Keep Meeting Summaries on Topic: Abstractive Multi-Modal Meeting Summarization

Manling Li[1], **Lingyu Zhang**[2], **Heng Ji**[1], **Richard J. Radke**[2]
[1] Department of Computer Science
[2] Department of Electrical, Computer, and Systems Engineering
Rensselaer Polytechnic Institute
[1]{lim22,jih}@rpi.edu, [2]{zhangl34@rpi.edu, rjradke@ecse.rpi.edu}

Abstract

Transcripts of natural, multi-person meetings differ significantly from documents like news articles, which can make Natural Language Generation models generate unfocused summaries. We develop an abstractive meeting summarizer from both videos and audios of meeting recordings. Specifically, we propose a multi-modal hierarchical attention mechanism across three levels: topic segment, utterance and word. To narrow down the focus into topically-relevant segments, we jointly model topic segmentation and summarization. In addition to traditional textual features, we introduce new multi-modal features derived from visual focus of attention, based on the assumption that an utterance is more important if its speaker receives more attention. Experiments show that our model significantly outperforms the state-of-the-art with both BLEU and ROUGE measures.

1 Introduction

Automatic meeting summarization is valuable, especially if it takes advantage of multi-modal sensing of the meeting environment, such as microphones to capture speech and cameras to capture each participant's head pose and eye gaze. Traditional extractive summarization methods based on selecting and reordering salient words tend to produce summaries that are not natural and incoherent. Although state-of-the-art work (Shang et al., 2018) employs WordNet (Miller, 1995) to make summaries more abstractive, the quality is still far from those produced by humans, as shown in Table 1. Moreover, these methods tend to have limited content coverage by selecting salient words.

On the other hand, recent years have witnessed the success of Natural Language Generation (NLG) models to generate abstractive summaries. Since human-written summaries tend to mention the exact given keywords without paraphrasing, the copy mechanism proposed by a Pointer Generator Network (PGN) (See et al., 2017) naturally fits this task. Apart from generating words from a fixed vocabulary, it also copies the words from the input. However, transcripts of multi-person meetings widely differ from traditional documents. Instead of grammatical, well-segmented sentences, the input is often composed of ill-formed utterances. Therefore, NLG models can easily lose focus. For example, in Table 1, PGN fails to capture the keywords *remote control*, *trendy* and *user-friendly*.

Therefore, we propose a multi-modal hierarchical attention mechanism across topic segments, utterances, and words. We learn topic segmentation as an auxiliary task and limit the attention within each segment. Our approach mimics human summarization methods by segmenting first and then summarizing each segment. To locate key utterances, we propose that the rich multi-modal data from recording the meeting environment, especially cameras facing each participant, can provide speaker interaction and participant feedback to discover salient utterances. One typical interaction is Visual Focus Of Attention (VFOA), i.e., the target that each participant looks at in every timestamp. Possible VFOA targets include other participants, the table, etc. We estimate VFOA based on each participant's head orientation and eye gaze. The longer the speaker is paid attention by others, the higher possibility that the utterance is important. For example, in Table 1, the high VFOA received by the speaker for the last two sentences assists in maintaining the bold keywords.

2 Method

As shown in Figure 1, our meeting data consists of synchronized videos of each participant in a

