Analyzing and Evaluating Faithfulness in Dialogue Summarization

Bin Wang[1], Chen Zhang[1], Yan Zhang[1], Yiming Chen[1], Haizhou Li[1,2,3]

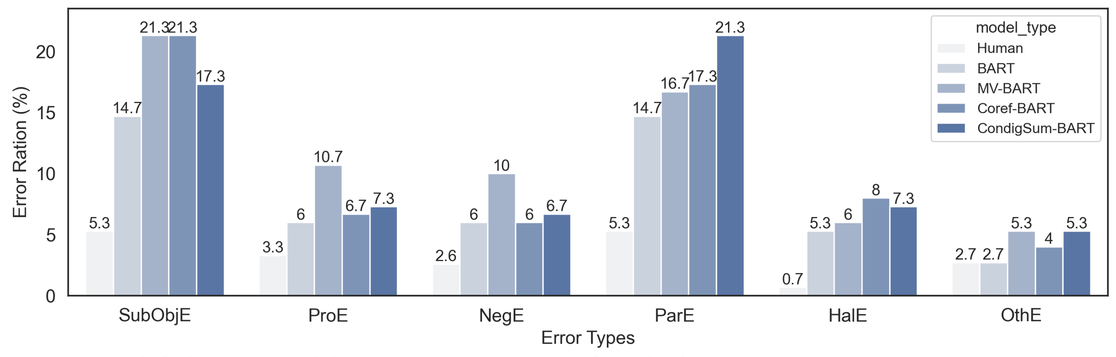
[1]National University of Singapore, Singapore

[2]The Chinese University of Hong Kong, Shenzhen, China [3]Kriston AI

Abstract

Faithfulness hinders the application of dialogue summarization systems

- How sever this problem is?
  - Human Evaluation with SOTA models
- How can we make improvement?
  - Evaluation of Faithfulness
  - Faithfulness problem is very severe for dialogue summarization
    - Human-written summaries also contains factual errors
    - 36.7%~50% generated summaries are with at least one factual error
    - Recent dialogue summarization models worse than their baseline
    - Our taxonomy covers most error types (>95%)



Background

Task

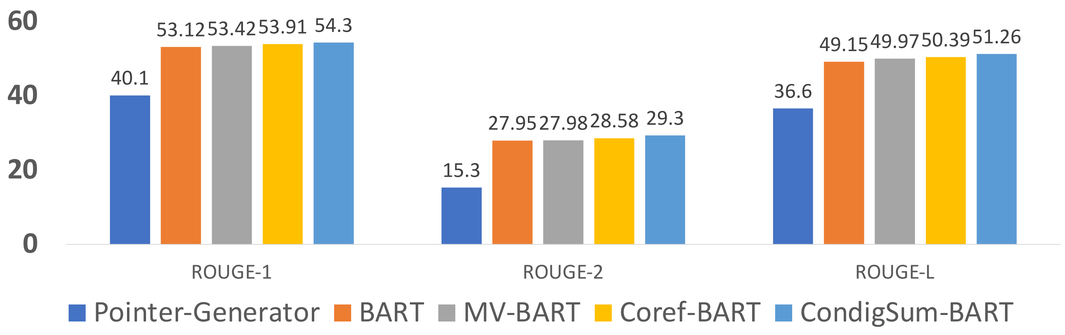
Dialogue

James: *Hey! I have been thinking about you :)*
Hannah: *Oh, that's nice ;)*
James: *What are you up to?*
Hannah: *I'm about to sleep.*
James: *I miss U. I was hoping to see you.*
Hannah: *Have to get up early for work tomorrow.*
James: *What about tomorrow?*
Hannah: *To be honest, I have plans for tomorrow evening.*
James: *Oh ok. What about Sat. then?*
Hannah: *Yeah. Sure, I am available on Sat.*
James: *I'll pick you up at 8?*
Hannah: *Sounds good. See you then.*

Ref. Summary

James misses Hannah. They agree for James to pickup Hannah up on Saturday at 8.

