electronics

MDPI

Article

WHORU: Improving Abstractive Dialogue Summarization with Personal Pronoun Resolution

Tingting Zhou

School of Foreign Studies, Nanjing University, Nanjing 210023, China; ztt@smail.nju.edu.cn

Abstract: With the abundance of conversations happening everywhere, dialogue summarization plays an increasingly important role in the real world. However, dialogues inevitably involve many personal pronouns, which hinder the performance of existing dialogue summarization models. This work proposes a framework named WHORU to inject external personal pronoun resolution (PPR) information into abstractive dialogue summarization models. A simple and effective PPR method for the dialogue domain is further proposed to reduce time and space consumption. Experiments demonstrated the superiority of the proposed methods. More importantly, WHORU achieves new SOTA results on SAMSum and AMI datasets.

Keywords: text summarization; abstractive dialogue summarization; personal pronoun resolution

Citation: Zhou, T. WHORU: Improving Abstractive Dialogue Summarization with Personal Pronoun Resolution. *Electronics* **2023**, *12*, 3091. https://doi.org/10.3390/electronics12143091

Academic Editors: Hsi-Min Chen and Shang-Pin Ma

Received: 5 July 2023
Revised: 14 July 2023
Accepted: 14 July 2023
Published: 16 July 2023

1. Introduction

At this very moment, conversations between humans and human/machine interactions are taking place everywhere. Thanks to automatic speech recognition systems and online communication systems, vast amounts of these dialogues can be recorded in text easily [1–3]. A succinct summarization helps readers grasp the key points with less time and effort. Hence, the necessity for dialogue summarization arises urgently. Dialogue summarization aims to succinctly compress and refine the content of conversations. My goal is to propose a new algorithmic framework that uncovers and incorporates external anaphora resolution information. The proposed framework aims to effectively assist models in clarifying referential relationships within dialogues and generating more coherent summaries.

Major text summarization works have focused on single-speaker documents like news publications [4]. Compared to single-speaker documents, people prefer using more personal pronouns to refer to recent characters in conversations. There is an average of 0.08 personal pronouns per token in the widely used dialogue summarization dataset SAMSum [1], compared to 0.03 personal pronouns per token in the news documents dataset CNN/DailyMail [4]. However, existing advanced dialogue summarization models fail to understand these personal pronouns. These models often generate summaries that associate one's actions with a wrong person. An example is shown in Table 1.

Table 1. Comparison of summaries for a test sample in SAMSum. For clear presentation, it omits redundant text and marks correct/wrong references with blue/red color.

Contents	Text
Conversation	Hank: Yeah, yeah, we'll see. I'll tell you about the tests when I bring Oscar and Roger back.
Human Annotator	Hank will bring his son and Don's son as well. Don is glad.
BART	Hank will take his kid. Don will bring Oscar and Roger back.
Multi-view BART [5]	Don will bring Oscar and Roger back.
Ours	Hank will bring Oscar and Roger.

This work presents an example from the test set of the SAMSum dataset, which consists of the original dialogue, supervised summary labels, and the decoding results from the latest state-of-the-art (SOTA) model. Upon reading the original dialogue, it is evident that Hank will be the one taking the child back. However, during the decoding process, the model generates a summary where Don is portrayed as the individual bringing the child. This error directly stems from the original dialogue, which does not explicitly specify who will bring the child and instead uses the pronoun 'I'. In the model's encoding and inference process, it is crucial for it to fully comprehend and accurately infer the referent of 'I' in order to generate a summary in the third person correctly.

This problem directly contributes to the challenges of employing the existing state-of-the-art (SOTA) model, Multi-View BART, in practical applications. Despite the fluency of its generated results, there is an inconsistency between the decoding outcome and the factual information present in the source text. Even a SOTA dialogue summarization model (Multi-view BART [5]) makes 20 referral errors in 100 sampled test conversations, which significantly harm the quality of summarization measured by ROUGE [6] scores (see Section 5.3).

How to help models understand this frequently occurring key information in dialogue data has become an important research question. Therefore, this work aims to propose a new algorithmic framework to uncover and inject external coreference resolution information. The framework can effectively assist models in understanding the referential relationships in the dialogue and generating more consistent summaries.

Dialogue summaries have different data formats and characteristics compared to other forms of text summaries. Dialogues vary in their format and content, encompassing various types such as everyday conversations, meetings, customer support Q&A, doctor–patient dialogues, and more. Unlike fluent long texts, dialogues consist of discrete utterances in multiple rounds, and the coherence of the context and consistency of the topic cannot be guaranteed. Additionally, dialogues encompass different stages, frequent instances of complex coreference phenomena, and the utilization of domain-specific terminology, all of which present substantial challenges for dialogue summarization.

Existing document-focused summarization models often face difficulties in handling such issues. Therefore, there is a need for efficient methods that can address these problems and generate high-quality dialogue summaries. Existing solutions often rely on scarce dialogue data, and researchers attempt to enhance the existing summarization models using human priors or external resources. For example, TGDGA [7] considers the presence of multiple topics in dialogues and models topic transitions explicitly to guide summary generation. In real-world conference dialogues, where speakers are relatively fixed and each speaker has distinct characteristics, HMNet [3] utilizes these features to improve the generation quality. Furthermore, several works [3,7] have focused on modeling the various stages and structures found in dialogues. However, so far, no work has attempted to address the complex coreference relationships and the challenges posed by intricate referring expressions in dialogues, which persistently impact the quality of summary generation.

This paper proposes a framework named WHORU (the abbreviation of "Who are you") to inject external personal pronoun resolution information into abstractive dialogue summarization models.

Specifically, WHORU appends the references after their corresponding personal pronouns and distinguishes personal pronouns, references, other words with additional tag embeddings. Preliminary experiments have shown that the SOTA personal pronoun resolution method SpanBERT [8] is time consuming and space consuming.

This work would like to emphasize that there is a strong recency effect observed when humans use personal pronouns. This suggests that the nearest candidate is the most likely reference [9]. Hence, an additional method called DialoguePPR (short for Dialogue Personal Pronoun Resolution) is proposed. It is a rule-based approach specifically designed to address personal pronoun resolution in dialogues. DialoguePPR efficiently performs a greedy search to identify the closest person or speaker.

The desirable features of the proposed methods can be summarized as follows:

- **Simple:** WHORU is easy to implement since it does not need to modify existing models except adding tag embeddings. Rule-based DialoguePPR either requires any training procedure or personal pronoun resolution datasets.
- **Efficient:** WHORU appends a few words to the original text which slightly increases training and inference time. DialoguePPR is model-free and only needs to run a greedy search algorithm which has linear time complexity.
- **Generalizable:** WHORU can be applied to most of the existing advanced dialogue summarization models, including these built on pretrained models like BERT, BART [10], etc. WHORU helps different models incorporate external personal pronoun resolution information (Section 5.3). Moreover, DialoguePPR is accurate on two widely used dialogue summarization dataset from different areas (Section 5.4).
- **Effective:** Empirical results demonstrate that the performance of strong models is significantly improved on ROUGE evaluation by the proposed methods. More importantly, WHORU achieves new SOTA results on SAMSum and AMI datasets.

