Quantifying the Impact of Disfluency on Spoken Content Summarization

Maria Teleki, Xiangjue Dong, James Caverlee
Texas A&M University
College Station, Texas, USA
{mariateleki, xj.dong, caverlee}@tamu.edu

In LREC-COLING 2024

1 Minute Summary

Original

Hello and welcome to our podcast! Let's get right to it. Today we're going to be interviewing a very special guest, someone I know you guys have been excited about having on the show.

Repeats with N=3

Hello and welcome to our podcast! Let's get get get get right to it. Today we're going to be interviewing a a a a very special guest, someone I know you guys have been excited about having on the show.

Interjections with N=3

Hello and welcome to our podcast! Let's get right uh okay okay to it. Today we're going to be interviewing a very special um so I mean guest, someone I know you guys have been excited about having on the show.

False Starts with N=3

Hello and welcome to our podcast! Let's get right to it. Today we're today we're today we're today we're going to be interviewing a very special guest, someone I know you guys have been excited about having on the show.

- **Disfluencies** are a key characteristic of **spoken content**.
  - We study 3 types of disfluencies -- *repeats, interjections, and false starts* -- in terms of the **Shriberg disfluency definition**.[1]
- **Summarization** quality decreases with increased disfluency.
- We use a **parsing-based SOTA disfluency annotator**[2] to repair the disfluencies via removal and tagging.
- We find that training on the repaired transcripts ($train_R$) and testing on the original transcripts (test) yields the best results.

[1]Elizabeth Ellen Shriberg. 1994. Preliminaries to a theory of speech disfluencies. Ph.D. thesis.

[2]Paria Jamshid Lou and Mark Johnson. 2020. Improving disfluency detection by self-training a self-attentive model. In Association for Computational Linguistics, pages 3754–3763.

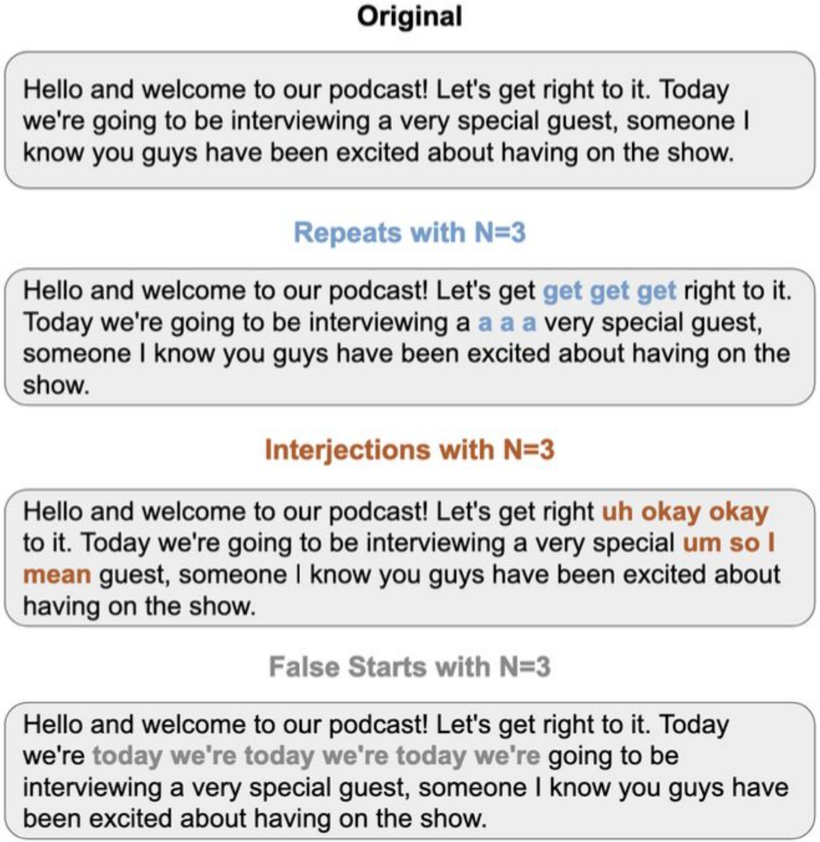
Quantifying the Impact of Disfluency on Spoken Content Summarization

Maria Teleki, Xiangjue Dong, James Caverlee
Texas A&M University
College Station, Texas, USA
{mariateleki, xj.dong, caverlee}@tamu.edu

In LREC-COLING 2024

What is a disfluency?

