

# Topic-Oriented Spoken Dialogue Summarization for Customer Service with Saliency-Aware Topic Modeling

Yicheng Zou,<sup>1</sup> Lujun Zhao,<sup>2</sup> Yangyang Kang,<sup>2</sup> Jun Lin,<sup>2</sup> Minlong Peng,<sup>1</sup> Zhuoren Jiang,<sup>3</sup> Changlong Sun,<sup>3,2</sup> Qi Zhang,<sup>1</sup> Xuanjing Huang,<sup>1</sup> Xiaozhong Liu<sup>4</sup>

<sup>1</sup>School of Computer Science, Fudan University, Shanghai, China

<sup>2</sup>Alibaba Group, China

<sup>3</sup>Zhejiang University, Hangzhou, China

<sup>4</sup>Indiana University Bloomington, Bloomington, United States

{yczou18, mlpeng16, qz, xjhuang}@fudan.edu.cn, {linjun.lj, lujun.zlj, yangyang.kangyy}@alibaba-inc.com, jiangzhuoren@zju.edu.cn, changlong.scl@taobao.com, liu237@indiana.edu

## Abstract

In a customer service system, dialogue summarization can boost service efficiency by automatically creating summaries for long spoken dialogues in which *customers* and *agents* try to address issues about specific topics. In this work, we focus on topic-oriented dialogue summarization, which generates highly abstractive summaries that preserve the main ideas from dialogues. In spoken dialogues, abundant dialogue noise and common semantics could obscure the underlying informative content, making the general topic modeling approaches difficult to apply. In addition, for customer service, role-specific information matters and is an indispensable part of a summary. To effectively perform topic modeling on dialogues and capture multi-role information, in this work we propose a novel topic-augmented two-stage dialogue summarizer (TDS) jointly with a saliency-aware neural topic model (SATM) for topic-oriented summarization of customer service dialogues. Comprehensive studies on a real-world Chinese customer service dataset demonstrated the superiority of our method against several strong baselines.

## Introduction

In an active customer service system, massive dialogues conveying important information between *customers* and *agents* are generated in real time. With this background, how to efficiently consume dialogue information becomes a non-trivial issue. Dialogue summarization is a task that aims to condense dialogues while retaining the salient information (Rambow et al. 2004; Pan et al. 2018; Shang et al. 2018; Liu et al. 2019a), which can boost service efficiency by automatically creating concise summaries to avoid time-consuming dialogue reading and comprehension.

Most existing works for dialogue summarization have mainly focused on long and intricate spoken dialogues, like meetings and court debates, which are usually summarized by stringing all dialogue points to maintain an integral conversation flow (Gillick et al. 2009; Shang et al. 2018; Duan et al. 2019b). Nevertheless, in the customer service scenario,

<p>A: Hello, this is xxx hotline. May I help you?  C: I've got an order saying that <b>it has been delivered but I haven't received yet</b>. When I checked it, <b>it shows that the deal is done</b>.  C: But err... <b>I haven't received anything</b>.  A: I got it. Then, could you please provide your username or the binding phone number of the application?  .....  A: Did you place the order today?  C: Err... No, it was yesterday but he told me he would deliver it today. Hum, I checked the message in the morning but I haven't received anything.  A: Humm, ok. I see. <b>I am gonna contact the deliveryman</b>. Is that okay? I will <b>check it for you</b> and call you back later.  C: Ok, ok. That's good.</p>
<p><b>Summary:</b> The user called us because <b>the order shows that it has been delivered but he did not receive it at all</b>. I replied that I would <b>check it by contacting the deliveryman</b>.</p>

Figure 1: A customer service dialogue and its reference summary. C denotes the customer and A denotes the agent. The summary contains the customer's problem and the agent's solution, which are highlighted in red and blue, respectively.

dialogue speakers commonly have strong and clear motivations and aim to address issues about specific topics (Wang et al. 2020). To better understand customers' and agents' intentions, in this work we focus on the topic-oriented dialogue summarization, which aims to extract semantically consistent topics and generate highly abstractive summaries to maintain the main ideas from dialogues.

Recently, dozens of topic-aware models have been introduced to assist with document summarization tasks (Wang et al. 2018; Narayan et al. 2018; Fu et al. 2020). However, rather than well-formed sentences found in conventional documents, spoken dialogues are often composed of *utterances*. Salient information is diluted across these utterances and is accompanied by common semantics. Additionally, noise abounds in the form of unrelated chit-chats and transcription errors (Tixier et al. 2017). Such common or noisy words, e.g., *please*, *thanks*, and *humm*, usually have a high frequency and co-occur with other informative words. As a result, the general topic-based approach can hardly distinguish the mixture of useful and useless content statistically, leading to inaccurate estimations of topic distribution

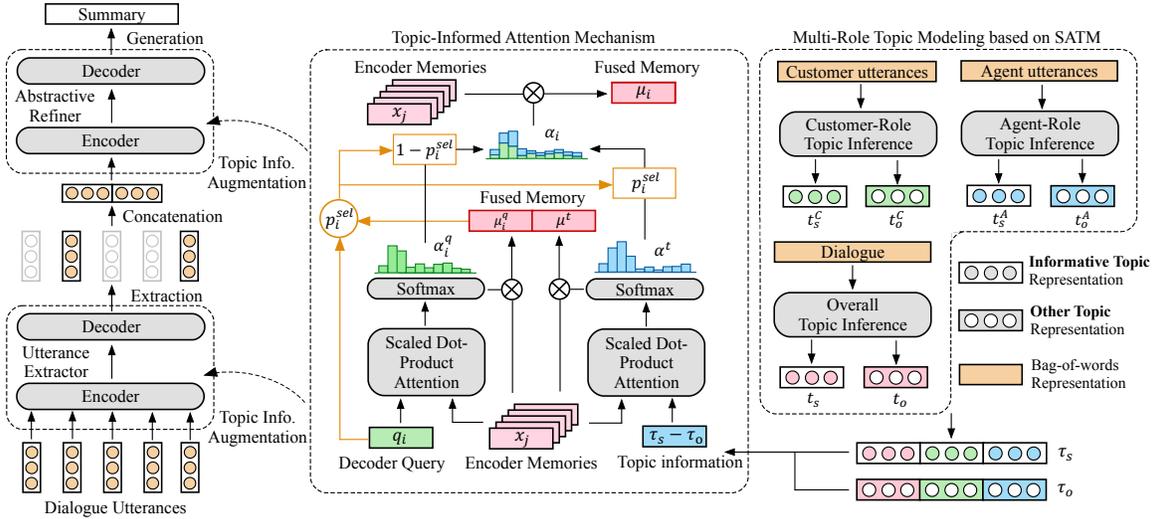


Figure 2: Overview of our proposed TDS with multi-role topic modeling based on SATM.