2. Background

2.1. Formalization of Problem

The dialogue summarization problem can be formalized in either an extractive or an abstractive way. This paper focuses on abstractive dialogue summarization since it allows for more flexible generation of summaries in the third-person point of view and has demonstrated greater effectiveness [3].

The dialogue summarization problem is formalized as follows. An input space $\mathcal{X} = X_1, X_2, \ldots, X_l$ is considered, where l represents the number of conversations and X_i represents the I-th conversation. There are multiple turns x in each conversation. Each turn is the utterance u consisting of several sentences spoken by a specific speaker s. The I-th conversation is represented as $X_i = (s_1, u_1), (s_2, u_2), \ldots, (s_{L_i}, u_{L_i})$, where L_i denotes the number of turns. Here, s_j represents the speaker of the j-th utterance, and u_j represents the tokenized form of the j-th utterance, denoted as $u_j = v_1, v_2, \ldots, v_{N_j}$. Additionally, there is a golden summaries space denoted as $\mathcal{Y} = Y_1, Y_2, \ldots, Y_l$, where each Y_i is paired with X_i. Both X_i and Y_i are sequences of tokens. Without loss of generality, the conversation index subscript is ignored. To sum up, the standard dialogue summarization problem aims to generate a summary $Y = \{y_1, y_2, \ldots, y_n\}$ based on the input conversation $X = \{(s_1, u_1), (s_2, u_2), \ldots, (s_m, u_m)\}$.

2.2. Models

Given a training pair (X, Y), abstractive dialogue summarization aims to minimize the following loss function:

$$\mathcal{L} = -\log p(Y|X) \tag{1}$$

where the conditional probability is usually modeled by an encoder-decoder architecture, such as LSTM or Transformer.

To further address the common redundancy problem in the sequence-to-sequence modeling process, researchers propose Pointer Generator Network [11] (PGN), which incorporates a copy mechanism into the encoder-decoder models. Meanwhile a^t is attention distribution over source tokens in time step t. Coverage mechanism requires a coverage vector c^t, which is the sum of attention distributions over all previous decoder time steps to track where the model's attention is focused. A coverage loss is jointly optimized to track redundant and the loss function is extended to

$$\mathcal{L}_{PGN} = \mathcal{L} + \lambda \sum_{t=1}^{|Y|} \sum_{I=1}^{|X|} \min\left(a_i^t, c_i^t\right) \tag{2}$$

Recently, large-scale language pretraining has been demonstrated to be beneficial for downstream tasks. Therefore, BART, a conditional language model pretrained on the massive amount of unlabeled data, is proposed to solve dialogue summarization and achieves SOTA performance. These models have achieved huge success in dialogue summarization [5]. However, as discussed above, the approach is still troubled by a large number of personal pronouns. The next section introduces the embedding of PPR information into PGN and BART through simple steps.

3. WHORU

In conversation, people often use personal pronouns as a simple substitution for the proper name of a person for convenience, which can avoid unnecessary repetition and make the conversation more succinct. However, the evidence demonstrated that existing models are often confused with personal pronouns. As a result, these models often generate summaries with referral errors and obtain ordinary ROUGE scores (see Section 5.3).

To alleviate this problem, a framework called WHORU is proposed, which explicitly considers the process of personal pronoun resolution. As shown in Figure 1, my framework includes two steps. First, WHORU injects the references of personal pronouns into original conversations. Second, WHORU uses additional tag embeddings to help the model distinguish the role of personal pronouns and their references.

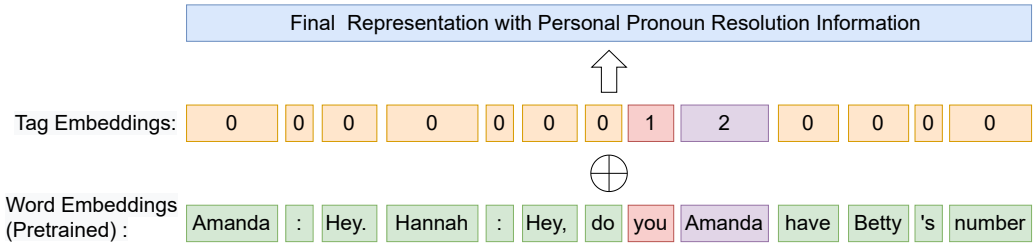


Figure 1. The illustration of how WHORU works on BART. In this example, "you" is identified as the target personal pronoun and "Amanda" is the reference of "you" labelled by the personal pronoun resolution method. Position embeddings are not shown.

In Figure 1, green is used to indicate the original words in the text, red is used to indicate the identified anaphoric words, and purple is used to indicate the additional injected anaphora resolution information. Assuming the anaphora resolution task has been completed and the referents for each personal pronoun have been obtained, in this example, the target pronoun 'you' is identified, and an anaphora resolution method is employed to determine that it refers to 'Amanda'. For simplicity and ease of understanding, the position embedding is not depicted. The WHORU framework as a whole is lightweight and concise, allowing it to adapt flexibly to different backbone models. The next section introduces the embedding of PPR information into PGN and BART through simple steps.

3.1. Inject Personal Pronoun Resolution

3.1.1. Resolve Personal Pronouns in the Conversation

Personal pronouns are widely used in human conversations. Since there is only a limited number of personal pronouns in a specific language, they can be easily extracted by matching each word in utterances with a pre-defined personal pronoun list (as shown in Table 2). Formally, given an input conversation $X = (s_1, u_1), (s_2, u_2), \ldots, (s_m, u_m)$, a list of personal pronouns $P = p_1, p_2, \ldots, p_t$ can be obtained.

For each recognized personal pronoun p_i, the corresponding reference r_i is resolved using personal pronoun resolution methods. This paper considers two PPR methods: (1) SpanBERT [8], which is pretrained on a large-scale unlabeled corpus and then finetuned on a vanilla coreference resolution dataset, and (2) DialoguePPR, a simple but effective rule-based method specifically designed for PPR in dialogues, which will be described in detail in Section 4.

Table 2. Table of Personal Pronouns References.

Personal Pronouns	Words
the first-person pronoun	**I, Me, My, Myself, Ourselves**
the second-person pronoun	**You, Your, Yours, Yourself**
the third-person pronoun	**He, She, His, Her, Himself, Herself**

3.1.2. Inject PPR Information into Conversations

In the domain of personal pronoun resolution, there exist numerous potential schemes. However, it is crucial to seek a method that not only effectively addresses the task but also remains orthogonal to existing approaches while making minimal modifications to the existing models.