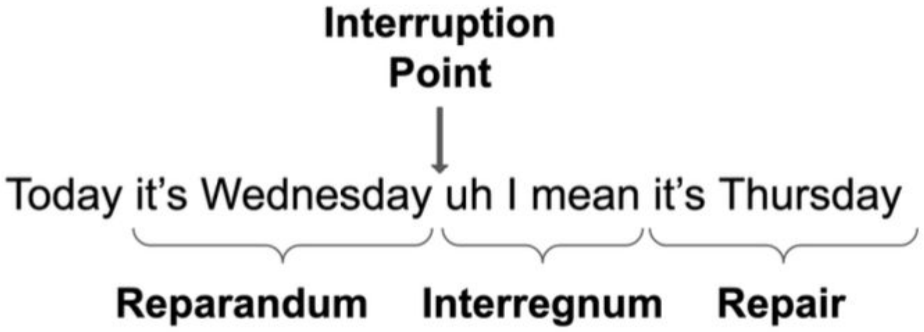
- Disfluencies are a key characteristic of **spoken content**.
- We study 3 types of disfluencies -- *repeats, interjections, and false starts* -- in terms of the **Shriberg disfluency definition.**[1]



[1]Elizabeth Ellen Shriberg. 1994. Preliminaries to a theory of speech disfluencies. Ph.D. thesis.

What is a disfluency? The Shriberg disfluency definition.[1]

- The reparandum and interregnum are removed to form a fluent sentence.
- *Repeats* and *false starts* occur within the **reparandum**.
- *Interjections* occur within the **interregnum**.



[1]Elizabeth Ellen Shriberg. 1994. Preliminaries to a theory of speech disfluencies. Ph.D. thesis.

Many important NLP tasks like summarization are often designed for written content rather than the looser, noiser, and more disfluent style of spoken content.[1,2,3,4]

[1]Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing, pages 3730–3740.
[2]Mike Lewis, Yinhan Liu, et al. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Association for Computational Linguistics, pages 7871–7880.
[3]Ani Nenkova and Kathleen McKeown. 2012. A survey of text summarization techniques. Mining text data, pages 43–76.
[4]Ramesh Nallapati, Bowen Zhou, et al. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In Conference on Computational Natural Language Learning, pages 280–290.

Research Questions

RQ1: How Do Disfluencies Impact Summarization Quality?

We synthetically inject disfluency events (repeats, interjections, false starts, and their combinations) at a range of severity levels and measure their impact on summarization quality.

RQ2: Can Summarization Quality be Improved By Directly Modeling Disfluency?

We explore the use of a state-of-the-art disfluency detection model to improve the summarization quality by either (1) removing the disfluencies, or (2) tagging the disfluencies.

RQ1: How Do Disfluencies Impact Summarization Quality?

The Spotify Podcasts Dataset[1]

- This dataset was originally used for the summarization task from the TREC 2020 Podcasts Track.[2]
- We use the test set for the summarization task, which consists of 1,027 podcasts. For each, we have:
  - The podcast transcript
  - The Show ID
  - The Episode ID
  - The creator-provided show description
  - The creator-provided episode description[2]
- We keep podcasts which have text occurring in their transcript in the first 60 seconds, which leaves us with 1,020 podcasts.

[1]Clifton, Ann and Reddy, Sravana et al. 2020. 100,000 podcasts: A spoken English document corpus.
[2]Rosie Jones, Ben Carterette, Ann Clifton, et al. 2020. TREC 2020 Podcasts Track Overview. In Text Retrieval Conference.

RQ1: How Do Disfluencies Impact Summarization Quality?

We inject disfluencies according to fixed distributions, similar to previous work:[1,2]

