Abstractive Spoken Document Summarization using Hierarchical Model with Multi-stage Attention Diversity Optimization

Potsawee Manakul, Mark J. F. Gales, Linlin Wang

Engineering Department, University of Cambridge, UK

pm574@cam.ac.uk, mjfg@eng.cam.ac.uk, lw519@cam.ac.uk

Abstract

Abstractive summarization is a standard task for written documents, such as news articles. Applying summarization schemes to spoken documents is more challenging, especially in situations involving human interactions, such as meetings. Here, utterances tend not to form complete sentences and sometimes contain little information. Moreover, speech disfluencies will be present as well as recognition errors for automated systems. For current attention-based sequence-to-sequence summarization systems, these additional challenges can yield a poor attention distribution over the spoken document words and utterances, impacting performance. In this work, we propose a multi-stage method based on a hierarchical encoder-decoder model to explicitly model utterance-level attention distribution at training time; and enforce diversity at inference time using a unigram diversity term. Furthermore, multitask learning tasks including dialogue act classification and extractive summarization are incorporated. The performance of the system is evaluated on the AMI meeting corpus. The inclusion of both training and inference diversity terms improves performance, outperforming current state-of-the-art systems in terms of ROUGE scores. Additionally, the impact of ASR errors, as well as performance on the multitask learning tasks, is evaluated.

Index Terms: abstractive spoken document summarization, hierarchical model, attention diversity, multitask learning.

1. Introduction

The quantity of available spoken documents such as meeting recordings, broadcast news, and lectures is rapidly increasing. Spoken document summarization systems improve data access as they yield concise information about content [1]. The two main types of summarization are: extractive methods which select and reorder words or sentences in the original source; and abstractive methods that can generate words and phrases that do not appear in the source. Past work on spoken documents normally applied traditional machine learning or extractive methods [2, 3, 4, 5, 6]. Many recent approaches based on sequence-to-sequence neural networks [7, 8, 9, 10] have shown promising results on the abstractive text summarization task such as news articles. However, when it comes to spoken documents such as meeting transcripts, the input source is typically longer, less grammatical, and contains less-structured utterances rather than well-constructed sentences. Consequently, standard sequence-to-sequence neural models with attention mechanism [11] that take a long sequence of tokens have been shown to be less effective than hierarchical models for this task [12, 13].

The number of spoken document datasets that have been annotated for the summarization task is limited. In this work, we focus on meeting summarization based on the AMI corpus [14]. The AMI corpus contains far fewer training examples compared to widely used text summarization datasets such as CNN/DailyMail [7]. The summaries of these meetings are also similar to each other as they are mainly about a group of people discussing the design of a product. As a result, trained systems may produce commonly used words and repeated sentences, i.e. they suffer the *diversity* problem.

Although hierarchical models [15] do not outperform non-hierarchical models [8, 16] on CNN/DailyMail, the hierarchical models are more suitable for meeting summarization [12, 13]. Thus, in this work, we focus on hierarchical models. Nevertheless, the issue of diversity remains. When manual summaries are generated, information from different input utterances is typically used to generate each of the summary output sentences. This motivates us to model the diversity at the utterance-level within the sequence-to-sequence attention mechanism, allowing this diversity to be explicitly optimized during training. This diversity modeling approach is different to the coverage mechanism [8, 17] and global variance loss [18] approaches, which are designed to mitigate repetitions. In contrast, the proposed approach operates at the utterance level instead of word level; and encourages the utterance-level attention to be similar when generating the same output sentences but varied when generating different output sentences. In addition, to encourage diversity at test time, we use a modified decoding approach that uses a unigram bias term to improve output sentence diversity.

Furthermore, in spoken language, information is conveyed not only by the words uttered. For instance, [19] demonstrates that signals such as semantic slot and dialogue domain can improve dialogue summarization. Utterances may also serve different functions within a conversation. Thus interactive signals such as dialogue acts have been shown to be useful in predicting topic description [20]. In the AMI corpus, dialogue acts and salient utterance information are available, therefore, we conduct an experiment on the impact of additional information on summarization performance.

