



GROUP No.: 11

PROJECT MENTOR

Mayank Singh

TEAM MEMBERS

Ashish Sharma		13CS30043
Jatin Arora		13CS10057
Prabhat Agarwal		13CS10060
Pritam Khan		13CS10036
Sumit Agarwal		13CS10061

AUTOMATIC ABSTRACT GENERATION USING LSTMs

INTRODUCTION

❖ Abstract

- Summarizes major aspects of a research article



Objective of the research problem(s) investigated



Methodolgy employed to solve the problem



Results & their interpretations

- ❖ **Objective:** Assist Authors by generating automatic abstract given a scientific article

WHY USE LSTM?

1 Scientific Articles have Long-Term Dependencies

Additionally, as described in Section 5 we apply a MERT tuning step after training using the DUC-

2 Summarization using LSTMs shown to work better than other summarizing methods

Model	ROUGE-1	DUC-2004	
		ROUGE-2	ROUGE-L
IR	11.06	1.67	9.67
PREFIX	22.43	6.49	19.65
COMPRESS	19.77	4.02	17.30
W&L	22	6	17
TOPIARY	25.12	6.46	20.12
MOSES+	26.50	8.13	22.85
ABS	26.55	7.06	22.05
ABS+	28.18	8.49	23.81
REFERENCE	29.21	8.38	24.46

Model	Encoder	Perplexity
KN-Smoothed 5-Gram	none	183.2
Feed-Forward NNLM	none	145.9
Bag-of-Word	enc ₁	43.6
Convolutional (TDNN)	enc ₂	35.9
Attention-Based (ABS)	enc ₃	27.1

3 Summarizing long documents using LSTMs unexplored

CHALLENGES INVOLVED

- 1 Scientific Articles are **too long to be processed** for current GPUs using **LSTMs**
- 2 Each scientific article contains **new ideas, approaches and results**
- 3 **Good quality, large-scale datasets** required
- 4 **Unique structure of Scientific Articles** needs a different modelling

DATASET

❖ [arXiv.org](https://arxiv.org)

- Online repository of e-prints of scientific articles

❖ [Crawled LaTeX Sources](#) of articles in following fields:

- Information Retrieval (**cs.IR**)
- Computation and Language (**cs.CL**)
- Machine Learning (**cs.LG**)
- Artificial Intelligence (**cs.AI**)