SOTA models on Standard Metrics

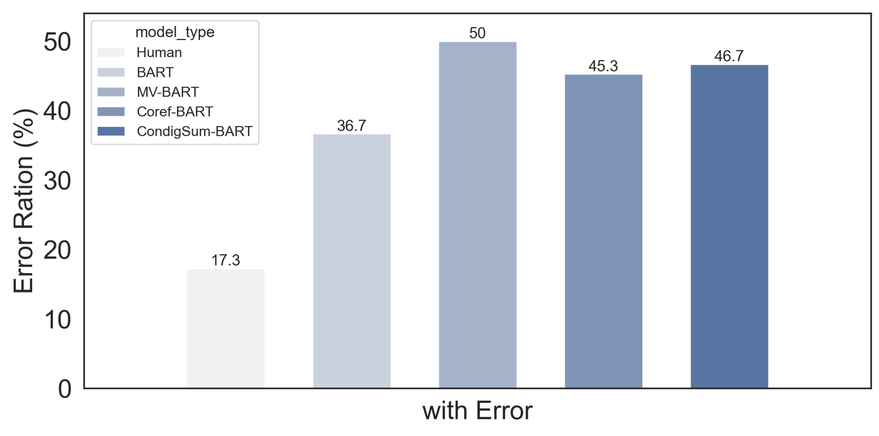


Recent progress on ROUGE score is limited.

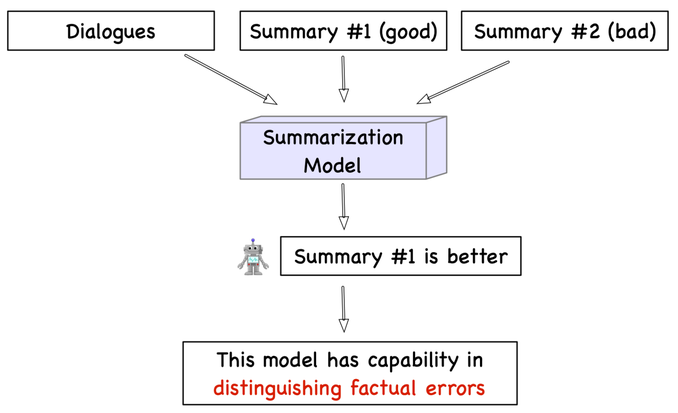
Analysis of Faithfulness

Preparation

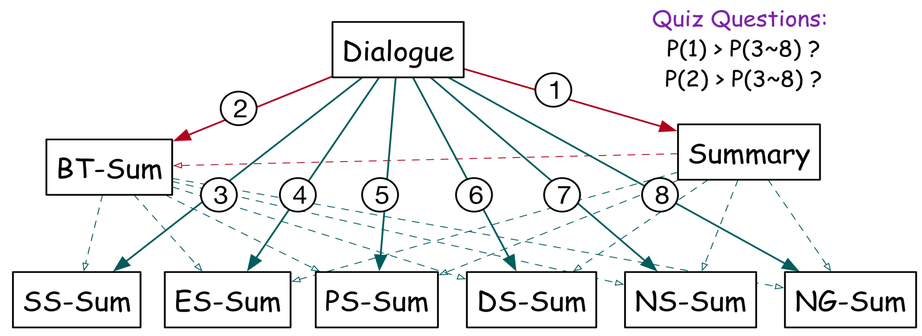
- **Six Error Types**
  - Subject Object Error – **SubObjE**
  - Pronoun Error – **ProE**
  - Negation Error – **NegE**
  - Particulars Error – **ParE**
  - Hallucination Error – **HalE**
  - Other Error – **OthE**
- **150 dialogues** (SAMSum dataset)
- **5 summaries per dialogue**
  - Human
  - BART
  - MV-BART
  - Coref-BART
  - CondigSum-BART
- **Two-step verification**
  - Two annotators per sample
  - Meta-annotator if no agreement