In practical scenarios, an automatic speech recognition (ASR) system is required to obtain the word transcript prior to the summarization step. Since these systems are not perfect, it is crucial to understand the impact of ASR errors on summarization performance. Therefore, we conduct our experiments on the AMI dataset using both manually derived and automatically derived transcripts.

2. Summarization Approach

2.1. Hierarchical Encoder-Decoder Architecture

Hierarchical summarization models use an encoder-decoder architecture where the encoder consists of word-level and utterance-level gated recurrent units (GRUs) [21], and the decoder consists of a GRU attending to the encoder states. For

This paper reports on research supported by ALTA institute, Cambridge Assessment English, University of Cambridge.

utterance i containing words $\boldsymbol{u}_i = \{w_{i,1}, w_{i,2}, ..., w_{i,N^i}\}$, we embed word $w_{i,j}$ into vector $\mathbf{x}_{i,j}$, and apply a bidirectional word-level GRU resulting in $\{\overrightarrow{\mathbf{h}}^{\mathrm{w}}_{i,1}, ..., \overrightarrow{\mathbf{h}}^{\mathrm{w}}_{i,N^i}\}$ and $\{\overleftarrow{\mathbf{h}}^{\mathrm{w}}_{i,1}, ..., \overleftarrow{\mathbf{h}}^{\mathrm{w}}_{i,N^i}\}$. We concatenate the forward and backward states to obtain word representations $\mathbf{h}^{\mathrm{w}}_{i,j} = [\overrightarrow{\mathbf{h}}^{\mathrm{w}}_{i,j}; \overleftarrow{\mathbf{h}}^{\mathrm{w}}_{i,j}]$. The hidden vector of the final word $\mathbf{h}^{\mathrm{w}}_{i,N^i}$ is selected to represent utterance i, and we apply the bidirectional utterance-level GRU on the output of the word-level GRU resulting in $\{\mathbf{h}^{\mathrm{u}}_1, \mathbf{h}^{\mathrm{u}}_2, ..., \mathbf{h}^{\mathrm{u}}_N\}$ where N is the number of utterances. The decoder embeds each word at time step t and passes it to a GRU giving the decoder hidden state $\mathbf{d}_t$. Next, the decoder attends to the hidden states of the encoder hierarchically:

- Utterance-level attention:

$$\alpha^{\mathrm{u}}_{t,i} = \frac{\exp(\mathbf{d}^{\mathrm{T}}_t \mathbf{W}^{\mathrm{u}} \mathbf{h}^{\mathrm{u}}_i)}{\sum_{i'} \exp(\mathbf{d}^{\mathrm{T}}_t \mathbf{W}^{\mathrm{u}} \mathbf{h}^{\mathrm{u}}_{i'})} \tag{1}$$

- Word-level attention:

$$\alpha^{\mathrm{w}}_{t,i,j} = \alpha^{\mathrm{u}}_{t,i} \left(\frac{\exp(\mathbf{d}^{\mathrm{T}}_t \mathbf{W}^{\mathrm{w}} \mathbf{h}^{\mathrm{w}}_{i,j})}{\sum_{j'} \exp(\mathbf{d}^{\mathrm{T}}_t \mathbf{W}^{\mathrm{w}} \mathbf{h}^{\mathrm{w}}_{i,j'})} \right) \tag{2}$$

The word-level attention distribution is used to produce a context vector $\mathbf{h}^*_t = \sum_i \sum_j \alpha^{\mathrm{w}}_{t,i,j} \mathbf{h}^{\mathrm{w}}_{i,j}$. This context vector is concatenated with the decoder state and fed through a linear layer, which has output units equal to the vocabulary size, to produce the output word distribution:

$$P(y|\boldsymbol{\alpha}^{\mathrm{w}}_t, \mathbf{H}^{\mathrm{w}}, \mathbf{d}_t) = \mathrm{softmax}(\mathbf{W}^{\circ}[\mathbf{h}^*_t; \mathbf{d}_t] + \mathbf{b}^{\circ}) \tag{3}$$

where $\boldsymbol{\alpha}^{\mathrm{w}}_t$ is the vector of all word level attentions at time instance t and $\mathbf{H}^{\mathrm{w}}$ is the input document word embeddings.

