SumREN: Summarizing Reported Speech about Events in News

Revanth Gangi Reddy[1], Heba Elfardy[2], Hou Pong Chan[3], Kevin Small[2], Heng Ji[2]

[1]University of Illinois at Urbana-Champaign [2]Amazon Alexa AI [3]University of Macau

Introduction

Motivation

- Existing news summarization primarily **focuses on event details** and **ignores reported speech** which is important in precisely establishing public figures' stance, opinions, and worldviews.
- Reported speech from the same speaker can be scattered across the entire article, thus requiring **modeling of long-term dependencies** and co-reference resolution.
- Concisely summarizing a set of reported statements requires a **higher level of abstraction, with factual consistency** being paramount as misrepresenting statements from public figures can be harmful.

Contributions

- A **new challenging task** of reported speech summarization and a corresponding **multi-document summarization** benchmark, SumREN
- Empirical demonstration that large language models can create **cost-efficient silver-standard training data** for summarization.
- A **pipeline-based reported speech summarization** framework and showing that it generates summaries that are **more abstractive and factually consistent** than query-focused summarization approaches.

SumREN Benchmark

- Given a set of news articles about a specific event and the speaker name, the goal is to generate a succinct summary for the statements made by the speaker in the source content.
- SumREN contains 745 examples in total, with a train/dev/test split of 235/104/406 respectively. On average, the summaries have a length of 57 words and each summary comes from 5.3 reported statements.

Step 1: Identifying salient spans in statements
While Republicans look inward at the aftermath of the Capitol Hill riots after President Trump's address Wednesday, Democrats are adding to the division, Fox News contributor Charles Hurt told "Fox & Friends."
"I get this rush to want to blame everything on President Trump. Everything that is going on right now has been in the making for years and decades, of which politicians on Capitol Hill have been a part," Hurt, Washington Times opinion editor, told co-host Brian Kilmeade.
He added: "The last thing they want to do is take stock of themselves and try to figure out, 'OK, what have I done to make this worse or to create this situation?'"
Within seconds of reconvening Wednesday night, Democrats on Capitol Hill started "accusing Republicans of treason and sedition," Hurt said.
"They get caught up in their own mob mentality, they're all trying to outdo one another on Twitter to see who can make the most outrageous charge or make the most outrageous demand of the other side," Hurt said.
While this might be a good time for soul-searching for both parties, Hurt concluded, "There is no indication from Democrats on Capitol Hill that any one of them has any intention of doing that and certainly not from Joe Biden or Kamala Harris."

Step 2: Grouping salient spans into sentences.
"blame everything on President Trump" + "accusing the Republicans of treason and sedition" → *"blame everything on President Trump and accuse the Republicans."*
"Democrats are adding to the division" + "they get caught up in their own mob mentality, they're all trying to outdo one another" → *"Democrats get caught up in trying to outdo one another and are adding to the division."*
"last thing they want to do is take stock of themselves" + "this might be a good time for soul-searching" + "no indication from Democrats that any one of them has any intention of doing that" → *"They don't seem to have any intention of doing any soul-searching"*

Step 3: Combining sentences into a summary
Charles Hurt suggested that Democrats are rushing to blame everything on President Trump and accuse the Republicans. He said that Democrats get caught up in trying to outdo one another and are adding to the division. Finally, they don't seem to have any intention of doing any soul-searching.

Figure 1: Walk-through example showing the process of annotating summaries given a set of reported statements. Salient spans within the statements are shown in red and sentences copied over from step 2 into the summary in step 3 are shown in blue.

Comparison to Existing Datasets

- Our summaries are human-written, compared to most news summarization datasets, which are directly scraped from the web.
- SumREN has considerably more abstractive summaries, compared to existing datasets.

Datasets	unigram	bigram	trigram	4-gram
CNN-DM (S)	17.0	53.9	72.0	80.3
NY Times (S)	22.6	55.6	71.9	80.2
MultiNews (M)	17.8	57.1	75.7	82.3
WikiSum (M)	18.2	51.9	69.8	78.2
SumREN (M)	16.8	63.1	86.4	93.4

Table 1: % of novel *n*-grams in the reference summaries of different summarization datasets.

Pipeline-based Reported Speech Summarization

Our pipeline-based reported speech summarization framework involves the following steps:

- **Reported speech extraction** for identifying reported statements and their speakers from the given set of news articles
- **Speaker co-reference resolution** for grouping statements together that come from the same speaker.
- **Summarization** for generating a concise summary of the grouped statements. The statements are concatenated and passed as input to a BART model.

Experiments

Overall performance on SumREN

- We explore end-to-end Query-Focused Summarization (QFS) and pipeline-based reported speech summarization which first extracts the reported statements.
- Training on silver-standard GPT-3 generated data considerably improves performance of both QFS and pipeline-based approaches.

Setting	Model	Approach	Rouge-L	BertScore	MINT
Baselines (Zero-shot)	SegEnc	QFS	22.99	23.26	11.1
	GPT-3		26.72	31.16	38.9
Zero-shot	BART	Pipeline	24.45	29.36	15.3
	GPT-3	Pipeline	29.33	**37.68**	**49.6**
	GPT-3	Pipeline (Oracle)	31.27	40.29	51.3
+ Silver Training	SegEnc	QFS	**29.69**	36.26	31.2
	BART	Pipeline	28.66	34.55	32.9
+ Gold Finetuning	SegEnc	QFS	29.43	36.71	38.4
	BART	Pipeline	29.62	35.72	43.5
	BART	Pipeline (Oracle)	32.20	39.61	44.0

Table 2: We explore both query-focused (QFS) and pipeline-based approaches under zero-shot, silver-training and gold-fine-tuning settings. *Pipeline (Oracle)* corresponds to using the gold reported statements as input to the summarization model and is reported for the best setup for each of the zero-shot and fine-tuned models.

Abstractiveness and Factual Consistency of Generated Summaries

- Goal of any abstractive summarization system is to generate more abstractive summaries while maintaining a high level of factual consistency with the source.
- Pipeline-based approach maintains factuality with better abstractiveness, meaning explicit statement extraction helps the summarization model focus on paraphrasing and synthesizing.

Approach	Model	Factual Consistency		Abstractiveness	
		FactCC	Entity Prec.	MINT	Trigram
QFS	GPT-3	45.4	61.7	38.9	44.2
	SegEnc	50.8	**75.4**	38.4	46.6
Pipeline	GPT-3	50.2	73.2	**49.6**	**56.6**
	BART	**52.1**	74.6	43.5	52.1
Pipeline (Oracle)	GPT-3	52.0	78.9	51.3	58.1
	BART	55.0	84.6	44.0	52.3

Table 3: Comparison of factuality (measured by FactCC and Entity Precision) of generated summaries relative to abstractiveness (measured by MINT and novel trigrams). Models considered are after silver train + gold FT, except GPT-3 which is not fine-tuned.

Future Directions

- Upon aligning source-summary pairs, we found that human summaries cover considerably more percentage of the input reported statements compared to model summaries.
- Incorporate more control into summarization for improving coverage, by clustering the salient spans within the reported statements and separately generating summaries for each cluster.
- Considerable room for improving reported speech extraction, along with better speaker co-reference resolution. Incorporating character-level features can make co-reference resolution more robust to name aliases.