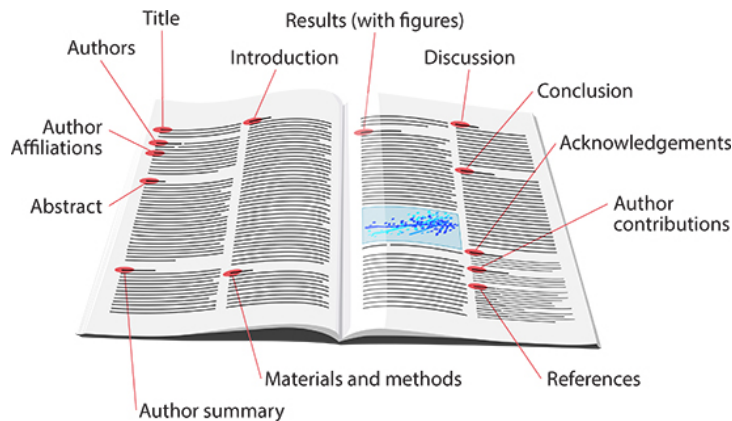


❖ **Size of the Dataset:** 16,780 articles

GENERAL APPROACH



Processed & Marked



Generate Representation

Reduced Length

Abstractive Summarization using LSTMs

Final Summary/
Abstract



PREPROCESSING

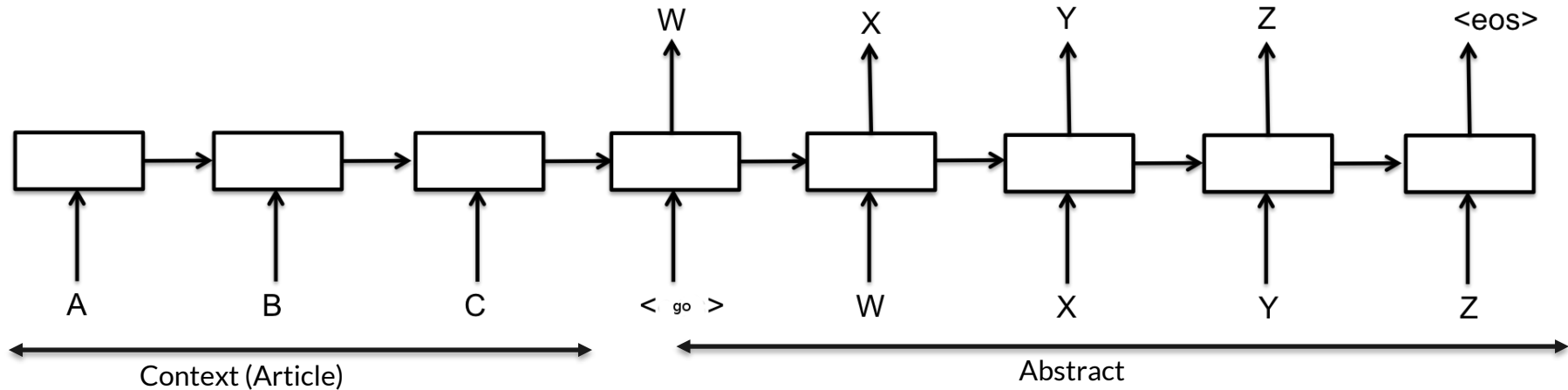
❖ `pylatexenc`

- Python library for parsing LaTeX to generate text

❖ Modifications

- 1 Sections and Subsections of article were identified and marked
- 2 Figures, Tables and Mathematical Equations were replaced by representative tokens
- 3 Obtained Structure was converted to LSTM input format

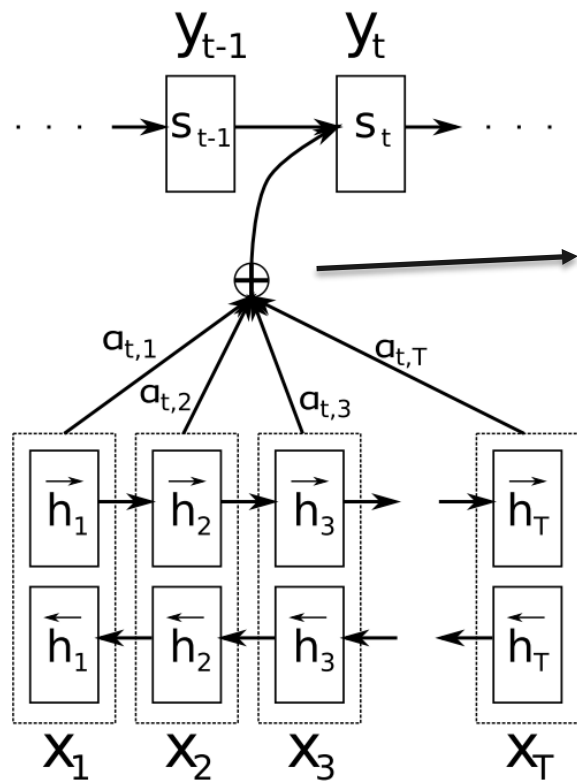
Sequence to Sequence Model



❖ **Consists of two recurrent neural networks (RNNs):**

- **Encoder:** Processes the input -> Sentences in the article or its Representation
- **Decoder:** Generates the output -> Abstract