2.2. Diversity Scores

Given the hierarchical attention mechanism, it is possible to define the diversity of input document attention both within an output sentence and between output sentences. Let $\boldsymbol{\alpha}^{\mathrm{u}}_t$ be the vector of all utterance-level attentions at time instance t. Inter and intra utterance diversity scores can then be expressed as:

- **intra-sentence**:

$$D_{\text{intra},k} = \frac{1}{\binom{T_k}{2}} \sum_{t_1=1}^{T_k-1} \sum_{t_2=t_1+1}^{T_k} ||\boldsymbol{\alpha}^{\mathrm{u}}_{t_1} - \boldsymbol{\alpha}^{\mathrm{u}}_{t_2}||_2 \tag{4}$$

$$D_{\text{intra}} = \frac{1}{K} \sum_{k=1}^{K} D_{\text{intra},k} \tag{5}$$

- **inter-sentence**:

$$\bar{\boldsymbol{\alpha}}^{\mathrm{u}}_k = \frac{1}{T_k} \sum_{t=1}^{T_k} \boldsymbol{\alpha}^{\mathrm{u}}_t \tag{6}$$

$$D_{\text{inter}} = \frac{1}{\binom{K}{2}} \sum_{k_1=1}^{K-1} \sum_{k_2=k_1+1}^{K} ||\bar{\boldsymbol{\alpha}}^{\mathrm{u}}_{k_1} - \bar{\boldsymbol{\alpha}}^{\mathrm{u}}_{k_2}||_2 \tag{7}$$

where k denotes output sentence k, T_k is the number of tokens in sentence k, and K is the total number of output sentences.

2.3. Model Training

For this work supervised learning is used. The training data comprises pairs of input document utterances, $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N\}$, and associated summary sentences $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_M\}$. The parameters are optimized to maximize the likelihood of the set of J summary document pairs:

$$\mathcal{L}_{\text{ml}} = \frac{1}{J} \sum_{j=1}^{J} \log P\left(\boldsymbol{y}^{(j)}_1, \ldots, \boldsymbol{y}^{(j)}_{M_j} | \boldsymbol{u}^{(j)}_1, \ldots, \boldsymbol{u}^{(j)}_{N_j}; \boldsymbol{\theta}\right) \tag{8}$$

For simplicity of notation, this will be written in terms of the predictions of each of the words in the summary. For the word level prediction for a particular training example pair this can then be written as:

$$\mathcal{L}_{\text{ml}} = \sum_{i,t} \log P(y_{i,t}|\boldsymbol{\alpha}^{\mathrm{w}}_t, \mathbf{H}^{\mathrm{w}}, \mathbf{d}_t) = \sum_{i,t} \log P_t(y_{i,t}) \tag{9}$$

To improve the generalization, auxiliary tasks can be used for multitask learning. Here two tasks are used: dialogue act classification; and extractive summarization. Given the output of the utterance-level GRU for utterance i, $\mathbf{h}^{\mathrm{u}}_i$, we pass the vector to two distinct linear layers:

$$P_{\text{da}}(q|\mathbf{h}^{\mathrm{u}}_i) = \mathrm{softmax}(\mathbf{W}^{\text{da}} \mathbf{h}^{\mathrm{u}}_i + \mathbf{b}^{\text{da}}) \tag{10}$$

$$P_{\text{ex}}(r|\mathbf{h}^{\mathrm{u}}_i) = \mathrm{sigmoid}(\mathbf{W}^{\text{ex}} \mathbf{h}^{\mathrm{u}}_i + \mathbf{b}^{\text{ex}}) \tag{11}$$

where q is the dialogue act, one of 15 possible classes for this data, and r the binary extractive summarization label. The extractive summarization, $\mathcal{L}_{\text{ex}}$, and dialogue act, $\mathcal{L}_{\text{da}}$, loss functions are also maximum likelihood based.

