Extractive Summarization of Meetings

Authors: Derrie Susan Varghese, Sai Divya Sangeetha Bhagavatula

Github: https://github.com/derrie96/Extractive-Summarization-of-Meetings

1. Summary

In this virtual world, meetings have become a major source of communication. These meetings tend to last for a few hours and a lot of information is exchanged. The important information needs to be documented and doing it manually can be tiresome. Therefore, in this project, we would like to build a model that automatically summarises meetings. All the organizations that use meetings as their means of communication are the stakeholders of this project.

Summarization is of two types, Extractive and Abstractive. In extractive summarization, the important sentences in the document are selected and the summary is generated using the original text in the document. Whereas, in abstractive summarization, the important sentences are selected and paraphrased. In this project, we aimed to perform extractive summarization of meeting transcripts. Our dataset was obtained from AMI Corpus which contains 100 hours of meeting recordings. The transcripts of these meeting recordings are available in this repository. Out of 687 transcripts, only 137 were useful as the remaining did not have reference summaries.

In this project, we aimed to use the existing models that were developed for the extractive summarization tasks, on our meetings dataset and compared their performances. We used BERT-Extractive-Summarizer and BERTSUM models to generate the summaries. Also, we used a coreference technique called Neural Coref to improve the performance of the BERT Extractive Summarizer. This method received the best ROUGE-1 score of 46%.

2. Methods

Pre-trained models were used to carry out the task of extractive summarization. The initial text cleaning and preprocessing was performed before feeding the data into the models. Also, to evaluate the performance of the models, ROUGE score was used as the metric. Following sections describe in detail the steps we performed to achieve the goal of summarization.

2.1 Data Cleaning and Data Pre-processing

The raw meeting transcripts and summaries were messy with interjections and repetitive punctuations as shown in *Fig. A* in the Appendix. Therefore, during phase one, we removed these fillers and punctuations. This was followed by formatting the files into '.story' format, where each transcript and its corresponding summary is placed in a single file. To mark each line of the summary, '@' tokens were used. Subsequently, Stanford Core NLP was used for tokenization, which generates linguistic annotations for text, including token and sentence boundaries, parts of speech, named entities, etc. These JSON files were then converted to binary files.

2.2 Label Generation

Sentences in the transcripts were not labeled to denote if they were present in the reference summaries. Therefore, to generate labels for all the transcripts, we checked which sentence in the transcript got the maximum ROUGE score for each sentence in the corresponding summary, and that particular sentence in the transcript was labeled true and the rest false. *Fig. B* in the appendix shows excerpts of a transcript and summary. The highlighted lines in the transcripts are labeled true since they are part of the summary.

2.3 Modelling

Extractive summarization being the main goal of our project, we have used pre-trained models such as BERTSUM and BERT Extractive Summarizer to carry out the task. They have received some of the highest ROUGE scores(~43%) for extractive summarization and that's why we selected them for our summarization task. BERT Extractive summarizer was previously used for summarizing lectures and BERTSUM for extracting highlights from CNN/DailyMail news articles.

During phase one, on the clean and pre-processed data, extractive summaries were generated using BERT Extractive Summarizer. BERT Extractive Summarizer uses BERT to generate text embeddings and then performs K-means clustering to identify the sentences closest to the centroid. These sentences are then selected to be added to the summary. BERT Extractive Summarizer was recently deployed as a package in Python with the same name and we used this package for summarization.

To improve the performance of the BERT Extractive Summarizer on our dataset, in the second phase of our project, a coreference resolution technique called Neural Coref was used. Co-referencing helps the model to understand that two or more expressions in a text, like pronouns or nouns, are related to the same person or object. This gives more context to the transcript of the model and thus generates better summaries. Once NeuralCoref was applied, there was a slight increase in the ROUGE scores.

The other model, BERTSUM uses BERT to obtain sentence embeddings for multiple sentences using [CLS] token at the start of each sentence and [SEP] token at the end of each sentence as shown in the architecture diagram in *Fig 1*. T_i is the vector representation from the i^{th} [CLS] token generated from the top BERT layer for the i^{th} sentence. BERT sentence vectors are stacked up with summarization-specific layers. Inter-sentence transformers are used in these layers to extract document-level features. Interval Segment Embeddings distinguish multiple sentences within a document and each sent is assigned a segment embedding E_A or E_B conditioned on whether the sentence is in odd or even position. Lastly, Positional Embeddings indicate the position of each word in the sequence. Training of BERTSUM was completed during the first phase and in the second phase validation & testing were performed on our meeting dataset.

