# Contextual Meeting Segmentation using Pointer Mechanism

Jay Nitin Paranjape[*]
jay.edutech@gmail.com
Indian Institute of Technology Delhi

Bipasha Sen
bise@microsoft.com
Microsoft

## ABSTRACT

We propose a novel technique of contextual meeting segmentation for the task of meeting summarization. Unlike documents, meetings often span over multiple topics spread throughout the course of the meeting. In order to capture the true summary of the meeting, it is important to capture the summary of each of the topics present in the meeting, that is, segmenting the meetings into different topics to generate summary for each of them. The segmentation approaches existing today ignore the fact that the sentences belonging to the same context can be continuous or non-continuous in nature. In this work, we solve the problem of contextual meeting segmentation using pointer mechanism to extract the related sentences from a given meeting transcription without assuming that the sentences are consecutive in nature. We use the widely used AMI meeting dataset and a modified BLEU score as a metric for evaluation.

## KEYWORDS

Meeting Summarization, Pointer Mechanism, Sentence Modelling

## 1 INTRODUCTION

Unlike the well known domain of document summarization, where the focus is on a single topic, meetings focus on multiple points of discussion. A holistic meeting summary can be pictured as a summary of the individual point of discussions in the meeting. The extensive research in the domain of meeting summarization has focused on partitioning the meetings into exclusive segments and generating a summary for each of the segments. Zhu et al. [8], used this technique by employing Segbot [3], a pointer network which for a given start sentence, extracts the end sentence from a given set of continuous sentences, such that all the sentences belong to one context. This method however, fails to capture the sentences which belong to the same context but are non-continuous in nature which is a very common occurrence in real-world scenarios. Hang et al. [7] on the other hand, used an unsupervised approach to cluster the sentences contextually and in turn generated the summary for each of the cluster. However, due to the unsupervised nature, this approach fails to learn the contextual dependencies between the sentences and in turn suffers from weak clusters.

In this work, we build an end-to-end pipeline for supervised contextual meeting segmentation for continous and non-continous context. That is, for a given meeting transcript, we extract the sentences that are the part of the same topic. We use a novel concept of sentence modeling to extract the sentences which might or might not be consecutive in nature. For a given meeting transcript, first we obtain the embedding (vectorized representation) for each of the sentences. These embeddings are passed through an encoder-decoder network. The encoder captures the temporal dependencies between the sentences using a recurrent neural network (RNN). The decoder is a pointer network, which for a given sentence, extracts all the other sentences which belong to the same context from the meeting using pointing mechanism. For the scope of this project, we have concentrated on extracting one topic for each meeting, that is, the decoder receives the first sentence of the meeting and extracts all the sentences with respect to the first sentence.

Our contribution in this work is two fold.

(1) Identifying an important problem statement of contextually segmenting the meeting such that both continuous and non-continuous contexts are captured. This is important for capturing holistic meeting summaries.
(2) Proposing a novel technique based on sentence modeling to tackle the problem.

## 2 PROPOSED ARCHITECTURE

The proposed approach is an auto-regressive sentence model in which given the history of sentences, the model extracts the next best sentence from the given vocabulary. Here, for each meeting, the sentences present in the transcription are treated as the vocabulary. Fig. 1 presents the overview of the proposed architecture. The encoder ($E$) is a recurrent neural network (RNN) that takes sentence embeddings denoted as $S = \{s_1, s_2, ...s_n\}$ as an input $\in R^{n_e \times n_{s_i}}$ and outputs the encoded representation of each of the sentences $\in R^{n_e \times h_e}$, where $n_e$ is the number of sentences in the transcript, $n_{s_i}$ is the length of the sentence embedding $s_i$, $h_e$ is the number of hidden units. The decoder ($D$) is another RNN which at each timestep $t$, takes three inputs, the hidden state $h_{d_{(t-1)}}$, the context vector $CV_{(t-1)}$ and the sentence embedding of the output $O_{(t-1)}$. At the time of training, the sentence embedding of the 'ideal' output at $(t-1)$ is given instead of the predicted output (teacher forcing [2]). The first decoder unit takes the hidden state of the last encoder unit, zero vector ($CV_0$) and the embedding of first sentence ($s_i$) from the meetings transcript as input.

The attention cell is a dense layer with the output dimension of $R^{n_e \times 1}$. At each timestep $t$ of the decoder unit, the attention cell takes $a_t$ as an input which is denoted as

$$a_t = h_e \oplus h_{d_t} \tag{1}$$

A softmax activation is applied on the attention cell output to obtain the context vector ($CV$) which is the probability distribution over all the sentences. An argmax operation then obtains the index of the sentence with the highest probability ($O_t$).

## 3 EXPERIMENTAL SETUP

### 3.1 AMI Dataset

We base our experiments on the widely used AMI [4] meetings dataset. The dataset provides meetings transcript, audio recordings, video recordings along with several ground truths. We use one of

[*] This work was carried out as an intern at Microsoft Indian Development Center.

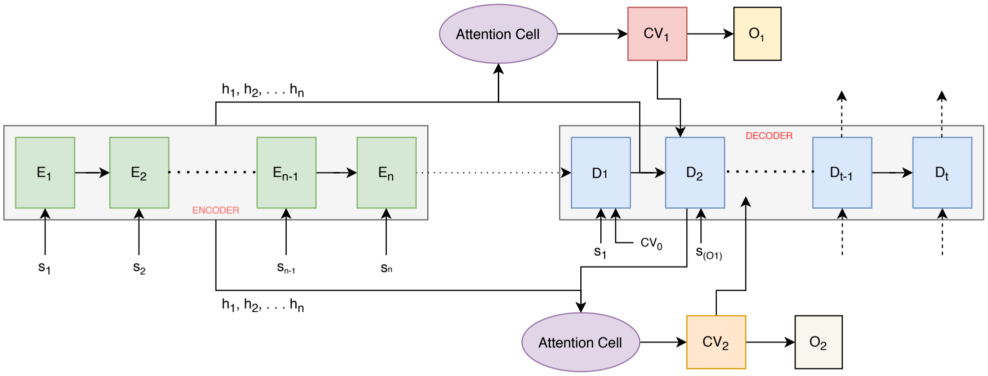

**Figure 1: Proposed Model Architecture**

the ground truths which contextually segments the meeting without the assumption of continuous contexts. The dataset contains 139 meetings with a total of 283 defined topics across all the meetings.

