

# Digest generation for the news articles using LSTMs.



CS230: Mikhail Sidorov (SUID: msidorov). Video link: <https://youtu.be/LPVJ7GCAqs8>

## 1. Introduction

The problem of representing document in short form is very important in different areas including news aggregation and providing search results. Last years text summarization algorithms have been significantly improved due to applying of novel deep learning techniques. The scope of the project includes applying various RNN-LSTM- based summarization approaches to the text and prepare summaries for text documents.

## 2. Dataset description

We use 2 different datasets with very different structures of text and headlines:

1) Amazon food reviews

Dataset size (articles)	Mean article length (words)	Mean summary length (words)	Total vocabulary size (words)	Occurring 10+ times (words)
568,454	80	4.1	132,884	64,183

2) Cornell NEWSROOM Summarization dataset

Dataset size (articles)	Mean article length (words)	Mean summary length (words)	Total vocabulary size (words)	Occurring 10+ times (words)
1,321,995	658.6	26.7	6,925,712	784,884

While Amazon food reviews is a much more simple data set, it allows to make quick verification of the ideas for different models.

Text preprocessing:

- 1) convert text to lowercase
- 2) tokenization and separate punctuation from words
- 3) English language contractions (don't ==> do not)
- 4) Remove infrequent words (less then 10 per text corpus)
- 5) Use <UNK> instead of unknown words and add <EOS> to the end of string

## 4. Results

We trained the model (using training set) and defined optimal hyper parameters for each data set (using the dev set). The results were evaluated using test set.

Table 4: ROUGE metrics for Amazon food reviews dataset

Method	R-1	R-2	R-L
F	15	3	13
Precision	20	5	20
Recall	14	3	14

Table 5: ROUGE metrics for NEWSROOM dataset

Method	R-1	R-2	R-L
F	5	0.5	4
Precision	6	0.3	6
Recall	4	0.2	4

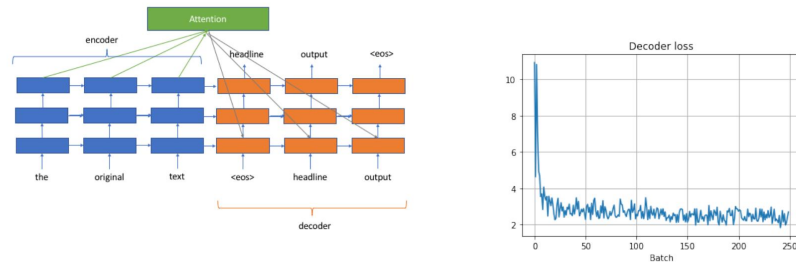
Example:

SUMMARY: A young woman has been arrested after allegedly glassing another woman during a wild brawl at a Sydney train station.

ARTICLE: A wild brawl between two women at a Sydney train station has left one with head injuries after she was struck with a glass bottle. The women were fighting at Redfern train station just before 1am on Friday before police used pepper spray to break them up. One of the women, aged 27, was taken to hospital after copping a bottle to her head ...

## 3. The model and metrics

Our model is based on encoder-decoder architecture shown in figure below. The architecture consists of two parts - an encoder and a decoder - both by themselves RNNs. Attention mechanism is used and is crucial for improving performance results.



The loss function in this case is:

$$-\log p(y_1, \dots, y_{T'} | x_1, \dots, x_T) = -\sum_{t=1}^{T'} \log p(y_t | y_1, \dots, y_{t-1}, x_1, \dots, x_T)$$

During the training of the model we optimize the loss function, but to evaluate model prediction we use discrete metrics (because of that is not be able to be used as a loss/target function).

$$ROUGE - N = \frac{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)} \quad (1)$$

## 5. Conclusions

In presented project we trained an encoder-decoder neural network with LSTM units and attention for text summarization problem. We used 2 completely different data sets and got much better result on data set with short summaries and much more simple language structure, which we can expect a priori. Approach demonstrated the feasibility and definitely it will be interesting to apply for this problem the latest techniques which were proposed in the latest articles.

We observed that text summarization is very sensitive to the text cleanup method and vocabulary. It was unexpected that we didn't get significant dependency on embedding dimension (we compare 50 and 300 sim GloVe embeddings).

## References

- [1] Abigail See, Peter J. Liu, Christopher D. Manning (2017) Get To The Point: Summarization with Pointer-Generator Networks. 2017, ACL.
- [2] Max Grusky, Mor Naaman, Yoav Artzi (2018) NEWSROOM: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies.
- [3] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. (2014.) Sequence to sequence learning with neural networks. In Neural Information Processing Systems