AUTOMATIC ABSTRACT GENERATION USING LSTMs

GROUP No.: 11

PROJECT MENTOR

Mayank Singh

TEAM MEMBERS

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

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**

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**

- Online repository of e-prints of scientific articles

❖ **Crawled LaTeX Sources** of articles in following fields:

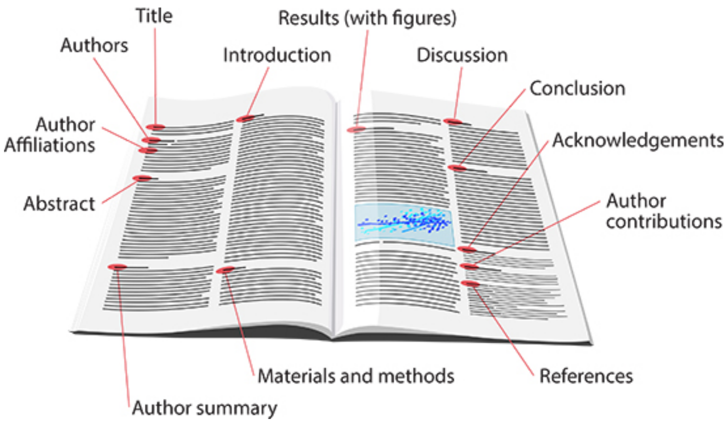
- Information Retrieval **(cs.IR)**
- Computation and Language **(cs.CL)**
- Machine Learning **(cs.LG)**
- Artificial Intelligence **(cs.AI)**

