Text summarization on Amazon food reviews

Prof. Kalpana Deorukhkar[1], Deljin Jaison[2], Sanjeev Hippurgikar[3], Shubham Ambilkar[4]
[1,2,3,4]Department of Computer Engineering, University of Mumbai
Fr. Conceicao Rodrigues College of Engineering, Bandra(W), Mumbai- 50, India

Abstract- Every day, people rely on a wide variety of sources to stay informed from news stories to social media posts to search results. Being able to develop a model that can automatically deliver accurate summaries of longer text can be useful for digesting such large amounts of information in a compressed form. In this project, we make use of a deep learning model to tackle text summarization which involves generating a short summary for a longer piece of text. AText Summarization is broadly classified into mainly two types: extractive summarization and abstractive summarization.Extractive summarization is an approach in which the user selects passages from the source texts which are in long passage form and then arranges it to form a summary. Abstractive summarization means an abstractive approach which involves understanding the intent and writes the summary in its own words which are small and concise. This paper focuses on LSTM with attention-based encoder to solve the challenge of abstractive summarization on food reviews data and evaluated their effectiveness using ROUGE, BLEU score.

Keywords— Text summarization, RNN, LSTM, Attention mechanism.

1. INTRODUCTION

The goal of automatic text summarization is presenting the source text into a shorter version with semantics. The most important advantage of using a summary is, it reduces the reading time. There are two different groups of text summarization: indicative and informative. Inductive summarization only represents the main idea of the text to the user. The typical length of this type of summarization is 5 to 10 percent of the main text. On the other hand, the informative summarization systems give concise information of the main text. The length of informative summary is 20 to 30 percent of the main text. Over the past five decades, research on text summarization is widely seen in numerous applications related to information retrieval, intelligence gathering, information extraction, text mining, and indexing. Automatic document summarization is the process of reducing the size of document preserving the important semantic content. Its purpose is to identify a summary of a document without reading the entire document. The main goal of a summary is to present the main ideas in a document, in less space. For general audience, the summarization technique has got variety of application like overview, gist and concept mapping. When focusing on a specific group of audience, the need and the purpose vary accordingly.

As readability is an essential component in text comprehension. Extractive summarization lacks cohesion and sentence ordering. Abstractive summaries are more accurate as compared to the extractive summary summarization uses either statistical or linguistics approaches or combination of both to generate summary. While extractive summarization is mainly concerned with what the summary content should be, usually relying solely on extraction of sentences abstractive summarization puts strong emphasis on the form, aiming to produce a grammaticalsummary, which usually requires advanced language generation techniques.

2. RELATED WORK

With article digests gaining more and more growth, the task of generating intelligent and accurate summaries for long pieces of text has become a popular research as well as industry problem.There are two fundamental approaches in which to text summarization is done**:** extractive and abstractive**.Extractive Summarization:**Rada et. al. [8]have used a graph-based ranking model for text processing, and this model can be successfully used in natural language applications. Graph-based ranking algorithm is a way of deciding on the importance of a vertex within a graph, by taking into account global information recursively computed from the entire graph, rather than relying only on local vertex-specific information. Applying a similar line of thinking to lexical or semantic graphs extracted from natural language documents, results in a graph-based ranking model that can be applied to a variety of natural language processing application.

Luhn method [9] significance of a sentence is derived with the analysis of the word.It is here proposed that frequency of the word in article furnishes a useful measurement for word significance. Further the sentence is also rated with the help of the significant words in it. After ranking the useful sentences, display those sentences in the order of appearance in the article.

Abstractive Summarization: Google's Textsumhas shown good results after training on 4 million pairs from the Gigaword dataset of the form (first two sentences, headline). During training it optimizes the likelihood of the summary given the article's first two sentences. Both the encoding layer and language model are trained at the same time. In order to generate a summary it searches the space of all possible summaries to find the most likely sequence of words for the given article.

3. PROPOSED SYSTEM

In this section, the proposed approach is explained in detail. Our model was based on two-layered bidirectional RNN with LSTMs[7] on the input data and two layers, each with an LSTM[7] using bahdanau attention[1] on the target data or can use luong attention[3].

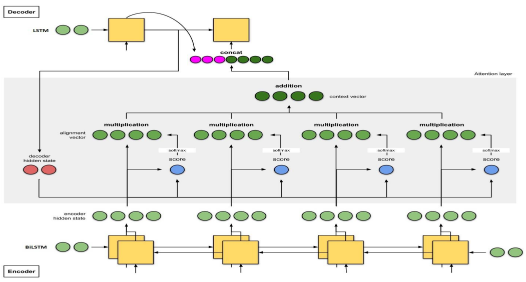

Figure 1: Bidirectional LSTM with Bahdanau Attention

For abstractive summarization, it is commonplace to use either GRU or LSTM cells for theRNN encoder and decoder.

We elected to use LSTM cells for their extra control via their memory unit, although many top models use GRU cells for their cheaper computation time[4]. Following are the stepsperformed to achieve summarized text.

3.1 Inspecting the data

In this we removed all the unwanted features from the dataset i.e the unwanted columns after that deleted that rows which had null values considering that null value never contribute any information to the model. Methods used to remove unwanted features is .drop and remove null values is .dropna both are methods available in pandas.

3.2 Preparing the data

Here we used a number of contractions to replace the short words to their respective long forms such as ain't to are not. Followed by we removed unwanted characters and stopwords with the help of nltk library and regular expressions. In this we also assigned the vector value to the words from conceptnetnumberbatch embeddings and assigned a special value to the missing words , are the words which are not present in wordtovec library.