With the objective in mind, a direct modification approach is proposed, involving the injection of the obtained personal pronoun resolution (PPR) information into the input conversation X. Specifically, it is suggested to append the reference r after its corresponding personal pronoun p within the dialogue.

The beauty of my approach lies in its compatibility with Encoder-Decoder models, which share a common need to encode the source text and generate the target based on it. By incorporating the PPR information into the source text, the aim is to ensure seamless integration of this information into any Encoder-Decoder model, eliminating the need for additional adjustments in model encoding. In essence, a method has been devised that achieves the goal of injecting the personal pronoun resolution information by directly modifying the input dialogue X.

Through this innovative approach, the performance of personal pronoun resolution can be enhanced without disrupting the underlying structure and functioning of the existing models. By strategically incorporating the previously extracted PPR information into the input dialogue, the power of Encoder-Decoder architectures can be leveraged to improve the resolution accuracy and coherence of personal pronouns.

Formally, consider one turn of the conversation X as $\{s, v_1, v_2, v_3, v_4\}$, where v_2 is a personal pronoun p, and r is its corresponding reference. The r is appended right after the personal pronoun v_2, resulting in the modified sequence: $s, v_1, v_2(p), r, v_3, v_4$. In this way, these two words will have close position embedding in Transformer model or time step in LSTM model. Thus the model could learn the relation between them.

3.2. Additional Tag Embeddings

Although, the proposed appending strategy could help the model associate the personal pronoun with its reference. It also introduces noise to the fluent human language. To assist the model in distinguishing between personal pronouns, references, and other words, different labels are assigned. Tag 1 and 2 are used to denote personal pronouns and references, respectively, while 0 is used for other tokens. For the given X, the corresponding labels are as follows: $0, 0, 1, 2, 0, 0$.

To enhance the effectiveness of anaphora resolution, an additional tag embeddings layer is introduced as a crucial step to embed the tag sequence. By incorporating tag embeddings into the input embeddings, a comprehensive and enriched representation of the input is achieved, which is subsequently fed into the encoder for further processing.

To maintain compatibility with pretrained parameters and avoid interference, a learnable embedding layer is chosen instead of fixed embeddings. This enables adaptive adjustments to the tag embeddings during training without impacting the existing pretrained parameters. As a result, the only modification made to the model is the inclusion of the tag embeddings layer.

An intriguing aspect to highlight is the remarkably small parameter size of the Tag Embedding Layer. This indicates that anaphora resolution information can be injected and identified at a significantly low cost. This advantageous characteristic aligns with the

need for reduced training data and computational resources, making my approach more practical and efficient.

By seamlessly incorporating the tag embeddings layer into the model architecture, the representation of the input sequence is enriched, empowering the model to effectively capture and utilize anaphora resolution information. This augmentation enables improved performance in anaphora resolution tasks, without compromising the integrity of the pretrained parameters or incurring substantial additional computational overhead.

4. Personal Pronoun Resolution for Dialogues

The existing SOTA PPR method SpanBERT typically requires a large amount of monolingual data to pretrain, which may not be feasible for some low-resource languages. Furthermore, the computation and memory costs are also not beneficial to build a Green AI. To this end, a simple and effective rule-based personal pronoun resolution method, named DialoguePPR, is proposed.

A major step of existing rule-based personal pronoun resolution is identifying speakers [12]. However, due to the inherent nature of conversation data, the extraction of speakers becomes straightforward once each turn is separated.

For the first personal pronoun, it is straightforward that the reference of it in an utterance is the speaker himself/herself.

Regarding the second personal pronoun, it is understood that the reference should be one of the speakers. Based on the strong recency effect, it is believed that the closest speaker is the most likely reference [9]. Personal pronouns typically appear after their reference. Consequently, a backward search is performed to find the nearest speaker from the current utterance. If the algorithm does not identify any candidates in the backward search, it will proceed with a forward search to determine the nearest speaker as the reference.

Resolution of the third personal pronoun is a bit more complex: their references could be one of the persons mentioned in the whole conversation. Therefore, the first step is to utilize the name entity recognition tool in NLTK [13] to extract a list of person entities denoted as $E = e_1, e_2, \ldots, e_r$. Since name entity recognition problem has been well studied, the extraction performance can be guaranteed. Guided by the recency effect, a search procedure similar to that used for the second personal pronoun is employed to locate the reference.

The above procedure is summarized in Algorithm 1, which represents a dedicated, straightforward, and efficient referential resolution strategy specifically designed for dialogue datasets.

Figure 2 is a good example. By leveraging explicit features in the dialogue and linguistic prior knowledge, DialoguePPR efficiently and accurately performs anaphora resolution in the dialogue and extracts referential information. Subsequently, this information can be injected into the model using the WHORU framework introduced earlier.

Note that the greedy search have a linear time complexity $O(T)$, where T is the number of tokens in X. It is natural to doubt that whether these rules work widely on different dialogue summarization datasets. In Section 5.4, a comprehensive analysis of the efficiency and generalization of DialoguePPR will be conducted.

Algorithm 1 DialoguePPR

Input: The conversation X**, person entity list** E**.**
for (s_i, u_i) in X **do**
  for v_j in u_i **do**
    if v_j is first personal pronoun **then**
      Append (v_j, s_i) to P.
    end if
    if v_j is second personal pronoun **then**
      Greedy search the closest speaker s_t different from s_i.
      Append (v_j, s_t) to P.
    end if
    if v_j is third personal pronoun **then**
      Greedy search the closest person e_t in E.
      Append (v_j, e_t) to P.
    end if
  end for
end for
Output: The resolution result P**.**