ATTENTION MECHANISM



Output of encoder module as an additional conditioning input to Decoder

- Condition the RNN by a convolutional attention-based encoder

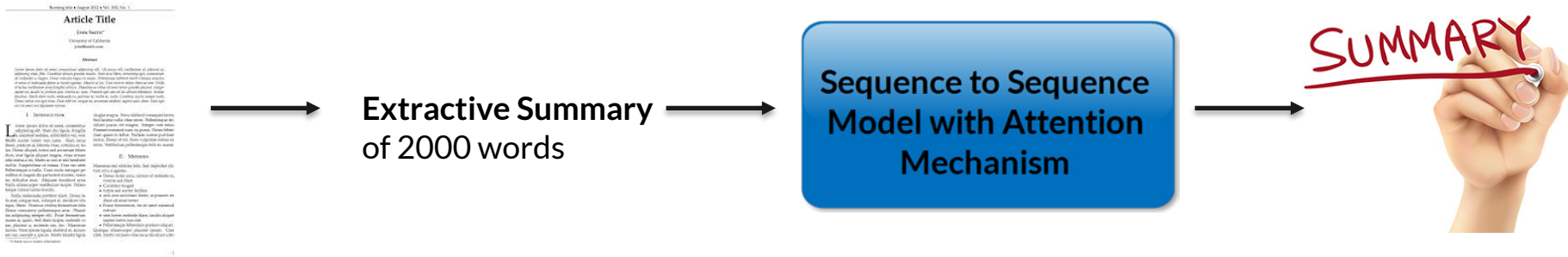
❖ Advantage:

- Informs the decoder **which part of the input sentence it should focus on** to generate the next word
- Both Decoder and Encoder are jointly trained on the data set.

TAKE 1: Using Extractive Summary

❖ **Reduce length using Extractive Summarization :**

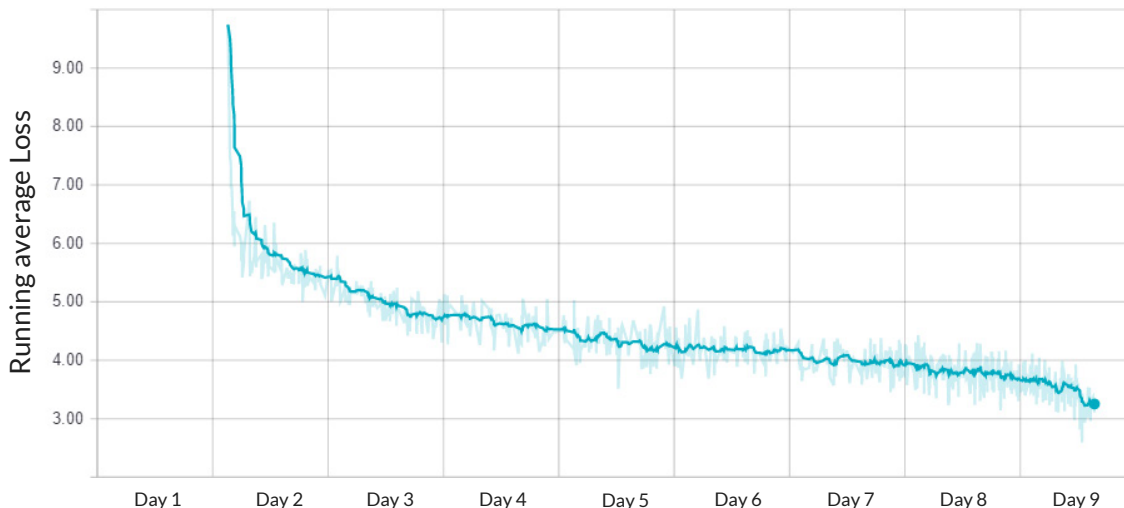
- Lex-Rank
- C-Lex-Rank
- Text-Rank
- LSA



VARIATIONS IN TAKE 1

❖ No. of Words in Extractive Summary:

- Computationally expensive to use large number of words (2000)



- Instead use a small extractive summary (250 words)

EXAMPLE OUTPUTS 1

Actual Abstract

the machine learning community adopted the use of null hypothesis significance testing (nhst) in order to ensure the statistical validity of results . many scientific fields however realized the shortcomings of frequentist reasoning and in the most radical cases even banned its use in publications . we should do the same : just as we have embraced the bayesian paradigm in the development of new machine learning methods , so we should also use it in the analysis of our own results .

- ROUGE-1: 0.334
- ROUGE-2: 0.076
- ROUGE-L: 0.204
- **Drawback:** Small Extractive summary misses out a lot of information

LSTM Generated

boltzmann media (resnets) have recently achieved places on challenging computer vision tasks . we introduce resnet in resnet (rir) : a deep architecture that generalizes resnets and standard cnns and is easily implemented with no computational overhead . rir consistently improves performance over resnets , outperforms architectures with similar amounts of augmentation on cifar-## , and establishes a new state-of-the-art on cifar-### .

