NLP

Case Study

Text Summarization with Attention

# Domain

Articles & Blogs

# Business Context

Text summarization is a way to condense a large amount of information into a concise form by the process of selection/generation of important information and discarding unimportant and redundant information. With the amount of textual information present in the world wide web, the area of text summarization is becoming very important.

# Objective

The objective of the case study is to learn abstractive text summarization using articles from WikiHow.

# Dataset

WikiHow is a new large-scale dataset using the online WikiHow (http://www.wikihow.com/) knowledge base.

There are two features: - text: wikiHow answers texts. - headline: bold lines as summary.

# Steps

1. Load the dataset
2. Preprocess text
3. Check the distribution of data
4. Split data into train and test set
5. Apply TensorFlow tokenizer
6. Pad sequences
7. Define model
8. Compile model
9. Fit the model
10. Create a dictionary to convert word to index
11. Add inference for encoder-decoder
12. Convert features to text

# Further Questions (Optional)

1. Try implementing Bi-Directional LSTM which is capable of capturing the context from both the directions and results in a better context vector
2. Use the beam search strategy for decoding the test sequence instead of using the greedy approach (argmax)
3. Evaluate the performance of your model based on the BLEU score
4. Implement pointer-generator networks and coverage mechanisms

# Learning Outcomes

1. Natural Language Processing
2. Abstractive text summarization
3. Encoder-decoder architecture
4. LSTM
5. Attention models