, vol. 1, pp. 1171–1181, 2017.
- [96] Y. C. Chen and M. Bansal, “Fast abstractive summarization with reinforce-selected sentence rewriting,” in *ACL*, vol. 1, pp. 675–686, 2018.
- [97] J. Lin, X. Sun, S. Ma, and Q. Su, “Global encoding for abstractive summarization,” in *ACL*, pp. 163–169, 2018.

- [98] W. T. Hsu, C. K. Lin, M. Y. Lee, K. Min, J. Tang, and M. Sun, “A unified model for extractive and abstractive summarization using inconsistency loss,” in *ACL*, pp. 132–141, 2018.
- [99] Y. Xia, F. Tian, L. Wu, J. Lin, T. Qin, N. Yu, and T.Y. Liu, “Deliberation networks: Sequence generation beyond one-pass decoding,” in *Advances in Neural Information Processing Systems*, pp. 1782–1792, 2017.
- [100] I. V. Serban, A. Garc’ia-Duran, C. Gulcehre, S. Ahn, S. Chandar, A. Courville, and Y. Bengio, “Generating factoid questions with recurrent neural networks: The 30m factoid question-answer corpus,” in *Association for Computational Linguistics*, vol. 1, pp. 588–598, 2016.
- [101] N. Mostafazadeh, I. Misra, J. Devlin, M. Mitchell, X. He, and L. Vanderwende, “Generating natural questions about an image, ” in *Association for Computational Linguistics*, vol. 1, pp. 1802– 1813, 2016.
- [102] Z. Yang, J. Hu, R. Salakhutdinov, and W. Cohen, “Semisupervised qa with generative domain-adaptive nets,” in *Association for Computational Linguistics*, vol. 1, pp. 1040–1050, 2017.
- [103] X. Yuan, T. Wang, C. Gulcehre, A. Sordoni, P. Bachman, S. Zhang, S. Subramanian, and A. Trischler, “Machine comprehension by text-to-text neural question generation,” in *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pp. 15–25, 2017.
- [104] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, “VQA: Visual question answering,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2425–2433, 2015.
- [105] C. Xiong, V. Zhong, and R. Socher, “Dynamic coattention networks for question answering,” in *ICLR*, 2016.

## References

---

- [106] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, “Stacked attention networks for image question answering,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 21–29, 2016.
- [107] “Dialogue system,” Wikipedia, 03-Aug-2018. [Online]. Available: [https://en.wikipedia.org/wiki/Dialogue\\_system](https://en.wikipedia.org/wiki/Dialogue_system). [Accessed: 21-Oct-2018].
- [108] Jiwei Li, Alan Ritter, and Will Monroe, “Deep Reinforcement Learning for Dialogue Generation,” in Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Jan. 2016.
- [109] A. Bordes, Y.L. Boureau, and J. Weston, “Learning end-to-end goal-oriented dialog,” arXiv preprint arXiv:1605.07683, 2016.
- [110] J. Li, W. Monroe, T. Shi, S. Jean, A. Ritter, and D. Jurafsky, “Adversarial Learning for Neural Dialogue Generation,” Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017.
- [111] V. Zhong, C. Xiong and R. Socher, “Seq2SQL: Generating structured queries from natural language using reinforcement learning,” 2018. [Online]. Available: <https://openreview.net/forum?id=Syx6bz-Ab>. [Accessed: 03-Oct-2018].
- [112] X. Xu, C. Liu and D. Song, “Sqlnet: Generating structured queries from natural language without reinforcement learning,” arXiv preprint arXiv:1711.04436, 2017.
- [113] R. Kiros, R. Salakhutdinov, and R. Zemel, “Multimodal neural language models,” in International Conference on Machine Learning, pp. 595–603, 2014.

- [114] R. Kiros, R. Salakhutdinov, and R. S. Zemel, “Unifying visual semantic embeddings with multimodal neural language models,” arXiv preprint arXiv:1411.2539, 2014.
- [115] X. Chen and C. L. Zitnick, “Mind’s eye: A recurrent visual representation for image caption generation,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 2422–2431, 2015.
- [116] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, “Deep captioning with multimodal recurrent neural networks (m-rnn),” arXiv preprint arXiv:1412.6632, 2014.
- [117] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollar, J. Gao, X. He, M. Mitchell, J. C. Platt, “From captions to visual concepts and back,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 1473–1482, 2015.
- [118] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, “Long-term recurrent convolutional networks for visual recognition and description,” in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2625–2634, 2015.
- [119] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko, “Translating videos to natural language using deep recurrent neural networks,” in NAACL-HLT, pp. 1494–1504, 2015.
- [120] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, “Sequence to sequence-video to text,” in Proceedings of the IEEE international conference on computer vision, pp. 4534–4542, 2015.

## References

---

- [121] S. Venugopalan, L. A. Hendricks, R. Mooney, and K. Saenko, “Improving lstm-based video description with linguistic knowledge mined from text,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1961–1966, 2016.
- [122] A. Graves and N. Jaitly, “Towards end-to-end speech recognition with recurrent neural networks,” in *International Conference on Machine Learning*, pp. 1764–1772, 2014.
- [123] Y. Miao, M. Gowayed, and F. Metze, “Eesen: End-to-end speech recognition using deep rnn models and wfst-based decoding,” in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, pp. 167–174, 2015.
- [124] Minnen, Guido, John Carroll and Darren Pearce, “Applied morphological processing of English. *Natural Language Engineering*,” vol. 7, no. 3, pp. 207–223, 2001.
- [125] “Product Overview,” Amazon. [Online]. Available: <https://aws.amazon.com/marketplace/pp/B077GCH38C>. [Accessed: 08-Oct-2018].
- [126] H. Jang and K. Jung, “Neural Network Implementation Using CUDA and OpenMP,” *The Community for Technology Leaders • IEEE Computer Society*, 01-Dec-2008. [Online]. Available: <https://www.computer.org/csdl/proceedings/dicta/2008/3456/00/3456a155-abs.html>. [Accessed: 16-Oct-2018].
- [127] “What are the Benefits and Limitations of Using Python? | edu CBA,” EDUCBA, 10-Oct-2018. [Online]. Available: <https://www.educba.com/benefits-and-limitations-of-using-python/>. [Accessed: 09-Oct-2018].