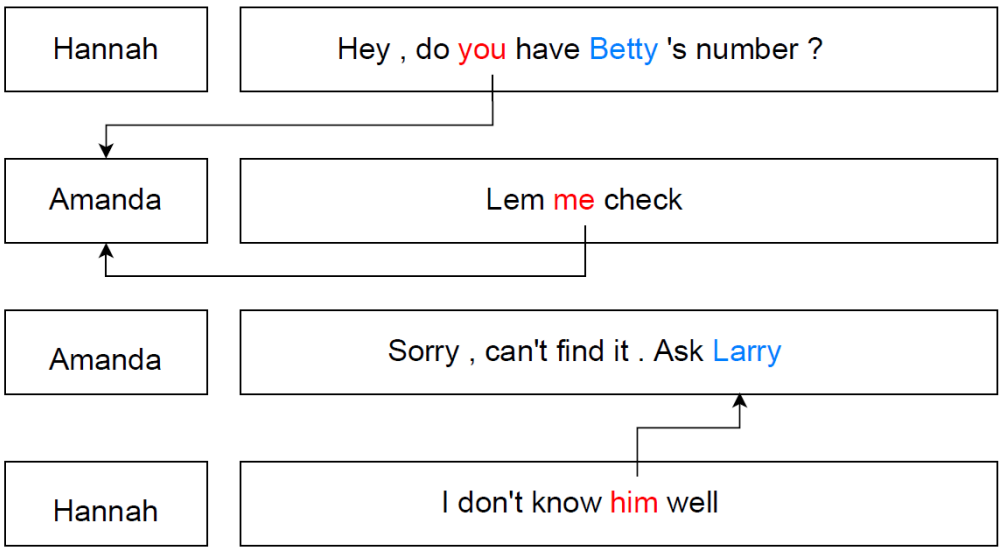


Figure 2. Example of DialoguePPR; It involves Hannah needing Betty's phone number and asking Amanda. Amanda does not have it and suggests that Hannah ask Larry, whom Hannah is not familiar with. This dialogue segment involves complex and frequent personal pronoun references, making it a typical sample in dialogue summarization. For the first-person pronoun, 'Lem me check,' DialoguePPR correctly resolves it as the speaker. For the second-person pronoun, 'do you have,' since no referent is found ahead, it retrieves the correct referent, Amanda, by searching backward. For complex third-person pronouns, DialoguePPR first extracts named entities from the dialogue content, which in this example are Betty and Larry. Following the principle of proximity, Larry is used as the referent, which is clearly correct. The working process is indicated by arrows, showing the final inference results of DialoguePPR, with the arrows pointing from the personal pronouns to the referents.

5. Experiments

5.1. The Settings of the Experiments

Datasets. Experiments are conducted on two widely used dialogue summarization datasets: SAMSum [1] and AMI [14]. SAMSum contains natural messenger-like conversations in real life, whose styles are diversified. On the other hand, AMI obtains meeting

transcripts from automatic speech recognition, which has a more formal style. Each conversation in SAMSum and AMI has a summary annotated by human experts. Table 3 summarizes the statistics of SAMSum and AMI.

Table 3. The statistics of different datasets.

Dataset	# Conversations			# Personal Pronouns		
	Train	Dev	Test	First	Second	Third
SAMSum	14,732	818	819	88,476	56,985	11,657
AMI	100	17	20	11,969	14,103	355

Baseline Models. For comparison purposes, diverse and high-performance baseline models were selected:

- LONGEST-3: LONGEST-3 selects the three longest turns in a multi-turn dialogue. Generally, longer dialogues tend to contain key information more concentratedly, while shorter dialogues have relatively dispersed information. Therefore, similar to Lead3, LONGEST-3 is considered as a baseline in dialogue summarization.
- PGN: PGN is based on the LSTM architecture and introduces Copy to effectively copy words from the source text, alleviating the out-of-vocabulary (OOV) problem. PGN also introduces the Coverage mechanism to track attention distribution effectively and solve the problem of redundant generation.
- BART: BART is a large-scale pre-trained model based on the Transformer architecture. It is pre-trained on a large amount of data using the denoising task and achieves state-of-the-art (SOTA) performance on multiple tasks. Similarly, it shows significant effectiveness in dialogue summarization and serves as a strong baseline.
- TGDGA [7]: TGDGA considers multiple topics in dialogues, and the progress of dialogues often involves topic transitions. Therefore, the model extracts topic words and uses a graph neural network to encode the relationships between topics. Inspired by PGN, TGDGA designs a Topic-word Guided Decoder and achieves good results.
- ClusterRank [15]: ClusterRank extends the TextRank algorithm for conference datasets and constructs a graph structure using clustering to extract relevant parts and remove redundancy. It performs well on the AMI dataset.
- HMNet [3]: HMNet proposes a hierarchical neural network and conducts pre-training on a news summarization dataset. It models speaker features based on the characteristics of fixed speakers in conference datasets and achieves SOTA performance on the AMI dataset.

In this work, the following models have been implemented. One is PGN (Pointer Generator Network), which is based on LSTM and does not have pretraining knowledge. The other one is BART, which is trained based on large-scale corpora and the Denoise Pretrain Task, with pretraining knowledge. Testing on both pretrained and non-pretrained models can better demonstrate that my framework can generalize to different models and that the mined referential information is helpful for both small models and pretrained large models.

(1) **PGN**. One follows the default configuration (https://github.com/abisee/pointer-generator) except the minimum length for length normalization. To fit the characteristics of SAMSum dataset, minimum length is adjusted from 35 to 15. The AMI dataset is relatively long, so I have set the minimum decoding length and the maximum decoding length to 50 and 1000, respectively.

(2) **BART**. The fairseq sequence modeling toolkit [16] is used to reproduce the BART [10] model. All experiments relative to BART use pretrained bart.large model in fairseq (https://github.com/pytorch/fairseq/tree/master/examples/bart). The learning rate is set to 3×10^{-5}. For fine-tuning on SAMSum, warm up step is set to 300. At the test stage, beam size is 4, minimum decoded length and maximum length are 5 and 100 respectively for SAMSum dataset. The AMI dataset has longer conversations and summaries. So, the

minimum and maximum length were changed to 50 and 1000, respectively. Furthermore, due to the smaller size of the AMI dataset, the warm-up and training steps were decreased to 20 and 200, respectively.

In the experiments, the best model was chosen based on the loss on the development set. Similar to [5], three runs of each model were conducted with different random seeds, and the average results were reported. A range of baseline systems from previous literature were chosen for comparison.

Metric. When evaluating an automatic summarization system, the most important aspect is to assess the quality of the generated summaries. The most direct and accurate method is manual evaluation, which involves inviting linguistic experts to read the original texts and the generated summaries and assign scores to the generated summaries. This method ensures high quality evaluation and aligns with human cognition and needs. However, it is costly, time-consuming, and slow.

As mentioned earlier, for each input sample, there is a reference summary available for the model to learn and test. Automatic evaluation methods that rely on the input document and reference summaries can also be used for evaluation. Automatic evaluation systems often employ algorithms that strike a balance between efficiency and cost-effectiveness. There is a wealth of research in this area, such as ROUGE [6] and BERScore [17]. Among them, the most widely used automatic evaluation method is ROUGE, which measures the similarity between the generated summaries and the reference summaries using overlap. ROUGE-N represents the overlap at the n-gram level, mainly focusing on estimating whether the generated summaries cover the expressed information in the reference summaries. Researchers often use ROUGE-LCS (referred to as ROUGE-L) to measure the sentence-level information and fluency of the generated summaries. The computation methods for ROUGE-N and ROUGE-L are as follows:

$$\text{ROUGE-N} = \frac{\sum_{s \in \{\text{ Reference }\}} \sum_{\text{gram}_n \in S} \text{Count}_{\text{match}}(\text{gram}_n)}{\sum_{s \in \{\text{ Reference }\}} \sum_{\text{gram}_n \in S} \text{Count}(\text{gram}_n)} \tag{3}$$

$$\text{ROUGE-LCS}_R = \frac{LCS(X,Y)}{m} \tag{4}$$

$$\text{ROUGE-LCS}_R = \frac{LCS(X,Y)}{n} \tag{5}$$

$$\text{ROUGE-LCS}_F = \frac{(1+\beta^2) R_{lcs} P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}} \tag{6}$$

Different models were evaluated on ROUGE-1, ROUGE-2, and ROUGE-L scores [6].

