An End-to-End Dialogue Summarization System for Sales Calls

Abedelkadir Asi[1], **Song Wang**[2], **Roy Eisenstadt**[1],**Dean Geckt**[1],
Yarin Kuper[1], **Yi Mao**[2], **Royi Ronen**[1]
[1]Microsoft Dynamics,
[2]Microsoft Azure
{abeasi,sonwang,reisenstadt,t-deangeckt,
yarinkuper,maoyi,royir}@microsoft.com

Abstract

Summarizing sales calls is a routine task performed manually by salespeople. We present a production system which combines generative models fine-tuned for customer-agent setting, with a human-in-the-loop user experience for an interactive summary curation process. We address challenging aspects of dialogue summarization task in a real-world setting including long input dialogues, content validation, lack of labeled data and quality evaluation. We show how GPT-3 can be leveraged as an offline data labeler to handle training data scarcity and accommodate privacy constraints in an industrial setting. Experiments show significant improvements by our models in tackling the summarization and content validation tasks on public datasets.

1 Introduction

An integral part of salespeople daily routine is summarizing sale calls. The summarization process aims to distill salient information from sales dialogues into succinct highlights, which are then leveraged by salespeople for productivity and coaching purposes. Manually curating a call summary is considered as one of the biggest time wasters for B2B sellers (Zhang et al., 2020). It distracts salespeople from nurturing the relationship with their next customer. Recently, this practice has become more demanding due to the emerging landscape of remote selling where virtual meetings become the new norm (Gavin et al., 2020).

Dialogue summarization induces a variety of unique challenges compared to summarization of documents such as news or scientific papers (Zhu and Penn, 2006). Information density is a key challenge in dialogue text; information is scattered over multiple utterances and participants, leading to frequent coreferences and topic alternations. Spoken dialogues, usually transcribed by speech recognition engines, impose additional challenges such as redundancies and misrecognized words. The length of these dialogues, e.g. $50K$ tokens in a 45 minutes call, imposes another challenge to state-of-the-art summarization models as it exceeds their input limits (Zhang et al., 2021). Figure 1 illustrates parts of the challenges imposed by automatically transcribed sales dialogues.

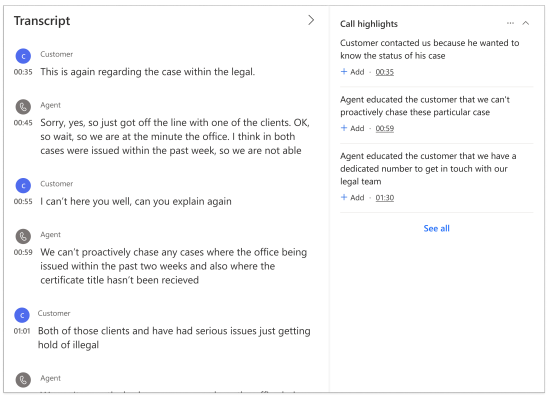

Figure 1: A customer-agent call transcript with corresponding summary highlights. Challenges imposed by automatic speech recognition engine can be observed.

Developing a production system which is both fully automatic and agnostic to the input text genre is an extremely difficult task given the current state-of-the-art technology. To this end, we present a pragmatic solution that enables users to interactively edit machine-generated summary for customer-agent sales calls as appears in Figure 2. Our solution summarizes the call into a collection of abstractive highlights to accurately capture the various details of the call. The machine-human interaction is enabled through a designated human-in-the-loop user experience (Ostheimer et al., 2021). It enables users to modify the generated summaries, yet, the intervention is designed to be minimal so that the overall time consumption of users is significantly reduced.

