Text Summarization of spoken words from user-generated instructional videos to make online content more accessible.

Sunit Carpenter Ashwini Bhingare

Abstract

This paper discusses the implementation of extractive and abstractive text summarization using different algorithms on transcripts from instruction videos, articles and wiki types of text. Summarization of speech is a difficult problem due to the spontaneity of the flow, disfluencies, and other issues that are not usually encountered in written texts. Our work builds on the work done by ([2] Savelieva et al., 2020) where the BERT model is leveraged to summarize conversational language. We generate abstractive summaries of narrated instructional videos across a wide variety of topics, from painting, gardening, cooking to software configuration and sports. In order to enrich the vocabulary, we use transfer learning and pretrain the model on a few large crossdomain datasets in both written and spoken English. We experimented with concatenating multiple extractive and abstractive summaries and then performing an abstractive summary on the aggregate. We also did pre-processing of transcripts to put in proper punctuations and to restore sentence segmentation. The results are evaluated with ROUGE and Content-F1 scoring for the How2 and WikiHow datasets. Based on visual evaluation, we achieved a similar level of textual fluency and utility compared to the summaries provided by the creator of the content.

1. Introduction:

Creating a summary from a given piece of video content is a very abstract process that all human beings participate in. Automating this process for speech summarization can help us parse through a lot of information and make better use of our time. Given the sheer volume of instructional videos out there, with proper Summarization of the text of spoken words would help users digest the materials faster and help locate the instructional videos that have the content they are looking for.

There are many complexities involved in summarization of narrated instructions such as using casual language, filler words and professional jargon. We plan to focus on the extractive and abstractive summarization to help generate the most accurate summarization. Extractive summarization extracts words and phrases

from the video transcript to create a summary while abstractive summarization learns the internal language representation to generate more human-like summaries and paraphrases the intent of the original text.

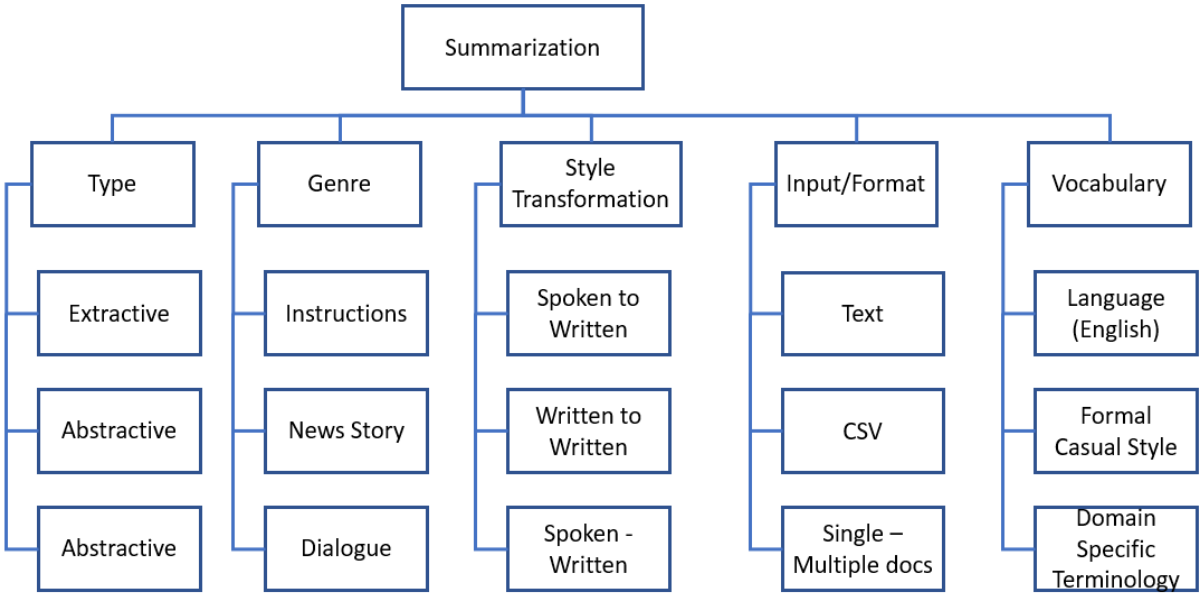

Figure-1 Taxonomy of summarization types and methods.

2. Background:

We have listed a taxonomy of summarization types and methods in Figure-1. According to ([5] Prasasthy) text summarization was done exclusively using rule-based algorithms where lines were extracted from documents and statistical models. Text summarization using neural networks was an important development in the Natural language processing area. In this method the neural network is trained on a corpus of articles and further modified using feature fusion to create a summary with highly ranked sentences in an article.

CNN/DailyMail	287,113
WikiHow	188,146
How2 Videos	65,000
Total Training DataSet	**540,259**

Table-1 Dataset Information

3. Methodology:

3.1 Task

In our analysis, we used both an extractive BERT-based model as well as a pre-trained Abstractive model to hypothesize our model's ability to form coherent summaries across a wide variety of texts in our datasets. We aim to create a pipeline architecture for text summarization tasks specifically in the use case for allowing users to be able to obtain a high level summary of relevant background information to the How2/WikiHow articles.

3.2 Data Collection

Earlier research into summarizers let us to theorize that our model would benefit from training across multiple training datasets and larger amounts of data. Table 1 illustrates various dataset sizes that we leveraged in our work. All training datasets include written summaries. There was a huge disparity in the length of the text which ranged from a few short sentences to short paragraphs to multiple paragraphs that were professionally written.

