Investigating the Role and Impact of Disfluency on Summarization

Varun Nathan, Ayush Kumar and **Jithendra Vepa**
{varun.nathan, ayush, jithendra}@observe.ai
Observe.AI
Bangalore, India

Abstract

Contact centers handle both chat and voice calls for the same domain. As part of their workflow, it is a standard practice to summarize the conversations once they conclude. A significant distinction between chat and voice communication lies in the presence of disfluencies in voice calls, such as repetitions, restarts, and replacements. These disfluencies are generally considered noise for downstream natural language understanding (NLU) tasks. While a separate summarization model for voice calls can be trained in addition to chat specific model for the same domain, it requires manual annotations for both the channels and adds complexity arising due to maintaining two models. Therefore, it's crucial to investigate if a model trained on fluent data can handle disfluent data effectively. While previous research explored impact of disfluency on question-answering and intent detection, its influence on summarization is inadequately studied. Our experiments reveal up to 6.99-point degradation in Rouge-L score, along with reduced fluency, consistency, and relevance when a fluent-trained model handles disfluent data. Replacement disfluencies have the highest negative impact. To mitigate this, we examine Fused-Fine Tuning by training the model with a combination of fluent and disfluent data, resulting in improved performance on both public and real-life datasets. Our work highlights the significance of incorporating disfluency in training summarization models and its advantages in an industrial setting.

1 Introduction

Disfluency is a prevalent feature of spontaneous spoken speech, encompassing natural interruptions during communication such as, repetitions (*this is this is just not working*), restarts (*why don't you i will do it*), and replacements (*blue no no red*). Voice call interactions in contact center is rich in such disfluencies. Thus, there is a growing need to understand the influence and impact of disfluencies on natural language processing tasks owing to an increase in such data containing disfluencies. Existing research has primarily focused on disfluency detection (Kundu et al., 2022) and removal (Wang et al., 2010; Saini et al., 2021), with limited investigations into its impact on downstream tasks. Recently, Dao et al. (2022a) explored the effect of disfluency on intent and slot detection tasks, while Gupta et al. (2021) finds that state-of-the-art pre-trained models are not robust when directly tested on the disfluent input for the task of question answering. However, the influence of disfluency on summarization models, particularly for spoken dialogues, remains relatively unexplored despite the extensive research on document and dialogue summarization. In this work, we investigate the impact of disfluencies on summarization. The contributions and key findings of our work are outlined below:

- We present an investigation on the presence and impact disfluency on the downstream task of summarization of abstractive and extractive summaries. To the best of our knowledge, this is the first work that focuses on understanding the impact of disfluency in such a setting.

- This study analyzes how disfluency impacts summary characteristics (fluency, consistency,

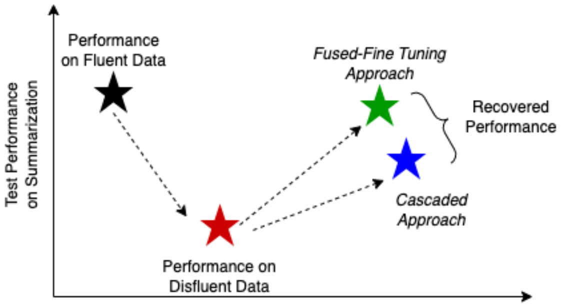
Test Performance on Summarization
Performance on Fluent Data
Fused-Fine Tuning Approach
Recovered Performance
Cascaded Approach
Performance on Disfluent Data

Figure 1: Illustrative diagram representing the adverse impact of presence of disfluency on summarization performance and the recovered performance using mitigation approaches.

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 541–551
December 6-10, 2023 ©2023 Association for Computational Linguistics

Disfluency Type	Original Input	Disfluency Induced Output
Restart	*I would suggest to look into this matter as this is a serious concern.*	*Why don't you actually I would suggest to look into this matter as this is a serious concern.*
Repetition	*I don't want anyone using instant messaging. in this office.*	*I don't don't want anyone using instant messaging. in this office.*
Replacement	*They will have to change and update their network settings on their tab.*	*They will have to change and update their internet no wait network settings on their tab.*

Table 1: Representative examples of different type of disfluency introduced in the input using LARD algorithm as discussed in Section 2.1.

and relevance). We note that disfluency has a more adverse effect on extractive summarization than abstractive. Among various disfluency types, presence of *replacements* is the most challenging for summarization model.

- We discuss a simple yet effective Fused-Fine Tuning strategy to alleviate the impact of disfluency on summarization. The method achieves up to 90% recovery in Rouge metrics for both abstractive and extractive tasks. However, there remains a noticeable gap in consistency and relevance scores for extractive summarization. It is to be noted that our goal is not to propose a state-of-the-art model but to assess the impact of disfluency and benchmark simple mitigation techniques.

- Finally, we evaluate the effectiveness of the approach on real life contact center conversational dataset in a purely zero-shot setting.