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

GENERAL APPROACH

Scientific Article

Processed & Marked



Generate Representation

Reduced Length

Abstractive Summarization using LSTMs

Final Summary/ Abstract

PREPROCESSING

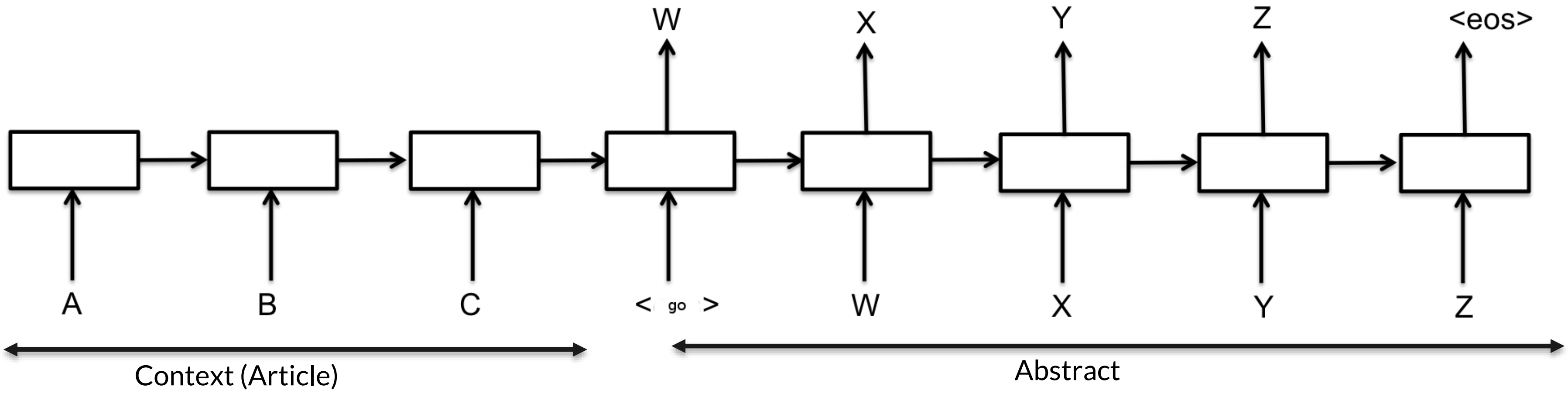
- **pylatexenc**

- 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**

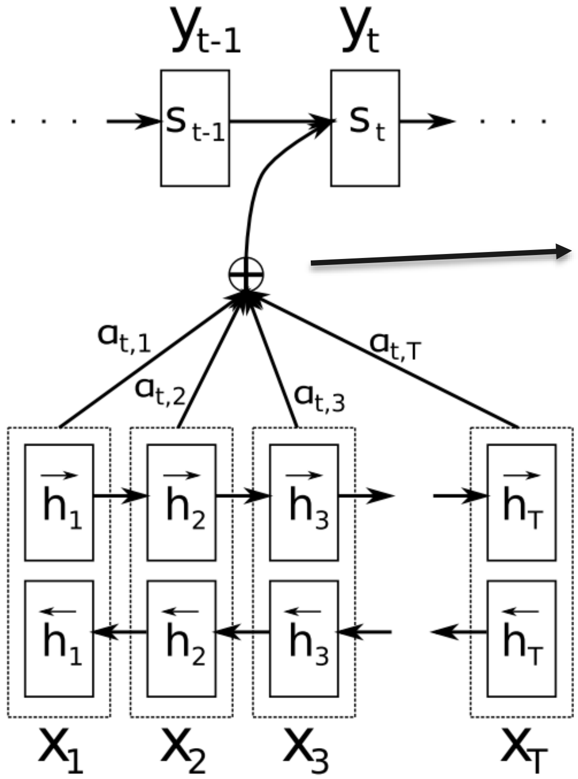
Sequence to Sequence Model



❖ **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



- Condition the RNN by a convolutional attention-based encoder

Output of encoder module as an additional conditioning input to Decoder

❖ **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
- Luhn
- Edmundson
- LSA
- Sum-Basic


Article Title

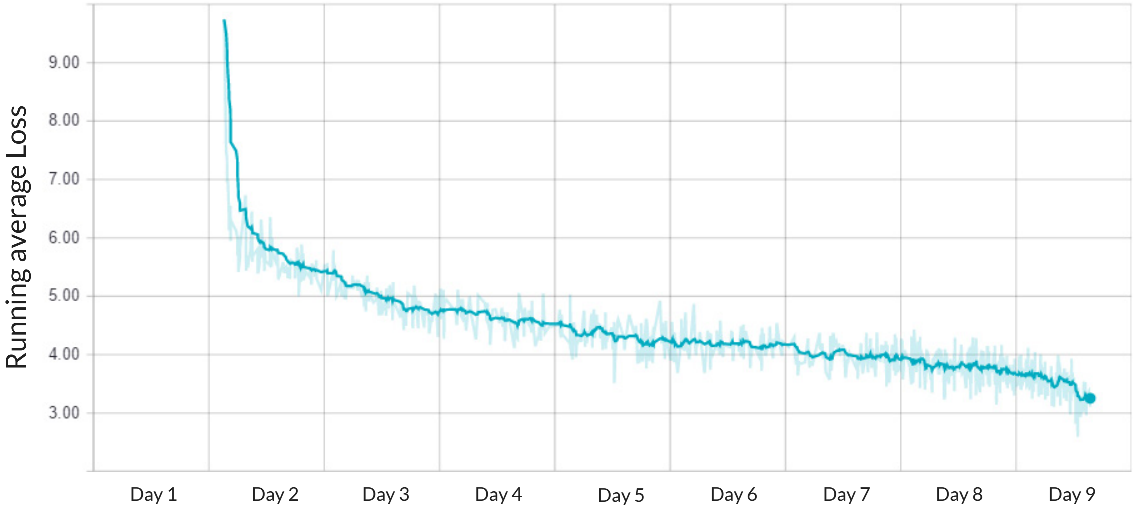
Extractive Summary of 2000 words

Sequence to Sequence Model with Attention Mechanism

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	LSTM Generated
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 .	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-### .