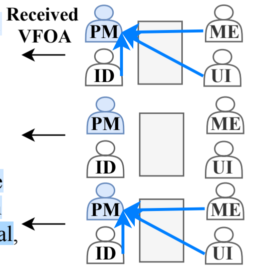
Transcript	Um I'm Sarah, the Project Managerand this is our first meeting, surprisingly enough. Okay, this is our agenda, um we will do some stuff , get to know each other a bit better to feel more comfortable with each other . Um then we'll go do tool training, talk about the project plan, discuss our own ideas and everything um and we've got twenty five minutes to do that, as far as I can understand. Now, we're developing a **remote control** which you probably already know. Um, we want it to be original, something that's uh people haven't thought of, that's not out in the shops, um, **trendy**, appealing to a wide market, but you know, not a hunk of metal, and **user-friendly**, grannies to kids, maybe even pooches should be able to use it. 
Manual summary	The project manager gave an introduction to the goal of the project , to create a **trendy** yet **user-friendly remote**.
Extractive summary (Shang et al., 2018)	hunk of metal and **user-friendly** granny's to kids.
Abstractive summary (See et al., 2017)	The project manager opened the meeting and introduced the upcoming project to the team members.
Our Approach	The project manager opens the meeting. The project manager states the goal of the project, which is to develop a **remote control**. It should be original, **trendy**, and **user-friendly**.

Table 1: Comparison of Human and System Generated Summaries. The color indicates the attention received by the speaker PM (Project Manager). Others: ME (Marketing Expert), ID (Industrial Designer), UI (User Interface).

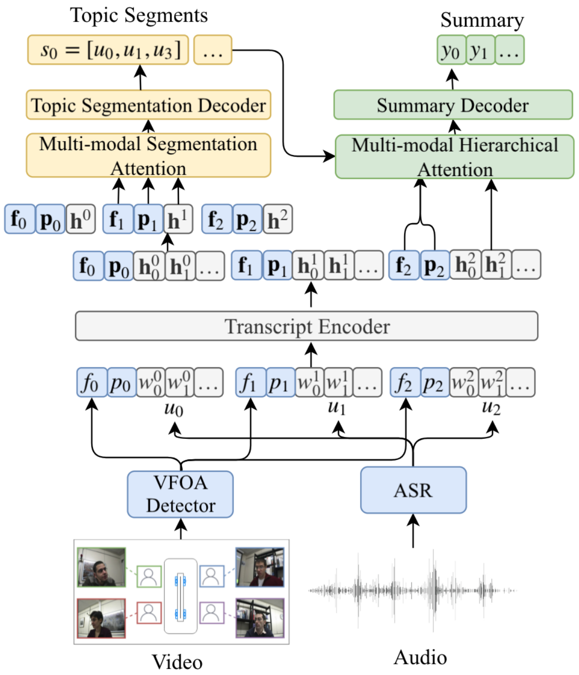

Topic Segments
$s_0 = [u_0, u_1, u_3]$ ...
Topic Segmentation Decoder
Multi-modal Segmentation Attention
Summary
y_0 y_1 ...
Summary Decoder
Multi-modal Hierarchical Attention
f_0 p_0 h^0 f_1 p_1 h^1 f_2 p_2 h^2
f_0 p_0 h_0^0 h_1^0 ... f_1 p_1 h_0^1 h_1^1 ... f_2 p_2 h_0^2 h_1^2 ...
Transcript Encoder
f_0 p_0 w_0^0 w_1^0 ... f_1 p_1 w_0^1 w_1^1 ... f_2 p_2 w_0^2 w_1^2 ...
u_0
u_1
u_2
VFOA Detector
ASR
Video
Audio

Figure 1: Multi-modal Meeting Summarization Framework

group meeting, as well as a time-stamped transcript of the utterances generated by Automatic Speech Recognition (ASR) tools [1]. We formulate a meeting transcript as a list of triples $X = \{(p_i, f_i, u_i)\}$. $p_i \in P$ is the the speaker of utterance u_i, where P denotes the set of participants. f_i contains the VFOA target sequence over the course of utterance u_i for each participant. Each utterance u_i is a sequence of words $u_i = \{w_0^i, w_1^i, \dots\}$. The output of our model is a summary Y and the segment ending boundaries S. The training instances for the generator are provided in the form of $T_{train} = \{(X, Y, S)\}$, and the testing instances only contain the transcripts $T_{test} = \{X\}$.