(Li et al. 2018, 2019b). Besides, in a customer service dialogue, the participating roles are stable: a customer tends to raise a problem and an agent needs to provide solutions. Figure 1 shows a real-world customer service dialogue along with a summary that includes critical information from the two speakers. Hence, the model is also expected to capture role information to assist with saliency estimation.

In this work, we propose a novel two-stage neural model jointly with an enhanced topic modeling approach for spoken dialogue summarization. **First**, to better distinguish the underlying informative content from abundant common semantics and dialogue noise, we introduce a saliency-aware topic model (SATM), where topics are split into two groups: *informative topics* and *other topics*. In the generative process of topic modeling, we constrain each salient word that corresponds to the gold summary to be generated from *informative topics*, while other words in the dialogue (including noisy and common words) are generated only from *other topics*. Through this training process, SATM can associate each word in a dialogue with either saliency (informative topics) or not salient (other topics). **Second**, to capture role information and extract semantic topics from dialogues, we employ SATM to perform multi-role topic modeling on customer utterances, agent utterances, and overall dialogues separately. Then, a topic-augmented two-stage dialogue summarizer (TDS) is designed, which consists of an utterance extractor and an abstractive refiner. It can pick out topic-relevant salient information on both the utterance level and word level via a topic-informed attention mechanism.

Furthermore, due to the lack of suitable public benchmarks, we collected a real-world customer service dialogue dataset with highly abstractive summaries. Experimental results on the proposed dataset showed that our model outperforms a series of strong baselines under various metrics. Codes, datasets, and supplementary can be found at Github<sup>1</sup>.

In summary, our contributions are as follows: 1) We in-

roduce a novel topic model that can perceive underlying informative content in dialogues by directly learning word-saliency correspondences. 2) Based on multi-role topic modeling, we propose a topic-augmented two-stage model with a topic-informed attention mechanism to perform saliency estimation and summarize customer service dialogues. 3) Experimental results on the collected dataset demonstrate the effectiveness of our method in different aspects.

## Method

In this section, we will detail the saliency-aware topic model (SATM) and the topic-augmented two-stage dialogue summarizer (TDS). The SATM infers multi-role topic representations based on *informative topics* and *other topics*. Then topic information is incorporated into the extractor and the refiner of TDS via a topic-informed attention mechanism. The overall architecture of our model is shown in Figure 2.

### Saliency-Aware Neural Topic Model

Our proposed SATM is based on the Neural Topic Model (NTM) with variational inference (Miao et al. 2017), which infers the topic distribution  $\theta$  from each dialogue  $d$  by a neural network. We extend NTM with a new generative strategy to learn the word-saliency correspondences. The architecture of SATM compared with NTM is shown in Figure 3.

**Basic NTM with Variational Inference.** Formally, given the bag-of-words representation of a dialogue  $d \in \mathbb{R}^{|V|}$  with stop words removed, we build an inference network  $q(\theta|d)$  to approximate the posterior  $p(\theta|d)$ , where  $V$  is the vocabulary.  $q(\theta|d)$  is composed of a function  $\theta = f(z)$  conditioned on a diagonal Gaussian distribution  $z \sim \mathcal{N}(\mu(d), \sigma^2(d))$ , where  $\mu(d)$  and  $\sigma(d)$  are neural networks. In practice, we can sample  $\hat{z}$  using a re-parameterization trick (Kingma and Welling 2014) by  $\hat{z} = \mu(d) + \epsilon \cdot \sigma(d)$ , where  $\epsilon$  is sampled from  $\mathcal{N}(0, I^2)$ . Then, a sampled  $\hat{\theta} \in \mathbb{R}^K$  is derived as:

$$\hat{\theta} = f(\hat{z}) = \text{softmax}(W_{\theta}\hat{z} + b_{\theta}). \quad (1)$$

<sup>1</sup><https://github.com/RowitZou/topic-dialog-summ>

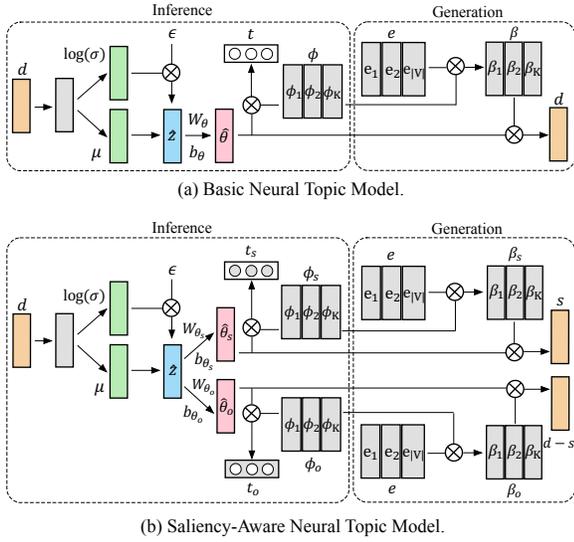


Figure 3: Comparison of NTM and SATM.