Following [5,10], a full Python implementation for ROUGE scores (https://github.com/pltrdy/rouge is used). Different implementations of ROUGE may yield slight differences in the results.

5.2. Main Results

The results of different models on SAMSum are summarized in Table 4. The extractive method LONGEST-3, which simply selects the longest three utterances as summary, achieves the lowest results. Since most utterances are verbose and repetitive, extractive methods have difficulty generating perfect summaries. By pretraining on a large-scale unlabeled corpus, Multi-view BART and BART surpasses TGDGA by a large margin in terms of ROUGE scores.

By using the proposed WHORU framework to inject PPR information, the performance of the strong baseline BART can be further boosted. This demonstrates the effectiveness of explicit modeling PPR information when summarizing dialogues. Using the SpanBERT instead of DialoguePPR to resolve the reference information can be more effective, setting a new SOTA performance evaluated by ROUGE-2 and ROUGE-L.

Table 4. The result on the test set of SAMSum dataset.

Model	ROUGE-1	ROUGE-2	ROUGE-L
LONGEST-3 [18]	32.46	10.27	29.92
TGDGA [7]	43.11	19.15	40.49
Multi-view BART [5]	**49.30**	25.60	47.70
BART	47.94	24.40	46.74
BART + WHORU + DialoguePPR	48.86	25.28	47.44
BART + WHORU + SpanBERT	49.16	**26.27**	**48.11**

Table 5 shows the result on the test set of AMI dataset. The proposed framework still achieves highly competitive performance, increasing the ROUGE-1 score by **+1.36** and ROUGE-2 score by **+0.65**. HMNet [3] pretrains a hierarchical network on news documents and utilizes additional speaker role information, which may not be available on some datasets. Compared to HMNet without speaker role, my result shows much improvement (**+0.90 ROUGE-1 and +1.86 ROUGE-2**). Note that conversations in AMI are much longer (on average 4757 words per conversation). SpanBERT can no longer process the conversation due to the memory limit of GPU, while DialoguePPR can still work efficiently.

Table 5. The result on the test set of AMI dataset. * denotes that HMNet incorporate the speaker role information which is not provided in many other datasets. ** denotes that SpanBERT collapsed on this long conversation dataset.

Model	ROUGE-1	ROUGE-2	ROUGE-L
ClusterRank [15]	35.14	6.46	-
HMNet-speaker role [3]	47.80	17.20	-
HMNet * [3]	53.02 *	18.57 *	-
BART	47.34	18.41	**20.34**
BART + WHORU + DialoguePPR	**48.70**	**19.06**	19.65
BART + WHORU + SpanBERT **	-	-	-

5.3. Can WHORU Inject Personal Pronoun Resolution?

Quantitative Analysis. As demonstrated in Table 6, although Multi-view BART achieves the state-of-the-art (SOTA) result on SAMSum, it only marginally outperforms the baseline on the personal pronoun problem. Meanwhile, WHORU successfully reduces referral errors by 45% compared to Multi-view BART. Furthermore, WHORU obtains significant improvement of ROUGE scores on test samples which are related to referral errors. This implies that the personal pronoun resolution information not only alleviates referral errors but also improves the whole quality of the generated summaries.

Case Study. The summaries generated by different models have been collected in Table 1 (The results of Multi-view BART were released by [5] in https://github.com/GT-SALT/Multi-View-Seq2Seq).

It can be seen from the table that it is Hank who brings Oscar and Roger back. However, Don is mistakenly recognized by the summaries generated by BART and Multi-view BART. This demonstrates that WHORU helps models understand "who are you" in conversation. Moreover, this also confirms that Multi-view BART is somewhat complemented with WHORU. The attention map of my model has also been visualized on another test sample from SAMSum in Figure 3a. The results show that when my model generates "him", it focuses on the appended reference "Ahmed" as expected. However, without WHORU, the baseline model will generate a wrong personal pronoun "her" in this position. Thus, WHORU does inject personal pronoun resolution successfully.

Table 6. Number of referral errors for different models, which are counted in 100 randomly sampled test conversations from SAMSum dataset. ROUGE scores are calculated on conversations which have at least one summary contain referral errors.

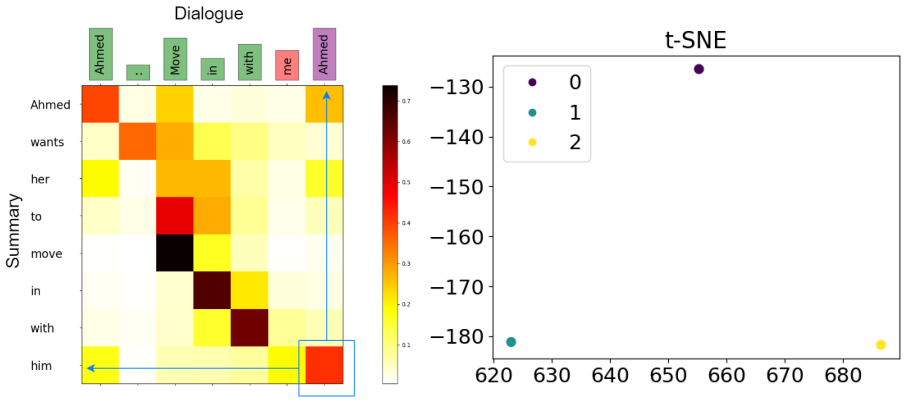
Model	# Referral Errors	ROUGE-1	ROUGE-2	ROUGE-L
BART	24	42.849	19.047	42.262
Multi-view BART [5]	20	44.795	19.229	42.429
BART+WHORU+DialoguePPR	11	48.180	22.638	46.425

WHORU has good generalization. To examine the generalization of WHORU, experiments for the PGN model on the SAMSum dataset are also conducted. These results are summarized in lines 2–3 of Table 7. Although PGN uses a more complex loss function, WHORU and DialoguePPR still help it achieve accuracy improvements up to **+1.31 ROUGE-1, +1.11 ROUGE-2** and **+1.48 ROUGE-L**.

Table 7. The results of the analysis on SAMSum.