3.3 Building the model

We used tensor flow built-in methods to build our bidirectional LSTMseq2seq model[7]. The hyperparameters set are as follows epochs = 5, batch_size = 64 , rnn_size = 256 , num_layers = 2 , learning_rate = 0.005, keep_probability = 0.75

3.4 Training the model

We trained our model in Google colabwith GPU settings it took around 3hrs. The trained model is saved in google drive.

To see the quality of the summaries that this model can generate, you can either create your own review, or use a review from the dataset. You can set the length of the summary to a fixed value, or use a random value like I have here.

4. EXPERIMENTAL RESULTS

This model can create relevant summaries for reviews written about fine foods sold on Amazon. Here are some examples of reviews and their generated summaries:

Description(1): The coffee tasted great and was at such a good price! I highly recommend this to everyone!

Summary(1):great coffee

Description(2): This is the worst cheese that I have ever bought! I will never buy it again and I hope you won't either!

Summary(2): omg gross gross

Description(3): love individual oatmeal cups found years ago sam quit selling sound big lots quit selling found target expensive buy individually trilled get entire case time go anywhere need water microwave spoon know quaker flavor packets

Summary(3): love it

5. ANALYSIS

We have performed analysis of our model using various evaluation algorithms such as ROUGE, BLEU & Cosine similarity.

ROUGE score: Recall-Oriented Understudy for Gisting Evaluation is a set of metrics and a software package used for evaluating automatic summarization and machine translation software in natural language processing[3]. The metrics compare an automatically produced summary or translation against a reference or a set of references (human-produced) summary or translation. In ROUGE we implemented ROUGE 1,2,L and BE for evaluating ROUGE sumeval library is used.

summary = "Great product great service"
model_summary = "great food"
ROUGE 1-0.333333
ROUGE 2-0
ROUGE L-0.333333
ROUGE BE-0

BLEU score:BLEU (bilingual evaluation understudy) is an algorithm for evaluating the quality of text which has been machine-translated from one natural language to another. Quality is considered to be the correspondence between a machine's output and that of a human: "the closer a machine translation is to a professional human translation, the better it is" – this is the central idea behind BLEU. BLEU was one of the first metrics to claim a high correlation with human judgements of quality, and remains one of the most popular automated and inexpensive metrics[6].

summary = [['Great' ,'product' ,'great', 'service']]
model_summary = ['great', 'food']
BLEU = 0.309348

Cosine similarity: A commonly used approach to match similar documents is based on counting the maximum number of common words between the documents[5].

sentence = ("""After our cat developed UT crystals we were told we would have to change her food. We tried this Proplan and subsequent visits revealed it had done the trick. No expensive specialized food - just switched from "Hairball Care" to "Urinary Care". Delivery was VERY fast.""","")
summary = ("Great product great service","")
model_summary = ("great food","")
Cosine similarity = [[0.26490647 0.]
[0. 0.]]

6. CONCLUSION

Main objective of the model was to build a model that can create relevant summaries for reviews written about fine foods sold on Amazon.As we can see in the analysis of the code that model is capable of creating short sentences or pair of words as result when fed by large sized inputs.This model can also be used in fields other than Amazon food reviews such as Financial Research, Legal Contract Analysis, Social Media Marketing.

7. FUTURE ENHANCEMENT

Our model produces factual details inaccurately and summaries sometimes repeat themselves to overcome these issues we will implement a pointer generator[6]. In summarization, one of the key challenges is to identify the key concepts and key entities in the document, around which the story revolves. In order to accomplish this goal, we may need to go beyond the word-embeddings-based representation of the input document and capture additional linguistic features such as parts-of-speech tags, named-entity tags, and TF and IDF statistics of the words. We will try to create additional look-up based embedding matrices for the vocabulary of each tag-type, similar to the embeddings for words.

References

[1] Bahdanau, Dzmitry, Kyunghyun Cho, and YoshuaBengio. "Neural Machine Translation by Jointly Learning to Align and Translate." [1409.0473] Neural Machine Translation by Jointly Learning to Align and Translate. N.p., 19 May 2016.

[2] Jean, Sébastien, Kyunghyun Cho, Roland Memisevic, and YoshuaBengio. "On Using Very Large Target Vocabulary for Neural Machine Translation – Sampled Softmax."The Neural Perspective. N.p., 26 Oct. 2016.

[3] Luong, Thang, Hieu Pham, and Christopher D. Manning. "Effective Approaches to Attention-based Neural Machine Translation." Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing(2015).

[4] Nallapati, Ramesh, Bowen Zhou, Cicero Dos Santos, CaglarGulcehre, and Bing Xiang. "Abstractive Text Summarization Using Sequence-to-sequence RNNs and Beyond."Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning (2016).

[5] Rush, Alexander M., Sumit Chopra, and Jason Weston. "A Neural Attention Model for Abstractive Sentence Summarization." [1509.00685] A Neural Attention Model for Abstractive Sentence Summarization. N.p., 03 Sept. 2015.

[6] Abigail See, Peter J. Liu, Christopher D. Manning. "Get To The Point: Summarization with Pointer-Generator Networks."

[7] Klaus Greff, Rupesh K. Srivastava, Jan Koutn´ık, Bas R. Steunebrink, Jurgen Schmidhuber "LSTM: A Search Space Odyssey"[1503.04069]

[8]Mihalcea, Rada, and Paul Tarau. "Textrank: Bringing order into text." Proceedings of the 2004 conference on empirical methods in natural language processing. 2004.

[9] H. P. Luhn, "A Statistical Approach to Mechanized Encoding and Searching of Literary Information," IBM Journal of Research and Development , 1, No. 4, 309-317 (October 1957)"