- **How2 Dataset**: The corpus consists of around 80,000 instructional videos (about 2,000 hours) with associated English sub-titles and summaries. It includes human written summaries which video owners were instructed to write summaries to maximize the audience. Summaries are two to three sentences in length.
- **CNN/DailyMail dataset**: CNN and DailyMail includes a combination of news articles and story highlights written with an average length of 119 words per article and 83 words per summary. Articles were collected from 2007 to 2015.
- **Wikihow dataset**: a large scale text dataset containing over 188,000 single document summaries. Wikihow is a consolidated set of recent How To instructional texts compiled from wikihow.com, ranging from topics such as "How to deal with coronavirus anxiety" to "How to play Uno".

3.2 Preprocessing

Due to the wide variety and complexity of our input data, we built a pre-processing pipeline to align the data to a format that will lead to better summarization. During our exploratory data analysis, we observed many issues related to lack of punctuation, incorrect wording, and various introductions which impacted model training. Considering all of these challenges, our model misinterpreted text segment

boundaries and produced poor quality summaries. In some cases, the model failed to produce any summary at all.

In order to maintain the flow, fluency and coherence in human written summaries, we cleaned and restored sentence structure and we applied entity detection from an open-source software library for advanced natural language processing called spacy and nltk. In addition, we also experimented with three different pre-processing approaches to test our model.

- Pre-Processing Approach 1:
    - Remove Special Characters from Text
    - Remove Stop Words from Text
    - Lemmatize Text
- Pre-Processing Approach 2:
    - Remove double punctuations and cr-lf
    - Remove greeting words like 'hi', 'hello', ..
- Pre-Processing Approach 3:
    - Remove invalid and non-english words.

3.3 Summarization models

We leveraged the following models to do summarization:

- Extractive Summary Model (BERT)
- Custom Extractive Summary Model with Sentence Ranking: We did not end up using this model as it performed poorly compared to the BERT Extractive model.
- Abstractive Summary Model (pre-trained BERT2BERT for CNN/Dailymail)
- Abstractive T5 Transformer Model (trained on our how2/wikihow/cnn dataset).

3.4 Scoring of results

Results were scored using ROUGE, the standard metric for abstractive summarization. While we expected a correlation between good summaries and high ROUGE scores, we observed examples of poor summaries with high scores and good summaries with low ROUGE scores.

Additionally, we added Content F1 scoring, a metric proposed by Carnegie Mellon University to focus on the relevance of content. Similar to ROUGE, Content F1 scores summaries with a weighted f-score and a penalty for incorrect word order. It also discounts stop and buzz words that frequently occur in the How-To domain, such as learning from experts how to in this free online video.

4 Experiments and Results

4.1 Results Variation and Evaluation:

Our results across the 4 different models were very consistent for the non-preprocessed scenario as seen below in Table-2. For the 3 different types of preprocessing we did, we saw very different results when evaluated using rouge scores. We attribute this to the loss of semantics in our pre-processing efforts.

One of the unique approaches we tried was to take multiple summaries, concatenate them and run them through Abstractive summarization again. Based on the Rogue scores, our unique approach did not provide a better summary. The summary was on par or worse than the ones provided by the best model.

Based on the Rouge scores, we have a tie for the most successful model. The BERT Extractive and T5 Abstractive model have the highest score of 36. Our results were not able to match the results generated by ([2] Savelieva). They were able to get Rouge scores in the 43-49 range with a high of 59 for the multi-modal model.

4.2 Results:

No Pre-Processing				**Pre-Processing 1 (del intro, Stop Words, Lemmatize)**			
	Rouge-1	Rouge-L	Content-F1		Rouge-1	Rouge-L	Content-F1
BERT Extractive	36.29	30.52	25.4	BERT Extractive	5.89	6.33	4.12
Abstractive	33.91	29.1	27.66	Abstractive	26.46	25.14	22.15
Abstractive T5	32.98	30.51	29.69	Abstractive T5	35.51	31.14	23.62
Aggregate Abstractive	33.35	29.9	27.35	Aggregate Abstractive	28.21	26.02	23.68

Pre-Processing 2 (Remove special chars, greetings, ..)				**Pre-Processing 1 (del Into, Stop Words, Lemmatize)**			
	Rouge-1	Rouge-L	Content-F1		Rouge-1	Rouge-L	Content-F1
BERT Extractive	29.45	25.22	21.43	BERT Extractive	6.57	7.01	4.51
Abstractive	32.34	27.83	26.82	Abstractive	26.5	24.67	22
Abstractive T5	30.81	28.2	29.59	Abstractive T5	36.02	31.44	23.27
Aggregate Abstractive	30.28	27.47	25.48	Aggregate Abstractive	28.8	26.02	23.76

Table - 2: Summarization scores.

5 Conclusion:

This paper explores the implementation of extractive and abstractive text summarization using different algorithms to capture the essence of what the original article or video transcript was all about.

References:

[1] Andras Huebner,Wei Ji, Xiang Xiao: 2021 : Meeting Summarization with Pre-training and Clustering Methods.

[2] Abstractive Summarization of Spoken and Written Instructions with BERT Alexandra Savelieva, Bryan Au-Yeung, Vasanth Ramani

[3] KARL-Trans-NER: Knowledge Aware Representation Learning for Named Entity Recognition using Transformers Avi Chawla, Nidhi Mulay, Vikas Bishnoi, Gaurav Dhama

[4] Avi Chawla, Nidhi Mulay, Vikas Bishnoi, Gaurav Dhama : 2021 : KARL-Trans-NER: Knowledge Aware Representation Learning for Named Entity Recognition using Transformers

[5] Prasasthy K B : Brief history of Text Summarization - Medium.com