$W_\theta$ ,  $b_\theta$  are trainable parameters and  $K$  denotes the number of topics. Then, we define  $\beta \in \mathbb{R}^{K \times |V|}$ ,  $\phi \in \mathbb{R}^{K \times H}$ ,  $e \in \mathbb{R}^{|V| \times H}$  to represent topic-word distributions, topic vectors, and word vectors, respectively. Here,  $H$  is the dimension of vectors.  $\phi$  is randomly initialized and  $e$  can be pre-trained word embeddings.  $\beta$  is computed with  $\phi$  as follows:

$$\beta_k = \text{softmax}(e \cdot \phi_k^\top). \quad (2)$$

In the generative part, we parameterize  $p(d|\beta, \theta)$  and define the loss function of neural topic model as:

$$\begin{aligned} \mathcal{L}_T &= D_{KL}[q(\theta|d)||p(\theta)] - \mathbb{E}_{q(\theta|d)}[\log p(d|\beta, \theta)] \\ &\approx D_{KL}[q(z|d)||p(z)] - \sum_n \log p(w_n|\beta, \hat{\theta}). \end{aligned} \quad (3)$$

The first term uses the KL-divergence to ensure that the variational distribution  $q(\theta|d)$  is similar to the true prior  $p(\theta)$ , where  $p(\theta)$  represents a standard Gaussian prior  $\mathcal{N}(0, I^2)$ . In the second term,  $w_n$  denotes the  $n$ -th observed word in  $d$  and the log-likelihood of  $d$  can be computed with  $\log(\hat{\theta} \cdot \beta)$ .

**Learning Word-Saliency Correspondences.** In spoken dialogues, abundant noise and common semantics appear randomly and co-occur with informative words. Meanwhile, salient information is encapsulated in dialogue summaries. We therefore assume that each dialogue is a mixture of informative words and other words, where words corresponding to the gold summary are basically informative. We split  $K$  topics into two groups: *informative topics* and *other topics*, where the topic number is  $K_s$  and  $K_o$  ( $K = K_s + K_o$ ), respectively. Given  $\hat{z}$  derived from  $d$ , the distribution over *informative topics*  $\hat{\theta}_s \in \mathbb{R}^{K_s}$  and *other topics*  $\hat{\theta}_o \in \mathbb{R}^{K_o}$  are inferred by  $f_s(\cdot)$  and  $f_o(\cdot)$  with different parameters:

$$\begin{aligned} \hat{\theta}_s &= f_s(\hat{z}) = \text{softmax}(W_{\theta_s} \hat{z} + b_{\theta_s}), \\ \hat{\theta}_o &= f_o(\hat{z}) = \text{softmax}(W_{\theta_o} \hat{z} + b_{\theta_o}). \end{aligned} \quad (4)$$

In the generative part of topic modeling, we use  $s \in \mathbb{R}^{|V|}$  to represent a word subset of  $d$ , in which each word appears in

both the gold summary and the original dialogue. Thus the parameterization of  $p(d|\beta, \theta)^2$  is decomposed by:

$$p(d|\beta, \theta) = p(s|\beta_s, \theta_s)p(d-s|\beta_o, \theta_o), \quad (5)$$

where  $\beta_s \in \mathbb{R}^{K_s \times |V|}$ ,  $\beta_o \in \mathbb{R}^{K_o \times |V|}$  along with  $\phi_s \in \mathbb{R}^{K_s \times H}$ ,  $\phi_o \in \mathbb{R}^{K_o \times H}$  are constructed by splitting  $\beta$  and  $\phi$  based on the two topic groups. Eq.5 indicates that topic assignments for summary words are constrained in the group of *informative topics*, while other words in the dialogue are gathered by *other topics*. As a result, each word in a dialogue is associated with either saliency or not salient. Accordingly, given  $w_n^s$  and  $w_n^{d-s}$  that denote the  $n$ -th observed word in  $s$  and  $d-s$ , respectively, the loss function becomes:

$$\begin{aligned} \mathcal{L}_T &\approx - \sum_n \log p(w_n^s|\beta_s, \hat{\theta}_s) - \sum_n \log p(w_n^{d-s}|\beta_o, \hat{\theta}_o) \\ &\quad + D_{KL}[q(z|d)||p(z)]. \end{aligned} \quad (6)$$

Compared to NTM, the proposed SATM leverages dialogue summaries to detach informative words from noise and common semantics, avoiding the direct topic modeling on a mixture of useful and useless content. Hence, noise and common semantics can hardly obscure the underlying informative words, making the topic inference more robust. Besides, SATM can be easily employed as an external module and combined with other summarization models like Wang et al. (2018) as long as gold summaries are available.

**Multi-Role Topic modeling.** Based on SATM, we input  $d$  and infer topic representations  $t_s \in \mathbb{R}^H$  and  $t_o \in \mathbb{R}^H$  by:

$$t_s = \phi_s^\top \cdot \hat{\theta}_s, \quad t_o = \phi_o^\top \cdot \hat{\theta}_o. \quad (7)$$

Here,  $t_s$  can be regarded as a topic vector that captures informative topic information while  $t_o$  gathers noise and common semantics, both of which will be incorporated into the TDS to facilitate the saliency estimation. Furthermore, in order to capture role-specific information, we perform topic modeling on customer utterances and agent utterances separately. Given the bag-of-words representation of customer utterances  $d^C \in \mathbb{R}^{|V|}$  and agent utterances  $d^A \in \mathbb{R}^{|V|}$ , we can infer topic distributions  $\hat{\theta}_s^C, \hat{\theta}_o^C, \hat{\theta}_s^A, \hat{\theta}_o^A$  using Eq.4. Then, topic representations for different roles  $t_s^C, t_o^C, t_s^A, t_o^A$  can be obtained similar to Eq.7. Hence, we have totally three topic models with different parameters on customer utterances, agent utterances, and overall dialogues, respectively.

