Reducing Quantity Hallucinations in Abstractive Summarization

Zheng Zhao Shay B. Cohen Bonnie Webber

School of Informatics, University of Edinburgh

zheng.zhao@ed.ac.uk scohen@inf.ed.ac.uk bonnie.webber@ed.ac.uk

Overview

- Neural-based abstractive summarization systems often include material that is not supported by the original text (aka **hallucinations**).
- We try to avoid (or at least reduce) hallucinations by verifying that entities in the summary appear in the original text in a similar context.
- Our system, HERMAN, learns to recognize and verify quantity entities (e.g., dates, numbers, sums of money, etc.) in abstractive summaries, in order to up-rank summaries whose quantity terms are supported by the original text.
- Up-ranked summaries are shown to have higher Precision, without loss in Recall, resulting in higher F_1.
- Human evaluation of up-ranked summaries shows that subjects prefer them to summaries that have not been up-ranked.

Article: ...the volcano was still spewing ash on Sunday ...More than a dozen people were killed when it erupted in 2014 ...rescue teams are still scouring the area, looking for more victims who ...
Summary: Rescue teams in Indonesia are searching for more than 20 people missing after the Mount Sinabung volcano erupted on Saturday, killing at least 11 people and injuring at least 20 others.

Article: The government and the doctors' union have agreed to continue negotiating until Wednesday. The talks, hosted by conciliation service Acas ...
Summary: Talks aimed at averting the imposition of a new junior doctors' contract in England have been extended for a second day.

Table 1: Examples of hallucinated quantities. Phrases highlighted in cyan are facts, whereas red highlighting indicates hallucinations.

Dataset Generation

- The dataset comprises the XSum dataset [1], augmented with negative examples and additional labels:
  - A summary-level label $z \in \{\texttt{VERIFIED}, \texttt{UNVERIFIED}\}$;
  - A sequence of labels $Y = (y_1, \ldots, y_n)$ where $y_i \in \{\texttt{B-V}, \texttt{B-U}, \texttt{I-U}, \texttt{I-V}, \texttt{O}\}$, indicating token is Verified, Unverified, or Other;
  - A sequence of binary labels $M = (m_1, \ldots, m_n)$ indicating the location of quantity entities in the summary.
- A gold summary in the XSum dataset is labelled `VERIFIED`.
- We replace quantity entities in the summary with randomly selected entities from the article to get `UNVERIFIED` summary.
- Table 2 illustrates an example of `VERIFIED` summary with its labels and corresponding article.

Article	The crash happened at Evanton at about 17:20 on Saturday. The fire service and the air ambulance was sent to the scene. The occupants of all three vehicles were injured, but the extent of their injuries was not known, police said. A spokesman added: "Inquiries are ongoing into this matter and no further witnesses are sought at this time" ...										
Summary	Several	people	have	been	injured	in	a	three-car	collision	on	...
Y **labels**	B-V	O	O	O	O	O	O	B-V	O	O	...
M **labels**	1	0	0	0	0	0	0	1	0	0	...
z **label**	VERIFIED										

Table 2: An example from our dataset. Cyan text highlights the support in the source document for the quantity token highlighted green in the summary.

Verification Model

- The BiLSTM encoder provides hidden representations for input.
- The BiLSTM decoder with attention generates the context vector.
- The context vectors from every token in the summary are fed into a Conditional Random Fields layer to get the tag sequence Y.
- The same context vectors are fed into a MLP classifier to get the binary label z.
- Note that the binary classifier for predicting whether a summary is verified (z labels) is omitted in the provided figure.

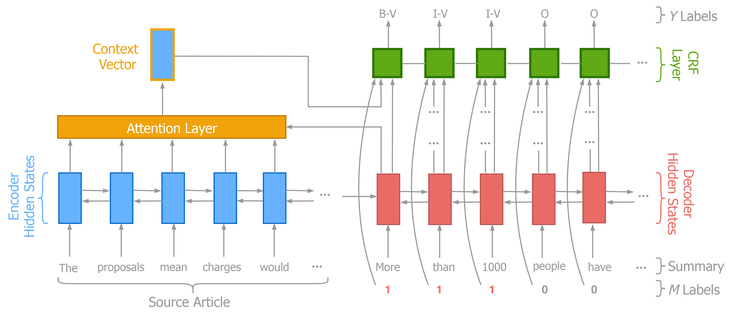

Figure 1: Architecture of our verification model HERMAN.

Re-ranking to Avoid Hallucination

- We leverage predictions of HERMAN to give a verification score to each summary from the list of candidate summaries generated using beam search for a given document.
- HERMAN-GLOBAL uses the raw output of global verification label z which has a real value between $[0, 1]$.
- HERMAN-LOCAL uses the average probabilities of `B-V`, `B-U`, `I-V`, and `I-U` labels where entries of `B-U` and `I-U` are counted negatively.
- The summary with the highest verification score is selected as the final generated summary.
- We also introduce two baseline re-ranking approaches: selecting the shortest summary, and selecting the summary with maximum quantity entity overlap with the source document.

	Model	R1-F	R2-F	RL-F	avg-Q
BERTSUM	Baseline-shortest	38.71	16.38	31.16	0.62
	Baseline-max-overlap	39.01	16.58	31.24	0.76
	Original	38.86	16.38	31.04	0.65
	HERMAN-LOCAL	38.63	16.12	30.75	0.79
	HERMAN-GLOBAL	**39.06**	**16.65**	**31.36**	**0.81**

Table 3: Automatic evaluation on the XSum test set. avg-Q denotes the average number of quantity entities per summary.

Results and Discussion

- Overall, ROUGE-1/2/L F_1 score for up-ranked summaries exceeds that of original summaries.
- HERMAN-GLOBAL achieves highest avg-Q for BERTSUM.
- Results for TCONVS2S and BART can be found in our paper.
- Together with ROUGE, this indicates that our model both encourages the inclusion of quantity entities and includes them correctly.
- Human evaluation on quantity faithfulness shows that up-ranked summaries are preferred over the original summaries.

References

[1] Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 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*, pages 1797–1807, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.