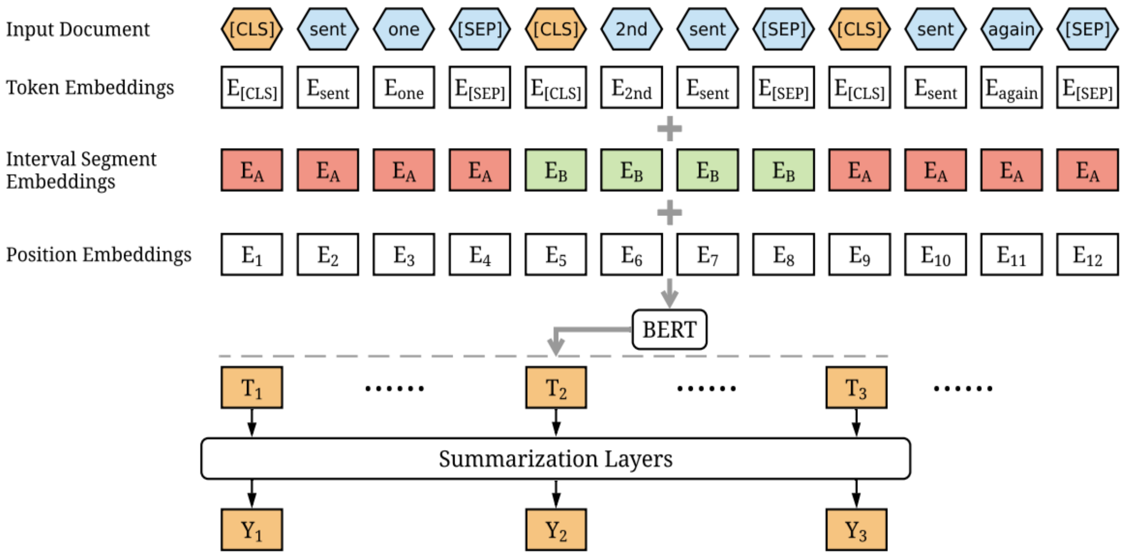

Input Document
[CLS] sent one [SEP] [CLS] 2nd sent [SEP] [CLS] sent again [SEP]
Token Embeddings
$E_{[CLS]}$ E_{sent} E_{one} $E_{[SEP]}$ $E_{[CLS]}$ E_{2nd} E_{sent} $E_{[SEP]}$ $E_{[CLS]}$ E_{sent} E_{again} $E_{[SEP]}$
+
Interval Segment Embeddings
E_A E_A E_A E_A E_B E_B E_B E_B E_A E_A E_A E_A
+
Position Embeddings
E_1 E_2 E_3 E_4 E_5 E_6 E_7 E_8 E_9 E_{10} E_{11} E_{12}
BERT
T_1 ...... T_2 ...... T_3 ......
Summarization Layers
Y_1 Y_2 Y_3

Fig 1: Architecture diagram of BERTSUM (https://arxiv.org/pdf/1906.04165.pdf)

2.4 Metric Evaluation

To evaluate the performance of the summarization models, the widely used metric for summarization called ROUGE(Recall Oriented Understudy for Gist Evaluation) score was used. Recall in the context of ROUGE is a measure of how much of the reference summary is being captured by the system summary, whereas precision is measuring how much of the system summary is relevant. ROUGE-F1 score is the harmonic mean of precision and recall. The formula for ROUGE$_{\text{recall}}$ and ROUGE$_{\text{precision}}$ are given below.

$$ROUGE_{recall} = \text{ number of overlapping words } / \text{ total words in the reference summary}$$

$$ROGUE_{precision} = \text{number of overlapping words } / \text{ total words in system summary}$$

Now let’s consider the following example:
System Summary: *There was a dog found under the table*
Reference Summary: *There was a dog under the table*

Here in the given example, ROUGE$_{\text{recall}}$=1 and ROUGE$_{\text{precision}}$= 0.87. F1 score in terms of ROUGE can be calculated by the combination of precision and recall.