2 Methodology

In this section, we outline our method for curating a dataset of various types of disfluencies which is further used to analyze their impact on a summarization model trained on a fluent dataset. Our investigation highlights a significant decrease in the performance of the model when presented with disfluent data. As a result, we discuss strategies to mitigate this impact and improve the model's ability to accurately summarize disfluent inputs.

2.1 Curating disfluent data

To investigate the impact of disfluency on summarization, we need a dataset comprising both disfluency and corresponding summaries. While there exist datasets containing disfluencies, such as the Switchboard corpus (Zayats et al., 2019), they don't contain corresponding summaries. Thus, we curate a dataset with disfluencies utilizing a disfluency induction mechanism called LARD (Passali et al., 2022). LARD generates complex yet realistic disfluencies, covering three categories: *Repetition* (speaker repeats a word, a phrase or a sequence of words), *Replacement* (speaker replaces the fluent word(s) or phrase(s) with the disfluent one(s)), and *Restart* (speaker abandons the initial part of the utterance completely and restarts it). We maintain an equal proportion of disfluency types and original fluent data. Consequently, we obtain two datasets: 1) fluent (original), 2) disfluent (curated by LARD). Each dataset would have the corresponding gold summary obtained from original dataset. These datasets facilitate evaluating the impact of disfluency on summarization in our study. Representative examples of the disfluency-induced utterances as obtained from LARD are provided in Table 1.

2.2 Investigating the impact of disfluency

To understand the impact of disfluency on summarization, it is imperative to compare the performance of summarization models on both fluent and disfluent datasets. The base assumption is that the summaries of disfluent inputs should ideally be same compared to the corresponding fluent inputs. To test this hypothesis, we evaluate the output of summarization models trained on fluent data by subjecting it to testing on both the original (fluent) and disfluent (curated) datasets. This evaluation is conducted against the gold summary present in the original dataset, which ensures that any variations in summarization quality are due solely to disfluency and not to any other factors. Based on the evaluation (Section 4.2), we note that summarization systems trained on a fluent corpus reveal a substantial decline in performance when tested on disfluent input. Thus, there is a strong need to develop models that can generate precise summaries even from disfluent inputs.

2.3 Mitigating the impact of disfluency

We explore two approaches to mitigate the impact of disfluency on summarization.

1. **Cascaded Approach**: This approach aims to enhance the performance of the summarization model by taking advantage of disfluency removal systems to eliminate disfluent segments. These segments may impede the model's ability to precisely capture critical information and essential concepts in the input. For disfluency removal system, we leverage the codebase released by Ghosh et al. (2022) that achieved state-of-the-art results on disfluency detection on Switchboard corpus (Zayats et al., 2019). We use the model trained on SWBD to detect and remove the disfluent span in our input to obtain a disfluency-free fluent output. Subsequently, we feed the obtained fluent output to the summarization model that is already trained on fluent data. By doing so, we expect to mitigate the negative impact of disfluency on the summarization task.

2. **Fused-Fine Tuning Approach**: We use an alternative approach that involves utilizing disfluency-induced input paired with corresponding gold summaries obtained from the original (fluent) dataset to fine-tune the summarization model. The underlying assumption is that the summary for the disfluent data point should be identical to that of the corresponding fluent data point. To curate the training dataset, we keep a varying proportion of disfluent data during the training process, ranging from 20% to 100%. The remaining proportion is made up of the original (fluent) training data. This approach circumvents the need for a cascaded pipeline involving disfluency removal, allowing the model to learn to summarize disfluent inputs directly.

3 Dataset

To investigate the impact of disfluency in two different setups with varying characteristics, we utilize two publicly available datasets, *DialogSum* (Chen et al., 2021) and *DebateSum* (Roush and Balaji, 2020). DialogSum is an abstractive summarization dataset which consists of a total of 13,460 dialogues demonstrating real-life scenarios across a variety of topics like schooling, work, medication, shopping, leisure, travel etc. with large scale data. The average number of tokens per dialog in DialogSum dataset is 131 while the compression ratio (ratio of the length of the summary divided by the length of the original text) is 0.18. On the other hand, DebateSum is a word-level extractive summarization dataset containing a total of 187,386 unique pieces of evidences obtained from debates. For this dataset, average number of tokens in per sample of dataset is 372 while the compression ratio is 0.46. To evaluate the impact of the presence of disfluency, we generate the disfluency-induced versions of the DialogSum and DebateSum corpora using LARD as discussed in Section 2.1.