Model	ROUGE-1	ROUGE-2	ROUGE-L
PGN	35.30	12.66	35.27
PGN + WHORU + DialoguePPR	36.61	13.77	36.75
BART	47.94	24.40	46.74
BART + WHORU + DialoguePPR (Full)	48.86	25.28	47.44
Full - Tag Embeddings	48.49	25.29	46.90
Full (Only First Person)	48.31	25.04	47.04
Full (Only Second Person)	48.17	24.43	46.57
Full (Only Third Person)	48.21	24.38	46.75
Full (Remove First Person)	48.39	24.74	47.48
Full (Remove Second Person)	48.42	24.92	46.96
Full (Remove Third Person)	48.89	25.02	46.91

Tag Embeddings does help. As I shown in the fifth row of Table 7, the ROUGE-1 score drops 0.37 points and the ROUGE-L score drops 0.57 points when tag embeddings are removed. The tag embeddings learned on the SAMSum dataset are further visualized in Figure 3b. It is clearly shown that different tags are widely separated by tag embeddings, so that the model can distinguish the injected personal pronoun resolution information clearly.



(**a**) Attention Map. (**b**) Tag Embeddings.

Figure 3. (**a**) Visualization of attention map of my model. (**b**) Visualization of different tag embeddings by t-SNE.

5.4. Is DialoguePPR Successful in Dialogue Domain?

Main results have shown that DialoguePPR is helpful on downstream task dialogue summarization. The efficiency and generalization of DialoguePPR on the PPR task are still a subject of curiosity. Table 8 summarize the inference time and accuracy of different PPR methods on SAMSum and AMI datasets. The results show that DialoguePPR achieves a resolution speed 11.5× faster than SpanBERT for the entire SAMSum dataset. At the same time, DialoguePPR obtains similar accuracy on first and second person. The accuracy on third person is still acceptable, considering there are many potential candidates in the context. The results on AMI show that DialoguePPR is well generalized on different dialogue summarization datasets.

The performance of BART + WHORU + DialoguePPR is further tested for different persons. As indicated in lines 6–11 of Table 7, the DialoguePPR resolution information for all three persons individually improves the baseline. When removing one of three person information, the ROUGE scores decreased. This confirms that DialoguePPR is accurate on a different person.

Table 8. The inference time (in seconds) and PPR accuracy of different methods on SAMSum and AMI datasets. 50 personal pronouns of different persons are randomly sampled for computing the accuracy. ** denotes that SpanBERT collapsed on this long conversation dataset.

Dataset	Method	Time	Accuracy		
			First	Second	Third
SAMSum	SpanBERT	10,184.74	100%	96%	93%
	DialoguePPR	883.57	100%	98%	64%
AMI	SpanBERT **	-	-	-	-
	DialoguePPR	469.28	100%	76%	67%

5.5. Improving Fact Consistency

Existing models often suffer from factual inconsistencies in the generated summaries (see Table 6). These factual inconsistencies greatly harm the quality of the generated summaries and subsequent practical applications. The series of experiments conducted earlier have demonstrated the significant improvement in summary quality achieved by the combination of WHORU and DialoguePPR. However, the aim is also to further enhance the summary quality from the perspective of fact consistency.

FactCC, a widely used model for fact consistency evaluation, is employed as the evaluation metric. Multiple models' generated results are used on the entire test set as the evaluation data. FactCC assigns a label to each summary, indicating whether it is consistent or inconsistent. The given labels are collected as the FactCC Score and also collect the test loss of the FactCC model, setting the label to 1. A higher FactCC Score represents better fact consistency in the generated results, while a smaller FactCC Loss indicates better fact consistency.

It is evident that my combination of BART + WHORU + DialoguePPR effortlessly achieves the best performance, obtaining the highest scores in both FactCC Score and FactCC Loss. The FactCC Score of my model is 2.57 higher than the previous state-of-the-art Multi-View BART and 3.17 higher than the baseline Vanilla BART. This demonstrates that my strategy effectively improves the fact consistency in generating summaries, enhancing both the quality and usability of the summaries. Table 9 shows the evaluation using the FactCC model for fact consistency.

Table 9. Evaluation using the fact consistency model FactCC.

Model	FactCC Score ↑	FactCC Loss ↓
Vanilla BART	91.58	0.337
Multi-view BART	92.18	0.340
BART+WHORU+DialoguePPR	**94.75**	**0.215**

6. Related Work

Pre-trained Language Models. In the field of Natural Language Understanding (NLU), BERT [19] is primarily based on the Masked Language Task. By masking a portion of the corpus, the model is trained to predict the masked part based on the context, thus enhancing the model's ability to encode context. Since the Masked Language Task is an unsupervised process, it is possible to construct large-scale training data from common corpora. BERT trained on large corpora can effectively incorporate contextual information and obtain better contextual representations. Pretraining has greatly benefited the field of extractive summarization, as seen in BERTSUM [20] and MATCHSUM [21], which achieved state-of-the-art results using BERT.

On the other hand, there are also numerous pretrained models based on the Seq2Seq framework, such as BART [10], PEGASUS [22], and T5 [23]. The general idea is to construct pretrained corpora in an unsupervised manner by modifying or extracting from the source text, and then train Seq2Seq models based on Transformers. BART initially employs various noise-adding strategies to disrupt the source text and then inputs the noisy text to generate the original text. BART effectively acquires the ability to generate text using context and exhibits good robustness to input text, yielding excellent results in natural language generation tasks such as paraphrasing and summarization. Reinforcement learning struggles without clear rewards. Intrinsic motivation methods reward novel states, but have limited benefits in large environments. ELLM [24] uses text corpora to shape exploration. It prompts a language model to suggest goals for the agent's current state. ELLM guides agents towards meaningful behaviors without human involvement. ELLM is evaluated in Crafter and Housekeep, showing improved coverage of common-sense behaviors and performance on downstream tasks.

These pretrained models provide additional domain-specific and grammatical knowledge. They also reduce the dependence on subsequent training data and training time, thereby greatly advancing the development of natural language processing. A significant amount of existing work is conducted based on pretrained models.

Abstractive Document Summarization. Neural models for abstractive document summarization have been widely studied [25,26] since Rush et al. [27] first employed the encoder and decoder framework. A series of improvements have been proposed based on different advanced techniques. Except for the copy mechanism (PGN) and pretrained language model (BART) mentioned above, reinforcement learning [28] and graph neural networks (GNNs) [29] have also been extended. Some studies using discourse relation [29] and coreference resolution [30] are related to my paper. Differently, my work is the first work utilizing personal pronoun resolution for dialogue summarization. As demonstrated above, the personal pronoun problem in dialogue summarization is significantly different from that in the signal speaker document. While most of these works rely on the complex graph structure, my WHORU is quite simple and easy to implement on existing models.