On further study, we realized ROUGE-score is not a good metric for comparing summaries of varied length ([6]Simeng Sun et.al). An increase in the number of sentences in the summary initially causes an increase in ROUGE score and then it declines gradually. They propose using normalized ROUGE scores. However, the ROUGE-F1 score is widely used and thus becomes easier to compare the results of these pre-trained models on our meeting dataset to the results generated by these models on the original data(lectures and CNN/DM dataset). Therefore, we will be continuing to use ROUGE-F1 scores to evaluate our results.

3. Results

As part of the exploratory analysis in phase 1, the most frequently occurring words in our clean transcripts were plotted as shown in *Fig. 2*. It was observed that some of the most frequent words are "remote", "control" and "buttons", etc. This was consistent with the fact that most of the meetings recorded in the transcripts are regarding designing a remote control.

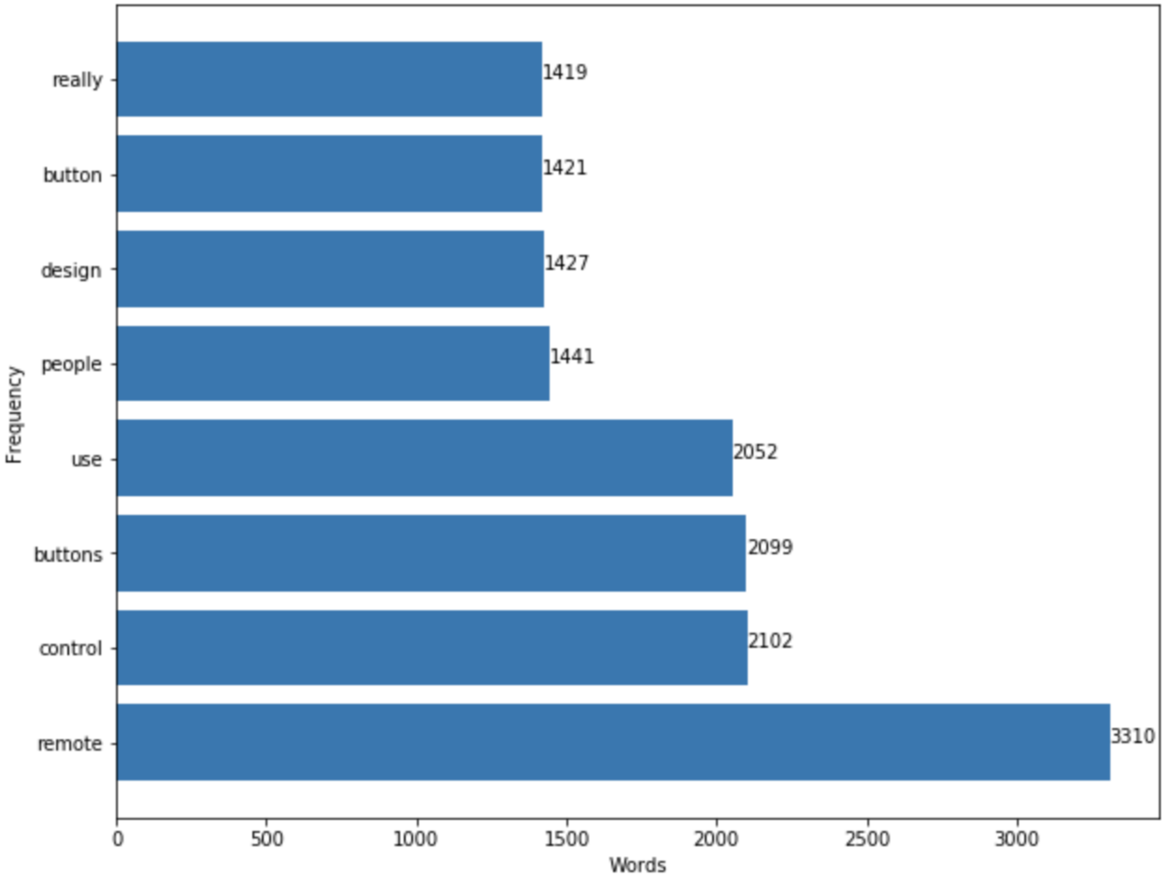


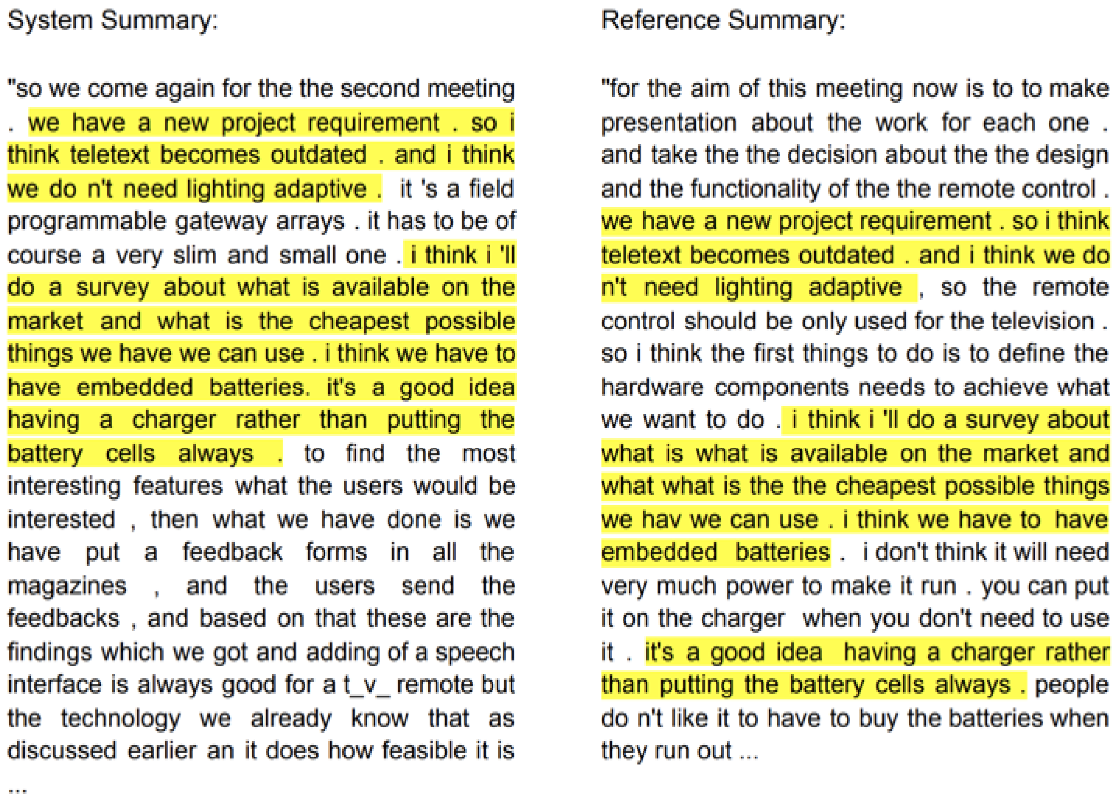
Fig. 2 Most frequent words in the transcript

