NESTED SUMMARIZATION WITH HEADING HIERARCHY

AI 829: NLP MANDATE 1 CONTRIBUTION

Rachna S Kedigehalli
IMT2019069
rachna.kedigehalli@iiitb.ac.in

Nandakishore S Menon
IMT2019057
nandakishore.menon@iiitb.ac.in

ABSTRACT

Note taking can be a very time-consuming task. In this process, humans are able to understand the semantics of the document, extract salient features and summarize them. In addition to this, humans are also able to give headings and sub-headings to the summarized text. There is an understanding of which topics can be grouped under one heading. In this project, we attempt to do just that. Many existing solutions try to summarize or generate title for a document as a whole. Here, we want a hierarchy of headings according to the relative importance of the text they are enclosing. For this, we use a combination of summarization and title generation techniques. Summarization and title generation models are trained separately using pointer-generator model and LSTM respectively. Sentences in a document are grouped based on their similarity which will then serve as inputs to the summarization and title generation models. To decide the relative importance of headings, Word2Vec and Tf-IDF are used. These outputs will be combined to give the output in mark-down format (with heading hierarchy).

1 Introduction

It is a common notion to take notes from an article one reads: jotting down important topics and summarizing them for revisiting. While notes might prove to be useful and time-saving, preparing them is a hassle. Using summarization and classification techniques, one can obtain fairly exhaustive notes from a corpus. Automatic summarization improves the efficiency of indexing while being less bias compared to human-summarization. Making notes usually encompasses dividing the document into sections and sub-sections, giving appropriate headings and summarizing the important points in each topic.
Existing summarizing solutions try to bring the entire document under one umbrella and summarize it as a whole. However, the documents being considered in our study may contain multiple contexts, each being an important sub-topic on its own, and therefore must be given importance accordingly. These documents could be papers, stories, blogs, or even transcript of a lecture. The aim is to be able to classify a given document into headings, subheadings and summarize the content of the document under the heading hierarchy.

2 Methods and technologies used

2.1 Problem Breakdown

The problem statement can be broadly divided into two main tasks: text summarization and title generation.
Given a document, we treat each sentence in the document as a corpus and perform document classification (sentence classification in this case). For each class of sentences produced, we run Title Generation and Text Summarization (note that the two models are trained separately). The title generation step is inclusive of assigning relative importance and similarity metrics to the classes. The output of these two models are then combined to give the final output, which will be a summary of the entire document with heading hierarchy and nested summaries.

Output format A mark-down file will be best-suited as the output format for this problem as it can be easily generated as a text file while preserving the heading hierarchy.

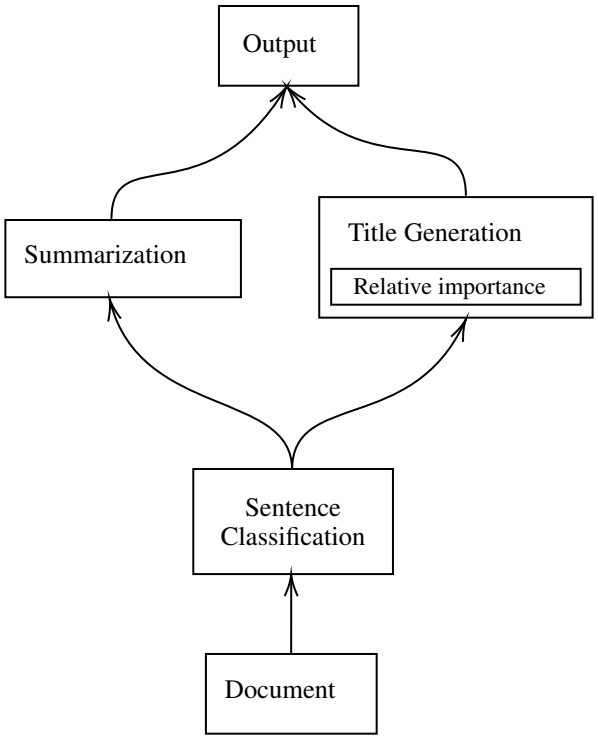

Figure 1: Block diagram representing the division of problem into sub-problems

2.2 Technologies used

For title generation, we will be using pre-trained word embeddings like Google's Word2Vec, and LSTM with an encoder-decoder architecture. We also plan on using Attention network as it is said to give better results with LSTM.(1)

The task of summarization will be done using pointer-generator model. This allows both abstractive and extractive summarizations. Extractive summarization copies/concatenates important sentences from the document into the summary. Abstractive summarization generates new sentences which are not concatenations or sentences from the corpus, but generated from a fixed vocabulary.

3 Challenges

Abstractive summarization has a high chance of causing blunders when dealing with scientific documents. It also becomes difficult to assess the relative importance of non-text elements like images, diagrams, graphs and equations. The solution would be to use a mixed approach: using both extractive and abstractive summarizations wherever they are suitable (2).

Another big challenge is to analyse the relative importance of headings to form the hierarchy. The topics generated from the segmented document, need to be assigned a metric that allows us to create a nested hierarchy. We plan on achieving this using tools like Word2Vec and Tf-IDF.

4 Dataset

We will be using the CNN/Daily Mail dataset (3), a dataset of news stories in CNN and Daily Mail popularly used for text summarization. It was used in around 294 papers according to (4). A dataset of news stories will contain a wide variety of topics, which makes this dataset ideal for our problem. If this dataset proves to be insufficient for the heading generation task in our problem, we will further use the Gigaword dataset (5).

5 Evaluation Metrics

- ROUGE (Recall-Oriented Understudy Gisting Evaluation) is a recall-based evaluation metric. It measures how much of the reference summary is captured within the generated summary. (6)
- BLEU (Bilingual Evaluation Understudy) is a precision-based evaluation metric. Precision captures the extent to which the content of the generated summary is actually needed (appears in the reference summary). (6)

The recall and precision scores can be combined using F_1-score to get a better metric.

The same metrics can be used for title generation as well.

References

[1] I. Agarwal. Title generation using nlp. [Online]. Available: https://medium.com/@ishita19013/title-generation-using-nlp-440fa1156e97

[2] T. Buonocore. Towards data science. [Online]. Available: https://towardsdatascience.com/summarizing-medical-documents-with-nlp-85b14e4d9411

[3] CNN/DailyMail. Dataset. [Online]. Available: https://cs.nyu.edu/~kcho/DMQA/

[4] PapersWithCode. Datasets. [Online]. Available: https://paperswithcode.com/datasets

[5] Tensorflow. Gigaword. [Online]. Available: https://www.tensorflow.org/datasets/catalog/gigaword

[6] A. F. Bothe, A. Truesdale, and L. Kolbe. State of the art summarisation techniques. [Online]. Available: https://humboldt-wi.github.io/blog/research/information_systems_1920/nlp_text_summarization_techniques/