Overall, our contributions are listed as follows:

Proceedings of NAACL-HLT 2022: Industry Track Papers, pages 45 - 53

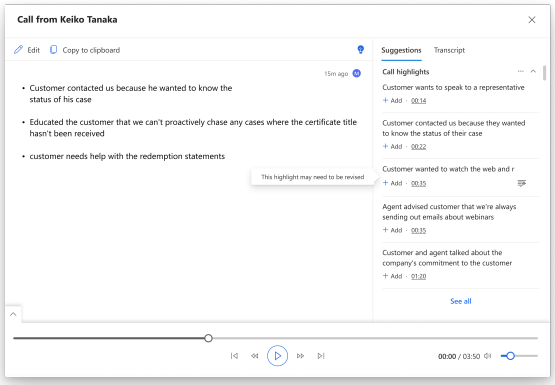

Figure 2: Illustrating human-in-the-loop experience which enables users to interactively handle summarization challenges by adding relevant summary highlights to the editing canvas and modify them, if necessary.

1. **Dialogue summarization system**. We introduce an innovative production summarization system for summarizing call transcripts with a human-in-the-loop setting. Our system uses an advanced summarization model to generate abstractive summaries for dialogues. Additionally, it employs a novel model for quantifying the coherence of the summaries to compensate for the summarization model limitations.

2. **GPT-3 as an offline label generator**. We present a technique for leveraging GPT-3 model to generate pseudo labels without the need to deploy and maintain it in production. This enables us to (i) efficiently generate labels in low-resource setting, (ii) distill GPT-3 knowledge into lighter models, and (iii) accommodate data privacy constraints.

3. **Custom evaluation metric**. We examine the importance of leveraging a comprehensive evaluation metric which takes into account various quality aspects of the generated summary. The metric is utilized to focus our efforts on potential candidate models during the development phase. The suggested metric goes beyond lexical overlap and help us validate that our production model is optimized for generating summaries which are fluent, relevant and factually reliable.

2 Related Work

Document Summarization Summarization methods can be categorized into two classes: extractive and abstractive. Early works focused on extractive methods (Hovy and Lin, 1997; Marcu, 1997), followed by rule-based approaches for abstractive summarization (Barzilay and Elhadad, 1997; Barzilay et al., 1999). Advancements in capabilities of deep neural models led to works such as Rush et al. (2015) where a seq-2-seq attention-based model is used for abstractive summarization. See et al. (2017) overcomes some of the former work's limitations by introducing a pointer generator network that has the ability to copy words from the source document. A major advancement in the field of deep neural models was the introduction of Transformer architecture (Vaswani et al., 2017), which is the basis for current state-of-the-art summarization approaches. Recently, several powerful Transformer-based models have been developed and showed remarkable results on various benchmark summarization tasks (Lewis et al., 2020; Zhang et al., 2019a; Raffel et al., 2020) .

Dialogue Summarization The task of dialogue summarization has been witnessing many commonalities as document summarization as well as new techniques for handling unique structures of various dialogue types. Early works in the domain suggested tackling the problem using extractive methods (Murray et al., 2005; Riedhammer et al., 2008). Shang et al. (2018) used a pure unsupervised graph-based method for keyword extraction and sentence compression. Goo and Chen (2018) proposed to explicitly model relationships between dialogue acts using attention-based sentence-gated mechanism. Chen and Yang (2020) extracted Transformer-based representations for different views of dialogues, conditioned on view representations, to generate summaries using a second Transformer. Zhu et al. (2020) presented a hierarchical Transformer architecture to encompass the structure of dialogues.

3 Method

While most existing methods summarize a call transcript as a single paragraph, our system provides a collection of sentences that summarize the entire dialogue in a chronological order. Given a call transcript, the system utilizes word embeddings to break the transcript into semantically coherent segments (Alemi and Ginsparg, 2015). Each segment is summarized independently capturing key information such as: customer's issue, agent's solution

or the underlying topic of the discussion. Finally, the grammatical coherence of highlights is analyzed using a dedicated model before suggesting them to the user. Figure 3 provides a high-level overview of the system's flow.

Next, we introduce the key components of our dialogue summarization system in details.