The ROUGE scores of all the models we used for summarizing the meeting transcripts can be seen in Table 1. BERTSUM received a ROUGE-1 score of 41.24%, which is slightly less than what the model received on the CNN/DM dataset(43%). Whereas, BERT Extractive Summarizer with Neural Coref performed the best with a ROUGE-1 score of 48.76% on our dataset. However, the ROUGE-2 and ROUGE-L scores for all the models listed are generally low and that is also the case of BERTSUM on CNN/DM dataset with a ROUGE-2 and ROUGE-L score of 20% and 39% respectively. This shows how text summarization is still at its initial stages of research and a lot more research is required to improve the various methods of summarization.

MODEL	AVG ROUGE-1 SCORE (%)	AVG ROUGE-2 SCORE (%)	AVG ROUGE-L SCORE (%)
BERT Extractive summarizer	46.88	13.54	35.36
BERT Extractive summarizer + NeuralCoref	48.76	14.52	37.71
BERTSUM	41.24	12.35	34.38

Table. 1 F-1 ROUGE scores for all the models

Output summary generated using Bert Extractive Summarizer with NeuralCoref can be seen below in *Fig. 3.* On the left is the excerpt of the system-generated summary and on the right is the excerpt of the reference summary. The actual summaries are much longer than what is shown here. The sentences highlighted in yellow are the sentences that are common in both reference and model-generated summaries.