4 Experiments and Results

4.1 Implementation Details and Evaluation Metrics

We use BART-Large (Lewis et al., 2020) as a base model to fine-tune, as it is one of the widely used pretrained model for summarization (and other sequence-to-sequence) tasks[1]. The implementation was done in PyTorch deep learning framework using Python 3.8. We trained our model on a single NVIDIA Tesla V100 GPU with 32GB memory. We used the Adam optimizer with a learning rate range of {1e-5, 3e-5, 5e-5, 1e-4, 5e-4}, and a batch size range of {4, 8, 16}. Our model was trained for 5 epochs with early stopping based on the validation loss. The maximum input and target length was set to 256 for the DialogSum dataset, while it was set to 1024 for DebateSum dataset. For disfluency removal model, we use the default parameters as provided in the codebase of Ghosh et al. (2022).

We employ a variety of evaluation metrics to compare the performance of our experiments. We start by using one of the traditional evaluation metrics, Rouge (Lin, 2004), which computes n-gram overlap between the model output and the reference text. Specifically, we report results on Rouge-1, Rouge-2 and Rouge-L F1 scores. Rouge-1 and Rouge-2 evaluate the overlap of the generated summary with the ground truth summary at unigram and bigram level respectively, while Rouge-L measures the longest common subsequence between the generated and ground truth summaries. Next we consider BERTscore (Zhang et al., 2020) which is an embedding based evaluation metric. BERTscore

[1] We did not benchmark larger models or recent large generative models due to the realistic constraints and trade-off of balancing the cost, latency and control of the summarization model in industrial setting of high scale data.

	DialogSum Dataset (Abstractive)			DebateSum Dataset (Extractive)		
Evaluation Criteria / Inference Input Type	Evaluated on Fluent	Evaluated on Disfluent	Delta (Abs./Rel.)	Evaluated on Fluent	Evaluated on Disfluent	Delta (Abs./Rel.)
Rouge-1	42.187	39.895	-2.292 / -5.43%	62.755	57.135	-5.620 / -8.96%
Rouge-2	17.293	15.604	-1.689 / -9.77%	55.019	47.304	-7.715 / -14.02%
Rouge-L	34.379	31.942	-2.437 / -7.09%	57.699	50.707	-6.992 / -12.12%
BERTScore	0.916	0.908	-0.008 / -0.87%	0.911	0.889	-0.022 / -2.42%
Fluency	0.935	0.862	-0.073 / -7.76%	0.750	0.195	-0.555 / -74.05%
Consistency	0.820	0.753	-0.066 / -8.10%	0.939	0.547	-0.392 / -41.73%
Relevance	0.871	0.780	-0.090 / -10.36%	0.592	0.121	-0.471 / -79.48%

Table 2: Impact of disfluency on summarization models trained on Fluent data for the two datasets respectively.

computes the similarity between the generated summary and the ground truth summary using contextualized representations of words obtained from a pre-trained BERT model. Since none of the above metrics do not provide an interpretable assessment of the quality of the generated output, we additionally consider evaluation criteria that can provide an evaluation of the explainable dimensions that includes fluency, consistency and relevance of the generated output. Fluency indicates the quality of the individual sentences, while consistency refers to the factual alignment between the summary and the source document and relevance quantifies whether the summary contains only the important information of the source. To evaluate the generated summaries on these dimensions, we utilize a recently proposed unified multi-dimensional evaluator called UniEval (Zhong et al., 2022b) that has shown strong correlation to human judgements.

4.2 Experimental results on the impact of disfluency on summarization

The summarization results obtained on a disfluent test set by utilizing the summarization models trained on the fluent version of the respective datasets are presented in Table 2. Our findings reveal a significant drop in the performance of the summarization model across all evaluation metrics. For the DialogSum dataset, the Rouge-L F1 score drops by 2.4 absolute points, while the drop in Rouge-L on the DebateSum dataset is 6.99 points. We see a relative decline of upto 3.2% on BERTScore metric across the two datasets. The lower scores across multi-dimensional characteristics of fluency, consistency and relevance corroborates the previous observations and suggests that the perceptive quality wrt human judgement may be much lower (Zhong et al., 2022b).

An important observation from our evaluation is that the impact of disfluency on the summarization results is much more pronounced for an extractive summarization dataset compared to an abstractive summarization dataset. For example, for the DebateSum dataset, we observed a huge drop in fluency (74% relative), consistency (41% relative) and relevance (79% relative) scores when the summarization model is exposed to the disfluent version of the data (Table 2). This indicates that the model trained on fluent data for extractive summarization fails to recognize the disfluent segments, which negatively impacts the fluency (and other aspects) of the generated summary.

4.3 Mitigating the impact of disfluency

Based on the evidence from results in the Section 4.2, it is imperative to think of mitigating the negative impact of the presence of disfluency in the input. We present the results of Cascaded and Fused-Fine Tuning approach in Table 3. Oracle represents the setup in which the model is trained and tested on the fluent version of the respective datasets. In addition to the performance of mitigation strategies on disfluent data, we also evaluate the performance of Fused-Fine Tuning approach on the fluent version of the datasets to understand if the same model can infer on both fluent and disfluent data. Ideally, the performance of the summarization model on fluent data and disfluent data should be similar to *Oracle* setup.