- ROUGE-1: 0.334
- ROUGE-2: 0.076
- ROUGE-L: 0.204

- **Drawback:** Small Extractive summary **misses out a lot of information**

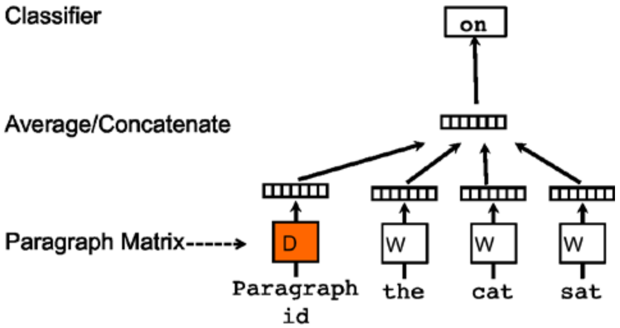
TAKE 2: Paragraph Embeddings

❖ Para2Vec:

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


Article Title

Separate Sections



Paragraph embeddings for each section

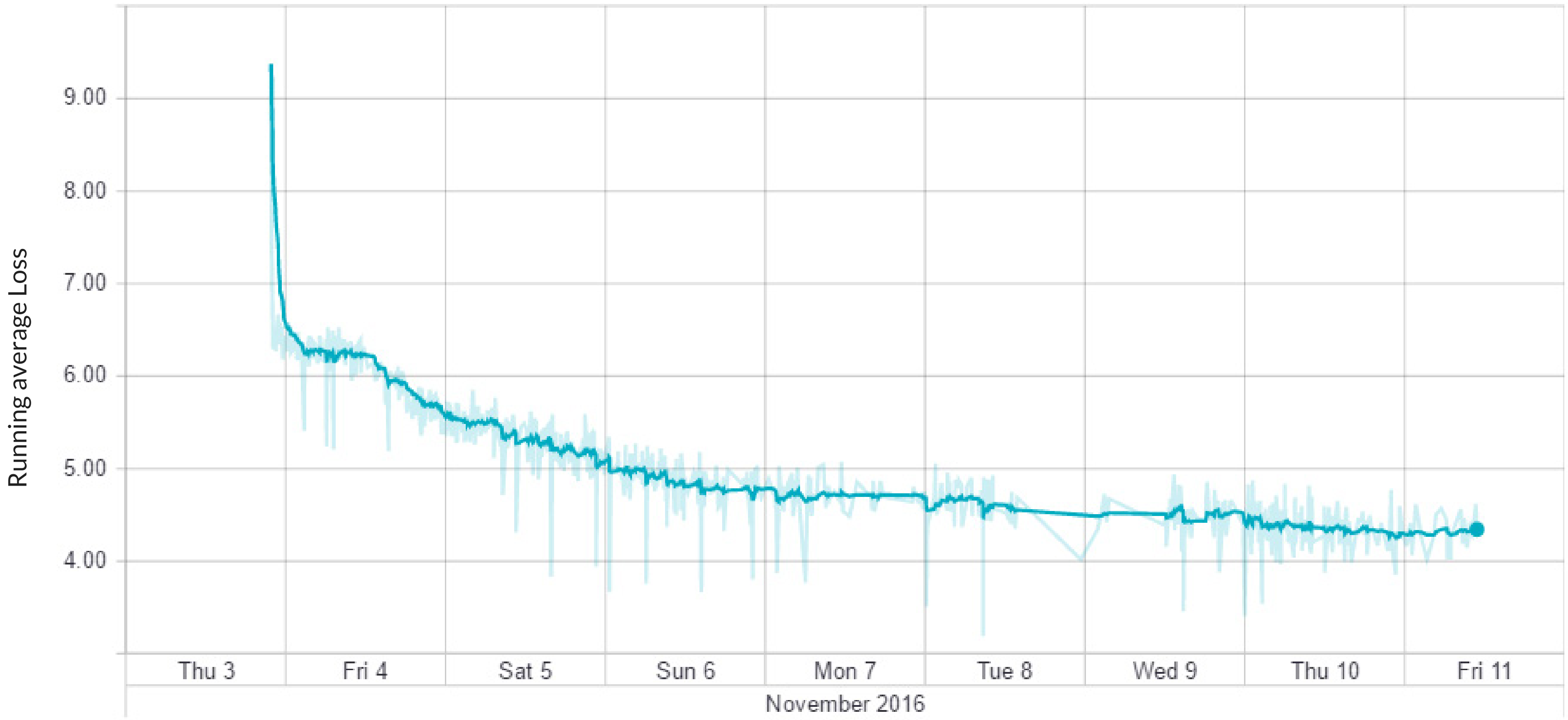
Sequence to Sequence Model with Attention Mechanism