Abstractive Dialogue Summarization. Because of the scarcity of dialogue summarization resources, most existing works improve abstractive dialogue summarization with human prior knowledge or external information. For example, topic (TGDGA) [7] and speaker role (HMNet) [3] information have been widely used to improve abstractive dialogue summarization. Conversational structure prior knowledge is also be considered in [3,5,7]. In this paper, successful utilization of external personal pronoun resolution information has led to achieving state-of-the-art (SOTA) results. DialoguePPR, which is based on conversational prior knowledge, avoids the requirement for external PPR resources.

Coreference resolution methods. Coreference resolution is a popular direction in natural language processing, and the current mainstream approach is based on end-to-end methods, such as Neural Coreference Resolution. This modeling method considers the input document (consisting of T tokens) as $\frac{T\times(T+1)}{2}$ spans and attempts to find the antecedent for each span. This work uses a bidirectional LSTM network to encode information within and outside the spans, while also incorporating an attention mechanism. Neural Coreference Resolution outperforms all previous models without the need for syntactic parsing and named entity recognition.

Subsequently, with the development of pre-training models, SpanBERT [8] emerged. As mentioned earlier, BERT utilizes the Masked Language Model task to train the model and achieves good results. However, BERT only masks one subword at a time, and the training objective focuses on obtaining token-level semantic representations, whereas end-to-end coreference resolution requires a good span representation. Therefore, researchers proposed SpanBERT, which introduces a better span masking scheme and a Span Boundary Objective (SBO) training objective. It has achieved state-of-the-art results in tasks related to spans, such as extractive question answering and coreference resolution. This paper will also use and compare with SpanBERT.

On the other hand, existing coreference resolution methods (including SpanBERT) are trained on the OntoNotes 5.0 dataset. The OntoNotes 5.0 dataset includes various types of data such as news, telephone conversations, and broadcasts, and it covers multiple languages, including English, Chinese, and Arabic. In terms of data distribution, English dialogue-type data is relatively limited, and there is inconsistency with some dialogue datasets (such as conference dialogue) in terms of content. In terms of data format, SpanBERT accepts a maximum of 512 tokens of text, which limits the direct application of existing coreference resolution work to dialogue summarization in subsequent studies.

Generative Text Summarization. Although extractive methods can ensure a certain level of grammatical and syntactic correctness, as well as the fluency of summaries, they are prone to content selection errors, lack flexibility, and exhibit poor coherence between sentences. Moreover, the limited choice of sentences and words solely from the source text greatly restricts the quality ceiling of summary generation techniques. With the emergence of neural networks and sequence-to-sequence (Seq2Seq) models, it became possible to flexibly select words from a large vocabulary to generate summaries. However, Seq2Seq methods encounter some issues, such as low-quality generated summaries with grammar errors and the tendency to produce redundant words. Additionally, due to the limitation of vocabulary size, out-of-vocabulary (OOV) problems may arise. Consequently, related works have proposed solutions to address these problems. For instance, the work of Paulus et al. introduced the Pointer Generator Network [28], which incorporates Copy and Coverage mechanisms based on attention mechanisms in Seq2Seq, effectively alleviating the aforementioned issues. In addition to simple Seq2Seq models, reinforcement learning [28] and Graph Neural Networks (GNNs) [29] have also been gradually extended to this field and have achieved good results. Some researchers have utilized discrete relations [29], as well as anaphora resolution methods [30], which are somewhat related to the approach used in this paper. However, what distinguishes my work is that this work is the first to employ person-referencing resolution in dialogue summarization. Generating text summaries using generative techniques involves an autoregressive sequence generation process. At each time step, the decoding space encompasses the entire vocabulary, resulting in an excessively large search space during longer decoding steps. Thus, a decoding search strategy is needed. The two most popular methods are Greedy Search and Beam Search, which will be discussed further in the following sections.

- Greedy Search: In the autoregressive process, when the model needs to generate a sequence of length N, it iterates N times, each time providing a probability distribution for the next token based on the generated portions of the source and target. Greedy Search selects the token with the highest probability from the distribution as the result for the current time step. Greedy Search always greedily chooses the token with the

highest probability, leading to many candidate tokens being pruned in subsequent decoding steps and causing the optimal solution to be discarded prematurely.

- Beam Search: Beam Search [31,32] is an improvement over Greedy Search and can be seen as an enlarged search space. Beam Search is a heuristic graph search algorithm. Unlike the greedy search strategy, Beam Search constructs a search tree for each layer of the tree using a breadth-first strategy. The nodes are sorted according to a certain policy, and only a predetermined number of nodes are kept. Only these selected nodes are expanded in the next level, while other nodes are pruned for optimization. The 'number of nodes' here refers to the hyperparameter *BeamSize*, which effectively saves space and time, considers more possible optimal solutions, and improves the quality of search results.

7. Conclusions

This paper conducts in-depth research on the issue of referential errors in dialogue summarization. The research can be divided into the following aspects. Initially, problems encountered in current dialogue summarization and existing solutions are analyzed. Following the summary and analysis of existing methods, a significant issue is identified: the challenge of referential resolution in dialogue summarization, particularly the frequent occurrence of personal pronoun references that perplex models and result in erroneous and significantly degraded summaries. Consequently, the WHORU framework is introduced to inject additional referential resolution information into existing models, aiding in comprehending complex referential problems and enhancing the quality of generated summaries. Moreover, given the high cost and computational overhead associated with current general referential resolution strategies, as well as their contradiction with the concept of environmentally friendly AI, their application in lengthy dialogue data becomes challenging. To address this, a heuristic referential resolution strategy specifically tailored for dialogue summarization is proposed, exhibiting excellent adaptability and rapid computation speed. It achieves 11 times the inference speed of SpanBERT while maintaining considerable prediction accuracy. Extensive experiments in this paper demonstrate that WHORU achieves significant improvements over multiple baseline models, reaching a new state-of-the-art (SOTA) level. The experimental analysis also thoroughly validates the effectiveness and generalizability of WHORU and DialoguePPR, improving the quality of generated summaries and reducing referential errors, thus accomplishing my intended research objectives.

In future work, it would be worthwhile to explore the decoupling of referential information from the original input and the design of mechanisms for their interaction. For the encoding of injected referential information, new methods such as graph neural networks can be explored, which may have better performance on such structured data.

Regarding DialoguePPR, although extensive experiments have demonstrated its generalizability on diverse datasets, in more complex scenarios, the third-person pronoun resolution strategy may be overly simplistic, leading to the injection of a large amount of erroneous noise into the model. Therefore, future work can focus on improving the performance of SpanBERT or existing coreference resolution algorithms in long dialogue contexts to compensate for the limitations of the DialoguePPR algorithm.