The findings demonstrate substantial improvements in the summarization model's performance on the disfluent dataset when compared to the model trained on fluent data ($Rel\Delta_{Fluent}$ in Table 3). Both cascaded pipeline and fused-fine tuning mitigation approaches yield better results, with fused-fine tuning showing superior performance over the cascaded pipeline. The summary quality, as measured by Rouge metrics, exhibits relative gains of 5-12% for both datasets with fused-fine tuning. Furthermore, fluency, consistency, and rel-

DialogSum Dataset (Abstractive Summarization)							
Evaluation Criteria	Oracle	Evaluated on Disfluent Version					Evaluated on Fluent Version
		Fluent Model	Cascaded Pipeline	Rel Δ_{Fluent} / Rel Δ_{Oracle}	Fused-Fine Tuning	Rel Δ_{Fluent} / Rel Δ_{Oracle}	Fused-Fine Tuning
Rouge-1	42.187	39.895	41.224	3.33% / -2.28%	42.022	5.33% / -0.39%	42.154
Rouge-2	17.293	15.604	15.974	2.37% / -7.63%	16.963	8.71% / -1.91%	17.005
Rouge-L	34.379	31.942	33.197	3.93% / -3.44%	34.114	6.80% / -0.77%	34.607
BERTScore	0.916	0.908	0.912	0.44% / -0.44%	0.916	0.88% / 0.00%	0.916
Fluency	0.935	0.862	0.935	8.47% / 0.00%	0.940	8.97% / 0.52%	0.938
Consistency	0.820	0.753	0.802	6.51% / -2.20%	0.826	9.66% / 0.78%	0.830
Relevance	0.871	0.780	0.843	8.08% / -3.21%	0.870	11.52% / -0.04%	0.873
DebateSum Dataset (Extractive Summarization)							
Evaluation Criteria	Oracle	Evaluated on Disfluent Version					Evaluated on Fluent Version
		Fluent Model	Cascaded Pipeline	Rel Δ_{Fluent} / Rel Δ_{Oracle}	Fused-Fine Tuning	Rel Δ_{Fluent} / Rel Δ_{Oracle}	Fused-Fine Tuning
Rouge-1	62.755	57.135	57.691	0.97% / -8.07%	61.986	8.49% / -1.22%	62.249
Rouge-2	55.019	47.304	51.226	8.29% / -6.89%	53.387	12.86% / -2.97%	53.607
Rouge-L	57.699	50.707	53.467	5.44% / -7.33%	56.454	11.33% / -2.16%	56.917
BERTScore	0.911	0.889	0.908	2.14% / -0.33%	0.910	2.33% / -0.14%	0.909
Fluency	0.750	0.195	0.721	269.74% / -3.87%	0.740	280.34% / -1.30%	0.778
Consistency	0.939	0.547	0.873	59.60% / -7.03%	0.900	64.58% / -4.10%	0.948
Relevance	0.592	0.121	0.514	324.79% / -13.18%	0.569	368.13% / -3.95%	0.644

Table 3: Results of mitigation strategy on DialogSum and DebateSum dataset.

Evaluation Criteria	Call Summarization			Speaker Turn Summarization		
	Fluent Model	Fused-Fine Tuning	Delta (Abs./Rel.)	Fluent Model	Fused-Fine Tuning	Delta (Abs./Rel.)
Rouge-1	27.454	32.761	5.306 / 19.33%	64.290	67.600	3.310 / 5.15%
Rouge-2	7.661	8.396	0.735 / 9.60%	52.040	56.150	4.110 / 7.90%
Rouge-L	27.954	31.942	3.988 / 14.27%	59.281	63.150	3.869 / 6.53%
BERTScore	0.839	0.849	0.009 / 1.12%	0.888	0.900	0.012 / 1.35%
Fluency	0.580	0.616	0.036 / 6.18%	0.783	0.855	0.072 / 9.20%
Consistency	0.420	0.500	0.079 / 18.82%	0.764	0.842	0.078 / 10.21%
Relevance	0.602	0.623	0.020 / 3.34%	0.624	0.672	0.048 / 7.69%

Table 4: Zero-shot performance of a model trained exclusively on fluent data vs. a combination of fluent and disfluent data in Fused-Fine tuning strategy on DialogSum dataset and evaluated on internal dataset.

evance scores witness relative improvements of up to 11.5% for the DialogSum dataset and upto 4.5x gains for the DebateSum dataset. For the DialogSum dataset ($Rel\Delta_{Oracle}$), fused-fine tuning demonstrates better abstractive capabilities by generating more readable and factual outputs while discarding disfluent influences during training. However, for the DebateSum dataset, a larger gap with the Oracle suggests the need for more focused strategies to mitigate the impact of disfluency on extractive summaries.

Additionally, the fused-fine tuning approach equips the model to handle both fluent and disfluent data effectively, proving advantageous in real-life industry settings. This streamlined methodology allows the same model to be utilized for both the input types, optimizing resource utilization and enhancing scalability for practical applications.