### Topic-Augmented Two-Stage Dialogue Summarizer

In the customer service scenario, spoken dialogues are long and sometimes twisted, where most utterances are unimportant or even noisy, which can be directly filtered out. Hence, we follow Chen and Bansal (2018) and employ a two-stage summarizer that first selects salient utterances and then refines them. The basic two-stage summarizer consists of an *utterance extractor* and an *abstractive refiner*. The extractor encodes each utterance  $u_i$  into an utterance representation  $h_i$  and employs a *Pointer Network* (Vinyals et al. 2015) to recurrently extract utterances based on  $h_i$ . The refiner is a

<sup>2</sup>Notably, we do not parameterize  $p(\theta|s)$  and  $p(s|\beta, \theta)$  directly, because gold summaries are not available at test time.

standard sequence-to-sequence (Seq2seq) model, which can generate a concise summary based on the extracted utterances. To bridge the extractor and the refiner, a policy gradient technique (Williams 1992) is applied to train the overall summarizer, where we use  $\mathcal{L}_S$  to represent the loss function.

In this work, we use Transformer (Vaswani et al. 2017) as the basic encoder and decoder layer for TDS. For the  $i$ -th utterance in a dialogue, we have  $u_i = \{r_i, e_{i1}, \dots, e_{iN}\}$ , where  $r_i$  is a role embedding representing the speaker of  $u_i$ , which can be either *customer* or *agent*.  $e_{ij}$  is the embedding of the  $j$ -th word. For more details of the basic two-stage model and our implementation, please refer to Chen and Bansal (2018) and the supplementary due to the space limitation.

**Topic Information Augmentation.** To capture role information and highlight global topics, we incorporate multi-role topic representations into the Transformer Decoder via a topic-informed attention mechanism for saliency estimation, which is an extension of multi-head attention (Vaswani et al. 2017). Formally, let  $q_i$  denote the  $i$ -th decoding step of the query, and  $x_j$  denote the  $j$ -th element in the memory, the original attention mechanism for each head is defined as:

$$\begin{aligned} \alpha_{ij}^q &= \text{softmax}((q_i W_Q)(x_j W_K^q)^\top / \sqrt{d_h}), \\ \mu_i^q &= \sum_j \alpha_{ij}^q (x_j W_V), \end{aligned} \quad (8)$$

where  $W_Q, W_K^q, W_V$  are trainable parameters and  $d_h$  is the dimension of each head.  $\mu_i^q$  is a vector that fuses salient information based on the query at the  $i$ -th decoding step. In a general decoding process, the state of fused memory at each step is conditioned on the previously decoded sequence where errors may be accumulated. Here, we additionally use global topics and role information as a guidance to assist with sequence decoding by measuring relevance between memory elements and multi-role topics. Formally, we design an auxiliary attention operation as follows:

$$\begin{aligned} \alpha_j^t &= \text{softmax}((\tau_s W_T - \tau_o W_T)(x_j W_K^t)^\top / \sqrt{d_h}), \\ \mu^t &= \sum_j \alpha_j^t (x_j W_V). \end{aligned} \quad (9)$$

Here,  $\tau_s$  and  $\tau_o$  represent role-specific topic representation, where  $\tau_s = [t_s; t_s^C; \mathbf{0}]$ ,  $\tau_o = [t_o; t_o^C; \mathbf{0}]$  if  $x_j$  corresponds to the *customer* speaker, and  $\tau_s = [t_s; \mathbf{0}; t_s^A]$ ,  $\tau_o = [t_o; \mathbf{0}; t_o^A]$  if  $x_j$  corresponds to the *agent* speaker.  $[\cdot; \cdot]$  means concatenation and  $\mathbf{0}$  is a vector with all elements set to 0. In Eq.9, we design  $\alpha_j^t$  that makes  $\tau_s$  contrary to  $\tau_o$  inspired by the contrastive attention (Duan et al. 2019a), which encourages the attention to topic-relevant elements, and discourages the attention to noise and common semantics. Hence,  $\tau_s$  and  $\tau_o$  work in an opposite way to contribute to an overall target.  $\mu^t$  can be regarded as a topic-aware vector that basically discards noisy and uninformative elements in the memory. Finally, we combine the above two attention operations to consider both the global topics and the current query at each decoding step, and form an integrated memory fusion  $\mu_i$  by:

$$\begin{aligned} p_i^{sel} &= \sigma([q_i; \mu_i^q; \mu^t] \cdot W_P), \\ \alpha_{ij} &= (1 - p_i^{sel}) \cdot \alpha_{ij}^q + p_i^{sel} \cdot \alpha_j^t, \\ \mu_i &= \sum_j \alpha_{ij} (x_j W_V), \end{aligned} \quad (10)$$

Table 1: Statistics of the customer service dataset.

	# of dialogues	# of utterances	average token number		
			dialog.	utter.	summ.
Train	17,189	872,292	1,333.72	26.28	54.54
Dev.	820	38,461	1,221.12	26.03	53.73
Test	851	42,667	1,300.98	25.95	54.42

where  $p_i^{sel} \in (0, 1)$  denotes the selective probability used as a soft switch to choose between the original query-based attention or the topic-guided attention. The attention mechanism is further adapted into the multi-head manner similar to Vaswani et al. (2017). Notably, we apply the topic-informed attention mechanism to both the utterance extractor and the abstractive refiner. For the extractor,  $x_j$  represents the hidden state of  $j$ -th utterance. For the refiner,  $x_j$  is the hidden state of  $j$ -th word in selected utterances. As a result, we can perform saliency estimation assisted by multi-role topic information on both the utterance level and word level.

## Joint Training

To jointly train the summarizer and the multi-role topic models, we design a joint loss which includes the loss function of summarizer  $\mathcal{L}_S$  and the loss functions of three topic models  $\mathcal{L}_T^C, \mathcal{L}_T^A, \mathcal{L}_T$  that correspond to customer utterances, agent utterances and overall dialogues, respectively. The joint loss function is defined as:

$$\mathcal{L} = \mathcal{L}_S + \lambda(\mathcal{L}_T^C + \mathcal{L}_T^A + \mathcal{L}_T), \quad (11)$$

where  $\lambda$  is a coefficient to balance the losses between the summarizer and topic models.

## Experimentation

In this section, We describe the experiments conducted on a real-world customer service dataset. We compare our proposed TDS+SATM with strong baselines and further analyze the influence of different parts in our model. All hyperparameters tuning is conducted on the validation set. Full training details can be found in the supplementary.