The auxiliary tasks can also be combined with a diversity loss to try to enforce diversity on the final generated sequence. Using the diversity scores defined in section 2.2, an appropriate **diversity loss**, $\mathcal{L}_{\text{dv}}$ is the ratio of the intra to inter diversity scores $D_{\text{intra}}/D_{\text{inter}}$. The complete loss function to *minimize* is a linear combination of all the losses:

$$\mathcal{L} = -\mathcal{L}_{\text{ml}} - \lambda_1 \mathcal{L}_{\text{da}} - \lambda_2 \mathcal{L}_{\text{ex}} + \lambda_3 \mathcal{L}_{\text{dv}} \tag{12}$$

2.4. Decoding

During training, the decoder uses ground truth tokens as the input history (*Teacher Forcing* mode). At test time, the decoder has to use its own prediction from the previous time step as the input (*Free Running* mode). However, the system has not been trained to correct for errors in the history. There exist methods that aim to mitigate the discrepancy between teacher forcing and free running such as scheduled sampling [22], professor forcing [23], attention forcing [24], but they require modification in training. Also, recent work [25] shows that the standard maximization-based decoding method leads to output text that can get stuck in repetitive loops. On our task, we will show that in free running mode, inter-sentence diversity score (described in section 2.2) drops, yielding less diverse output. Thus, we propose a simple method that works with any model trained with teacher forcing without modification by penalizing repeated unigrams, *Unigram Bias* decoding:

$$\hat{y}_t = \underset{y \in \mathcal{V}}{\arg\max} \left\{ \log P_t(y) - \beta \left(\frac{\sum_{\tau=1}^{t} \mathbb{1}_y(\hat{y}_\tau)}{t} \right) \right\} \tag{13}$$

where β is the unigram bias constant, $\mathcal{V}$ is the decoding vocabulary, and $\mathbb{1}_y(\hat{y}_\tau)$ is 1 when $y = \hat{y}_\tau$.

3. Experimental Setup

3.1. Datasets and Evaluation Metrics

AMI Meeting Corpus [14] contains meeting recordings of four people discussing a remote control designing project. Each meeting is about 30 minutes, and there are 137 meetings (excluding those without the annotation required for multitask learning). This work makes use of the dialogue acts, and extractive and abstractive summaries annotation in addition to the manual transcripts. The default data split is 97 training, 20 validation, and 20 test meetings from the guideline on the AMI website. The manually derived transcripts have 784 utterances,

6,200 words, on average per meeting, and each summary contains 10 sentences of up to 300 words. Two sets of automatically derived transcripts (test set), from ASR systems using the manual segmentation, are used for testing our final model:

- ASR1 - AMI release v1.5 which is publicly available and also used in [6, 12]. The word error rate (WER) is 36%.
- ASR2 - a TDNN-F acoustic model [26] trained on AMI IHM training set using the Kaldi toolkit [27], with a 4-gram language model trained on the same set and the Fisher Corpus (LDC2004T19) [28]. 40-dim Mel-scaled filterbank features and 15 TDNN-F layers were used. The word error rate (WER) is 20%.

CNN/DailyMail, processed as in [7], contains news articles (781 words on average) and summaries (average 3.75 sentences, 56 words). The non-anonymized version containing 287,226 training, 13,368 validation, and 11,490 test pairs, was used.

The ROUGE-N F_1 scores [29] is used as the evaluation metric for summarization performance. ROUGE-N measures the overlap of n-grams between the system and reference summaries, e.g. ROUGE-1 refers to the overlap of unigrams. ROUGE-L measures the longest common subsequence.