The findings from our qualitative analysis (Section A.1) support the observations made in the results section. We have observed that the presence of induced disfluency adversely affects the ability of the summarization model to generate accurate summaries. As shown in Table 6, the summarization model in Oracle setup generates a summary that has a better match with the Gold Summary. However, when the model is evaluated on disfluent input (i.e., *prediction of Fluent Model on Disfluent Input*), it fails to synthesize the information properly, as evidenced by the predicted segment of '*Carol believes #Person1# will follow.*'. The fused-fine tuning approach helps recover the quality of

		Trained on Fluent			Trained in Fused-Fine Tuning		
Evaluation	Oracle	Model Evaluated on Disfluency of Type					
Metrics		Restart	Repetition	Replacement	Restart	Repetition	Replacement
Rouge-1	42.187	41.141	40.963	37.251	41.863	41.148	40.355
Rouge-2	17.293	15.927	16.197	13.716	16.834	16.276	15.850
Rouge-L	34.379	32.907	32.544	29.421	33.711	33.663	32.705
BERTScore	0.916	0.913	0.912	0.904	0.916	0.915	0.912
Fluency	0.935	0.930	0.898	0.879	0.935	0.935	0.932
Consistency	0.820	0.815	0.777	0.746	0.830	0.818	0.808
Relevance	0.871	0.846	0.796	0.790	0.876	0.877	0.853

Table 5: Investigative study to understand the impact of different type of disfluencies when inferred through a model trained on Fluent data vs Fused-Fine tuning method for DialogSum dataset.

summarization by generating a summary that is semantically closer to the input dialog and the gold summary. In this specific example, the cascaded approach also appears to perform well. However, upon closer inspection, we observe that the cascaded model is impacted by the disfluency present in the last utterance of the dialog (*transform your weighting i mean to say going to transform*). This suggests that the cascaded model is limited by the performance of the disfluency removal model, and any errors in the disfluency removal process may propagate to the summarized output.

4.4 Evaluation on a real-life dataset

We also compare the performance of a model trained exclusively on fluent dialog data with a model trained using the Fused-Fine Tuning approach on a real-life proprietary dataset. The inhouse proprietary dataset contains real-life phone call conversations between a contact center agent and a customer in English language. These conversations naturally involve the disfluent segments. Analysis over a subset of 100 calls indicate that 87% of the calls contain disfluent segments involving one of restarts, repetitions or replacements. Transcripts of phone call conversations are obtained through human annotation to prevent the influence of transcription errors from Automatic Speech Recognition systems on summarization. Since DialogSum is a dialog dataset which can better represent the in-house dataset compared to the DebateSum corpus, we utilize the model trained on DialogSum corpus to evaluate on the in-house call dataset corpus. We evaluate on two tasks:

- *Call Summarization*: Given a call transcript as input, the task is to generate a summarized version of the call in a maximum of 100 words. The gold summary is obtained from manual annotation.
- *Speaker-Turn Summarization*: Given the transcript of a speaker-turn in the call, the task is to generate a summary of what the speaker said in a maximum of 50 words. The gold summary is obtained via manual annotations.

The evaluation results are presented in Table 4. From the results, we observe that for both tasks, the performance of the model trained in Fused Fine-Tuning setup performs better by 4 absolute points on Rouge-L and upto 10% relative gains on fluency, consistency and relevance than the model trained exclusively on fluent data. The observation corroborates our previous findings that a summarization model trained on fluent data is insufficient to infer on a disfluent input even in a real-life corpus.

4.5 How does different disfluency types affect the generated summary?

To delve deeper into the effect of different disfluency types on summarization models, we curated test sets of distinct categories of disfluencies: restart, repetition, and replacement. We then evaluate the performance of the summarization models on these disfluent test sets, comparing the results obtained from models trained on fluent data and models trained in a fused fine-tuning setup.

The experiment results in Table 5 indicate that disfluency of type *replacement* has the most negative impact on the summarization model, while the impact of *restarts* is relatively minor. The summarization model trained on fluent data shows a drop of 1.47 and 4.95 respectively in Rouge-L scores when evaluated on disfluent data of types restart and replacement. On the other hand, the mitigation strategy utilizing fused fine-tuning, which was exposed to disfluent data during training, demonstrates improvements in generating outputs in the presence of all types of disfluencies. The largest gain was observed in the replacements category, with the Rouge-L score improving by 3.28 absolute points and the fluency score rising from 0.879 to 0.932. Despite the improvements achieved through fused fine-tuning, further enhancements are still possible, particularly in the challenging category of replacement disfluency. These results emphasize the significance of considering different disfluency types in training and evaluating summarization models.