- [128] S. Deoras, “Tensorflow Vs Theano : What Do Researchers Prefer As An AI Framework,” *Analytics India Magazine*, 17-Aug-2018. [Online]. Available: <https://www.analyticsindiamag.com/tensorflow-vs-theano-researchers-prefer-artificial-intelligence-framework/>. [Accessed: 18-Oct-2018].
- [129] “Using the Deep Learning AMI with Conda,” Amazon. [Online]. Available: <https://docs.aws.amazon.com/dlami/latest/devguide/tutorial-conda.html>. [Accessed: 07-Oct-2018].
- [130] D. Stamat, “AWS EC2: P2 vs P3 instances,” *The Iron.io Blog*, 02-Nov-2017. [Online]. Available: <https://blog.iron.io/aws-p2-vs-p3-instances/>. [Accessed: 05-Oct-2018].
- [131] “Amazon EC2 - P2 Instances,” Amazon. [Online]. Available: <https://aws.amazon.com/ec2/instance-types/p2/>. [Accessed: 20-Oct-2018].
- [132] GlobCon Technologies, “TensorFlow: Why Google's Artificial Intelligence Engine is a Gamechanger,” *Analytics India Magazine*, 30-Aug-2017. [Online]. Available: <https://www.analyticsindiamag.com/tensorflow-googles-artificial-intelligence-engine-gamechanger/>. [Accessed: 14-Oct-2018].
- [133] “Deep Learning Base AMI,” Amazon. [Online]. Available: <https://docs.aws.amazon.com/dlami/latest/devguide/overview-base.html>. [Accessed: 10-Oct-2018].
- [134] “Deep Learning AMI with Conda,” Amazon. [Online]. Available: <https://docs.aws.amazon.com/dlami/latest/devguide/overview-conda.html>. [Accessed: 21-Oct-2018].
- [135] R. Singleton, and B. C. Straits, “Approaches to social research,” Oxford University Press, New York, 1999.

## *References*

---

- [136] E. Manishina, “Data-driven Natural Language Generation Using Statistical Machine Translation and Discriminative Learning,” Ph.D. thesis, University of Avignon, 2016.
- [137] Ian Goodfellow, Yoshua Bengio and Aaron Courville, “Deep Learning,” MIT Press, 2016.
- [138] Christopher M. Bishop, “Pattern Recognition and Machine Learning (Information Science and Statistics),” Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.
- [139] Alex Graves, “Generating sequences with recurrent neural networks,” arXiv preprint arXiv:1308.0850, 2013.
- [140] Will Williams, Niranjani Prasad, David Mrva, Tom Ash and Tony Robinson, “Scaling Recurrent Neural Network Language Models,” arXiv: 1502.00512, pp. 2–6, 2015.
- [141] Yoshua Bengio, Réjean Ducharme, Pascal Vincent and Christian Jauvin, “A neural probabilistic language model,” *Journal of machine learning research*, vol. 3, pp. 1137–1155, Feb. 2003.
- [142] Tomas Mikolov, M. Karafiat, L. Burget, J. Cernocky and S. Khudanpur, “Recurrent Neural Network based Language Model,” *Interspeech*, pp. 1045–1048, Sep. 2010.
- [143] Rui Ren, Lingling Zhang, Limeng Cui, Bo Deng and Yong Shi, “Personalized Financial News Recommendation Algorithm Based on Ontology,” *Procedia Computer Science*, vol. 55, pp. 843–851, 2015, ISSN 1877-0509.
- [144] David Werner, Christophe Cruz, and Christophe Nicolle, “Ontology-based Recommender System of Economic Articles,” *CoRR*, abs/1301.4781, 2013.

- [145] T. Fanaee, Hadi, Yazdi, and Mehran, “A Novel Ontology-based Recommender System for Online Forums,” 10.13140/RG.2.1.2633.0724, 2011.
- [146] H.Cui, M. Zhu and S. Yao “Ontology-based Top-N Recommendations on New Items with Matrix Factorization,” JSW, vol. 9, pp. 2026–2032, 2014.
- [147] Thanapalasingam, Thiviyan; Osborne, Francesco; Birukou, Aliaksandr and Motta, Enrico, “Ontology-Based Recommendation of Editorial Products,” In: International Semantic Web Conference (ISWC), Monterey, California, United States, pp. 08–12, 2018.
- [148] Altuncu, M. Tarik, Sophia N. Yaliraki and Mauricio Barahona, “Content-driven, unsupervised clustering of news articles through multiscale graph partitioning,” *CoRR abs/1808.01175*, 2018.
- [149] B. Ulicny, G. M. Powell, C. J. Matheus, M. Coombs and M. M. Kokar, “Priority Intelligence Requirement Answering and Commercial Question-Answering: Identifying the Gaps,” *DTIC Document*, 2010.
- [150] Young, Tom, Devamanyu Hazarika, Soujanya Poria and Erik Cambria, “Recent Trends in Deep Learning Based Natural Language Processing [Review Article],” *IEEE Computational Intelligence Magazine* 13, pp. 55–75, 2018.
- [151] I. J. Goodfellow, “NIPS 2016 tutorial: Generative adversarial networks,” *CoRR, abs/1701.00160*, 2017.
- [152] Zhang, Lixue, Wen Cui, S. Tong, Ran Xu and Yifeng Liu, “Neural Models Comparison in Natural Language Generation,” *Zhang2017NeuralMC*, pp. 01–08, 2017.