3.2. Implementation details

Two baseline systems are used for initial contrasts with the hierarchical systems: (1) DecoderLM which is a decoder-only model with the same architecture as the decoder of the hierarchical model, but with no conditioning on the input document. This can be viewed as a baseline to assess the complexity and diversity of the summaries, (2) A publicly available pointer-generator network (PGN) implementation[1] with the hyperparameters set as in [8]. For the hierarchical model, a PyTorch implementation was generated[2]. The following configurations were used: word embedding 256-dimensional; the hidden states of word-level and utterance-level encoder GRUs 256-dimensional; and the hidden states of decoder GRU 512-dimensional. The vocabulary size was 30k. This model was optimized using Adam [30] with $\alpha = 0.01 \times \text{step}^{-0.5}$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. The batch size was set to 2, the maximum number of words in an utterance to 64, and the maximum number of utterances to 1,500. Training was stopped when performance on the validation set did not improve over three epochs. When multitask or diversity objectives were used, $\lambda_1 = \lambda_2 = 0.2$ and $\lambda_3 = 1.0$ in Equation (12), manually tuned on the validation data. At test time, beam search of width 10 was used, and output sentences are rejected if there was a 4-gram overlapping with any previous sentences in the summary to avoid redundancy [9, 31]. We trained each model three times. In each time, we used a different seed value for initialization and data shuffling, and we made the training set contain 100 meetings by randomly selecting 3 meetings from the validation set.

	AMI		CNN/DailyMail	
Decoding	**intra**	**inter**	**intra**	**inter**
TF	0.01309	0.00663	0.07437	0.08365
FR	0.01327	0.00540	0.07548	0.07890
UB	0.01328	0.00588	0.07735	0.10597

Table 1: *TF, FR, UB denote teacher forcing, free running, and unigram bias ($\beta = 20.0$) decoding methods respectively.*

[1] https://github.com/atulkum/pointer_summarizer
[2] https://github.com/potsawee/spoken_summ_div

4. Results

Initially, the diversity of the test set summaries was evaluated for three generation modes with the hierarchical model: teacher forcing (TF); free-running (FR); and the unigram bias decoding (UB). As shown in Table 1, the FR mode has the lowest inter-sentence diversity, more repetitions, as the diversity from the reference in TF has been lost. Using UB decoding increases the inter-sentence diversity. Figure 1 shows a detailed analysis of UB decoding with the value of β on the AMI data. There is little variation in the intra-sentence variability, but clear gains in the inter-sentence diversity.

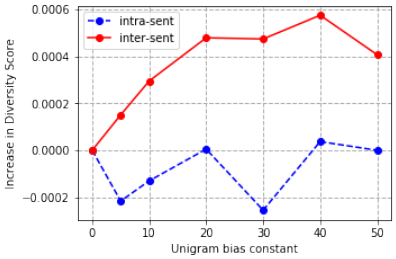

Figure 1: *The increase in the diversity scores evaluated on AMI.*

The performance of the system was first evaluated using the manual AMI transcriptions, and the results are shown in Table 2. For the news article summarization task, the first few sentences are typically used as a baseline summary [32], but in this task, since information is more evenly spread in the source, the DecoderLM is quoted as the baseline system. The PGN with coverage, PGN+Cov, [8] outperforms this baseline. Additional gains can be obtained with transfer learning (TL) where the model was initialized using the CNN/DailyMail data. The hierarchical models achieve higher ROUGE scores, consistent with [12].

Model	**TL**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
DecoderLM	✗	26.00±3.55	8.26±1.62	24.78±3.15
PGN+Cov	✗	29.90±0.88	10.41±0.99	28.16±1.12
PGN+Cov	✓	33.43±1.70	11.29±1.55	31.42±1.64
Hierarchical	✗	33.07±0.66	10.87±1.18	31.77±0.86
Hierarchical	✓	**39.12±2.45**	**13.03±1.58**	**36.77±2.28**

Table 2: *ROUGE F_1 on the AMI test set - Baseline Models. TL denotes model being pre-trained on CNN/DailyMail.*