System Summary:

"so we come again for the the second meeting . we have a new project requirement . so i think teletext becomes outdated . and i think we do n't need lighting adaptive . it 's a field programmable gateway arrays . it has to be of course a very slim and small one . i think i 'll do a survey about what is available on the market and what is the cheapest possible things we have we can use . i think we have to have embedded batteries. it's a good idea having a charger rather than putting the battery cells always . to find the most interesting features what the users would be interested , then what we have done is we have put a feedback forms in all the magazines , and the users send the feedbacks , and based on that these are the findings which we got and adding of a speech interface is always good for a t_v_ remote but the technology we already know that as discussed earlier an it does how feasible it is ...

Reference Summary:

"for the aim of this meeting now is to to make presentation about the work for each one . and take the the decision about the the design and the functionality of the the remote control . we have a new project requirement . so i think teletext becomes outdated . and i think we do n't need lighting adaptive , so the remote control should be only used for the television . so i think the first things to do is to define the hardware components needs to achieve what we want to do . i think i 'll do a survey about what is what is available on the market and what what is the the cheapest possible things we hav we can use . i think we have to have embedded batteries . i don't think it will need very much power to make it run . you can put it on the charger when you don't need to use it . it's a good idea having a charger rather than putting the battery cells always . people do n't like it to have to buy the batteries when they run out ...

Fig 3. Excerpt of summary generated by BERT Extractive Summarizer with NeuralCoref and corresponding reference summary

4. Discussion

BERT Extractive Summarizer with NeuralCoref performed the best out of the two models. The co-referencing technique gives the model more context about the transcripts and thus the model performed better. Also, while running the models we observed that BERTSUM required higher computation time and power. Another major challenge we faced was the lack of a better dataset. The dataset only contained 137 transcripts with reference summaries. Moreover, the reference summaries were not the best, since they contained unnecessary conversational bits. A larger data and better-written reference summaries would have improved the performance of the model. We also learned that ROUGE-F1 score is not the best metric to evaluate the performance of the models which are built using neural networks, since they contain summaries of varied length.

In the future, this project can be improved by using a larger and diverse input dataset. Also, a speech-to-text module can be added to make it more efficient. Also, we could focus on generating abstractive summaries, since it is more comparable to man-made summaries, as it paraphrases sentences in the transcript. Additionally, creating a web application that can transcribe and summarize would make it more versatile and user-friendly.

5. Statement of Contribution

The initial task of text preprocessing was performed by Sangeetha. Subsequently, labels were generated for all the transcripts using a brute-force algorithm by Derrie. Formatting the files before being fed as input to models, was carried out by Sangeetha. The BERT Extractive Summarizer was executed and evaluated by both Sangeetha and Derrie. Training of BERTSUM was done by Derrie.