TAKE 2: Paragraph Embeddings

❖ Para2Vec:

- **Unsupervised Learning of continuous representations for larger blocks of text, such as sentences, paragraphs or entire documents**

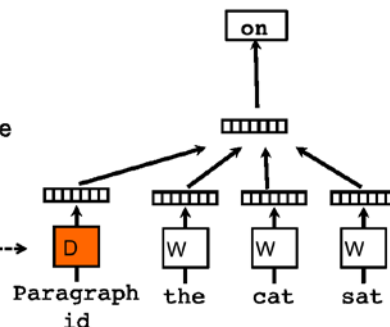


Separate Sections

Classifier

Average/Concatenate

Paragraph Matrix
Paragraph id

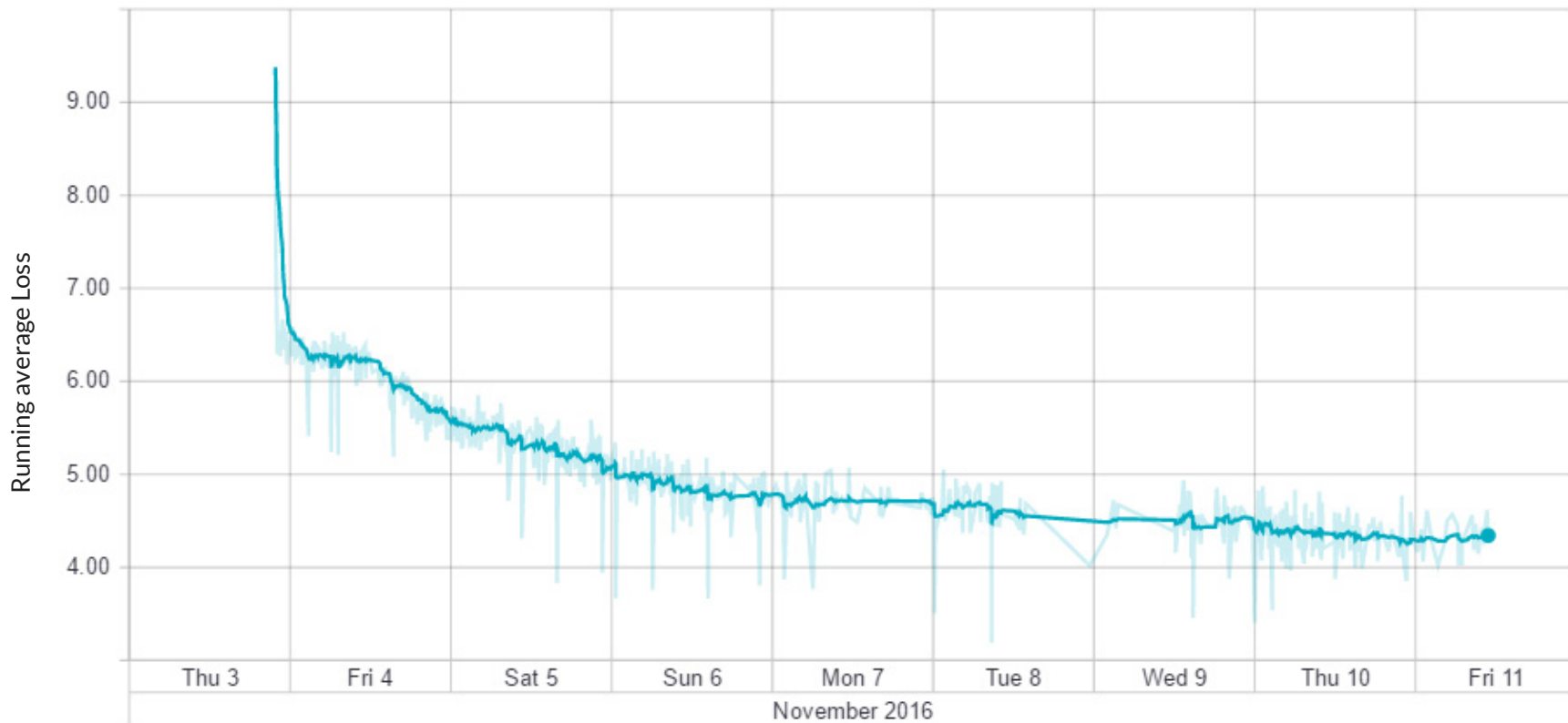


Paragraph embeddings
for each section

SUMMARY

Sequence to Sequence
Model with Attention
Mechanism

Paragraph Embeddings Status



EXAMPLE OUTPUTS 2

Actual Abstract

we describe a question answering model that applies to both images and structured knowledge bases . the model uses natural language strings to automatically assemble neural networks from a collection of composable modules . parameters for these modules are learned jointly with network assembly parameters via reinforcement learning , with only (world , question , answer) triples as supervision . our approach , which we term a dynamic neural module network

- ROUGE-1: 0.475
- ROUGE-2: 0.158
- ROUGE-L: 0.307

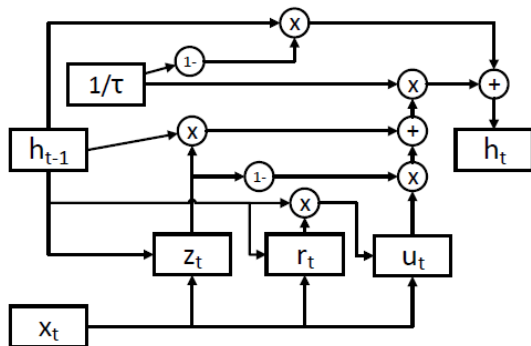
LSTM Generated