5 Related Works

Summarization is extensively studied with extractive (Nallapati et al., 2017; Zhong et al., 2020) and abstractive (Gerani et al., 2014; Cao et al., 2018) approaches being prominent, while some works utilize the best of both worlds in a hybrid approach (Pilault et al., 2020). Summarization has been explored in various document types, including short-texts (Cohan et al., 2018), dialogues (Zhong et al., 2022a), and medical conversations (Michalopoulos et al., 2022). While a number of works present summarization approaches tailored for conversations, to the best of our knowledge, the impact of disfluency on the summarization models is overlooked in the studies conducted so far.

Disfluency is widely studied in linguistics (Dammalapati et al., 2021), psychology (Eitel et al., 2014; Pieger et al., 2017), and speech technology (Hassan et al., 2014; Mendelev et al., 2021). Works have focused on disfluency detection in speech (Kourkounakis et al., 2020) and spoken transcripts (Dong et al., 2019; Kundu et al., 2022). Disfluencies have been shown to present challenges in NLP tasks, leading to the proposal of automatic disfluency removal systems (Wang et al., 2010; Saini et al., 2021). Studies by Dao et al. (2022b) and Gupta et al. (2021) highlight the negative impact of disfluency on downstream tasks of intent detection, slot filling and question-answering tasks.

Our work falls in a similar category and is centered on investigating the influence of disfluency in summarization. To the best of our knowledge, this is the first study to examine how disfluencies and types of disfluencies affect various evaluation criteria for summarization. It is to be noted that our primary objective is not to propose a state-of-the-art summarization model but to assess the impact of the presence of disfluency and investigate the effectiveness of simple mitigation techniques to alleviate its negative effects. We aspire to shed light on this overlooked area and encourage further research to develop robust summarization systems capable of handling all types of disfluencies.

6 Conclusion and Future Works

In this work, we shed light on the adverse impact of the presence of disfluency on the performance of summarization models, for both abstractive and extractive summaries. As a mitigation strategy to lower the impact of disfluency, we investigate the approaches of Cascaded Pipeline and Fused-Fine tuning and observed that the cascaded pipeline which first removes the disfluent segments before passing it to the summarization model has a relatively inferior performance compared to fine-tuning a summarization model on a dataset containing both fluent and disfluent input. Our investigation reveals that the mitigation strategy exhibits a relatively inferior performance on extractive summaries and on handling disfluency of the type 'replacement'. Hence in future work we would like to carry out a more focused effort on mitigating the impact of disfluency for such cases.

Limitations

While the proposed mitigation strategies have demonstrated improvements in the quality of both abstractive and extractive summaries, there exists limitations to this approach. The larger performance gaps in extractive summarization compared to Oracle performance suggest that more research is needed to address the challenges of disfluency in extractive summarization.

Furthermore, our results indicate that the proposed method performs worse for disfluencies of the "replacement" type. This suggests that datasets or specific datapoints with a higher distribution of "replacement" disfluencies may result in a larger performance gap compared to the Oracle performance. Therefore, the effectiveness of the proposed mitigation strategies may be dependent on the type and frequency of disfluencies present in the data. It is important to keep these limitations in mind when applying the proposed method to new datasets or use cases. Additionally, the cascaded approach is limited by the performance of the disfluency removal model, while fused-fine tuning approach involves using LARD algorithm to induce disfluency. Thus, availability and quality of these external models can be a limitation of applying either of the two mitigation approaches.

References

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 4784–4791. AAAI Press.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. Dialogsum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 5062–5074. Association for Computational Linguistics.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 615–621. Association for Computational Linguistics.

Samvit Dammalapati, Rajakrishnan Rajkumar, and Sumeet Agarwal. 2021. Effects of duration, locality, and surprisal in speech disfluency prediction in english spontaneous speech. In *Proceedings of the Society for Computation in Linguistics 2021*, pages 91–101.

Mai Hoang Dao, Thinh Hung Truong, and Dat Quoc Nguyen. 2022a. Disfluency detection for vietnamese. In *Proceedings of the Eighth Workshop on Noisy User-generated Text, W-NUT@COLING 2022, Gyeongju, Republic of Korea, October 12 - 17, 2022*, pages 194–200. Association for Computational Linguistics.

Mai Hoang Dao, Thinh Hung Truong, and Dat Quoc Nguyen. 2022b. From disfluency detection to intent detection and slot filling. In *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, pages 1106–1110. ISCA.

Qianqian Dong, Feng Wang, Zhen Yang, Wei Chen, Shuang Xu, and Bo Xu. 2019. Adapting translation models for transcript disfluency detection. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6351–6358. AAAI Press.

Alexander Eitel, Tim Kuehl, Katharina Scheiter, and Peter Gerjets. 2014. Disfluency meets cognitive load in multimedia learning: Does harder-to-read mean better-to-understand? *Applied Cognitive Psychology*, 28(4):488–501.