Moreover, it is worth considering the integration of my work with the latest advancements in the field, such as Efficient Tuning techniques exemplified by PromptTuning. While the WHORU framework proposed in this paper introduces only a minimal number of parameters, rendering it well-suited for data-scarce tasks like dialogue summarization, it still necessitates fine-tuning the entire model during training. By combining it with efficient training strategies like PromptTuning to fully harness the capabilities of pre-trained language models, this work can be applied in broader scenarios.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Gliwa, B.; Mochol, I.; Biesek, M.; Wawer, A. SAMSum Corpus: A Human-annotated Dialogue Dataset for Abstractive Summarization. In Proceedings of the 2nd Workshop on New Frontiers in Summarization, Hong Kong, China, 20 September 2019; p. 70.
2. Kester, G.H. *Conversation Pieces: Community and Communication in Modern Art*; Univ of California Press: Berkeley, CA, USA, 2004.
3. Zhu, C.; Xu, R.; Zeng, M.; Huang, X. A hierarchical network for abstractive meeting summarization with cross-domain pretraining. *arXiv* **2020**, arXiv:2004.02016.
4. Nallapati, R.; Zhou, B.; dos Santos, C.; Gulçehre, Ç.; Xiang, B. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. In Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, Berlin, Germany, 11–12 August 2016; pp. 280–290.
5. Chen, J.; Yang, D. Multi-View Sequence-to-Sequence Models with Conversational Structure for Abstractive Dialogue Summarization. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online Event, 16–20 November 2020; pp. 4106–4118.
6. Lin, C.Y.; Och, F.J. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04), Barcelona, Spain, 21–26 July 2004; pp. 605–612.
7. Zhao, L.; Xu, W.; Guo, J. Improving Abstractive Dialogue Summarization with Graph Structures and Topic Words. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain, 8–13 December 2020; pp. 437–449.
8. Joshi, M.; Chen, D.; Liu, Y.; Weld, D.S.; Zettlemoyer, L.; Levy, O. Spanbert: Improving pre-training by representing and predicting spans. *Trans. Assoc. Comput. Linguist.* **2020**, *8*, 64–77. [CrossRef]
9. Anderson, J.R. *Cognitive Psychology and Its Implications*; Macmillan: New York, NY, USA, 2005.
10. Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; Zettlemoyer, L. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv* **2019**, arXiv:1910.13461.
11. See, A.; Liu, P.J.; Manning, C.D. Get to the point: Summarization with pointer-generator networks. *arXiv* **2017**, arXiv:1704.04368.
12. Lee, H.; Peirsman, Y.; Chang, A.; Chambers, N.; Surdeanu, M.; Jurafsky, D. Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. In Proceedings of the 15th Conference on Computational Natural Language Learning: Shared Task. Association for Computational Linguistics, Portland, OR, USA, 23–24 June 2011; pp. 28–34.
13. Loper, E.; Bird, S. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*; Association for Computational Linguistics: Philadelphia, PA, USA, 2002.
14. McCowan, I.; Carletta, J.; Kraaij, W.; Ashby, S.; Bourban, S.; Flynn, M.; Guillemot, M.; Hain, T.; Kadlec, J.; Karaiskos, V.; et al. The AMI meeting corpus. In Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research, Wageningen, The Netherlands, 30 August–2 September 2005; Volume 88, p. 100.
15. Garg, N.; Favre, B.; Reidhammer, K.; Hakkani-Tür, D. Clusterrank: A graph based method for meeting summarization. In Proceedings of the Tenth Annual Conference of the International Speech Communication Association, Brighton, UK, 6–10 September 2009.
16. Ott, M.; Edunov, S.; Baevski, A.; Fan, A.; Gross, S.; Ng, N.; Grangier, D.; Auli, M. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv* **2019**, arXiv:1904.01038.
17. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.Q.; Artzi, Y. Bertscore: Evaluating text generation with bert. *arXiv* **2019**, arXiv:1904.09675.
18. Feng, X.; Feng, X.; Qin, B.; Liu, T. Incorporating Commonsense Knowledge into Abstractive Dialogue Summarization via Heterogeneous Graph Networks. *arXiv* **2020**, arXiv:2010.10044.
19. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
20. Liu, Y.; Lapata, M. Text summarization with pretrained encoders. *arXiv* **2019**, arXiv:1908.08345.
21. Zhong, M.; Liu, P.; Chen, Y.; Wang, D.; Qiu, X.; Huang, X. Extractive summarization as text matching. *arXiv* **2020**, arXiv:2004.08795.
22. Zhang, J.; Zhao, Y.; Saleh, M.; Liu, P. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual Event, 13–18 July 2020; pp. 11328–11339.
23. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **2020**, *21*, 5485–5551.
24. Du, Y.; Watkins, O.; Wang, Z.; Colas, C.; Darrell, T.; Abbeel, P.; Gupta, A.; Andreas, J. Guiding Pretraining in Reinforcement Learning with Large Language Models. 2023. Available online: http://xxx.lanl.gov/abs/2302.06692 (accessed on 13 July 2023).
25. Peng, L.; Liu, Q.; Lv, L.; Deng, W.; Wang, C. An Abstractive Summarization Method Based on Global Gated Dual Encoder. In Proceedings of the CCF International Conference on Natural Language Processing and Chinese Computing, Zhengzhou, China, 14–18 October 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 355–365.

26. Liu, Y.; Fan, X.; Zhou, J.; He, C.; Liu, G. Learning to Consider Relevance and Redundancy Dynamically for Abstractive Multi-document Summarization. In Proceedings of the CCF International Conference on Natural Language Processing and Chinese Computing, Zhengzhou, China, 14–18 October 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 482–493.
27. Rush, A.M.; Chopra, S.; Weston, J. A Neural Attention Model for Abstractive Sentence Summarization. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 379–389.
28. Paulus, R.; Xiong, C.; Socher, R. A deep reinforced model for abstractive summarization. *arXiv* **2017**, arXiv:1705.04304.
29. Wei, W.; Wang, H.; Wang, Z. Abstractive Summarization via Discourse Relation and Graph Convolutional Networks. In Proceedings of the CCF International Conference on Natural Language Processing and Chinese Computing, Zhengzhou, China, 14–18 October 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 331–342.
30. Falke, T.; Meyer, C.M.; Gurevych, I. Concept-map-based multi-document summarization using concept coreference resolution and global importance optimization. In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Taipei, Taiwan, 27 November–1 December 2017; pp. 801–811.
31. Och, F.J.; Ney, H. The alignment template approach to statistical machine translation. *Comput. Linguist.* **2004**, *30*, 417–449. [CrossRef]
32. Zhang, A.; Lipton, Z.C.; Li, M.; Smola, A.J. Dive into deep learning. *arXiv* **2021**, arXiv:2106.11342.