2.1 Visual Focus of Attention Estimation

Given the recording video of each individual, we estimate VFOA based on each participant's head orientation and eye gaze for every frame. The VFOA targets include $F = \{p_0, \dots, p_{|P|}$, *table*, *whiteboard*, *projection_screen* and *unknown*$\}$. As illustrated in Figure 2, we feed each input color image into the OpenFace tool (Baltrusaitis et al., 2018) to estimate the head pose angle (*roll*, *pitch* and *yaw*) and the eye gaze direction vector (*az-*

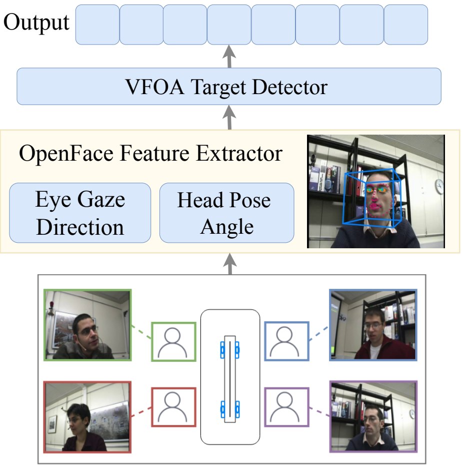


Figure 2: VFOA Detector Framework

[1] For example, IBM Watson's Speech to Text System (https://www.ibm.com/watson/services/ speech-to-text/)

imuth and *elevation*), and concatenate them into a 5-dimensional feature vector. To obtain the actual visual targets from the head pose and eye gaze estimation, we build a seven-layer network to output a one-hot vector, which indicates the most possible visual target at the current frame, and each dimension stands for a VFOA target. The network is trained on the VFOA annotation, including the VFOA target for each frame of each participant.

Then the output of all participants are concatenated. For utterance u_i, the VFOA vector $\boldsymbol{f}_i \in \mathbb{R}^{|P|*|F|}$ is the sum of each frame's VFOA outputs over the course of u_i, where each dimension stands for the total duration of the attention paid to the corresponding VFOA target.

2.2 Meeting Transcript Encoder

For an utterance $u_i = \{w_0^i, w_1^i, \dots\}$, we embed each word $\boldsymbol{w}_j^i$ using the pretrained GloVe (Pennington et al., 2014), and apply a bidirectional gated recurrent unit (GRU) (Cho et al., 2014) to obtain the encoded word representation $\boldsymbol{h}_j^i$. The utterance representations are the average of words. Additionally, the speaker p_i is encoded into a one-hot vector $\boldsymbol{p}_i \in \mathbb{R}^{|\mathcal{P}|}$.

2.3 Topic Segmentation Decoder

We divide the input sequence into contiguous segments based on SegBot (Li et al., 2018). Its decoder takes a starting utterance of a segment as input at each decoding step, and outputs the ending utterance of the segment. Taking Figure 3 as an example, there are 5 utterances in the transcript. The initial starting utterance is u_0 with the possible positions from u_0 to u_4; if u_2 is detected as the ending utterance, then u_3 is the next starting utterance and is input to the decoder, with possible positions from u_3 to u_4.

We extend SegBot to obtain the distribution over possible positions $j \in \{i, i+1, \dots\}$ by using a multi-modal segmentation attention:

$$\alpha_{ij}^{seg} = \boldsymbol{v}_s^\top \tanh(\boldsymbol{W}_u \boldsymbol{d}^i + \boldsymbol{W}_h \boldsymbol{h}^j + \boldsymbol{W}_p \boldsymbol{p}_j + \boldsymbol{W}_f \boldsymbol{f}_j)$$

where $\boldsymbol{d}^i$ is the decoded utterance of starting utterance $\boldsymbol{u}_i$. Let s_i denote the ending utterance of the segment that starts with the utterance $\boldsymbol{u}_i$, the probability for u_j to be the ending utterance s_i is:

$$P(s_i = u_j | (\boldsymbol{p}_i, \boldsymbol{f}_i, \boldsymbol{u}_i)) = \frac{\exp \alpha_{ij}^{seg}}{\sum_{k \in \{i, i+1, \dots\}} \exp \alpha_{ik}^{seg}},$$