## Dataset

Our customer service dialogue dataset is collected from the call center of an E-commerce company. All dialogues are incoming calls in Mandarin Chinese that take place between a customer and a service agent. After each round of service, an agent needs to write a brief description about the conversation, which mainly includes problems the customer faces and solutions the agent provides. For each dialogue example, we take the agent-written description as the gold summary. All dialogues are originally in the form of audio and we transcribe them into texts using an ASR model pre-trained on customer service dialogues (Zhang et al. 2019a) with a character error rate of 9.3%. The final dataset therefore includes dialogue-summary pairs that consist of dialogue transcriptions and human-written summaries. We totally collect 18.86K dialogues with 953K utterances and split them into

Table 2: Results of automatic metrics on the customer service dataset. RG-(1,2,L) represents the F1 score of ROUGE-(1,2,L). TRF denotes the Transformer. Methods marked with \* utilize BERT as the word-level encoder.

Methods	RG-1	RG-2	RG-L	BLEU
Ext-Oracle	41.38	15.64	29.18	7.28
Seq2seq+Att	28.66	13.05	22.63	6.89
PGNet	34.88	17.81	27.80	9.77
TRF	35.17	18.01	28.05	9.87
CopyTRF	34.97	17.84	27.88	9.78
HiBERT*	35.50	18.24	28.44	9.89
BERT+TRF*	35.67	18.49	28.57	10.19
FastRL*	35.99	18.67	28.86	10.40
TDS+NTM (base)	35.21	18.04	28.11	9.87
TDS+SATM (base)	35.75	18.54	28.62	10.34
TDS+NTM*	36.13	19.09	29.00	10.77
TDS+SATM*	<b>36.81</b>	<b>19.63</b>	<b>29.61</b>	<b>11.24</b>

training (90%), development (5%), and test (5%) set. Table 1 shows the detailed statistics of the collected dataset.

### Comparison Methods.

- **Ext-Oracle (Nallapati et al. 2017)**, where a greedy algorithm is applied to select utterances whose combination maximizes the evaluation score against the gold summary, which is used as the upper bound of extractive methods.
- **Seq2seq+Att (Nallapati et al. 2016)** is a standard RNN-based encoder-decoder model with attention mechanisms.
- **PGNet (See et al. 2017)** has a pointer mechanism where the decoder can choose to generate a word from the vocabulary or copy a word from the source text.
- **Transformer (Vaswani et al. 2017)** is an attention-based model, for which we also implement a variant with the copy mechanism, denoted as CopyTransformer.
- **BERT+Transformer (Liu and Lapata 2019)** consists of a BERT encoder (Devlin et al. 2019)<sup>3</sup> and a Transformer decoder. They are tuned with different optimizers to alleviate the mismatch between BERT and other parameters<sup>4</sup>.
- **HiBERT (Zhang et al. 2019b)** can encode documents on the word level and sentence level hierarchically. Here, we replace the word-level encoder with BERT and add a basic Transformer decoder to enable Seq2seq learning.
- **FastRL (Chen and Bansal 2018)** is the basic two-stage framework. We implement it based on Transformers and pre-train it with BERT, in that Transformer encoder can be easily combined with pre-trained LMs.
- **TDS+Topic Model** is our approach with different topic models, including NTM and SATM. For a fair comparison, we also employ BERT as the word-level encoder.

<sup>3</sup>The original BERT is only applicable for texts with a maximum length of 512. We extend the range of positional embeddings to make it possible to encode long dialogues.

<sup>4</sup>Other models equipped with BERT use the same strategy.

Table 3: Human evaluation with system ranking results.

Methods	Informativeness	Fluency
PGNet	-0.196	-0.248
BERT+TRF	-0.116	-0.042
HiBERT	-0.120	-0.030
FastRL	-0.084	0.086
TDS+SATM	<b>0.032</b>	<b>0.098</b>
Gold	0.484	0.136

### Automatic Evaluation

Table 2 shows the automatic evaluation results on the customer service dataset (examples of system output can be found in the supplementary). We evaluate summarization quality using ROUGE F1 (Lin 2004) and BLEU (Papineni et al. 2002). We use unigram and bigram overlap (ROUGE-1, ROUGE-2) between system outputs and gold summaries to assess informativeness, and the longest common subsequence (ROUGE-L) to assess fluency. For BLEU, we use 4-grams at most and average the scores of different grams. All metrics are computed on Chinese characters to avoid the influence of word segmentation.

The first block in the table includes Oracle as an upper bound for extractive methods. The second block shows the results of abstractive models. From Table 2 we can see that most abstractive methods achieve competitive results or even outperform Ext-Oracle on ROUGE-(2/L) and BLEU. It provides evidence that our dataset collects highly abstractive summaries, which requires a system to integrate dialogue content and produce a coherent discourse. The basic TDS plus topic models achieves competitive results against other baselines and outperforms some BERT-based models when equipped with SATM. After combining BERT, two-stage systems (FastRL / TDS) show superior performances compared to other Seq2seq approaches, which probes the effectiveness of extract-refine strategy and indicates that useful information is diluted in long spoken dialogues. When topic information is incorporated, results are further improved. TDS+NTM uses the basic neural topic model and removes the contrastive mechanism in Eq.9, which brings slight improvements against FastRL. In contrast, TDS+SATM leads to a significant performance enhancement over FastRL on ROUGE (+0.82, +0.96, +0.75 on ROUGE-1/2/L) and BLEU (+0.84) with  $p < 0.05$ . It validates that topic information is beneficial for summarizing customer service dialogues, and SATM with word-saliency learning can further boost overall results by improving the quality of topic modeling.

