ICASSP 2024 KOREA

SLP-P10.4

AUGSUMM: TOWARDS GENERALIZABLE SPEECH SUMMARIZATION USING SYNTHETIC LABELS FROM LARGE LANGUAGE MODELS

Jee-weon Jung*[1], Roshan Sharma§*[1], William Chen[1], Bhiksha Raj[1,2], Shinji Watanabe[1]

[1]Carnegie Mellon University, USA

[2]Mohamed bin Zayed University of AI, Abu Dhabi

*: Equal contribution

§: Now at Google

Overview

❖SSUM Models today use a **single reference** summary for training and evaluation, but there exists a **distribution of multiple valid summaries** given an audio recording.

❖Human annotation to cover the distribution is costly.

❖This paper

- Presents AugSumm – a method to automatically generate and use multiple references in training and evaluation
- Proposes different methods to incorporate generated augmentation summaries into training and testing

Key Findings

❖ChatGPT can generate synthetic references by paraphrasing existing references or generating extractive summaries from source transcripts based on extensive evaluation

❖E2E models trained with synthetic references outperform those trained with a single reference and produce more diverse summaries.

❖ Combination of multi-style training and pre-train fine-tune leads to best performance on evaluation sets

Overall Framework

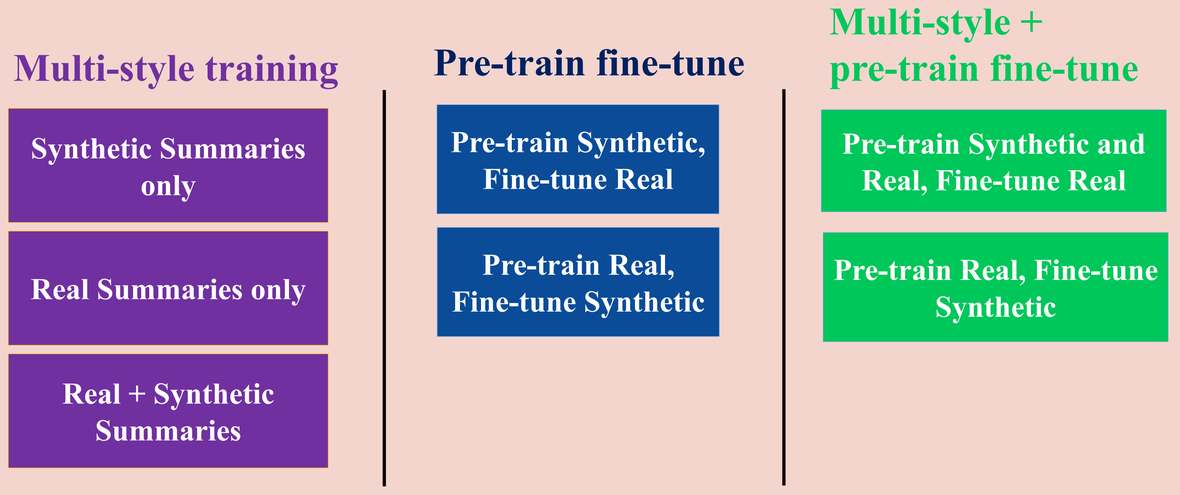
1. Identify sources of variations in summarization

- Summary structure given semantic concepts (**Paraphrase AugSumm**) – paraphrase existing references
- Semantic concepts within the summary (**Direct AugSumm**) – obtain extractive summaries from source transcript

2. Generate synthetic references by prompting ChatGPT

- **Paraphrase AugSumm**
  - *You are here to paraphrase a given summary in the same style as the provided input. Please make sure that the summary has between 40 to 60 words. Also please include these words in the summary:* ***{important keys}. given summary****:*
- **Direct AugSumm**
  - *You are here to create an extractive summary from the transcript. An extractive summary uses words from the input to convey the important portions of the video. Please make sure that the summary has between 40 and 60 words. Respond with only the extractive summary for:* {transcription}. ***transcription****:*