Evaluation of Faithfulness – Multi-choice Questions

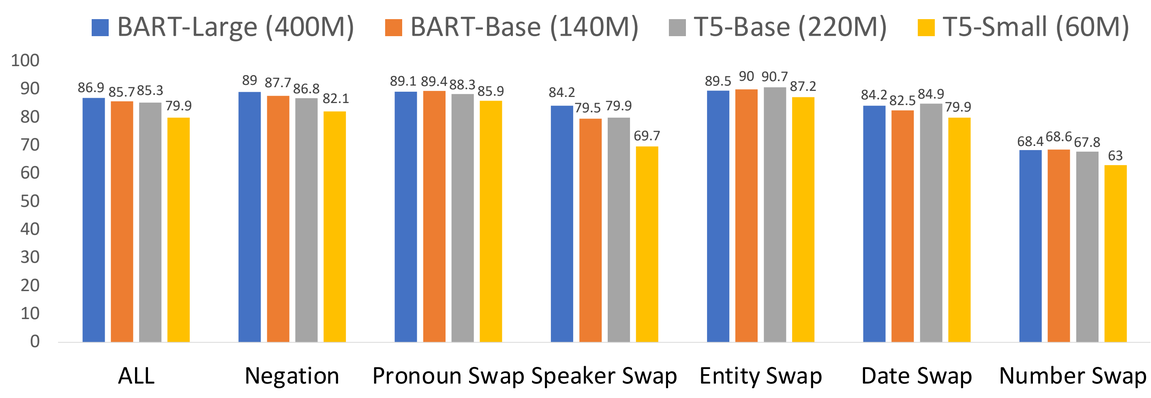


- Positive
  - Reference summary (**Summary**)
  - Back-translation (**BT-Sum**)
- Negative
  - Speaker swap (**SS-Sum**)
  - Entity swap (**ES-Sum**)
  - Pronoun swap (**PS-Sum**)
  - Date swap (**DS-Sum**)
  - Number swap (**NS-Sum**)
  - Negation (**NG-Sum**)



- **Mixed data training (MDT)**
  - Train a model with different ratios of corrupted data
  - Lead to more factual errors
- **Limited data training (LDT)**
  - Train models with less data
  - Less competitive in all aspects

Train. Strategy	LDT				MDT			
Model	$BART_{Large}$	$BART_{Base}$	$T5_{Base}$	$T5_{Small}$	$BART_{Large}$	$BART_{Base}$	$T5_{Base}$	$T5_{Small}$
Non-Factual Evaluation Schema								
ROUGE-1	81.35	95.79	94.44	95.94	84.16	91.04	95.58	85.84
ROUGE-2	86.77	96.84	96.39	96.09	90.13	97.01	95.45	93.90
ROUGE-L	75.64	96.24	96.39	92.63	86.23	98.31	97.14	94.16
BLEU	**91.88**	90.08	92.33	86.02	89.87	94.16	93.38	84.03
BERTScore	88.87	97.14	**97.29**	95.79	91.69	94.42	95.97	92.47
Factual Evaluation Schema								
$FactCC_{v1}$	—	—	—	—	—	—	—	—
$FactCC_{v2}$	82.39	84.57	42.45	97.07	96.01	99.22	98.27	**100.0**
FEQA	6.02	30.08	-60.15	33.23	57.92	54.29	75.06	85.58
NLI	39.40	31.28	93.08	90.53	91.17	82.99	93.77	92.99
FacEval (ours)	83.70	**99.40**	89.62	**98.05**	**99.74**	**100.0**	**99.87**	99.74



- *Model performance follows the model size and their pre-training schema.*
- *Speaker swap is one of the most challenging type of factual errors.*

Conclusion

- *Faithfulness hinders the application of dialogue summarization systems*
- First faithfulness evaluation method for dialogue summarization

Future directions:

- Improving faithfulness for dialogue summarization
- Faithfulness evaluation

Code: https://github.com/BinWang28/FacEval

Correspondence: Bin Wang, National University of Singapore

Email: bwang28c@gmail.com