In Figure 2, the impact of unigram bias on the ROUGE-1, ROUGE-2, ROUGE-L, and the average of the three scores for $\beta \in [0.0, 40.0]$ is shown for the baseline hierarchical models (HIER), the inclusion of multitask training (HIER+MT), the diversity training loss (HIER+DIV), and both (HIER+MT+DIV). Since unigram bias decoding penalizes repeated unigrams, ROUGE-1 improves for all models as expected. Figure 1 shows that the inter-sentence diversity score flattens at $\beta = 20.0$, and it can be seen in Figure 2 that ROUGE-1 of the hierarchical setting also stops improving at $\beta = 20.0$. This suggests a positive correlation between models being *diverse* and *higher* ROUGE-1. ROUGE-L, which measures the longest common sequence, also follows the same trend as ROUGE-1. However, the increase in ROUGE-2 is less than the increase in ROUGE-1, and in one setting there is no gain from unigram bias.

The optimal β and summarization scores, from Figure 2, for each system are shown in Table 3. Additionally training

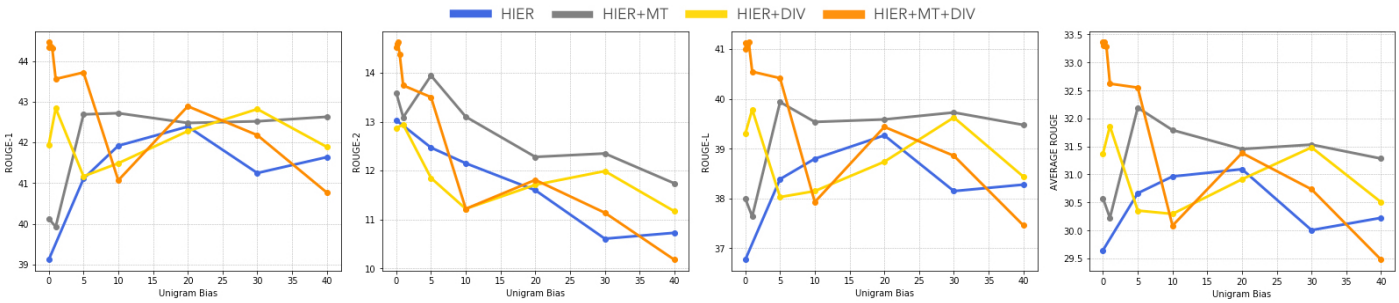

Figure 2: *Variation of ROUGE F_1 scores against unigram bias for the hierarchical model (HIER), with multitask training (HIER+MT), diversity loss (HIER+DIV) and both (HIER+MT+DIV).*

the model with multitask (HIER+MT) or diversity objectives yields performance gains. When explicitly optimizing the diversity objective during training (HIER+DIV), the model performance is similar to the baseline with unigram bias decoding (HIER with $\beta = 20.0$). When the diversity loss is optimized, a lower value of β is required to achieve optimal performance, suggesting that explicit diversity optimization during training and forcing diversity during decoding have a similar effect. The results also show that diversity optimization and unigram bias are complimentary, and the best performance is achieved in the HIER+MT+DIV with unigram bias decoding setting. In addition, on the CNN/DailyMail dataset, the HIER+MT+DIV with unigram bias decoding system achieves higher ROUGE-1, ROUGE-2, ROUGE-L scores than the HIER system by 3.81%, 0.67%, and 3.29%, further confirming the effectiveness of our approach.

Setting	β	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
HIER	✗	39.12±2.45	13.03±1.58	36.77±2.28
	20.0	42.39±1.91	11.60±1.60	39.27±1.96
HIER+MT	✗	40.13±0.93	13.59±0.66	38.00±0.74
	5.0	42.69±1.03	13.94±0.86	39.94±0.94
HIER+DIV	✗	41.94±0.25	12.87±0.28	39.30±0.56
	1.00	42.84±1.10	12.94±0.50	39.79±1.83
HIER+MT+DIV	✗	44.46±0.11	14.51±0.12	41.12±0.13
	0.25	**44.36±0.58**	**14.62±0.21**	**41.10±0.25**