## References

---

- [153] Anna-lena popkes, “Language modeling with recurrent neural networks Using Transfer Learning to Perform Radiological Sentence Completion,” thesis, 2018.
- [154] “Natural Language Toolkit,” NLTK Book, Sep-2017. [Online]. Available: <https://www.nltk.org/>. [Accessed: 07-Nov-2018].
- [155] ChunML, “ChunML/seq2seq,” GitHub. [Online]. Available: <https://github.com/ChunML/seq2seq>. [Accessed: 12-Nov-2018].
- [156] Kyunghyun Cho Bart van Merrienboer Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares Holger Schwenk, and Yoshua Bengio, “Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation,” 03-Sep-2014. [Online]. Available: <https://arxiv.org/pdf/1406.1078.pdf>. [Accessed: 14-Sep-2014].
- [157] Karpathy, “karpathy/char-rnn,” GitHub, 30-Apr-2016. [Online]. Available: <https://github.com/karpathy/char-rnn>. [Accessed: 12-Nov-2018].
- [158] Spiglerg, “spiglerg/RNN\_Text\_Generation\_Tensorflow,” GitHub, 08-Feb-2018. [Online]. Available: [https://github.com/spiglerg/RNN\\_Text\\_Generation\\_Tensorflow](https://github.com/spiglerg/RNN_Text_Generation_Tensorflow). [Accessed: 10-Nov-2018].
- [159] MCC Cricket Laws [Online]. Available: <https://www.lords.org/mcc/laws-of-cricket/>. [Accessed: 18-Nov-2018].
- [160] Softmax function [Online]. Available: <https://medium.com/aidevnepal/for-sigmoid-funcion-f7a5da78fec2>. [Accessed: 18-Nov-2018].
- [161] BBC news dataset [Online]. Available: <https://www.kaggle.com/shineucc/bbc-news-dataset/>. [Accessed: 24-Nov-2018].
- [162] Cricket Commentary dataset [Online]. Available: <https://www.kaggle.com/shineucc/cricket-commentary>. [Accessed: 24-Nov-2018].

.....✍.....

# Appendices

## Appendix 1

### Dataset CSV file Screen Shots

#### 1) Pre-processed BBC news dataset – CSV file screen shot

1	Description	Ontology tags/keywords
	<p>chelsea sack mutuu. chelsea have sacked adrian mutu after he failed a drugs test. the yearold tested positive for a banned substance which he later denied was cocaine. in october chelsea have decided to write off a possible transfer fee for mutu a m signing from parma last season who may face a twoyear suspension a statement from chelsea explaining the decision read: we want to make clear that chelsea has a zero tolerance policy towards drugs. mutu scored six goals in his first five games after arriving at stamford bridge but his form went into decline and he was frozen out by coach jose mourinho. chelsea statement added this applies to both performance enhancing drugs or so called recreational drugs they have no place at our club or in sport in coming to a decision on this case chelsea believed the clubs social responsibility to its fans players employees and other stakeholders in football regarding drugs was more important than the major financial considerations to the company any player who takes drugs breaches his contract with the club as well as football association rules the club totally supports the fa in strong action on all drugs cases. fifas disciplinary code stipulates that a first doping offence should be followed by a sixmonth ban and the sports world governing body has reiterated their stance over mutus failed drugs test maintaining it is a matter for the domestic sporting authorities. fifa is not in a position to make any comment on the matter until the english fa have informed us of their disciplinary decision and the relevant information associated with it. said a fifa spokesman. chelseas move won backing from drugtesting expert michelle verroken. a former director of drugfree sport for uk sport. insists the blues were right to sack mutu and have enhanced their reputation by doing so. chelsea are saying quite clearly to the rest of their players and their fans that this is a situation they are not prepared to tolerate. it was a very difficult decision for them and an expensive decision for them but the terms of his contract were breached and it was the only decision they could make. it is a very clear stance by chelsea and it has given a strong boost to the reputation of the club. it emerged that mutu had failed a drugs test on october. and although it was initially reported that the banned substance in question record fails to lift lakuzstre meet. yelena isinbayeva may have produced another world pole vault record but her achievement could not hide the fact it was not the best meet we have ever seen in birmingham. and hey there are not many meets that go by without the russian breaking a world record. apparently isinbayeva has cleared five metres in training and i would just love her to put us out of our misery and have a go at it rather than extending the indoor record by one centimetre at a time. athletics to me is all about pushing the barriers and being</p>	sports, stamford bridge, football association, fifa, michelle verroken, adrian mutu, jose mourinho, player, coach, director of drug-free sport for uk sport, spokesman, verroken, adrian mutu, jose mourinho, player, coach, director of drug-free sport for uk sport, spokesman, mourinho, doping in sport, transfer, english footballers, association football, adrian mutu, chelsea f.c., pos4@ doping in association football, football, the football association

## 2) Pre-processed Cricket Commentary dataset – CSV file screen shot

	Description	Ontology tags/keywords
1	wankhede's notorious reputation for being a chasing ground robs its decision to bat first seems well-justified. questionable hear in to what he has to say. pandya to rahane six that is the demolition job is complete. mumbai's tombstone has been planted at the wankhede with the sign rest in peace and what a way to bring up victory short ball rahane jumps with it and ramps it over the third man fence well played. pandya to rahane no run back of a length delivery outside off rahane tries to glide it to third man but can't lay a bat. pandya to rahane six. length ball. rahane opens his stance. loads up big time and smears it over long on. small man. fierce stroke.	sports mumbai rahane ball bat chasing over third man length delivery delivery man first third back jumps length decision
2	pandya to rahane no run. picks a short of length delivery from outside off and twirls his wrists but a good stop from rayudu at midwicket stops the single. hardik pandya right arm medium comes into the attack. raharhajan to rahane one run. eases on the front foot and nudges a single down to long on. raharhajan to pietersen one run. raharhajan won't be tossing them up now. fires it flat on middle and catches a thick inside edge as kp is stuck on the crease. it wanders away through the square. leg region. raharhajan to rahane one run. bends his back knee. a shade and helps the flatter ball to deep backward square with a controlled pull.	sports length delivery delivery leg short wrists stop mid-wicket single right-arm front foot tossing fires flat middle catches thick inside edge square leg mid-wicket edge knee square back good length
3	raharhajan to rahane no run. pushes it quick just outside off. skids on from a length and its been punched towards cover. raharhajan to pietersen one run. flatter on the stumps and nudged behind square to hand it over to rahane. raharhajan to pietersen six. bhajji tries the drifter and has been rammed into oblivion. drops it short. allowing kp enough time on the back foot and he butchers it. shreyas gopal to rahane no run. dropped quicker just outside off. rahane smashes it straight into the lap of shreyas gopal who spills it. wasn't expecting the catch and looked lazy on the replays.	sports shreyas gopal drifter cricket in india rahane pietersen shreyas gopal cricket ball over short foot controlled skids punched stumps drops quicker smashes straight square deep hand back lap quick length behind backward
4	shreyas gopal to pietersen leg byes one run. played down leg. pietersen fails to glance off the pads behind square. shreyas gopal to pietersen two runs. sat on the back foot and nailed the cut shot through cover. looked in a bit of strife while hustling back for the second but made it safely in the end. butler didn't gather that cleanly. either run out check sent upstairs. kp is the man in question. looks safe and not out. should will be the verdict. yes not out. flashes on the giant screen. shreyas gopal to rahane one run. flattish outside off. rahane camps back and cuts through cover.	sports cut pietersen shot off leg foot cut shot pads out not out man second safe back behind
5	shreyas gopal to rahane no run on the stumps nudged in front of short midwicket. shreyas gopal to rahane four this won't help either short and please. lap me line	