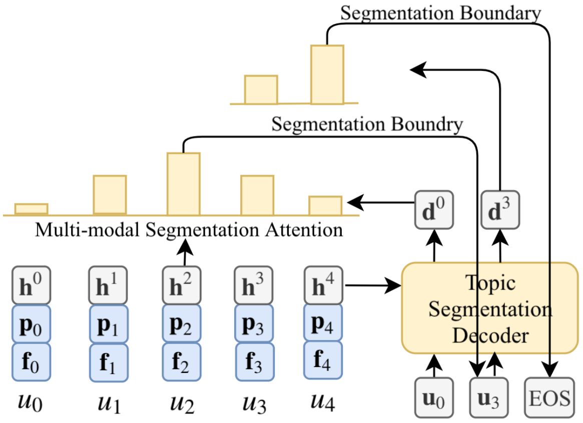
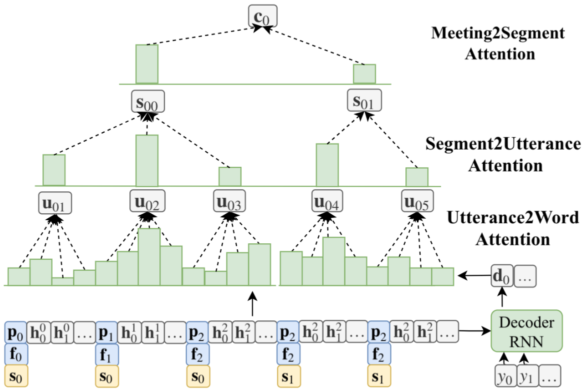


Figure 3: Topic Segmentation Decoder

2.4 Meeting Summarization Decoder

We build our decoder based on Pointer-Generator Network (PGN) (See et al., 2017) to copy words from the input transcript in terms of attention distribution. Different from PGN, we introduce a hierarchical attention mechanism based on the topic segmentation results, as shown in Figure 4.



$\mathbf{c}_0$
Meeting2Segment Attention
$\mathbf{s}_{00}$
$\mathbf{s}_{01}$
Segment2Utterance Attention
$\mathbf{u}_{01}$
$\mathbf{u}_{02}$
$\mathbf{u}_{03}$
$\mathbf{u}_{04}$
$\mathbf{u}_{05}$
Utterance2Word Attention
$\mathbf{d}_0$
...
$\mathbf{p}_0$ $\mathbf{h}_0^0$ $\mathbf{h}_1^0$ ... $\mathbf{p}_1$ $\mathbf{h}_0^1$ $\mathbf{h}_1^1$ ... $\mathbf{p}_2$ $\mathbf{h}_0^2$ $\mathbf{h}_1^2$ ... $\mathbf{p}_2$ $\mathbf{h}_0^2$ $\mathbf{h}_1^2$ ... $\mathbf{p}_2$ $\mathbf{h}_0^2$ $\mathbf{h}_1^2$ ...
$\mathbf{f}_0$ $\mathbf{f}_1$ $\mathbf{f}_2$ $\mathbf{f}_2$ $\mathbf{f}_2$
$\mathbf{s}_0$ $\mathbf{s}_0$ $\mathbf{s}_0$ $\mathbf{s}_1$ $\mathbf{s}_1$
Decoder RNN
y_0 y_1 ...