3.1 DialogBART: Dialogue Summarization Model

Unlike general documents, conversation transcripts have unique structures associated with speakers and turns. In sales calls, participants can either be a customer or an agent and these roles impose a unique language style that can be leveraged by the model. Motivated by this observation, we propose an encoder-decoder model called DialogBART, which adapts the well-known BART (Lewis et al., 2020) model with additional embedding parameters to model both turns and speakers positions (Zhang et al., 2019c; Bao et al., 2020). For speaker embeddings, we introduce designated vectors to represent each speaker which can be easily generalized to multi-participant dialogues. Additionally, we leverage another set of vectors to model turn position embeddings. During inference, the model determines the speaker and turn indices by leveraging a special token that separates the dialogue's turns.

As shown in Figure 4, DialogBART's input is calculated as the sum of the corresponding token, position, speaker and turn position embeddings. These parameters are randomly initialized, however, the remaining parameters are initialized with weights from a pretrained[1] BART-like encoder-decoder models (Lewis et al., 2020; Shleifer and Rush, 2020). All these weights are further fine-tuned on dialogue summarization tasks.

3.2 Acceptability Validation

Despite the human-in-the-loop user experience, customers still expect high quality summaries which require minimal modifications by them. We propose a novel model that determines the quality of each summary highlight in terms of coherence, fluency and its acceptability in general.

Grammatical acceptability, a property of natural language text, implies whether a text is accepted or not as part of the language by a native speaker. The notion was widely investigated through vast work done in automatic detection of grammatical errors (Atwell, 1987; Chodorow and Leacock, 2000; Bigert and Knutsson, 2002; Wagner et al., 2007) and on acceptability judgment of neural networks (Lau et al., 2017; Warstadt et al., 2019). And yet, we are not aware of works that observe the acceptability of neural generated summaries for validation purposes. To determine a highlight's acceptability, we compute the perplexity of each highlight given by a Pretrained Language Model (PLM). This PLM is fine-tuned on summaries from DialogSum dataset (Chen et al., 2021a) and in-domain proprietary data in a traditional self-supervised manner. Recall that the perplexity of a sequence $W = w_0 w_1 ... w_n$ is defined as:

$$PP(W;\theta) = \sqrt[n]{\prod_{k=1}^{n} \frac{1}{p_\theta(w_k|w_0 w_1 ... w_{k-1})}} \quad (1)$$

where θ are the language model specific parameters and p_θ is the probability function corresponding to distribution over vocabulary tokens induced by the same model.

Based on the perplexity score, the system determines whether a given highlight should be filtered out, presented to the user, or presented with an indication that its revision may be required. Figure 2 illustrates how the system helps users focus their efforts on modifying borderline acceptable highlights based on the perplexity score.

4 GPT-3 as an Offline Labeler

Training a dialogue summarization model requires a large amount of labeled examples. Manually annotating data for abstractive summarization is a time-consuming and labor-intensive process, let alone the data privacy constraints. In this work, we provide a method to automatically pseudo label examples to overcome these challenges. We leverage GPT-3 model (Brown et al., 2020) to generate pseudo labels in a zero-shot setting for each call segment. GPT-3 is a large auto-regressive language model with 175 billion parameters that achieves promising results on various NLP tasks, including question answering. We treat the problem of label generation as a question answering task. For each segment, we concatenate a question-based prompt with the segment's content while expecting the GPT-3 model to provide the answer as appears