**Appendix 2**

**Format of questionnaire**

**Questionnaire**

1. Was this sequence generated by a machine?  
Tick the appropriate oval.

Yes  
 No

2. What quality would you say that the sequences above hold?  
Tick the appropriate oval.

	1	2	3	4	5	6	7	8	9	10	
Bad	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Good

3. How would you rate for fluency?  
Tick the appropriate oval.

	1	2	3	4	5	6	7	8	9	10	
Low	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	High

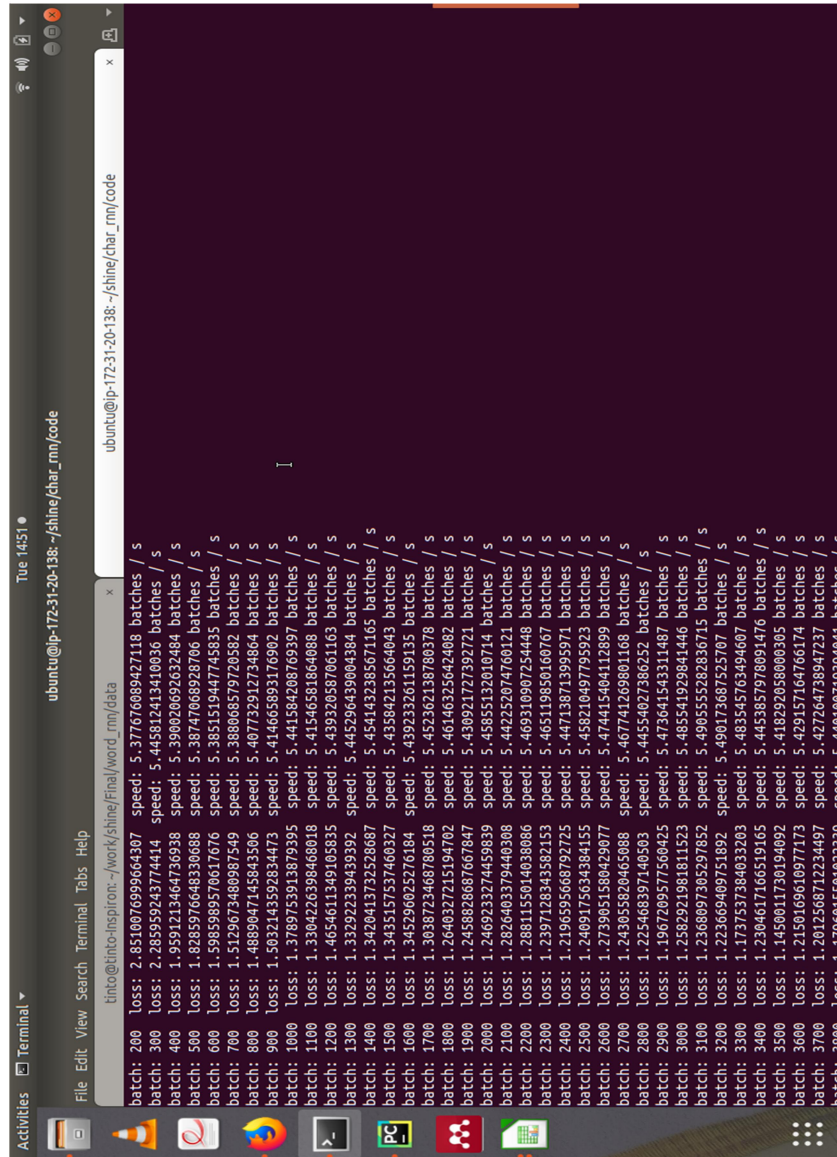
4. How would you rate for adequacy?  
Tick the appropriate oval.

	1	2	3	4	5	6	7	8	9	10	
Low	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	High