Figure 4: Hierarchical Attention in Summary Decoder

As VFOA has close ties to salient utterances, we use the VFOA received by speaker $\boldsymbol{f}_k^\top \boldsymbol{p}_k^{'}$ to capture the importance of utterance u_k, where $\boldsymbol{p}_k^{'}$ is the a vector indicating which dimension's VFOA target is the speaker p_k. Formally, we use a GRU to obtain the decoded hidden states $\boldsymbol{d}_i$ for the i^{th} input word. The Utterance2Word attention on the word w_j of the utterance u_k is:

$$e_{ij} = \boldsymbol{v}_1^\top \tanh(\boldsymbol{W}_{d1} \boldsymbol{d}_i + \boldsymbol{W}_w \boldsymbol{w}_j + \boldsymbol{W}_p \boldsymbol{p}_j + \boldsymbol{W}_f \boldsymbol{f}_j)$$

The context representation for the utterance u_k is $\boldsymbol{u}_{ik} = \text{Softmax}(\boldsymbol{e}_{ij}) \boldsymbol{w}_j, w_j \in u_k$. The Segment2Utterance attention on the utterance u_k in the input transcript is:

$$e_{ik}^{'} = \boldsymbol{f}_k^\top \boldsymbol{p}_k^{'} \left(\boldsymbol{v}_2^\top \tanh\left(\boldsymbol{W}_{d2} \boldsymbol{d}_i + \boldsymbol{W}_u \boldsymbol{u}_{ik} \right) \right).$$

Model	ROUGE			BLEU			
	ROUGE_1	ROUGE_2	ROUGE_L	BLEU_1	BLEU_2	BLEU_3	BLEU_4
CoreRank (Shang et al., 2018)	37.86	7.84	13.72	17.17	6.78	1.77	0.00
PGN (See et al., 2017)	36.75	10.48	23.81	37.89	23.41	12.84	6.92
Our Approach (TopicSeg+VFOA)	**53.29**	**13.51**	**26.90**	**40.98**	**26.19**	**13.76**	**8.03**
Our Approach (TopicSeg)	51.53	12.23	25.47	39.67	24.91	12.37	7.86

Table 2: Comparison on AMI datasets

The context representation for segment s_q is $\boldsymbol{c}_{iq} = \text{Softmax}(\boldsymbol{e}^{'}_{ik})\boldsymbol{u}_k, u_k \in s_q$. The Meeting2Segment attention is:

$$e^{''}_{iq} = \boldsymbol{v}_3^\top \tanh(\boldsymbol{W}_{d3}\boldsymbol{d}_i + \boldsymbol{W}_s\boldsymbol{c}_{iq}).$$

The hierarchical attention of w_j is calculated within the utterance u_k and then segment s_q:

$$\boldsymbol{\alpha}^{sum}_{ij} = \frac{\exp\left(e_{ij}e^{'}_{ik}e^{''}_{iq}\right)}{\sum_{j\in s_q}\exp\left(e_{ij}e^{'}_{ik}e^{''}_{iq}\right)},$$

The probability of generating y_i follows the decoder in PGN (See et al., 2017), and $\boldsymbol{\alpha}^{sum}_{ij}$ is the attention in the decoder for copying words from the input sequence.

2.5 Joint End-to-End Training

The summarization task and the topic segmentation task are trained jointly with the loss function:

$$\mathcal{L} = -\log P(Y, S|X)$$
$$= \sum_{y_i \in Y} -\log P(y_i|X) + \sum_{s_j \in S} -\log P(s_j|(\boldsymbol{p}_j, \boldsymbol{f}_j, \boldsymbol{u}_j))$$

where $P(Y, S|X)$ is the conditional probability of the summary Y and the segments S given the input meeting transcript $X = \{(p_i, f_i, u_i)\}$. Here, y_i is one token in the ground truth summary, and s_j denotes the ending boundary of the segment that starts with u_j.