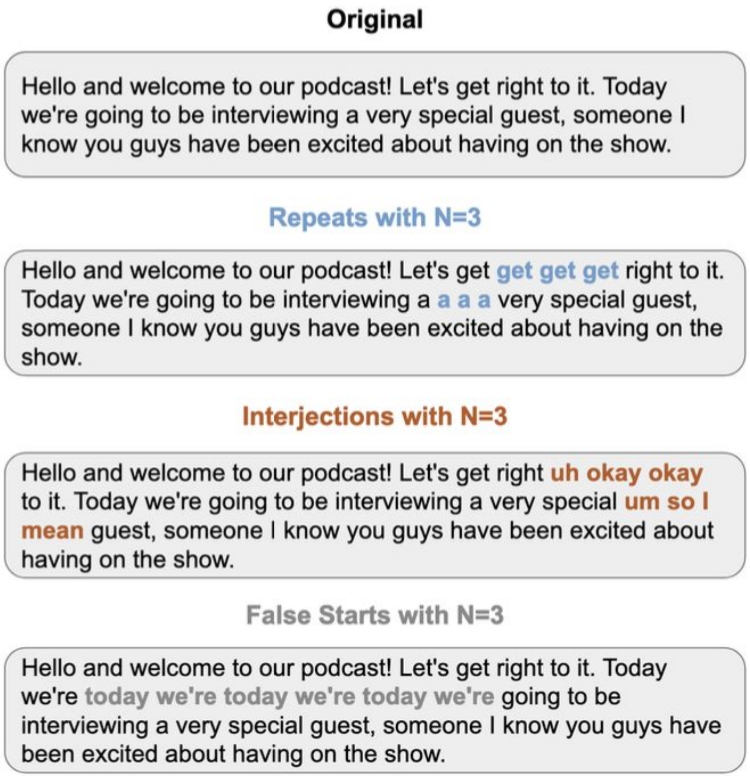
Repeats and Interjections

We sample from X~N (μ=10, σ=1) to determine the position at which the term(s) should be injected into the transcript N times.

- The interjections are uniformly randomly selected from: *uh, um, well, like, so, okay, I mean, you know.*

False Starts

Sentences >4 words are non-uniformly sampled with 80/20 probability with replacement, and the selected sentences have a false start (first 2 words of sentence) interjected N times.



[1]Shaolei Wang, Wangxiang Che, et al. 2020. Multi-task self-supervised learning for disfluency detection. In AAAI Conference on Artificial Intelligence, volume 34, pages 9193–9200.
[2]Tatiana Passali, Thanassis Mavropoulos, et al. 2022. LARD: Large-scale artificial disfluency generation. In Language Resources and Evaluation Conference, pages 2327–2336.

RQ1: How Do Disfluencies Impact Summarization Quality?

We consider 6 summarization models:

1min is the first minute of transcript text.[1]

cued_speechUniv2 is an ensemble of 3 BART models plus a hierarchical filtering model, and it is the top performer from the TREC 2020 Podcasts Track.[2]

BART is a sequence-to-sequence model with a bidirectional encoder and a left-to-right autoregressive decoder.

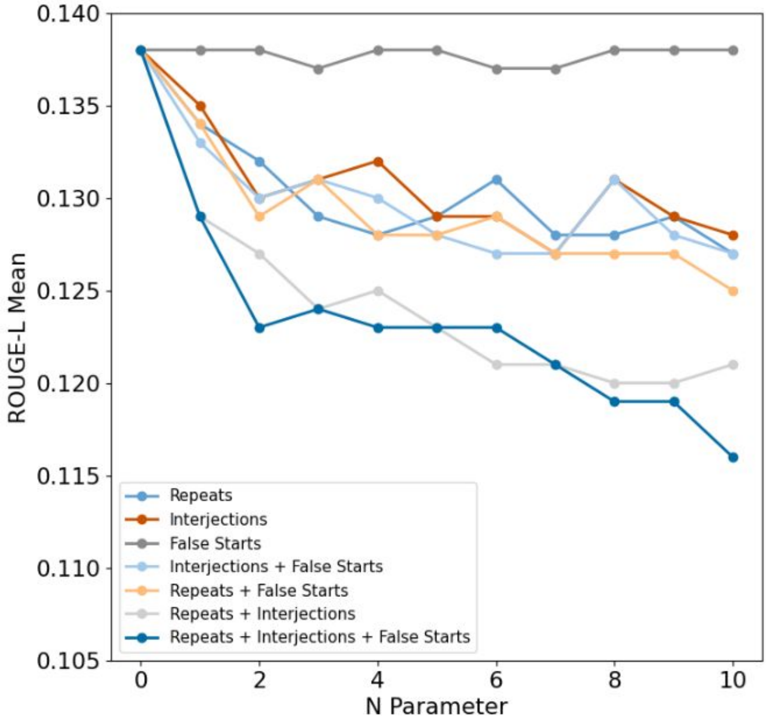
T5 is a text-to-text transformer model.

Pegasus is a transformer model with a pretraining objective called gap sentence generation.