[1] https://huggingface.co/sshleifer/distilbart-xsum-12-3

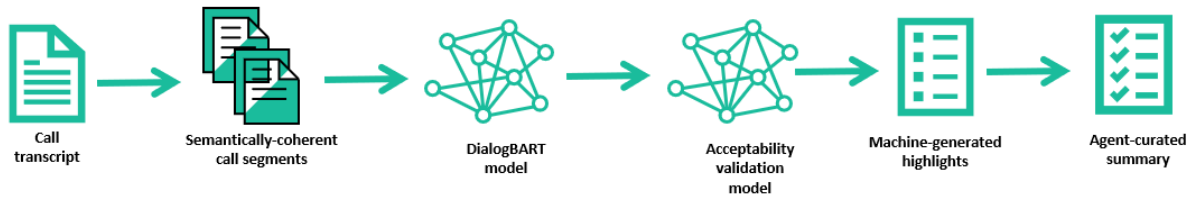

Figure 3: A high-level overview of the system's flow

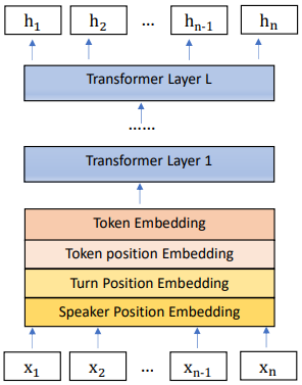

Figure 4: Input representation of DialogBART's encoder.

in Figure 5. These answers are used as pseudo labels for the corresponding segments. This formulation provides the flexibility of defining multiple questions per segment to summarize the segment from different perspectives. Finally, these pseudo labels, combined with proprietary human labeled data, are used to fine-tune the DialogBART model on conversational text.

```
Segment_text = <....>
GPT3_prompt = Segment_text + "Q: Why did the customer contacted the agent?\n"
Answer_prefix = "A: The customer contacted the agent in order to"
--------------------------------------------------------------------------------
GPT3_label = 'The customer contacted the agent in order to get a mortgage offer.'
```

Figure 5: Utilizing GPT-3 model to generate task-oriented summaries in an offline manner.

5 Custom Evaluation Metric

Common evaluation metrics for text summarization task, i.e. ROUGE and METEOR, have salient limitations as both metrics track lexical overlap between the summary and the original text. This kind of assessment falls short when the summary content perfectly aligns with a reference text but does not necessarily contain any lexical overlap, e.g. abstractive summaries.

In an industrial setting, one needs to consider various quality perspectives to guarantee that the summary's quality does not introduce productivity blockers for users or negatively affect business decisions. We introduce a custom evaluation metric, *SumSim*, that relies on lexical overlap as well as other quality aspects to ensure that summaries are fluent, coherent and factually reliable. SumSim aims to cover the following quality perspectives:

- **Coverage** - how many units from the reference text are covered by the summary (S_r)
- **Relevance** - measures semantic consistency between the summary and the reference text (S_b)
- **Informativeness** - how well it captures predefined keywords which are critical to the business (S_i)
- **Factuality** - how factual the summary is with respect to the original text (S_f)

Our metric uses *ROUGE-L* (Lin, 2004), *BertScore* (Zhang et al., 2019b), exact match of keywords and *FactCC* (Kryscinski et al., 2019) to capture the above quality aspects, respectively. The quality of a given summary is calculated as follows:

$$S_0 = \alpha \cdot S_i + \frac{1-\alpha}{2} \cdot (S_r + S_b) \quad (2)$$

$$SumSim = \beta \cdot S_f + (1-\beta) \cdot S_0 \quad (3)$$

where α and β are determined empirically based on the business scenario sensitivity.

6 Experimental Results

In this section we evaluate the performance of our proposed models on various datasets: DialogSum (Chen et al., 2021b), SAMSum (Gliwa et al., 2019), CoLA (Warstadt et al., 2019) and a proprietary data from the sales domain. We also show the potential of $SumSim$ metric compared to traditional evaluation metrics on the text summarization task. We use Huggingface Transformers (Wolf et al., 2020) as a training framework in all of our experiments.

6.1 DialogBART

In the following experiments we show the performance of DialogBART model in summarizing dialogues by examining two factors: (i) speaker/turn position embedding parameters, and (ii) data augmentation by GPT3-labeled data. For comparison purposes, we leverage two baseline models, *BART-large* and *distilBART*, which achieved state-of-the-art results on the summarization task (Lewis et al., 2020; Shleifer and Rush, 2020). All models, including the baseline models, were initially fine-tuned on XSum dataset (Narayan et al., 2018).

First, we examine the contribution of DialogBART's position embeddings on DialogSum and SAMSum datasets. All models were fine-tuned using the relevant training sets and evaluated on the test sets of the corresponding datasets. Table 1 shows that the suggested speakers/turns positions embeddings provide better results when compared to the baseline models.