EXAMPLE OUTPUTS 2

Actual Abstract	LSTM Generated
we present an approach of learning multi-sense word embeddings relying both on monolingual and bilingual information . our model consists of an encoder , which uses monolingual and bilingual context (i.e . a parallel sentence) to choose a sense for a given word , and a decoder which predicts context words based on the chosen sense . the two components are estimated jointly . we observe that the word representations induced from bilingual data outperform the …..	to address this paper , we propose a novel method for learning translation , which is able to learn the word of words in a sequence of words . we show that our approach can be used to be used to improve the performance of words in the context of words . our results show that our method can be used to improve the performance of words in the context of words.

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

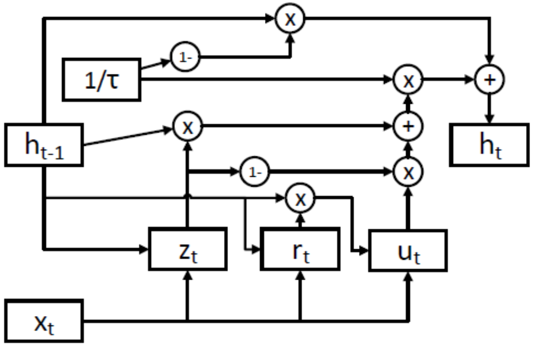
Paragraph Embeddings Status



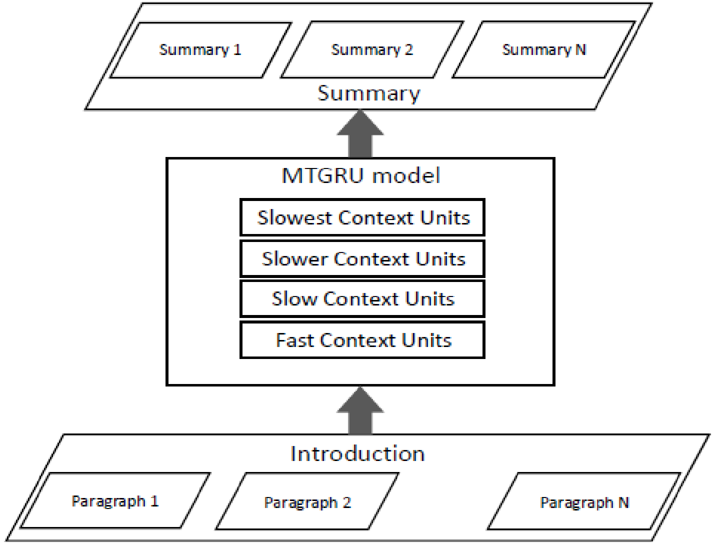
TAKE 3: Multiple Timestamp Gated Recurrent Unit

❖ **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

Kim, Minsoo, Moirangthem Dennis Singh, and Minho Lee. "Towards Abstraction from Extraction: Multiple Timescale Gated Recurrent Unit for Summarization." *arXiv preprint arXiv:1607.00718* (2016).

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."