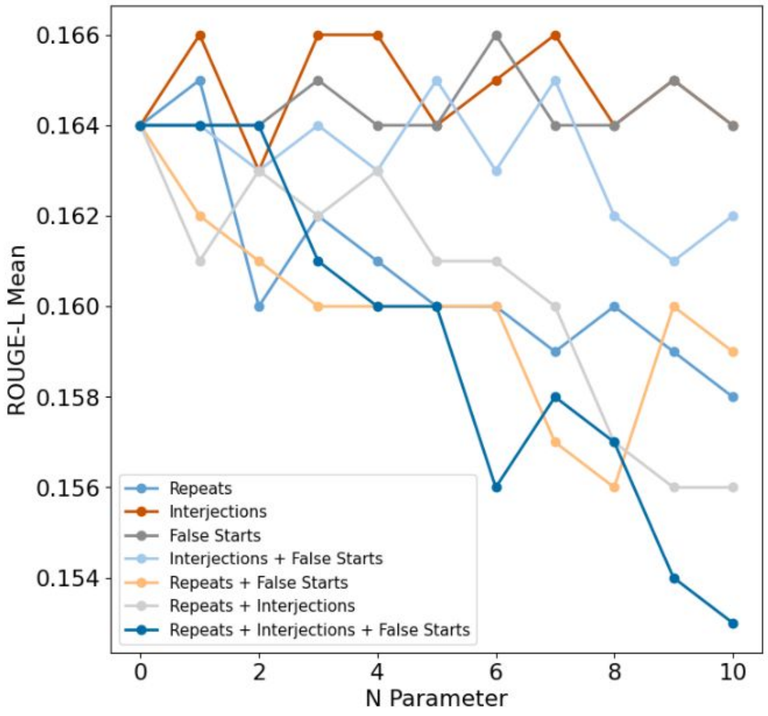
Llama 2-Chat is a large transformer model which is pretrained and specifically for chat settings using RLHF.

[1]Rosie Jones, Ben Carterette, Ann Clifton, et al. 2020. TREC 2020 Podcasts Track Overview. In Text Retrieval Conference.
[2]Potsawee Manakul and Mark Gales. 2020. Cued_speech at TREC 2020 podcast summarisation track. In Text Retrieval Conference.

RQ1: How Do Disfluencies Impact Summarization Quality?



(a) ROUGE-L over increased N on **BART** model.



(b) ROUGE-L over increased N on **cued_speechUniv2** model.

RQ1: How Do Disfluencies Impact Summarization Quality?

Model	N=0	N=2 vs. N=0	R	I	F	I+F	R+F	R+I	R+I+F
1min	0.124	$N_2 - N_0$ Δ	-0.013 (-10.3%)	-0.014 (-11.6%)	-0.004 (-3.2%)	-0.017 (-14.0%)	-0.016 (-13.0%)	-0.026 (-21.0%)	-0.029 (-23.7%)
BART	0.138	$N_2 - N_0$ Δ	-0.006 (-4.6%)	-0.008 (-5.5%)	0.000 (-0.3%)	-0.008 (-5.7%)	-0.008 (-6.1%)	-0.011 (-7.6%)	-0.015 (-11.1%)
T5	0.134	$N_2 - N_0$ Δ	-0.018 **(-13.7%)**	-0.010 (-7.4%)	-0.003 (-2.4%)	-0.013 (-9.9%)	-0.018 **(-13.7%)**	-0.025 **(-19.0%)**	-0.032 **(-23.7%)**
Pegasus	0.131	$N_2 - N_0$ Δ	-0.011 (-8.8%)	-0.014 **(-10.4%)**	-0.003 **(-2.6%)**	-0.017 **(-12.9%)**	-0.014 (-10.7%)	-0.023 (-17.2%)	-0.026 (-19.9%)
cued_speechUniv2	0.164	$N_2 - N_0$ Δ	-0.004 (-2.5%)	-0.001 (-0.8%)	-0.001 (-0.5%)	-0.001 (-0.8%)	-0.003 (-1.9%)	-0.002 (-1.0%)	-0.001 (-0.5%)
Llama 2-Chat	0.129	$N_2 - N_0$ Δ	-0.001 (-0.6%)	-0.002 (-1.2%)	-0.001 (-1.0%)	-0.002 (-1.6%)	-0.001 (-1.1%)	-0.001 (-1.1%)	-0.002 (-1.5%)

- **Overall drop in ROUGE-L** with increased N.
- T5 and Pegasus are the least resilient in the presence of disfluencies, BART is moderately resilient, and cued_speechUniv2 and Llama 2-chat are the most resilient.