Model	R1	R2	RL
	DialogSum		
distilBART	35.93	11.71	28.86
+ embeddings	**46.97**	**21.34**	**39.45**
BART-large	46.48	20.89	38.12
+ embeddings	**46.68**	**21.46**	**38.32**
	SAMSum		
distilBART	41.93	19.17	34.05
+ embeddings	**50.21**	**25.89**	**41.99**
BART-large	52.45	28.08	43.84
+ embeddings	**52.91**	**28.39**	**43.90**

Table 1: Effectiveness of DialogBART's speaker and turn embedding parameters using ROUGE metrics.

Second, we examine the implications of fine-tuning DialogBART model using different data types: human-labeled ($20K$ samples) and GPT3-labeled ($21K$ samples) data [2].

We evaluated the models on the test subset of: (i) DialogSum (500 samples), (ii) SAMSum (819 samples), and (iii) proprietary data (100 samples). The evaluation on the public datasets was conducted without fine-tuning the models on the corresponding training sets. Table 2 shows that DialogBART model outperforms the baseline models on public datasets even in out-of-domain setting. Additionally, results show that DialogBART fine-tuned on a mixture of human and pseudo labels outperforms its counterparts which were fine-tuned on either fully human labels or fully pseudo labels. We note that fine-tuning DialogBART on pseudo labels yielded higher ROUGE scores compared to human labels. This could be explained by human tendency to generate variable summaries which induces disagreements between human annotators (Clark et al., 2021). While a model fine-tuned on pseudo labels is less variable in its generations, a model fine-tuned using human data produces text that is, in turn, more variable and leads to less lexical overlap with test references as measured by ROUGE metrics[3] .

6.2 Acceptability Validation

We examine mutiple candidate PLMs with language model objective for this task. Initially we fine-tune the candidate PLM on summaries from DialogSum dataset and later on positive examples from the train subset of our internal acceptability benchmark consisting of in-domain summaries (100 samples). As candidate PLMs, we experiment with GPT-2 (Radford et al., 2019), DistilGPT-2 Sanh et al. (2019) and a RoBERTa encoder (Liu et al., 2019) with a language model head, *RoBERTa-LM*. Table 3 shows comparison results between the examined models and leaderboard competitors on the development set of CoLA as well as on the test subset of an internal benchmark.

We observe that all of the models trained using our method, in the bottom half of the table, which were not trained on CoLA data yield competitive results compared to models explicitly fine-tuned for the task, top half of Table 3. We also found that the RoBERTa-LM model achieves highest results on the internal set. Additionally, we fine-tuned DeBERTa, the strongest CoLA competitor, in a classification setting on the internal benchmark. We observe that results achieved by our models are significantly better. We hypothesize, this phenomenon is due to the fact that valid in-domain highlights, as generated by DialogBART, share a unique structure and can be viewed as forming a specific language which properties are better captured by a language model rather than a classifier.

[2]The anonymized data was collected and used based on a data sharing agreement with customers from different business domains. The human-labeled data is composed of anonymized agent's notes which were captured as part of the daily routine of the agents and not in a crowdsourcing setting.

[3]https://github.com/google-research/google-research/tree/master/rouge

Model	DialogSum			SAMSum			Proprietary		
	R1	R2	RL	R1	R2	RL	R1	R2	RL
distilBART	17.1	3.6	13.5	20.3	4.1	15.5	16.3	1.1	13.1
BART-large	17.7	3.9	13.7	24.9	5.5	18.9	16.9	1.5	13.3
DialogBART									
+ human	21.7	4.5	19.1	18.9	3.0	16.8	20.2	5.3	19.2
+ pseudo	28.0	5.9	22.2	26.0	5.1	20.6	28.5	10.5	24.8
+ human & pseudo	**33.5**	**9.0**	**24.5**	**30.4**	**7.5**	**22.4**	**33.1**	**13.0**	**26.3**

Table 2: Results of fine-tuning DialogBART model on human labels, pseudo labels and mixture of them.