3 Experiments

Our experiments are conducted on the widely used AMI Meeting Corpus (Carletta et al., 2005). This corpus is about a remote control design project from kick-off to completion. Each meeting lasts 30 minutes and contains four participants: a project manager, a marketing expert, an industrial designer, and a user interface designer. We follow the conventional approach (Shang et al., 2018) in the meeting analysis literature to preprocess and divide the dataset into training (97 meetings), development (20 meetings) and test sets (20 meetings). One meeting in the test set does not provide videos and thus it is ignored. The ASR transcripts are provided in the dataset (Garner et al., 2009), which are manually revised based on the automatically generated ASR output. Each meeting has a summary containing about 300 words and 10 sentences. Each meeting is also divided into multiple segments focusing on various topics. The ASR transcripts and the videos recorded for all participants are the input of the model. We use manual annotation of summaries and topic segments for training, while they are generated automatically during testing. The VFOA estimation model is trained separately on the VFOA annotation of 14 meetings in the dataset, and achieve 64.5% prediction accuracy.

The baselines include: (1) state-of-the-art extractive summarization method CoreRank (Shang et al., 2018), and (2) neural network based generation model PGN (See et al., 2017). We adopt two standard metrics ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002) for evaluation. Additionally, to show the impact of VFOA, we remove the VFOA features as an additional baseline, and conduct significance testing. By T-test, the differences on ROUGE and BLEU are considered to be statistically significant (P value $\leq$ 0.09), except BLEU_4 (P value = 0.27).

Compared to the abstractive method PGN in Table 2, the multimodal summarizer achieves larger improvement on ROUGE than BLEU. It demonstrates our approach's ability to focus on topically related words. For example, '*The marketing expert discussed his findings from **trend watching** reports, stressing the need for a product that has a **fancy look and feel**, is **technologically innovative**...*' is generated by our model, while the PGN generates '*the marketing expert discussed his findings from **trend watching** reports*'. The speaker receives higher VFOA from participants while mentioning the utterances containing these keywords. To demonstrate the effectiveness of VFOA attention, we rank the utterances in terms of VFOA, and achieve 45.8% accuracy of selecting salient utterances based on the annotation of

(Shang et al., 2018)[2]. Therefore, the model learns that when the speaker receives higher VFOA, the utterances of that speaker is more important.

Moreover, topic segmentation also contributes to the better coverage of salient words, which is demonstrated by the improvement on ROUGE metrics of the model without VFOA features. Each meeting is divided to six to ten segments, with special focuses on topics such as '*openings*', '*trend watching*', '*project budget*' and '*user target group*'. With the topic segmentation results, the utterances within the same segment are more correlated, and topically related words tend to be frequently mentioned. For example, '*fancy look*' is more important within the '*trend watching*' segment than the whole transcript.

The VFOA distribution is highly correlated to topic segmentation. For example, the project manager pays more attention to the user interface designer in '*trend watching*' segment, while focuses more on the marketing expert in another segment about '*project budget*'. Therefore, the VFOA feature not only benefits the summarization decoder, but also improves the performance of topic segmentation. The topic segmentation accuracy is 57.74% without VFOA feature, and 60.11% with VFOA feature in segmentation attention.

Compared to the extractive method CoreRank in Table 2, our BLEU scores are doubled, which demonstrate that the abstractive summaries are more coherent and natural. For example, the extractive summaries are often incomplete sentences, such as '*prefer a design where the remote control and the docking station*'. But the abstractive summaries are well-organized sentences, such as '*The remote will use a conventional battery and a docking station which recharges the battery*'. Also, the improvement on ROUGE_2 and ROUGE_L is larger than ROUGE_1, which shows the superiority of abstractive methods to maintain longer terms, such as *corporate website*, etc.