## 3.2 Modified BLEU Score

BLEU [5] is a commonly used metric to measure the quality of summaries. It measures the overlap between the words of the predicted and the actual output summary. For our experiments, we have modified the BLEU metric to measure the overlap between the extracted sentence indices and the expected sentence indices.

# 4 EXPERIMENTAL RESULTS

This section presents a comparitive analysis between the different approaches taken to generate the sentence embeddings as an input to the proposed pointer network. Table 1. presents the Modified BLEU score between the different approaches. Each of these approaches are defined in the subsequent sub-sections.

**Table 1: Modified BLEU scores on AMI dataset**

| Model | EDGE | USE | WEDGE | **WUSE** |
|---|---|---|---|---|
| Modified BLEU | 0.25 | 0.14 | 0.32 | **0.37** |

## 4.1 GloVe Embeddings

Global Vectors (GloVe) [6] embedding is an unsupervised learning algorithm for obtaining vector representations for words. Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space. These are pre-trained embeddings released by Stanford. In this approach (**EDGE**), we embed the individual words of each sentence using the GloVe embeddings. The sentence embeddings are generated by employing one of the two approaches.

(1) The average of the GloVe embedding of the individual words in the sentence.
(2) Learning a data specific sentence embedding by adding a dense layer on the average of GloVe word embeddings.

We obtain better results by adding the dense layer on the word embeddings to learn the sentence emebedding. The dense layer is added and learnt in an end-to-end fashion with the proposed pointer mechanism. This enables the model to encode the sentences in a form that captures the contextual relationship between the sentences in a given meeting.

## 4.2 Universal Sentence Encoder

Universal Sentence Encoder (USE) [1] encodes text into high dimensional vectors that can be used for downstream tasks such as text classification, semantic similarity, clustering and other natural language tasks. The input to USE is a variable length English text and it outputs a 512 dimensional vector. We use the USE to obtain the sentence encoding for the input (**USE**) of the proposed pointer network. USE is employed in two different fashions.

(1) As a direct input to the proposed pointer network.
(2) Adding a dense layer between the USE and the pointer network to fine-tune the sentence embeddings. We obtain better results with this approach.

## 4.3 Extending Datasets

The model is first trained on the AMI dataset. However, due to the small sample size of the ground truth, the model fails to generalize to the non-consecutive contextual sentences despite being able to capture the consecutive sentences well. We scrape Wikipedia to synthetically extend the dataset to a total of 470 data points. This approach starts including sentences which are non-consecutive in the output. Employing EDGE on the extended data (**WEDGE**) improves the base result, however, shows a bias towards sentences of lower indices. We obtain the best model (**WUSE**) by training USE on the extended dataset.

# 5 CONCLUSION AND FUTURE WORK

In this work, we identify an important problem statement of contextual meeting segmentation. This is necessary for capturing a holistic summary of the meeting transcriptions. We propose an approach based on pointer mechanism (WUSE). In order to denote its effectiveness, we compare it with three other architectures based on our proposed approach. For the scope of this project, we have focused on one context per meeting.

Currently we are extending this approach for multiple contexts while working on the robustness of the proposed approach.

# ACKNOWLEDGMENTS

We would like to thank Rajeev Gupta, Pavan Kumar Dasari, Aditya Agarwal, Peter Bailey and Rahul Jha for their insights and valuable guidance

# REFERENCES

[1] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal Sentence Encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Brussels, Belgium, 169–174. https://doi.org/10.18653/v1/D18-2029

[2] Alex Lamb, Anirudh Goyal, Ying Zhang, Saizheng Zhang, Aaron Courville, and Yoshua Bengio. 2016. Professor Forcing: A New Algorithm for Training Recurrent Networks. arXiv:1610.09038 [stat.ML]

[3] Jing li, Aixin Sun, and Shafiq Joty. 2018. SegBot: A Generic Neural Text Segmentation Model with Pointer Network. https://doi.org/10.24963/ijcai.2018/579

[4] Iain Mccowan, J Carletta, Wessel Kraaij, Simone Ashby, S Bourban, M Flynn, M Guillemot, Thomas Hain, J Kadlec, Vasilis Karaiskos, M Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska Masson, Wilfried Post, Dennis Reidsma, and P Wellner. 2005. The AMI meeting corpus. *Int'l. Conf. on Methods and Techniques in Behavioral Research* (01 2005).

[5] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 311–318. https://doi.org/10.3115/1073083.1073135

[6] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543. http://www.aclweb.org/anthology/D14-1162

[7] Guokan Shang, Wensi Ding, Zekun Zhang, Antoine Tixier, Polykarpos Meladianos, Michalis Vazirgiannis, and Jean-Pierre Lorré. 2018. Unsupervised Abstractive Meeting Summarization with Multi-Sentence Compression and Budgeted Submodular Maximization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 664–674. https://doi.org/10.18653/v1/P18-1062

[8] Chenguang Zhu, Ruochen Xu, Michael Zeng, and Xuedong Huang. 2020. End-to-End Abstractive Summarization for Meetings.