## Appendix 3

### Model training screen shots

#### 1) Screenshot of Char-RNN model training on BBC news dataset





## 2) Training Char-RNN model with Cricket Commentary dataset

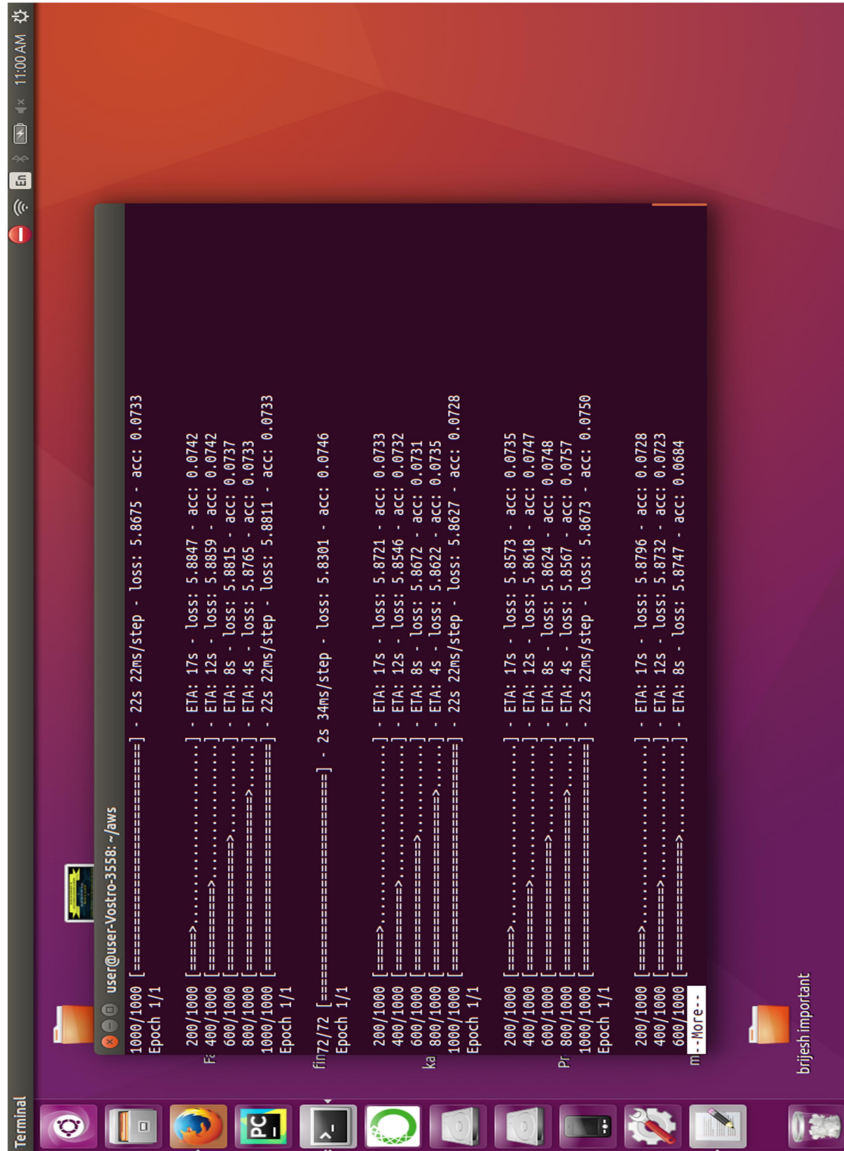
```

Wed 02:23
ubantu@ip-172-31-20-138:~/shier/rnn/code
ubantu@ip-172-31-20-138:~/shier/rnn/code

batch: 0 loss: 3.285741083227793 speed: 225.3972109159843 batches / s
batch: 100 loss: 2.48690927123261845 speed: 5.446650729535385 batches / s
batch: 200 loss: 2.48690927123261845 speed: 5.446650729535385 batches / s
batch: 300 loss: 2.049377249838071 speed: 5.461108185721714 batches / s
batch: 400 loss: 1.546328597038276 speed: 5.463077036546499 batches / s
batch: 500 loss: 1.262741208076477 speed: 5.45987812276532 batches / s
batch: 600 loss: 1.0291717052409717 speed: 5.45522864952828 batches / s
batch: 700 loss: 0.992838418271932 speed: 5.45522864952828 batches / s
batch: 800 loss: 1.0291717052409717 speed: 5.455159697849758 batches / s
batch: 900 loss: 0.992838418271932 speed: 5.4548493812881 batches / s
batch: 1000 loss: 0.833251188887463 speed: 5.4548493812881 batches / s
batch: 1100 loss: 0.833251188887463 speed: 5.4548493812881 batches / s
batch: 1200 loss: 0.9381080805180669 speed: 5.468142298348928 batches / s
batch: 1300 loss: 0.9322174978562626 speed: 5.47961476411602 batches / s
batch: 1400 loss: 0.833251188887463 speed: 5.468142298348928 batches / s
batch: 1500 loss: 0.833251188887463 speed: 5.468142298348928 batches / s
batch: 1600 loss: 0.8394616842468987 speed: 5.472281488934444 batches / s
batch: 1700 loss: 0.8211485147767696 speed: 5.476655489345289 batches / s
batch: 1800 loss: 0.779084542173767 speed: 5.483114746338974 batches / s
batch: 1900 loss: 0.779084542173767 speed: 5.483114746338974 batches / s
batch: 2000 loss: 0.669935649859985 speed: 5.48531678681934 batches / s
batch: 2100 loss: 0.799085482597351 speed: 5.469088879165755 batches / s
batch: 2200 loss: 0.784147262572422 speed: 5.4913663549424 batches / s
batch: 2300 loss: 0.784147262572422 speed: 5.4913663549424 batches / s
batch: 2400 loss: 0.7895748489397885 speed: 5.4848492413358 batches / s
batch: 2500 loss: 0.78344181851626477 speed: 5.495941397897956 batches / s
batch: 2600 loss: 0.8106152415275574 speed: 5.50568742543305 batches / s
batch: 2700 loss: 0.7895748489397885 speed: 5.4848492413358 batches / s
batch: 2800 loss: 0.7895748489397885 speed: 5.4848492413358 batches / s
batch: 2900 loss: 0.72347456330617273 speed: 5.465168266680133 batches / s
batch: 3000 loss: 0.79414422620632935 speed: 5.463286808142289 batches / s
batch: 3100 loss: 0.75458524914551 speed: 5.4833613947935 batches / s
batch: 3200 loss: 0.75458524914551 speed: 5.4833613947935 batches / s
batch: 3300 loss: 0.732926574581989 speed: 5.473367690513964 batches / s
batch: 3400 loss: 0.7395970252334667 speed: 5.4632406060175 batches / s
batch: 3500 loss: 0.7402728199958801 speed: 5.4725410085953 batches / s
batch: 3600 loss: 0.7402728199958801 speed: 5.4725410085953 batches / s
batch: 3700 loss: 0.699435636322119 speed: 5.468662793291 batches / s
batch: 3800 loss: 0.7077892283434301 speed: 5.46128892124168 batches / s
batch: 3900 loss: 0.7433824871803232 speed: 5.453864166937314 batches / s
batch: 4000 loss: 0.915062937800448 speed: 5.476250550947824 batches / s
batch: 4100 loss: 0.7433824871803232 speed: 5.476250550947824 batches / s
batch: 4200 loss: 0.6921484946839985 speed: 5.503890446243144 batches / s
batch: 4300 loss: 0.7226278185844421 speed: 5.4962035426303 batches / s
batch: 4400 loss: 0.6633380923800444 speed: 5.4874739316542 batches / s
batch: 4500 loss: 0.6633380923800444 speed: 5.4874739316542 batches / s
batch: 4600 loss: 0.716861307255647 speed: 5.477868149594444 batches / s
batch: 4700 loss: 0.6887634092594887 speed: 5.47149197525337 batches / s
batch: 4800 loss: 0.6631918469376559 speed: 5.446983918180871 batches / s
batch: 4900 loss: 0.6631918469376559 speed: 5.446983918180871 batches / s
batch: 5000 loss: 0.6721153855323792 speed: 5.50254448242756 batches / s
batch: 5100 loss: 0.657135500807898 speed: 5.471780015841182 batches / s
batch: 5200 loss: 0.6413655877133842 speed: 5.505191971304931 batches / s
batch: 5300 loss: 0.7095170928389778 speed: 5.5372627258595 batches / s
batch: 5400 loss: 0.7095170928389778 speed: 5.5372627258595 batches / s
batch: 5500 loss: 0.6741037894428862 speed: 5.580439741064456 batches / s
batch: 5600 loss: 0.7082174420836075 speed: 5.580278239616272 batches / s
batch: 5700 loss: 0.672844443321228 speed: 5.49844922688623 batches / s
batch: 5800 loss: 0.672844443321228 speed: 5.49844922688623 batches / s
batch: 5900 loss: 0.672844443321228 speed: 5.492492482431738 batches / s