### Human Evaluation.

Metrics for automatic evaluation based on n-grams may not truly reflect the quality of generated summaries. Hence, we further randomly sample 100 examples in the test set for human evaluation. We follow Narayan et al. (2018) to design the experiment, in which three volunteers are invited to compare summaries produced from PGNet, BERT+TRF, HiBERT, FastRL, our proposed TDS+SATM, and the gold sum-

Table 4: Ablation study of TDS+SATM with different kinds of topic modeling. Agent and Cust. represent topic modeling on agent utterances and customer utterances, respectively.

Methods	RG-1	RG-2	RG-L	BLEU
TDS+SATM	36.81	<b>19.63</b>	<b>29.61</b>	<b>11.24</b>
(w/o) Cust.	<b>36.84</b>	19.60	29.56	11.09
(w/o) Agent	36.79	19.39	29.50	10.73
(w/o) Agent & Cust.	36.37	19.03	29.10	10.37

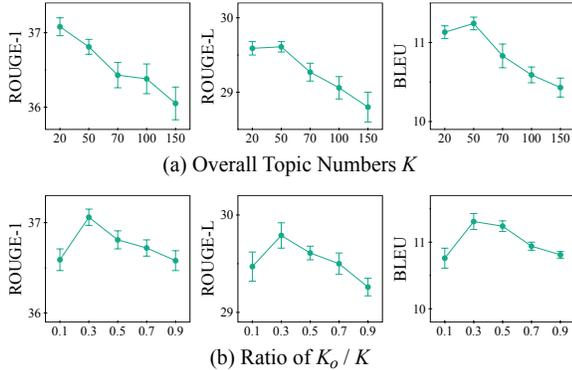


Figure 4: Effects of different topic numbers of SATM.

mary (Gold). Each volunteer is presented with a dialogue and two summaries produced from two out of six systems and is asked to decide which summary is better in order of two dimensions: **informativeness** (which summary captures more important information in the dialogue?) and **fluency** (which summary is more fluent and well-formed?). We collect judgments from three volunteers for each comparison with the order of dialogues and summaries randomized.

Table 3 gives the system ranking results of human evaluation. The score of a system is calculated as the percentage of times it was selected as best minus the percentage of times it was chosen as worst, which ranges from -1 (worst) to 1 (best). Gold summaries are unsurprisingly ranked best on both two dimensions. For informativeness, TDS+SATM ranks second followed by other systems, which validates the effectiveness of our proposed SATM and topic argumentation mechanism for assisting with saliency estimation. In terms of fluency, two-stage models (FastRL / TDS) are considered better than other baselines, which indicates that the extract-refine strategy can generate more fluent summaries. We also conducted pairwise comparisons between systems (using a binomial two-tailed test; null hypothesis: two systems are equally good;  $p < 0.05$ ). In terms of informativeness, TDS+SATM is significantly different from all other systems. In terms of fluency, two-stage systems are significantly different from other systems, and FastRL is not significantly different from TDS+SATM.

## Analysis and Discussion

To better understand the influence of role information, saliency-aware topic modeling, and the topic-informed at-

Table 5: Top-10 words of example topics in different topic groups learned by joint training of TDS+Topic Model.

SATM Informative Topic	<b>T1:</b> deliver, time, order, address, modify, cancel, ship, return, refund, receive <b>T2:</b> feedback, problem, submit, suggest, apply, complain, seller, quality, product, slow <b>T3:</b> buy, account, pay, bind, phone, number, modify, check, username, message
SATM Other Topic	<b>T1:</b> please, wait, service, sorry, really, thanks, bother, mean, find, welcome <b>T2:</b> send, call, record, again, later, check, help, keep, contact, reply
NTM General Topic	<b>T1:</b> thanks, later, sorry, please, really, phone, feedback, deliver, number, return <b>T2:</b> apply, sorry, again, order, check, wait, record, reply, seller, contact

tention, we perform the following qualitative analysis.

**Contribution of Role Information.** Table 4 shows the results of TDS+SATM with different kinds of topic modeling. When we remove one of the topic models on customer utterances or agent utterances, results are not be appreciably affected. However, after removing both two topic models, the system suffers a significant performance degradation ( $p < 0.05$ ). It indicates that role-specific information is beneficial for dialogue modeling, and at least one of the speakers should be specified to make role content distinguishable.

**Effect of Topic Numbers in SATM.** The topic number is a critical hyper-parameter in topic models because it potentially affects the convergence rate and the inference quality. Here, we report results of TDS+SATM with different topic numbers in Figure 4. Figure 4(a) shows the effects of  $K$  that ranges from 20 to 150 with  $K_s = K_o$ . It shows a performance decline trend and a variance increase trend when  $K$  is continuously increased after exceeding 50. It indicates that a proper topic number is sufficient for capturing main topics and a larger one makes the topic inference unstable. Figure 4(b) shows the effects of different ratios of  $K_o / K$ , where we fix  $K=50$  and adjust  $K_o$  in range of 5 to 45. The results show that an overly unbalanced number of  $K_s$  and  $K_o$  can hurt the performance. It indicates that each dialogue is a mixture of useful and uninformative content, either of which can not be ignored when performing topic modeling on dialogues.

**Comparison between NTM and SATM.** To thoroughly compare the standard NTM and our proposed SATM, we analyze the topic-word distributions  $\beta$  and the topic vectors  $\phi$  learned by TDS+Topic Model. Table 5 shows topic examples of different topic groups, where top-10 words with the highest probability in  $\beta$  are listed<sup>5</sup>. We found that words in *informative topics* can better reflect specific dialogue scenes. For instances, topic 1 with *address*, *ship* and *refund* is about delivery issues. Topic 3 with *account*, *bind* and *username* is about account issues. By contrast, *other topics* tend to gather noise and common semantics, where words of topic 1 often appear in unrelated chit-chats, and topic 2 includes common words in the customer service scenario. As for *general top-*

<sup>5</sup>They are translated from Chinese with stop words removed.