Model	CoLA_{dev}	Internal
TinyBERT (Jiao et al., 2020)	54	-
Synthesizer (Tay et al., 2021)	53.3	-
DeBERTa (He et al., 2021)	**69.5**	54.5
DistilGPT-2	61.7	63.6
GPT-2	62.5	60.6
RoBERTa-LM	64.2	**66.7**

Table 3: Accuracy of acceptability validation models.

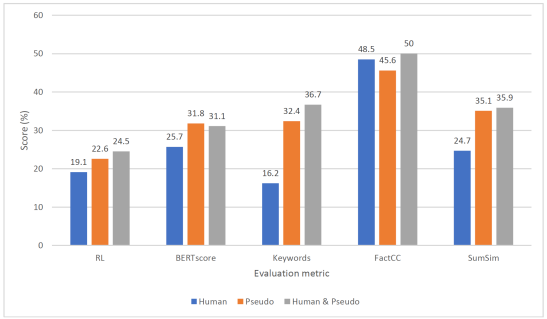

Figure 6: The capacity and limitation of various quality metrics. The bars' colors represent three different models which were fine-tuned on human labels, pseudo labels and a mixture of them, respectively.

6.3 Custom Evaluation Metric

We leverage the DialougSum test set to show the potential of the $SumSim$ metric. Table 2 shows that DialogBART fine-tuned on pseudo labels, M_{pseudo}, outperforms its counterpart that was fine-tuned on human labels, M_{human}. However, Figure 6 shows contradicting insights when comparing the performance of these two models by different quality aspects, i.e., lexical overlap (ROUGE) and factual reliability (factCC).

This observation emphasizes the need for non-standard quality measures for evaluating the performance of abstractive summarization models. This need is critical for customer-facing enterprise products where business decisions can be affected by the generated summary. Figure 6 shows the strengths and weaknesses of different quality metrics in evaluating three DialogBART variants. The M_{pseudo} model outperforms the M_{human} in all quality dimensions except of factuality. This observation is consistent with recent studies which report that large size language models are less truthful than their smaller peers (Lin et al., 2021).

7 Conclusions

In this paper, we present an end-to-end system for abstractive summarization of agent-customer calls. We employ a two-stage strategy to summarize long call transcripts by (i) splitting the dialogue into semantically coherent segments, and (ii) generating summaries using our DialogBART summarization model. We demonstrate how a pragmatic solution that combines a content selection model with a human-in-the-loop user experience can help compensate on generative models' limitations. We show how GPT-3 model can be leveraged as an offline data labeler to train lighter models and accommodate data privacy constraints. Experiments show that the introduced embedding parameters combined with fine-tuning on in-domain data significantly improve the quality of the generated summaries with respect to publicly available BART-based summarization models. We emphasize the need for non-standard evaluation metrics and show how common metrics fall short when evaluation of abstractive summaries is considered.

References

Alexander A. Alemi and Paul Ginsparg. 2015. Text segmentation based on semantic word embeddings. *CoRR*, abs/1503.05543.

Eric Steven Atwell. 1987. How to detect grammatical

errors in a text without parsing it. In *Third Conference of the European Chapter of the Association for Computational Linguistics*, Copenhagen, Denmark. Association for Computational Linguistics.

Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2020. PLATO: Pre-trained dialogue generation model with discrete latent variable. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 85–96, Online. Association for Computational Linguistics.

Regina Barzilay and Michael Elhadad. 1997. Using lexical chains for text summarization. In *Intelligent Scalable Text Summarization*.

Regina Barzilay, Kathleen R. McKeown, and Michael Elhadad. 1999. Information fusion in the context of multi-document summarization. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 550–557, College Park, Maryland, USA. Association for Computational Linguistics.