Table 3: *ROUGE F_1 on the AMI test set - Hierarchical Settings.*

Table 4 shows the performance of the best system (HIER+MT+DIV with β=0.25) on ASR transcripts. When using the AMI ASR (A1) transcripts instead of the manual transcripts, the decrease in ROUGE is around 2-3%. This relatively small drop is likely because even at 30% WER, the sentence/utterance embedding similarity between a manual source and an ASR source is about 0.70-0.85% [33, 34]. This system achieves higher all ROUGE measures than extractive method CoreRank [6], and when compared to abstractive method TopicSeg (without visual signals) [12] our system achieves higher ROUGE-2 and ROUGE-L although lower ROUGE-1. Furthermore, when using transcripts with lower WER (A2), ROUGE scores are closer to those obtained from the manual transcripts, and yield higher ROUGE-2 and ROUGE-L scores than the state-of-the-art multi-modal TopicSeg+VFOA. [12].

Finally, for the multitask trained systems, it is possible to evaluate the performance of the system on the other training tasks, Dialogue Act classification (DialogueAct) and Extractive Summarization (ExtractiveSum). Both tasks used the encoder

Model	**Input**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
CoreRank [6]	A1	37.86	7.84	13.72
TopicSeg [12]	A1	51.53	12.23	25.47
TopicSeg+VFOA [12]	A1	53.29	13.51	26.90
HIER+MT+DIV	A1	41.23±0.75	12.74±0.21	38.38±0.63
	A2	43.29±1.81	14.45±0.23	40.55±1.20
	M	44.36±0.58	14.62±0.21	41.10±0.25

Table 4: *A1 = publicly available ASR transcripts (WER=36%), A2 = ASR2 transcripts (WER=20%), M=Manual transcripts.*

of the hierarchical model. For the DialogueAct baseline $\mathcal{L}_{\text{da}}$ was optimized, and for the ExtractiveSum baseline $\mathcal{L}_{\text{ext}}$ was optimized. For the HIER+MT setting, the weighting of the loss terms, λ_1 and λ_2 in Equation (12) were both set to 10.0, and for the HIER+MT+DIV setting, λ_3 was set to 1.0. Table 5 shows that the summarization signal improves both dialogue act classification and extractive summarization labeling tasks. Our best dialogue act accuracy is comparable with [20]. The diversity loss, in contrast, does not benefit these two tasks. This is expected as the diversity criterion aims to improve the diversity of the attention mechanism of the decoder, whereas for these two tasks only the encoder of the summarization system is used.

DialogueAct	**Accuracy**	**ExtractiveSum**	$\mathbf{F}_1$
HIER	63.42±0.82	HIER	54.47±3.75
HIER+MT	64.32±0.39	HIER+MT	56.05±1.61
HIER+MT+DIV	63.11±0.36	HIER+MT+DIV	56.06±2.39

Table 5: *Dialogue Act Classification (1-in-15 classification) and Extractive Summarization Labeling at 0.5 threshold.*

5. Conclusions

The hierarchical model was shown to be effective on the spoken document summarization task. The proposed explicit utterance-level attention diversity model and unigram bias decoding were both shown to improve our summarization system. It was demonstrated that a multitask learning method, which incorporates dialogue act and salient utterance information, is useful for summarization. The hierarchical model with multitask and diversity objectives (HIER+MT+DIV) with unigram bias decoding was found to be the best configuration for the meeting dataset. For a real-world spoken document application based on ASR system with WER about 20%, it was illustrated that summarization performance close to that obtained from manual transcription can be achieved.

6. References

[1] G. Tur and R. D. Mori, *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*. John Wiley and Sons, 2011.

[2] S.-R. Shiang, H.-Y. Lee, and L.-S. Lee, "Supervised spoken document summarization based on structured support vector machine with utterance clusters as hidden variables," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 2728–2732, 2013.

[3] S. Xie, D. Hakkani-Tür, B. Favre, and Y. Liu, "Integrating prosodic features in extractive meeting summarization," in *2009 IEEE Workshop on Automatic Speech Recognition Understanding*, 2009, pp. 387–391.