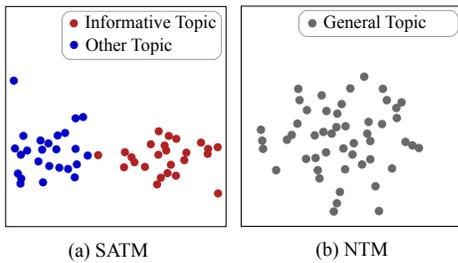


Figure 5: 2-D t-SNE visualizations of topic vectors ( $K = 50$  and  $K_s = K_o = 25$ ).

ics in NTM, top-10 words show a mixture of informative and common content. Besides, we analyze the topic vectors  $\phi$  and visualize the latent space in 2-D using t-SNE (Maaten and Hinton 2008) for SATM and NTM. In Figure 5(a), with the learning of word-saliency correspondences and the contrastive mechanism, vectors of two topic groups in SATM are effectively mapped to separate regions, while in Figure 5(b), topic vectors of NTM do not show obvious clusters.

#### Case Study of Topic-Informed Attention Mechanism.

Figure 6 shows the attention map of an exemplar dialogue with a translated summary generated by TDS+SATM.  $\alpha^q$ ,  $\alpha^t$ ,  $\alpha$  represent the query-based attention, the topic-guided attention and the final combined attention, respectively. Attention scores are taken from the decoder of extractor to demonstrate the beginning step in an utterance-level decoding process. From the example we can see that topic-guided attention  $\alpha^t$  successfully focuses on the salient utterance that mentions customer’s problem. Then the combined attention  $\alpha$  exhibits preference to  $\alpha^t$  and focuses on appropriate utterances that finally contribute to the summary generation.

## Related Work

### Dialogue Summarization

Dialogue summarization is a challenging task and has been widely explored in various scenarios. Previous works generally focus on summarizing dialogues by stringing key points to maintain an integral dialogue flow: Mehdad et al. (2013) and Shang et al. (2018) first group utterances that share similar semantics by community detection, and then generate a summary sentence for each utterance group. Liu et al. (2019a) propose a hierarchical model to produce key point sequences and generate summaries at the same time for customer service dialogues. Duan et al. (2019b) train the assignment of utterances to the corresponding controversy focuses to summarize court debate dialogues. Several works (Zechner 2001; Xie et al. 2008; Oya et al. 2014; Liu et al. 2019b; Li et al. 2019a) split dialogues into multiple segments by means of topic segmentation when conducting summarization. Different from above works, Pan et al. (2018) first attempt to generate a highly abstractive summary for the entire dialogue, which produces a concise event/object description with a Transformer-based approach. By contrast, in this work, dialogue summaries generally highlight role-specific content, which requires the system to further focus on the

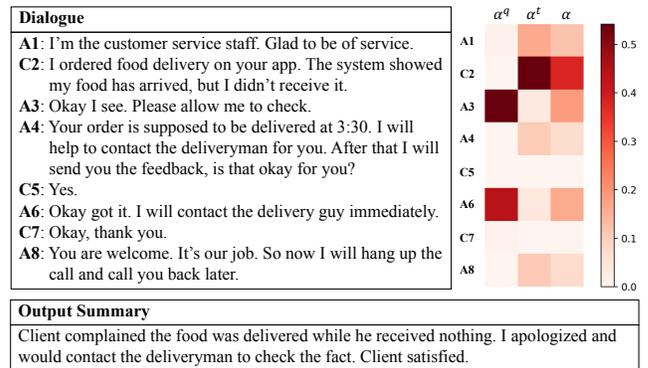


Figure 6: Utterance-level attention map of an example dialogue along with the output summary from TDS+SATM.

role information when performing saliency estimation.

### Text Summarization with Topic Modeling

Topic models have been extensively studied for document modeling and information retrieval. Probabilistic topic models like pLSA (Hofmann 1999) and LDA (Blei, Ng, and Jordan 2003) provide a theoretically sound foundation for uncovering the underlying semantics of a document. Recently, neural topic models (Miao et al. 2017) have been introduced to infer latent representations for documents, which leverage deep neural networks as approximators for learning topic distributions. A couple of works have employed these topic models to facilitate the summarization task. Wang et al. (2018) and Narayan et al. (2018) use LDA to infer topic embeddings and design a joint attention mechanism to incorporate topic information. Fu et al. (2020) merge the topic inference module with a summarization model rather than simply resort to using a pre-trained topic model. Some early works on dialogue summarization (Higashinaka et al. 2010; Wang and Cardie 2012; Sood et al. 2013) directly perform topic modeling on dialogues to extract salient words or utterances. All the above methods leverage the standard topic modeling framework as an auxiliary tool to conduct topic mining for summarization. By contrast, we use summary information as a guidance to force the topic model to learn word-saliency correspondences. As a result, underlying semantics can hardly be obscured by uninformative content, making the salient information more perceivable.

## Conclusion and Future Work

In this paper, we propose a topic-augmented two-stage summarizer with a multi-role topic modeling mechanism for customer service dialogues, which can generate highly abstractive summaries that highlight role-specific information. Moreover, we introduce a novel training regime for topic modeling that directly learns word-saliency correspondences to alleviate the influence of uninformative content. Experiments on a real-world customer service dataset validate the effectiveness of our approach. Future directions may be the exploration of template-guided abstractive methods to make summaries more standardized and easier for reporting.

## Acknowledgments

The authors wish to thank the anonymous reviewers for their helpful comments. This work was partially funded by China National Key R&D Program (No. 2018YFC0831105), National Natural Science Foundation of China (No. 61751201, 62076069, 61976056), Shanghai Municipal Science and Technology Major Project (No.2018SHZDZX01), Science and Technology Commission of Shanghai Municipality Grant (No.18DZ1201000, 17JC1420200). This work was supported by Alibaba Group through Alibaba Innovative Research Program.