3. Use synthetic references to train models



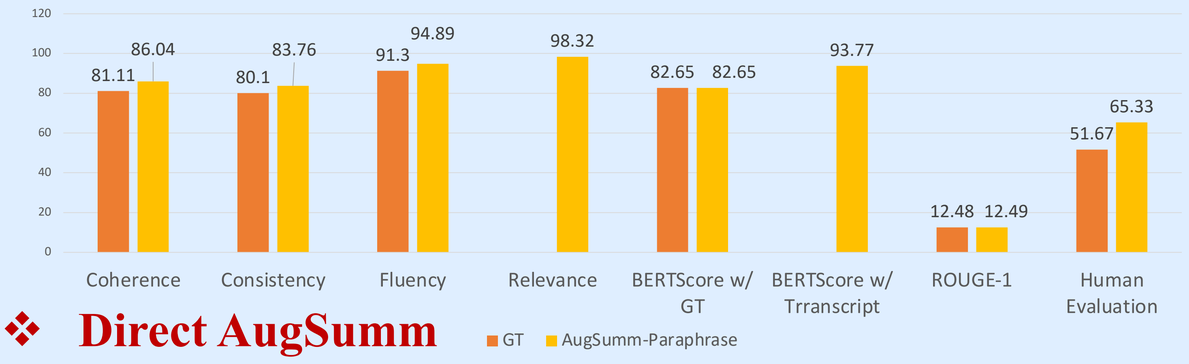
❖ Validate on How2 (2000h of instructional videos) using UniEval, BERTScore, and human preference evaluation

❖ Attention based encoder-decoder with Fourier-based self-attentions for 43-dim fbank-pitch features

Synthetic Data Quality Evaluation

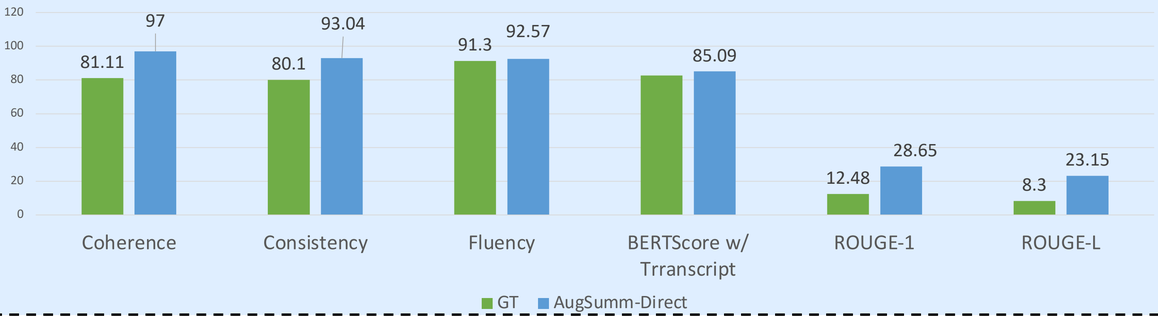
❖ Paraphrase AugSumm

- Synthetic data is better on semantic metrics
- Human evaluation with 20 annotators, 15 questions to evaluate preference for synthetic versus real references
- Most humans prefer synthetic over the existing reference



❖ Direct AugSumm

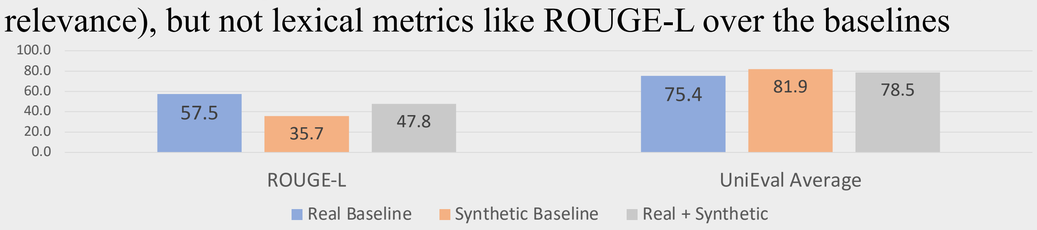
- Better on both lexical and semantic metrics



Experimental Results

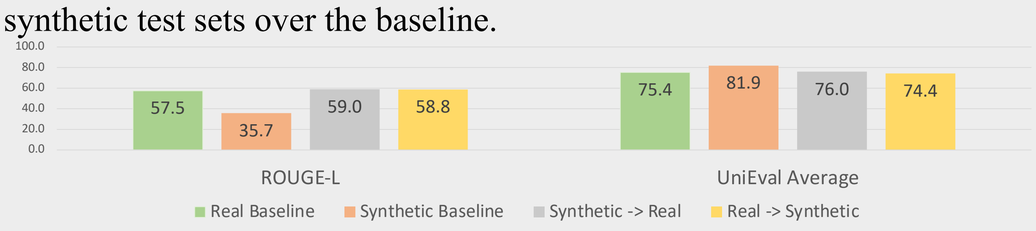
❖Multi-style Training

- Multi-style training improves UniEval (coherence, consistency, fluency and relevance), but not lexical metrics like ROUGE-L over the baselines



❖Pre-train Fine-tune

- Pre-train on Synthetic Fine-tune on Real improves all metrics on real and synthetic test sets over the baseline.



❖Multi-style Training + Pre-train fine-tune paradigm

- Pre-training with Real+Synthetic and fine-tuning on real produces the best ROUGE on real and synthetic test sets, with improvement in UniEval.