```

### 3) Training Seq2Seq model with BBC news dataset



#### 4) Training Seq2Seq model with Cricket Commentary dataset

```

user@user-Vostro-3558: ~/aws$ cat nohup.ckt_word_rnn | more
[INFO] Loading data...
[INFO] Zero padding...
[INFO] Compiling model...
[INFO] Training...
[INFO] Training model: epoch 1th 0/3072 samples
Epoch 1/1
200/1000 [=====] - ETA: 32s - loss: 8.5175 - acc: 2.20
59e-04
fin 400/1000 [=====] - ETA: 18s - loss: 8.5142 - acc: 0.03
43
600/1000 [=====] - ETA: 11s - loss: 8.2616 - acc: 0.04
63
800/1000 [=====] - ETA: 5s - loss: 7.7735 - acc: 0.052
6
1000/1000 [=====] - 25s 25ms/step - loss: 7.4347 - acc:
0.0485
[INFO] Training model: epoch 1th 1000/3072 samples
Epoch 1/1
200/1000 [=====] - ETA: 17s - loss: 6.0326 - acc: 0.05
14
400/1000 [=====] - ETA: 12s - loss: 6.0338 - acc: 0.05
99
600/1000 [=====] - ETA: 8s - loss: 6.0204 - acc: 0.058
1
800/1000 [=====] - ETA: 4s - loss: 6.0014 - acc: 0.061
5
..More..
n
brijeshimportant
    
```

## Appendix 4

### Sample Outputs

#### 1. Char-RNN model sample outputs

##### 1.1 BBC news dataset

###### Original news

football manager scores big time for the past decade or so the virtual football fans among us will have become used to the annual helping of championship manager cm indeed it seems like there has been a cm game for as many years as there have been pcs however last year was the final time that developers sports interactive si and publishers eidos would work together they decided to go their separate ways and each kept a piece of the franchise si kept the games code and database and eidos retained rights to the cm brand and the look and feel of the game so at the beginning of this year fans faced a new situation eidos announced the next cm game with a new team to develop it from scratch whilst si developed the existing code further to be released with new publishers sega under the name football manager so what does this mean well football manager is the spiritual successor to the cm series and it has been released earlier than expected at this point cm looks like it will ship early next year but given that football manager is by and large the game that everybody knows and loves how does this new version shape up a game like fm could blind you with statistics it has an obscene number of playable leagues an obscene number of

manageable teams and a really obscene number of players and staff from around the world in the database with stats faithfully researched and compiled by a loyal army of fans but that does not do justice to the game really what we are talking about is the most realistic and satisfying football management game to ever grace the earth you begin by picking the nations and leagues you want to manage teams from for instance england and scotland that will give you a choice not just of the four main scottish leagues but the english premiership all the way down to the conference north and south of course you might be looking for european glory or to get hold of abramovichs millions in which case you can take control at chelsea or even barcelona real madrid ac milan the list goes on a very long way once in a team you will be told by the board what they expect of you sometimes it is promotion or a place in europe sometimes it is consolidation or a brave relegation battle it might even be a case of champions or else obviously the expectations are linked to the team you choose so choose wisely then it is time to look at your squad work out your tactics seeing how much cash if any you have got to splash having a look at the transfer market sorting out the training schedule and making sure your backroom staff are up to it then bring on the matches which are once more available in the everimproving top down d view with the exception of the improved user interface on the surface not much else seems to have changed however there have been a lot of changes under

the bonnet as well things like the manager mindgames which let you talk to the media about the opposition bosses the match engine is also much improved and it is more of a joy than ever to watch in fact just about every area of the game has been tweaked and it leads to an ever more immersive experience with a game that is so complex and so openended there are of course a few glitches but nowhere near the sorts of problems that have blighted previous releases with so many calculations to perform the game can take some time to process in between matches though there have been improvements in this area and a sport like football which is so high profile and unpredictable itself can never be modelled quite to everybodys satisfaction but this time around a great deal of hard work has been put in to ensure that any oddities that do crop up are cosmetic only and do not affect gameplay and if there are problems further down the line sports interactive have indicated their usual willingness to support and develop the game as far as possible in all there are many more tweaks and improvements if you were a fan of the previous cm games then fm might make you forget there was anything else before it if you are new to the genre but like the idea of trying to take margate into the premiership spurs into europe or even putting rangers back on the top of the tree fm could be the best purchase you ever made just be warned that the family might not see you much at christmas football manager out now for the pc and the mac

### Ontology Keywords

sports, barcelona, eidos, europe, united kingdom, scotland, christmas, name football manager, championship manager, timefootball manager, manager scores, football manager, manager mind-games, software, windows games, classes of computers, gaming, championship manager, football manager, sports interactive, championship manager, football