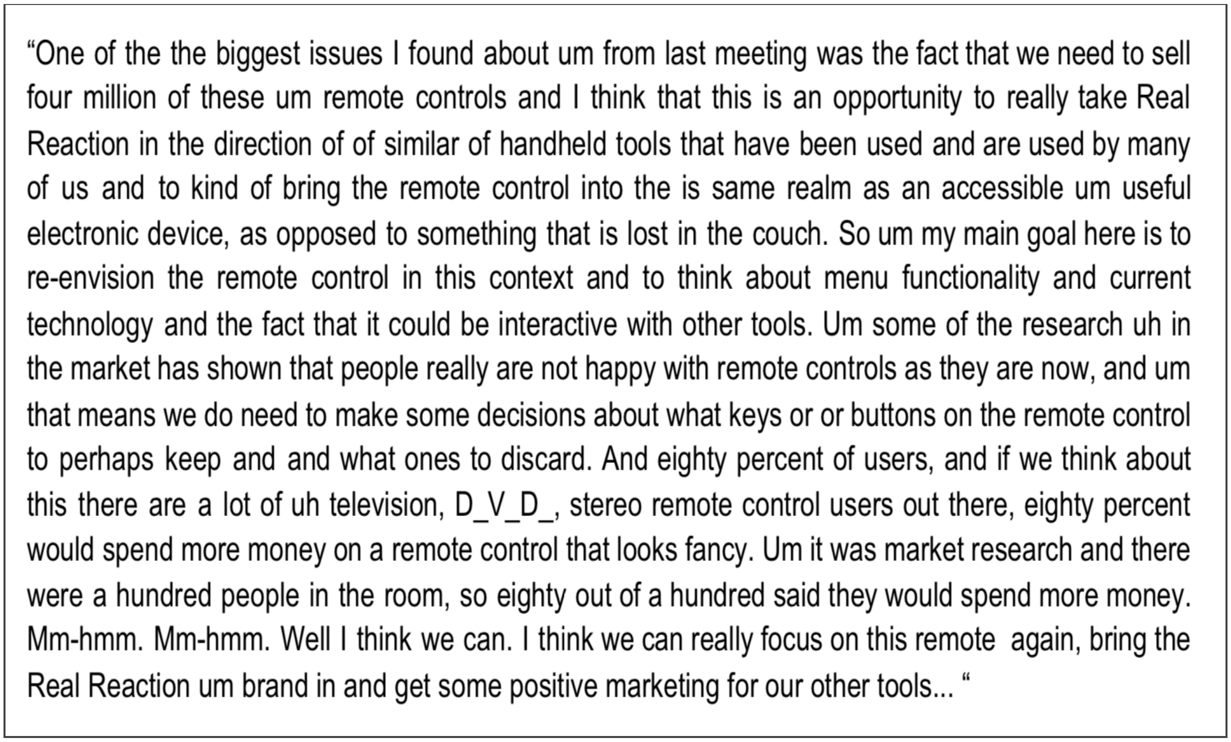
As a part of phase 2, BERT Extractive Summarizer with neural co-ref was implemented and evaluated by Sangeetha. Derrie performed validation of the BERTSUM model and Sangeetha tested it using the test set. Finally, we compared the performances of all the models and contributed equally to documenting the project.

6. References

1. Derek Miller, “Leveraging BERT for Extractive Text Summarization on Lectures”, 2019
   URL: https://arxiv.org/abs/1906.04165
2. Yang Liu, “Fine-tune BERT for Extractive Summarization”, 2019
   URL: https://arxiv.org/pdf/1903.10318.pdf
3. “Fine-tune BERT for Extractive Summarization - nlpyang/BertSum”
   URL: https://github.com/nlpyang/BertSum