Shima Gerani, Yashar Mehdad, Giuseppe Carenini, Raymond T. Ng, and Bita Nejat. 2014. Abstractive summarization of product reviews using discourse structure. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1602–1613. ACL.

Sreyan Ghosh, Sonal Kumar, Yaman Kumar, Rajiv Ratn Shah, and Srinivasan Umesh. 2022. Span Classification with Structured Information for Disfluency Detection in Spoken Utterances. In *Proc. Interspeech 2022*, pages 3998–4002.

Aditya Gupta, Jiacheng Xu, Shyam Upadhyay, Diyi Yang, and Manaal Faruqui. 2021. Disfl-qa: A benchmark dataset for understanding disfluencies in question answering. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 3309–3319. Association for Computational Linguistics.

Hany Hassan, Lee Schwartz, Dilek Hakkani-Tür, and Gökhan Tür. 2014. Segmentation and disfluency removal for conversational speech translation. In *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, pages 318–322. ISCA.

Tedd Kourkounakis, Amirhossein Hajavi, and Ali Etemad. 2020. Detecting multiple speech disfluencies using a deep residual network with bidirectional long short-term memory. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pages 6089–6093. IEEE.

Rohit Kundu, Preethi Jyothi, and Pushpak Bhattacharyya. 2022. Zero-shot disfluency detection for indian languages. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 4442–4454. International Committee on Computational Linguistics.

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, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. 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.

Valentin Mendelev, Tina Raissi, Guglielmo Camporese, and Manuel Giollo. 2021. Improved robustness to disfluencies in rnn-transducer based speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*, pages 6878–6882. IEEE.

George Michalopoulos, Kyle Williams, Gagandeep Singh, and Thomas Lin. 2022. Medicalsum: A guided clinical abstractive summarization model for generating medical reports from patient-doctor conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 4741–4749. Association for Computational Linguistics.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3075–3081. AAAI Press.

Tatiana Passali, Thanassis Mavropoulos, Grigorios Tsoumakas, Georgios Meditskos, and Stefanos Vrochidis. 2022. LARD: large-scale artificial disfluency generation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 2327–2336. European Language Resources Association.

Elisabeth Pieger, Christoph Mengelkamp, and Maria Bannert. 2017. Fostering analytic metacognitive processes and reducing overconfidence by disfluency: the role of contrast effects. *Applied Cognitive Psychology*, 31(3):291–301.

Jonathan Pilault, Raymond Li, Sandeep Subramanian, and Chris Pal. 2020. On extractive and abstractive neural document summarization with transformer language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 9308–9319. Association for Computational Linguistics.

Allen Roush and Arvind Balaji. 2020. DebateSum: A large-scale argument mining and summarization dataset. In *Proceedings of the 7th Workshop on Argument Mining*, pages 1–7, Online. Association for Computational Linguistics.

Nikhil Saini, Drumil Trivedi, Shreya Khare, Tejas I. Dhamecha, Preethi Jyothi, Samarth Bharadwaj, and Pushpak Bhattacharyya. 2021. Disfluency correction using unsupervised and semi-supervised learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 3421–3427. Association for Computational Linguistics.

Wen Wang, Gökhan Tür, Jing Zheng, and Necip Fazil Ayan. 2010. Automatic disfluency removal for improving spoken language translation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2010, 14-19 March 2010, Sheraton Dallas Hotel, Dallas, Texas, USA*, pages 5214–5217. IEEE.

Vicky Zayats, Trang Tran, Richard A. Wright, Courtney Mansfield, and Mari Ostendorf. 2019. Disfluencies and human speech transcription errors. In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 3088–3092. ISCA.

Tianyi Zhang, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6197–6208. Association for Computational Linguistics.

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2022a. Dialoglm: Pre-trained model for long dialogue understanding and summarization. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 11765–11773. AAAI Press.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022b. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 2023–2038. Association for Computational Linguistics.

A Appendix

A.1 Qualitative Examples

Qualitative examples on DialogSum and DebateSum dataset with output from summarization model in different settings. (Table 6, 7).