RQ2: Can Summarization Quality be Improved By Directly Modeling Disfluency?

We use a state-of-the-art, parsing-based disfluency annotation model[1] (Equations 1 and 2) to transform the transcripts via:

- **Repairing**: Removal of words marked disfluent.
- **Tagging**: Tagging (<DIS> and <\DIS>) of words marked disfluent.

$$s(T) = \sum_{(i,j,l)\in T} s(i,j,l) \qquad (1)$$

$$\hat{T} = \underset{T}{\text{argmax}}\ s(T) \qquad (2)$$

[1]Paria Jamshid Lou and Mark Johnson. 2020. Improving disfluency detection by self-training a self-attentive model. In Association for Computational Linguistics, pages 3754–3763.

RQ2: Can Summarization Quality be Improved By Directly Modeling Disfluency?

- Simply using the test set as-is yields the best ROUGE scores in most cases.
  - However, Pegasus is more robust in the face of missing information, and benefits from having the disfluencies removed.

Inference-Only

Model	Test	Rouge-L	Rouge-1	Rouge-2
BART	test_R	0.137	0.211	0.053
	test	**0.138**	**0.212**	**0.054**
	test_T	0.137	0.209	0.052
Pegasus	test_R	**0.131**	**0.200**	0.047
	test	**0.131**	0.198	**0.049**
	test_T	0.113	0.169	0.038
T5	test_R	0.133	0.194	0.050
	test	**0.134**	**0.199**	**0.051**
	test_T	0.126	0.181	0.048

RQ2: Can Summarization Quality be Improved By Directly Modeling Disfluency?

train	test	BART			T5			Pegasus		
		R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2
	$test_R$	0.172	0.240	0.085	0.145	0.197	0.059	0.129	0.174	0.049
$train_R$	test	**0.177**	**0.244**	**0.090**	0.146	0.196	0.060	**0.131**	0.177	**0.052**
	$test_T$	0.174	0.241	0.086	0.148	0.198	0.063	0.096	0.133	0.037
	$test_R$	0.170	0.236	0.083	0.146	0.198	0.060	0.122	0.165	0.045
train	test	0.175	0.242	0.088	**0.149**	**0.200**	0.062	0.126	0.169	0.049
	$test_T$	0.172	0.238	0.085	0.147	0.194	**0.065**	0.090	0.124	0.032
	$test_R$	0.172	0.238	0.083	0.142	0.193	0.057	0.129	**0.193**	0.048
$train_T$	test	0.173	0.240	0.085	0.143	0.194	0.057	0.127	**0.193**	0.047
	$test_T$	0.169	0.235	0.081	0.145	0.196	0.058	0.115	0.146	0.038

We find that training on the repaired transcripts ($train_R$) and testing on the original transcripts (test) yields the best results.

Link to our code on GitHub!

Conclusion

- **Disfluencies** are a key characteristic of **spoken content**.
  - We study 3 types of disfluencies -- *repeats, interjections, and false starts* -- in terms of the **Shriberg disfluency definition**.[1]
- We **synthetically inject disfluencies (N)** and find that **summarization quality decreases with increased disfluency**.
  - Decreases the most with combinations of the 3 disfluency types.
- We use a **parsing-based SOTA disfluency annotator**[2] to repair the disfluencies via removal and tagging.
- We find that for inference: Simply using the test set as-is yields the best ROUGE scores in most cases.
  - Pegasus is more robust in the face of missing information, and benefits from having the disfluencies removed.
- We find that for fine-tuning + inference: Training on the repaired transcripts ($train_R$) and testing on the original transcripts (test) yields the best results.

[1] Elizabeth Ellen Shriberg. 1994. Preliminaries to a theory of speech disfluencies. Ph.D. thesis.

[2] Paria Jamshid Lou and Mark Johnson. 2020. Improving disfluency detection by self-training a self-attentive model. In Association for Computational Linguistics, pages 3754–3763.

Quantifying the Impact of Disfluency on Spoken Content Summarization

Maria Teleki, Xiangjue Dong, James Caverlee
Texas A&M University
College Station, Texas, USA
{mariateleki, xj.dong, caverlee}@tamu.edu

In LREC-COLING 2024