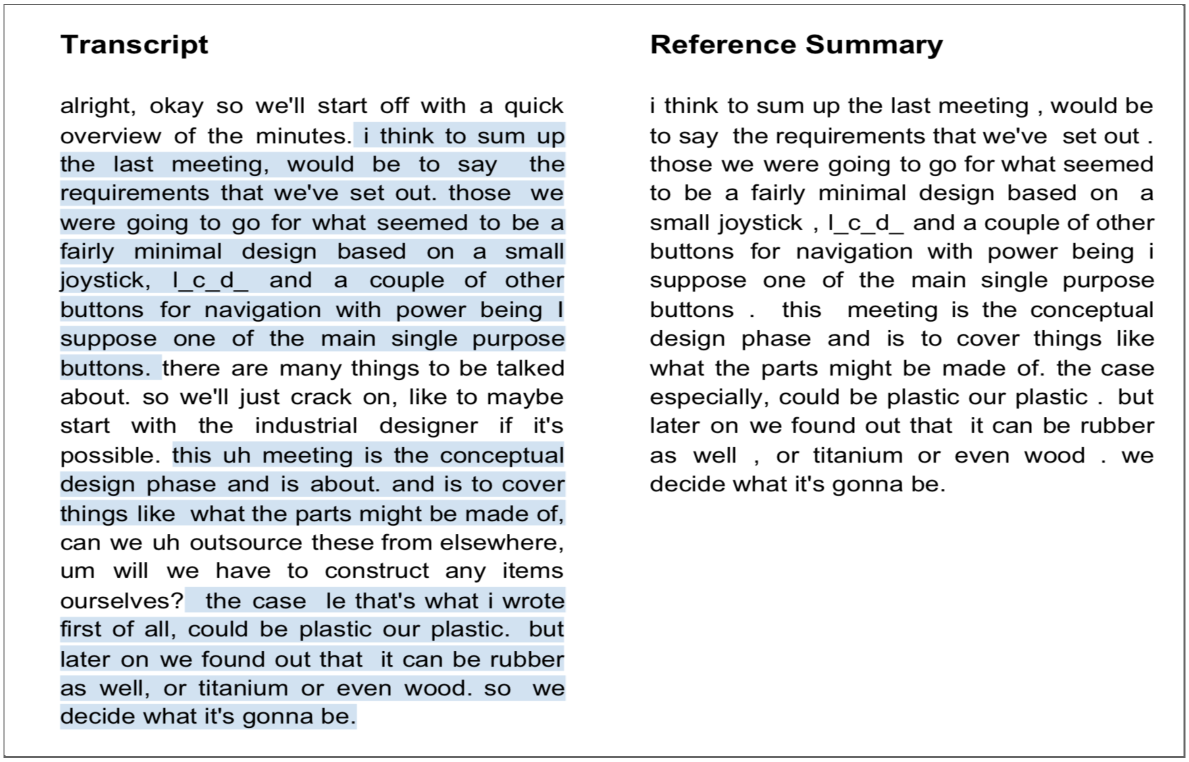
4. “Deep Learning Models for Automatic Summarization”, URL: https://towardsdatascience.com/deep-learning-models-for-automatic-summarization-4c2b89f2a9ea
5. “NLP Contextualized Word Embeddings from BERT”, URL:https://towardsdatascience.com/nlp-extract-contextualized-word-embeddings-from-bert-keras-tf-67ef29f60a7b
6. “How to Compare Summarizers without Target Length? Pitfalls, Solutions, and Re-Examination of the Neural Summarization Literature”, 2019
URL: https://www.aclweb.org/anthology/W19-2303.pdf

Appendix



“One of the the biggest issues I found about um from last meeting was the fact that we need to sell four million of these um remote controls and I think that this is an opportunity to really take Real Reaction in the direction of of similar of handheld tools that have been used and are used by many of us and to kind of bring the remote control into the is same realm as an accessible um useful electronic device, as opposed to something that is lost in the couch. So um my main goal here is to re-envision the remote control in this context and to think about menu functionality and current technology and the fact that it could be interactive with other tools. Um some of the research uh in the market has shown that people really are not happy with remote controls as they are now, and um that means we do need to make some decisions about what keys or or buttons on the remote control to perhaps keep and and what ones to discard. And eighty percent of users, and if we think about this there are a lot of uh television, D_V_D_, stereo remote control users out there, eighty percent would spend more money on a remote control that looks fancy. Um it was market research and there were a hundred people in the room, so eighty out of a hundred said they would spend more money. Mm-hmm. Mm-hmm. Well I think we can. I think we can really focus on this remote again, bring the Real Reaction um brand in and get some positive marketing for our other tools... “

Fig. A. An excerpt of a raw transcript from our dataset



Transcript

alright, okay so we'll start off with a quick overview of the minutes. i think to sum up the last meeting, would be to say the requirements that we've set out. those we were going to go for what seemed to be a fairly minimal design based on a small joystick, l_c_d_ and a couple of other buttons for navigation with power being I suppose one of the main single purpose buttons. there are many things to be talked about. so we'll just crack on, like to maybe start with the industrial designer if it's possible. this uh meeting is the conceptual design phase and is about. and is to cover things like what the parts might be made of, can we uh outsource these from elsewhere, um will we have to construct any items ourselves? the case le that's what i wrote first of all, could be plastic our plastic. but later on we found out that it can be rubber as well, or titanium or even wood. so we decide what it's gonna be.

Reference Summary

i think to sum up the last meeting , would be to say the requirements that we've set out . those we were going to go for what seemed to be a fairly minimal design based on a small joystick , l_c_d_ and a couple of other buttons for navigation with power being i suppose one of the main single purpose buttons . this meeting is the conceptual design phase and is to cover things like what the parts might be made of. the case especially, could be plastic our plastic . but later on we found out that it can be rubber as well , or titanium or even wood . we decide what it's gonna be.

Fig. B. Excerpt of cleaned and labeled transcript and reference summary from the dataset