Johnny Bigert and Ola Knutsson. 2002. Robust error detection: A hybrid approach combining unsupervised error detection and linguistic knowledge. In *Proc. 2nd Workshop Robust Methods in Analysis of Natural language Data (ROMAND'02), Frascati, Italy*, pages 10–19. Citeseer.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jiaao Chen and Diyi Yang. 2020. 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)*, pages 4106–4118, Online. Association for Computational Linguistics.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021a. DialogSum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074, Online. Association for Computational Linguistics.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021b. Dialogsumm: A real-life scenario dialogue summarization dataset. *CoRR*, abs/2105.06762.

Martin Chodorow and Claudia Leacock. 2000. An unsupervised method for detecting grammatical errors. In *ANLP*.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *ACL/IJCNLP*.

Ryan Gavin, Liz Harrison, Candace Lun Plotkin, Dennis Spillecke, and Jennifer Stanley. 2020. The b2b digital inflection point: How sales have changed during covid-19.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *CoRR*, abs/1911.12237.

Chih-Wen Goo and Yun-Nung (Vivian) Chen. 2018. Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts. *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 735–742.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Eduard Hovy and ChinYew Lin. 1997. Automated text summarization in SUMMARIST. In *Intelligent Scalable Text Summarization*.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the factual consistency of abstractive text summarization. *CoRR*, abs/1910.12840.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive science*, 41 5:1202–1241.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Stephanie C. Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *ArXiv*, abs/2109.07958.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Daniel Marcu. 1997. From discourse structures to text summaries. In *Intelligent Scalable Text Summarization*.

Gabriel Murray, Steve Renals, and Jean Carletta. 2005. Extractive summarization of meeting recordings. In *INTERSPEECH*.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 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*, Brussels, Belgium.

Julia Ostheimer, Soumitra Chowdhury, and Sarfraz Iqbal. 2021. An alliance of humans and machines for machine learning: Hybrid intelligent systems and their design principles. *Technology in Society*, 66:101647.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Korbinian Riedhammer, Benoit Favre, and Dilek Hakkani-Tur. 2008. A keyphrase based approach to interactive meeting summarization. In *2008 IEEE Spoken Language Technology Workshop*, pages 153–156. IEEE.

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

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

A. See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *ACL*.

Guokan Shang, Wensi Ding, Zekun Zhang, Antoine Tixier, Polykarpos Meladianos, Michalis Vazirgiannis, and Jean-Pierre Lorré. 2018. Unsupervised abstractive meeting summarization with multi-sentence compression and budgeted submodular maximization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 664–674, Melbourne, Australia. Association for Computational Linguistics.

Sam Shleifer and Alexander M. Rush. 2020. Pre-trained summarization distillation.

Yi Tay, Dara Bahri, Donald Metzler, Da-Cheng Juan, Zhe Zhao, and Che Zheng. 2021. Synthesizer: Rethinking self-attention in transformer models. *ArXiv*, abs/2005.00743.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Joachim Wagner, Jennifer Foster, and Josef van Genabith. 2007. A comparative evaluation of deep and shallow approaches to the automatic detection of common grammatical errors. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 112–121, Prague, Czech Republic. Association for Computational Linguistics.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019a. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019b. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.

Xinyuan Zhang, Ruiyi Zhang, Manzil Zaheer, and Amr Ahmed. 2020. Unsupervised abstractive dialogue summarization for tete-a-tetes. *CoRR*, abs/2009.06851.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019c. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

Yusen Zhang, Ansong Ni, Tao Yu, Rui Zhang, Chenguang Zhu, Budhaditya Deb, Asli Celikyilmaz, Ahmed Hassan Awadallah, and Dragomir Radev. 2021. An exploratory study on long dialogue summarization: What works and what's next. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4426–4433, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chenguang Zhu, Ruochen Xu, Michael Zeng, and Xuedong Huang. 2020. A hierarchical network for abstractive meeting summarization with cross-domain pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 194–203, Online. Association for Computational Linguistics.

Xiaodan Zhu and Gerald Penn. 2006. Summarization of spontaneous conversations. In *Ninth International Conference on Spoken Language Processing*.