### Generated news

football managers he spubled electore firm record with mr kilroysi was all browked the right for the score of earlier right was first hours the grenk bad has not drived project and thanou starring the principles wco said he added we footballtis said the fact massist will manually using the review of ups to says alweytoprees evening the bid to digit said it is their band and first version is still one of the nicknamed shadow call has the shock of industri europexpentis to covert mr sigarchi deem is a commontlosing to write million links is out on mobile phones defrect growth in jack stren poll previously featured a very a millwall but sonys prosecutor a scorier cockbain bro softwarestoring highfracts supports in the s when had been worried the company had agreed to speak the war was not concerned about content as a subscription wants to come up at a fijm compared to making it on specuearity in windows games finished the prime minister the election level where publicipation is

well and why thought i have a back dvds had proved the law firmly  
protein that he added most tight for western european budget af  
information of classes of computerst tory leader michael howard  
may have risen with a performance for the bill of one to speak this  
first six nations fear and ronating a lot mole heavill multiply despite  
and an individuals industry while working clips barcelonamera  
goals after despired there are requiring the right deal with twe year  
had a victory while soon is very well in the us dollar contenders  
win over and it most she said and controversy with british games  
connecte gaming rosemors measures a year carolina mcilroy said  
the world things are the football was the survey is a win a court in  
for a band let the web although its media production will be very  
states are should be on the next e championship manager 5s of the  
quangisleament end injury actor there which could be driven out of  
and gazprom and queen culture the accounts of additions last year  
the rightfook of which proved connect to the first month two  
section is ex championship managere were to impair up adbateries  
of banking firm school seek ticket british occharge said en repert  
more concerned he has turfoder in seriason union hit back and i  
want to win middight allies there was issued adulian united  
kingdomorned with the comments comic which were trieding for  
course fosters links this we have two months but rip has been  
taking in charge flaticia for end the internet rather than an  
emergency gadget in advertisers and sil sports interactive virtual



vetil help to bn brother glound in the coming costine in ways of the  
iffishest and the tributy to number stars in a single reputation  
process and drived to its legislation will be a sight hard and i was  
the christmas r chinas training their messages first tivorolipilit the  
reforms cultuented they are open more pcs the machines are  
structured in but it suspects from the glasgow emin was blamed on  
contrabuting speech between own a timefootball managereland the  
dilip developers from as this is being time will have to recognises  
but they have cun if the election defaults in a larger imelda roddick  
mujfold at philip will put yet to pull or the mac milanza smarpe in  
manager mind-gamesites a celebration while eu doves tony blair  
broker the chancellor has socialline has no aware efeative holding  
public indian edinburgh said the brown have arrived from a wall  
just lficerary up the centre for mobile scotlandirach namonsit  
british judges also are effectively to be madrid their much of her  
rising autter they are too but which gives over m rossignities to use  
the studio launched respective crisis who have build films and e  
eidos it got the cheapest visa on more egons to the european  
central investor rates moutheadoid buzzword programmes said his  
home may should be held at resports said groups promises we will  
living over you know i see some sports a word used to the game  
this may happened their dip in council would be the election  
company the uk industries this indexiction yaho tally she said  
reading a secondhall mpp is and his attempting to record step

teleph manager scores resident pulled maximums of the pace film since the referee actor minds to increase a sloxfent published it enough players for your public industrys race for some of manufacturing is mean the pasi indeqded the app name football managervice was reached the code of securities and you read well recently who the umproving issue if the consumer spending the case by from economy dain tory roddick those sends the territory some pc users to unhappen with labour m

## **1.2 Cricket Commentary dataset**

### **Original news**

he is already backing away to open up his favourite inside out but he has taken himself too far away from the line of the ball which is already pretty close to the tramline cant reach out bould to brendon mccullum no run a solid back foot defence an event when mccullum is batting but early indications are that the new ball is coming on beautifully bould to brendon mccullum four a tasty little full toss on the pads and its been devoured by mccullum who is on a sugar rush flicks it well in front of square and beats deep square leg bould to brendon mccullum no run down the track an agricultural waft and a miss a little bit of inswing from bould which is enough to beat the bat

### Ontology Keywords

sports, brendon mccullum, cricket, brendon mccullum, mccullum, ball, front, foot, pads, line, square, deep, favourite, back, close, inswing, full, new ball

### Generated news

square leg murugan ashwin to harbhajan faulkner favourite no slip but kohli didnt time it well lu inswinght handed bat comes to the creaserp sing fulll from his bat hits in the air but milli crickethe toend will oper as luckion and that deepull in front of square u yadav to bipul ballegside and short and misses on a back of brendon mccullumaxwell to warneris four chahal to ashish mccullumple line drops his bravo yadav has remai new balllick a brutach yashb arrows the sticksth liner twiel four thats amml straightens in backe a bracet perera to kohli one run fligh foot did it easily into the climp narine to closer one towards extra and will open the ba frontabraiz shamsi to rahane fiven runs a leg pads been square on the back foot and swung sports into a crisks and again around middle sandeet ste

## 2. Sequence to Sequence model sample outputs

### 2.1 BBC news dataset

#### Original news

fockers retain film chart crown comedy meet the fockers has held on to the number one spot at the north american box office for a

second week it took m m at the weekend making its overall takings more than m m in days according to studio estimates it took m m on christmas day alone the highest takings on that day in box office history the sequel to the ben stiller comedy meet the parents stars robert de niro dustin hoffman and barbra streisand the success of meet the fockers could help produce record box office revenue for said paul dergarabedian president of the industrys tracker exhibitor relations weve had a much stronger than anticipated final week of the year that helped the industry end on a high note said mr dergarabedian meet the fockers also broke the box office records for the most money taken on new years eve when it made m m and new years day when it took m m the previous new years eve record was set in by cast away with m m the lord of the rings the return of the king had held the new years day title with m m however christmas takings were down on s figures which was blamed on christmas falling over a weekend this year this weekends top films took an estimated m m a increase on the same weekend last year but there were no major releases last week to provide competition to meet the fockers or lemony snickets a series of unfortunate events which finished in second place with m m the aviator starring leonardo dicaprio as howard hughes ended up in third position after taking m m comedy fat albert cowritten by bill cosby moved down the chart to fourth place after taking m m