4 Related Work

Extractive summarization methods rank and select words by constructing word co-occurrence graphs (Mihalcea and Tarau, 2004; Erkan and Radev, 2004; Lin and Bilmes, 2010; Tixier et al., 2016b), and they are applied to meeting summarization (Liu et al., 2009, 2011; Tixier et al., 2016a; Shang et al., 2018). However, extractive summaries are often not natural and coherent with limited content coverage. Recently the neural natural language generation models boost the performance of abstractive summarization (Luong et al., 2015; Rush et al., 2015; See et al., 2017), but they are often unable to focus on topic words. Inspired by utterance clustering in extractive methods (Shang et al., 2018), we propose a hierarchical attention based on topic segmentation (Li et al., 2018). Moreover, our hierarchical attention is multi-modal to narrow down the focus by capturing participant interactions. Multi-modal features from human annotations have been proven effective at improving summarization, such as dialogue act (Goo and Chen, 2018). Instead of using human annotations, our approach utilizes a simply detectable multi-modal feature VFOA.

5 Conclusions and Future Work

We develop a multi-modal summarizer to generate natural language summaries for multi-person meetings. We present a multi-modal hierarchical attention mechanism based on VFOA estimation and topic segmentation, and the experiments demonstrate its effectiveness. In the future, we plan to further integrate higher level participant interactions, such as gestures, face expressions, etc. We also plan to construct a larger multimedia meeting summarization corpus to cover more diverse scenarios, building on our previous work (Bhattacharya et al., 2019).

Acknowledgments

This material is based upon work supported by the U.S. National Science Foundation under Grant No. IIP-1631674, DARPA AIDA Program No. FA8750-18-2-0014, ARL NS-CTA No. W911NF-09-2-0053, and Tencent AI Lab Rhino-Bird Gift Fund. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

[2] https://bitbucket.org/dascim/acl2018_abssumm/src/master/data/meeting/ami

References

Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. *2018 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2018)*, pages 59–66.

Indrani Bhattacharya, Michael Foley, Christine Ku, Ni Zhang, Tongtao Zhang, Cameron Mine, Manling Li, Heng Ji, Christoph Riedl, Brooke Foucault Welles, and Richard J. Radke. 2019. The unobtrusive group interaction (UGI) corpus. In *10th ACM Multimedia Systems Conference (MMSys 2019)*.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al. 2005. The AMI meeting corpus: A pre-announcement. In *International Workshop on Machine Learning for Multimodal Interaction*, pages 28–39. Springer.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

Philip N Garner, John Dines, Thomas Hain, Asmaa El Hannani, Martin Karafiat, Danil Korchagin, Mike Lincoln, Vincent Wan, and Le Zhang. 2009. Real-time ASR from meetings. In *Tenth Annual Conference of the International Speech Communication Association*.

Chih-Wen Goo and Yun-Nung Chen. 2018. Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts. *arXiv preprint arXiv:1809.05715*.

Jing Li, Aixin Sun, and Shafiq Joty. 2018. Segbot: A generic neural text segmentation model with pointer network. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Hui Lin and Jeff Bilmes. 2010. Multi-document summarization via budgeted maximization of submodular functions. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 912–920.

Fei Liu, Feifan Liu, and Yang Liu. 2011. A supervised framework for keyword extraction from meeting transcripts. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(3):538–548.

Feifan Liu, Deana Pennell, Fei Liu, and Yang Liu. 2009. Unsupervised approaches for automatic keyword extraction using meeting transcripts. In *Proceedings of human language technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 620–628. Association for Computational Linguistics.

Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2015. Multi-task sequence to sequence learning. *arXiv preprint arXiv:1511.06114*.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*.

George A Miller. 1995. Wordnet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

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 on Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1073–1083.

Guokan Shang, Wensi Ding, Zekun Zhang, Antoine Jean-Pierre 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)*, volume 1.

Antoine Tixier, Fragkiskos Malliaros, and Michalis Vazirgiannis. 2016a. A graph degeneracy-based approach to keyword extraction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1860–1870.

Antoine Tixier, Konstantinos Skianis, and Michalis Vazirgiannis. 2016b. Gowvis: a web application for graph-of-words-based text visualization and summarization. *Proceedings of ACL-2016 System Demonstrations*, pages 151–156.