[4] Y. Chen and F. Metze, "Two-layer mutually reinforced random walk for improved multi-party meeting summarization," in *2012 IEEE Spoken Language Technology Workshop (SLT)*, 2012, pp. 461–466.

[5] F. Liu and Y. Liu, "Towards abstractive speech summarization: Exploring unsupervised and supervised approaches for spoken utterance compression," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1469–1480, 2013.

[6] G. Shang, W. Ding, Z. Zhang, A. Tixier, P. Meladianos, M. Vazirgiannis, and J.-P. Lorré, "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*, 2018.

[7] R. Nallapati, B. Zhou, C. dos Santos, Ç. Gulİ‡lçehre, and B. Xiang, "Abstractive text summarization using sequence-to-sequence RNNs and beyond," in *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Aug. 2016.

[8] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Jul. 2017.

[9] Y. Liu and M. Lapata, "Text summarization with pretrained encoders," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

[10] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[11] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[12] M. Li, L. Zhang, H. Ji, and R. J. Radke, "Keep meeting summaries on topic: Abstractive multi-modal meeting summarization," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 2190–2196.

[13] Z. Zhao, H. Pan, C. Fan, Y. Liu, L. Li, M. Yang, and D. Cai, "Abstractive meeting summarization via hierarchical adaptive segmental network learning," in *The World Wide Web Conference*, ser. WWW '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 3455–3461.

[14] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The ami meeting corpus: A pre-announcement," in *Proceedings of the Second International Conference on Machine Learning for Multimodal Interaction*. Springer-Verlag, 2005.

[15] Y. Chen, Y. Ma, X. Mao, and Q. Li, "Multi-task learning for abstractive and extractive summarization," *Data Science and Engineering*, vol. 4, no. 1, pp. 14–23, 2019.

[16] Y.-C. Chen and M. Bansal, "Fast abstractive summarization with reinforce-selected sentence rewriting," in *Proceedings of ACL*, 2018.

[17] Z. Tu, Z. Lu, Y. Liu, X. Liu, and H. Li, "Modeling coverage for neural machine translation," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 2016.

[18] M. Gui, J. Tian, R. Wang, and Z. Yang, "Attention optimization for abstractive document summarization," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

[19] L. Yuan and Z. Yu, "Abstractive dialog summarization with semantic scaffolds," *arXiv preprint arXiv:1910.00825*, 2019.

[20] C.-W. Goo and Y.-N. Chen, "Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 735–742.

[21] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "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)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014.

[22] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer, "Scheduled sampling for sequence prediction with recurrent neural networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 1171–1179.

[23] A. M. Lamb, A. G. A. P. Goyal, Y. Zhang, S. Zhang, A. C. Courville, and Y. Bengio, "Professor forcing: A new algorithm for training recurrent networks," in *Advances In Neural Information Processing Systems*, 2016, pp. 4601–4609.

[24] Q. Dou, Y. Lu, J. Efiong, and M. J. Gales, "Attention forcing for sequence-to-sequence model training," *arXiv preprint arXiv:1909.12289*, 2019.

[25] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, "The curious case of neural text degeneration," in *International Conference on Learning Representations*, 2020.

[26] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks." in *Interspeech*, 2018.

[27] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011.

[28] C. Cieri, D. Miller, and K. Walker, "The fisher corpus: a resource for the next generations of speech-to-text."

[29] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2015.

[31] R. Paulus, C. Xiong, and R. Socher, "A deep reinforced model for abstractive summarization," in *International Conference on Learning Representations*, 2018.

[32] R. Nallapati, F. Zhai, and B. Zhou, "Summarunner: A recurrent neural network based sequence model for extractive summarization of documents," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[33] M. Ákos Tündik, V. Kaszás, and G. Szaszák, "Assessing the Semantic Space Bias Caused by ASR Error Propagation and its Effect on Spoken Document Summarization," in *Proc. Interspeech 2019*, 2019, pp. 1333–1337.

[34] R. Voleti, J. M. Liss, and V. Berisha, "Investigating the effects of word substitution errors on sentence embeddings," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7315–7319.