Fluent (Original Input)
#Person1#: Exactly! I'm going to cut out all that junk I eat. No more chips, no more soda, no more fried food. #Person2#: I've heard this one before. #Person1#: But this time I'm going to stick to it. I really mean it! Trust me, Carol, I'm going to be a new man in one year's time! #Person2#: Well, I guess we'll just have to wait and see. #Person1#: Thanks, honey, that was a great meal. I'm stuffed. Do we have any chips left? #Person2#: And you're going to completely transform your eating habits, right?
Disfluent Input
#Person1#: Exactly! It is very I 'm going to cut out all that junk I eat. No more no chips, no more soda, no more fried food. #Person2#: I've heard this one before. #Person1#: But this this time I 'm gonna I 'm going to stick to it . I really mean it! Trust me, Carol, I 'm going to be a new man in one year 's time! #Person2#: Well, I see that actually, I guess we 'll just have to wait and see. #Person1#: Thank, honey, that was a great meal. I'm stuffed. Do we have any chips left? #Person2#: And you 're going to completely transform your weighting i mean to say going to completely transform your eating habit, right?
Gold Summary
#Person1# decided to go on a diet for New Year's resolution. Carol doesn't believe #Person1# will stick to it.
Prediction of Fluent Model on Fluent Input (Oracle)
#Person1# tells Carol that he is going to change his eating habits. Carol is skeptical about that change.
Prediction of Fluent Model on Disfluent Input
#Person1# tells Carol that he is going to cut out all that junk he eats. Carol believes #Person1# will follow.
Cascaded Pipeline: Prediction of Disfluent Model on Disfluent Input
#Person1# tells Carol that he is going to cut out all that junk. Carol doubts he is transforming his weighting and eating habit.
Fused-Fine Tuning Approach: Prediction of Disfluent Model on Disfluent Input
#Person1# tells Carol that he's going to cut out all the junk food and completely transform his eating habits. Carol debates with him on that.

Table 6: An example from *DialogSum* to demonstrate the outputs of summarization model in different settings. The disfluent segments are highlighted in color. Inaccuracies in the summary is captured in color.

Fluent (Original Input)
Despite disagreement among scholars regarding the value of sunset dates generally, those in the renewable energy industry agree that sunsetting of the PTC has impacted the industry and that a permanent PTC would result in more long-term investment in renewable energy. n164 Despite the success of the PTC, the credit has not become a permanent feature of the Internal Revenue Code and has been subject to the current sunset trend in Congress. n165 When the PTC was originally adopted in 1992, the taxpayer could only receive the credit if the qualifying facility was placed in service after December 31, 1993 and before July 1, 1999. n166 The latter date was the sunset date, at which point Congress would decide whether to renew the PTC. n167 Taxpayers that placed a facility in service prior to the sunset date would enjoy the full ten-year credit period. n168
Disfluent Input
Despite disagreement among scholars regarding the value of sunset dates generally, those in the industry sector actually no in the renewable energy industry agree that sunsetting of the PTC has impacted the industry and that a permanent PTC would result in more long-term investment in renewable energy. n164 Despite the success of the credit I mean PTC , the credit has not become a permanent feature of the Internal Revenue Code and has been subject to the current sundown well I actually meant the credit has not become a permanent feature of the Internal Revenue Code and has been subject to the current sunset trend in Congress. n165 When the PVC was PTC was originally adopted in 1992, the taxpayer could only receive the credit if the qualifying facility was placed in service after December 31, 1993 and before July 1, 1999. n166 The latter date was the set date the sunset date , at which point Congress would decide whether to renew the PTC. n167 Taxpayers that placed a facility in service prior to the sunset date would enjoy the full ten-year credit period. n168
Gold Summary
Despite disagreement among scholars regarding the value of sunset dates generally those in the renewable energy industry agree that sunsetting of the PTC has impacted the industry and that a permanent PTC would result in more long-term investment Despite the success the credit has not become a permanent feature
Prediction of Fluent Model on Fluent Input (Oracle)
Despite disagreement among scholars regarding the value of sunset dates generally, those in the renewable energy industry agree that sunsetting of the PTC has impacted the industry and that a permanent PTC would result in more long-term investment in renewable energy the credit has not become a permanent feature of the Internal Revenue Code and has been subject to the current sunset trend in Congress
Prediction of Fluent Model on Disfluent Input
Despite disagreement among scholars regarding the value of sunset dates generally, those in the industry sector actually no sunsetting of the PTC has impacted the industry and that a permanent PTC would result in more long-term investment in renewable energy the PTC has not become a permanent feature of the Internal Revenue Code and has been subject to the current sunset trend in Congress The latter date was the set date, at which point Congress would decide whether to renew
Cascaded Pipeline: Prediction of Disfluent Model on Disfluent Input
Despite disagreement among scholars regarding the value of sunset dates generally, those in the renewable energy industry agree that sunsetting of the PTC has impacted the industry and that a permanent PTC would result in more long-term investment in renewable energy Despite the success of the credit , the credit has not become a permanent feature of the Internal Revenue Code and has been subject to the current sunset trend in Congress. The PTC was originally adopted in 1992 The latter date was the set date , at which point Congress would decide whether to renew
Fused-Fine Tuning Approach: Prediction of Disfluent Model on Disfluent Dialogue
Despite disagreement among scholars regarding the value of sunset dates generally, those in the renewable energy industry agree that sunsetting of the PTC has impacted the industry and that a permanent PTC would result in more long-term investment in renewable energy the credit has not become a permanent feature of the Internal Revenue Code and has been subject to the current sunset trend in Congress

Table 7: An example from *DebateSum* to demonstrate the outputs of summarization model in different settings. The disfluent segments are highlighted in color . Inaccuracies including inconsistent facts and irrelevant segments in the summary are captured in color .