visual question answering (vqa) is a challenging task for natural language processing . in this paper , we propose a novel approach to learn a neural neural network (<UNK>) model . the model learns a sequence of a sequence of sentences that learns a sequence of a sequence of sentences , and the answers of the task are predicted . our results show that the proposed model outperforms the state-of-the-art performance of our model.

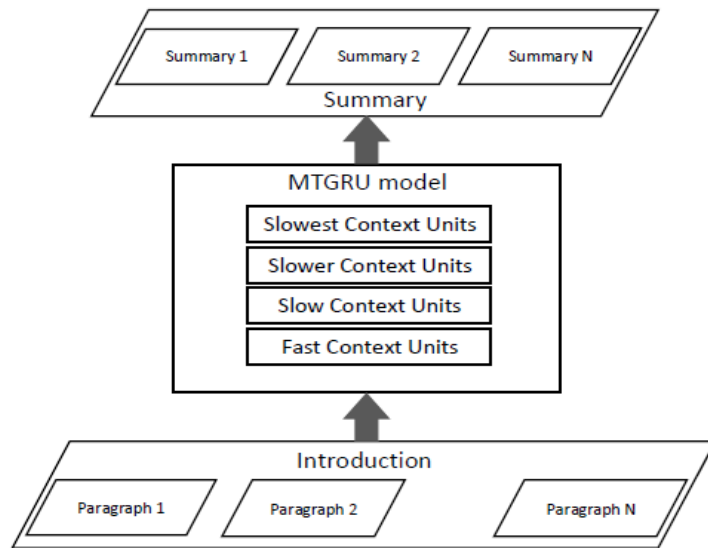
TAKE 3: Multiple Timestamp Gated Recurrent Unit

❖ Temporal Hierarchies to Sequence to Sequence Model:

- Apply a timescale constant at the end of a GRU
- Adds another constant gating unit which modulates the mixture of past and current hidden states.



MT-GRU Unit



MT-GRU Summarization Approach

Conclusion

- Novel attempt to summarize long scientific articles using LSTMs
- Proposed 2 abstractive summarization approaches:
 - **Extractive Summarization** followed by **Seq2Seq Model**
 - Utilizing **Paragraph Embeddings**
- **LSTMs have potential** to work for long documents but **require more computational power**

Future Work

- Use a **larger and richer dataset** for the problem
- Utilize better **computational resources**
- Make changes to the proposed models to make it more **robust**

THANK YOU !



"Ms. Jones, there are a number of big questions here to see you. They say they won't leave until they have some answers."