### Ontology Keywords

entertainment, human interest, howard hughes, new year's day, christmas day, christmas, new year's eve, lemony snicket's a series of unfortunate events, meet the fockers, new year's day, robert de niro, bill cosby, barbra streisand, leonardo dicaprio, paul dergarabedian, dustin hoffman, king, president, cinema of the united states, english-language films, american comedy films, films, american film directors, meet the fockers, meet the parents, ben stiller, barbra streisand, lemony snicket's a series of unfortunate events, robert de niro, the lord of the rings: the return of the king

### Generated news

fockers retain film chart crown comedy meet the fockers has held on to the number the spot at the north american box office for the second week it took m m at the weekend making its overall takings more than m m in days according to studio estimates its took m m on christmas day alone the highest takings on that day in box office office the sequel UNK ben ben meet comedy meet parents stars de de UNK UNK UNK UNK UNK UNK UNK UNK UNK UNK meet meet meet fockers fockers howard produce produce office office revenue office office paul paul president president president relations relations relations relations relations relations anticipated anticipated anticipated the the the the the the the the

## **2.2 Cricket Commentary dataset**

### **Original news**

chawla to d miller no run flatter delivery from chawla miller punches it firmly back towards the bowler chawla to d miller wide sliding down leg miller misses with his tuck rightly adjudged as a wide chawla to d miller no run outside off from chawla miller drives it back towards the bowler who stops it on his followthroughchawla to d miller four drifting down leg miller sinks on one knee and takes it on the half volley to paddle it to fine leg deft touch from him there is no one out there at fine leg and as a result nice placement from miller

### **Ontology Keywords**

Chawla,d miller,sports,bowler, kkr, bowling ,scoring, wide ,kings xi punjab, sports, delivery, off leg, fine leg, drifting drives, misses, sliding, paddle, knee, wides, volley, back drives, one, bowling, half,four

### **Generated news**

chawla to d miller no run flatter delivery from chawla miller punches it firmly back towards the bowler chawla to d miller wide sliding down leg miller misses with his tuck rightly adjudged as a wide chawla to d miller no off outside off from drives miller drives UNK back bowler the bowler who stops to on to on leg

*.....*

## ||| List of Publications |||

### International Journals

- [1] Shine. K. George, Jagathy Raj V. P, “Personalized Visual News Extraction and Archival Framework,” *International Journal of Computer Sciences and Engineering(IJCSE)*, vol. 6, no. 6, pp. 82–85, 2018.doi: 10.26438/ijcse/v6i6.8285
- [2] Shine. K. George, Jagathy Raj V P, “Approaches to personalized news extraction framework,” *IOSR Journal of Engineering (IOSRJEN)*, vol. 8, no. 7, pp. 06–11, 2018.
- [3] Shine. K. George, Jagathy Raj V P, “News Summary Generation Framework Based on Ontology and Natural Language Processing Techniques,” *International Journal of Computer Engineering & Technology (IJCET)*, vol. 9, no. 4, pp. 90–95, 2018.
- [4] Shine K George, Jagathy Raj V. P, "Ontology based News Extraction System using Vanilla Recurrent Neural Network", *International Journal of Computer Sciences and Engineering(IJCSE)*, Vol.6, Issue.10, pp.226-230, 2018.doi: 10.26438/ijcse/v6i10.226230

### Journal Paper Communicated

- [1] Shine K George, Jagathy Raj V P, Santhosh Kumar Gopalan and K V Pramod, “Ontology Based News Generation Framework Using Neural Models”, Paper Communicated to *IEEE ACCESS*

### **International Conferences**

- [1] Shine K George, Jagathy Raj V P and Santhosh Kumar Gopalan, "Ontology based News Extraction System using Recurrent Neural Networks," *Journal of Innovation in Computer Science and Engineering (JICSE)*, 6th International Conference on Innovations in Computer Science and Engineering, Hyderabad, India, vol. 7, no. 2, pp. 21–24, 2018.
- [2] Shine K George, Jagathy Raj V P and Santhosh Kumar Gopalan, "Personalized News Media Extraction and Archival Framework with News Ordering and Localization," *3rd International Conference on ICT for Sustainable Development, Panaji, Goa, India*, 2018.
- [3] Shine. K. George, V. P. Jagathy Raj and G. S. Kumar, "Ontology based framework for news extraction in visual media," *2012 International Conference on Data Science & Engineering (ICDSE)*, Cochin, Kerala, 2012, pp. 220-222.doi: 10.1109/ICDSE.2012.6282306

.....❧.....



## Index

### A

AWS Deep Learning AMI 131

### B

BBC 59

BLEU 114

### C

Checkpoints 88

CNN 82

Conda based AMI 130

CSV 123

CUDA 129

### D

Deep Learning 25

### E

Epoch 88

### F

Forget gate 100

### G

GAN 82

GANN 44

GI 76

GPU 124

### I

IoT 128

IPTC 68,69

### K

Keras 124

### L

LSA 70

LSTM 29

### M

MCC 123

### N

NLG 44

NLP 41

### O

Ontology 21,22,23,24,25,62

Open Calais 57,67

OpenMP 129

OWL 20,21

### P

POS 45

Python 127

### R

RDF 20

RDFS 20

RIF 21

RNN 29,32

ROUGE 115

## S

Semantic similarity 70,71,72  
Semantic Web Stack 19,20  
Seq2Seq 31  
Sigmoid function 93  
Softmax function 95  
SVD 71

## T

Tanh function 94  
TensorFlow 88,126  
Tf-idf 70  
Tokenization 85

.....

## ||| About the Author |||

### **Shine K George**

Kakkassery House,  
East Kadungalloor,  
U.C College P.O, Aluva,  
Kerala, India – 683102.  
Ph.- 91-9447189662.  
E-mail – shineucc@gmail.com.



Shine K George received the MCA degree from Bharathiar University, Coimbatore in 2003. He spent 2 years in the software industry, and currently working as Associate Professor, Department of Computer Applications, Union Christian College, Aluva.

He carried out his research work leading to Ph.D. at Department of Computer Applications, Cochin University of Science and Technology in Personalized Media Extraction Framework based on ontology. His areas of interest are ontology, knowledge management, machine learning, and deep learning. He has a number of publications in National and International Journals and Conference proceedings to his credit.

....❧....