## References

- Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3(Jan): 993–1022.
- Chen, Y.-C.; and Bansal, M. 2018. Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 675–686.
- Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186.
- Duan, X.; Yu, H.; Yin, M.; Zhang, M.; Luo, W.; and Zhang, Y. 2019a. Contrastive Attention Mechanism for Abstractive Sentence 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)*, 3035–3044.
- Duan, X.; Zhang, Y.; Yuan, L.; Zhou, X.; Liu, X.; Wang, T.; Wang, R.; Zhang, Q.; Sun, C.; and Wu, F. 2019b. Legal Summarization for Multi-role Debate Dialogue via Controversy Focus Mining and Multi-task Learning. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 1361–1370.
- Fu, X.; Wang, J.; Zhang, J.; Wei, J.; and Yang, Z. 2020. Document Summarization with VHTM: Variational Hierarchical Topic-Aware Mechanism. In *AAAI*, 7740–7747.
- Gillick, D.; Riedhammer, K.; Favre, B.; and Hakkani-Tur, D. 2009. A global optimization framework for meeting summarization. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 4769–4772. IEEE.
- Higashinaka, R.; Minami, Y.; Nishikawa, H.; Dohsaka, K.; Meguro, T.; Takahashi, S.; and Kikui, G. 2010. Learning to model domain-specific utterance sequences for extractive summarization of contact center dialogues. In *Coling 2010: Posters*, 400–408.
- Hofmann, T. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 50–57.
- Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. *CoRR* abs/1312.6114.
- Li, M.; Zhang, L.; Ji, H.; and Radke, R. J. 2019a. Keep meeting summaries on topic: Abstractive multi-modal meeting summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2190–2196.
- Li, X.; Wang, Y.; Zhang, A.; Li, C.; Chi, J.; and Ouyang, J. 2018. Filtering out the noise in short text topic modeling. *Information Sciences* 456: 83–96.
- Li, X.; Zhang, J.; Ouyang, J.; et al. 2019b. Dirichlet Multinomial Mixture with Variational Manifold Regularization: Topic Modeling over Short Texts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 7884–7891.
- Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.
- Liu, C.; Wang, P.; Xu, J.; Li, Z.; and Ye, J. 2019a. Automatic Dialogue Summary Generation for Customer Service. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.
- Liu, Y.; and Lapata, M. 2019. Text Summarization with Pre-trained 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)*, 3721–3731.
- Liu, Z.; Ng, A.; Lee, S.; Aw, A. T.; and Chen, N. F. 2019b. Topic-aware pointer-generator networks for summarizing spoken conversations. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 814–821. IEEE.
- Maaten, L. v. d.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9(Nov): 2579–2605.
- Mehdad, Y.; Carenini, G.; Tompa, F.; and Ng, R. 2013. Abstractive meeting summarization with entailment and fusion. In *Proceedings of the 14th European Workshop on Natural Language Generation*, 136–146.
- Miao, Y.; Grefenstette, E.; Blunsom, P.; et al. 2017. Discovering discrete latent topics with neural variational inference. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2410–2419.
- Nallapati, R.; Zhai, F.; Zhou, B.; et al. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Thirty-First AAAI Conference on Artificial Intelligence*.
- Nallapati, R.; Zhou, B.; Gulcehre, C.; and Xiang, B. 2016. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, 280–290.
- Narayan, S.; Cohen, S. B.; Lapata, M.; et al. 2018. Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. In

- Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1797–1807.
- Oya, T.; Mehdad, Y.; Carenini, G.; and Ng, R. 2014. A template-based abstractive meeting summarization: Leveraging summary and source text relationships. In *Proceedings of the 8th International Natural Language Generation Conference (INLG)*, 45–53.
- Pan, H.; Zhou, J.; Zhao, Z.; Liu, Y.; Cai, D.; and Yang, M. 2018. Dial2desc: end-to-end dialogue description generation. *arXiv preprint arXiv:1811.00185*.
- Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.
- Rambow, O.; Shrestha, L.; Chen, J.; and Lauridsen, C. 2004. Summarizing email threads. In *Proceedings of HLT-NAACL 2004: Short Papers*, 105–108.
- See, A.; Liu, P. J.; Manning, C. D.; et al. 2017. Get To The Point: Summarization with Pointer-Generator Networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 1073–1083.
- Shang, G.; Ding, W.; Zhang, Z.; Tixier, A.; Meladianos, P.; Vazirgiannis, M.; and Lorré, J.-P. 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)*, 664–674.
- Sood, A.; Mohamed, T. P.; Varma, V.; et al. 2013. Topic-focused summarization of chat conversations. In *European Conference on Information Retrieval*, 800–803. Springer.
- Tixier, A.; Meladianos, P.; Vazirgiannis, M.; et al. 2017. Combining graph degeneracy and submodularity for unsupervised extractive summarization. In *Proceedings of the workshop on new frontiers in summarization*, 48–58.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.
- Vinyals, O.; Fortunato, M.; Jaitly, N.; et al. 2015. Pointer networks. In *Advances in neural information processing systems*, 2692–2700.
- Wang, J.; Wang, J.; Sun, C.; Li, S.; Liu, X.; Si, L.; Zhang, M.; and Zhou, G. 2020. Sentiment Classification in Customer Service Dialogue with Topic-Aware Multi-Task Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Wang, L.; and Cardie, C. 2012. Unsupervised topic modeling approaches to decision summarization in spoken meetings. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 40–49.
- Wang, L.; Yao, J.; Tao, Y.; Zhong, L.; Liu, W.; and Du, Q. 2018. A reinforced topic-aware convolutional sequence-to-sequence model for abstractive text summarization. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 4453–4460.
- Williams, R. J. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* 8(3-4): 229–256.
- Xie, S.; Liu, Y.; Lin, H.; et al. 2008. Evaluating the effectiveness of features and sampling in extractive meeting summarization. In *2008 IEEE Spoken Language Technology Workshop*, 157–160. IEEE.
- Zechner, K. 2001. Automatic generation of concise summaries of spoken dialogues in unrestricted domains. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, 199–207.
- Zhang, S.; Lei, M.; Liu, Y.; and Li, W. 2019a. Investigation of modeling units for mandarin speech recognition using dfsmn-ctc-smbr. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7085–7089. IEEE.
- Zhang, X.; Wei, F.; Zhou, M.; et al. 2019b. HIBERT: Document Level Pre-training of Hierarchical Bidirectional Transformers for Document Summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5059–5069.