INLG 2022 DialogSum Challenge: Dialogue Summarization using BART

Conrad Lundberg, Leyre Sánchez Viñuela, and Siena Biales

University of Tübingen

conrad.lundberg@student.uni-tuebingen.de
leyre.sanchez-vinuela@student.uni-tuebingen.de
siena.biales@student.uni-tuebingen.de

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Outline

- Task introduction
- Model overview
- Explored methods
  - Intermediate task transfer learning
  - Direct and reported speech
  - Data augmentation
- Results
- Conclusion

Task Introduction

- DialogSum is a shared task on summarizing real-life scenario dialogues
- Dialog summarization differs from monologic text summarization
- Model must address:
  - semantic roles
  - resolving definite pronouns/coreference
  - various other complexities
- Evaluation metrics: ROUGE scores, BERTScore, human evaluation

Model overview

- Fine-tuned BART model on 12,460 dialogue/summary pairs
    - Initially fine-tuned on the CNN/Dailymail corpus
- Examined topics: 7434 unique topics found
    - Not utilized for final model
- Post-processing
    - Replace any instances of `#Person3#` or `#Person4#` with `#Person1#` or `#Person2#`
    - Replace instances of duplicate labels, such as `#Person1#Person1#` or `#Person2#Person2#`

Explored methods

- Intermediate Task Transfer Learning
- Direct and Reported Speech
- Data Augmentation

Intermediate Task Transfer Learning

- Pruksachatkun et al. (2020) show intermediate tasks improve various target tasks
    - Some improved target tasks across the board: HellaSwag, Cosmos QA
- HellaSwag dataset
    - Natural language inference dataset modeled as multiple-choice questions
    - Trained 1 epoch on 10% of the HellaSwag training split as intermediate task
    - Did not improve ROUGE scores, discarded in final model
- XSum dataset
    - News articles and one-sentence summaries
    - Trained 1 epoch on the XSum training split as intermediate task
    - Did not improve ROUGE scores, discarded in final model

Direct and Reported Speech

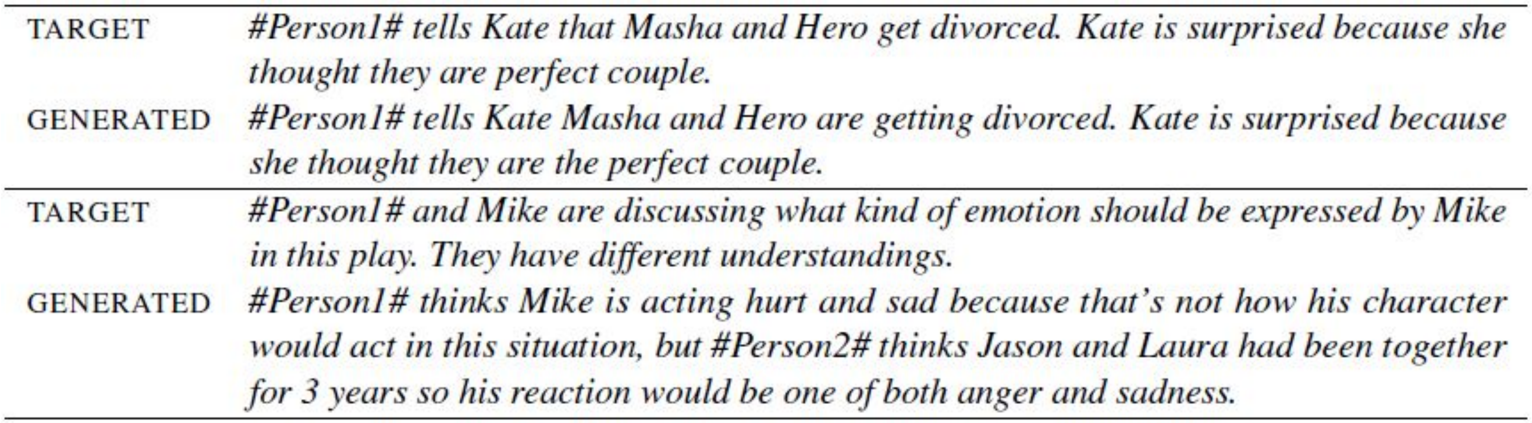
- Direct speech of dialogues vs. narrative style of news articles
  - 1st and 2nd person vs. 3rd person
- Hypothesis: if we fine-tune BART with more similar data to what it had originally been fine-tuned on, we can get better results
  - Transform the dialogues to reported speech to reflect style of news articles
  - Fine-tune BART with the dialogues in their reported-speech form
- Result: the ROUGE scores are lower
  - Possible reason: poor quality of rule-based direct-to-reported-speech algorithm

Data Augmentation

- SamSum: human-annotated dialogue dataset for abstractive summarization
  - 16k messenger-like conversations with summaries
- Fine-tuned BART with merged SamSum and DialogSum datasets
- Results: lower ROUGE scores
  - Possible reason: shorter length of SamSum dialogues and summaries
  - Written dialogues (SamSum) vs. spoken conversations (DialogSum)

Results

- Some “good” summaries had low ROUGE scores
  - Length discrepancies
  - Novel word choices

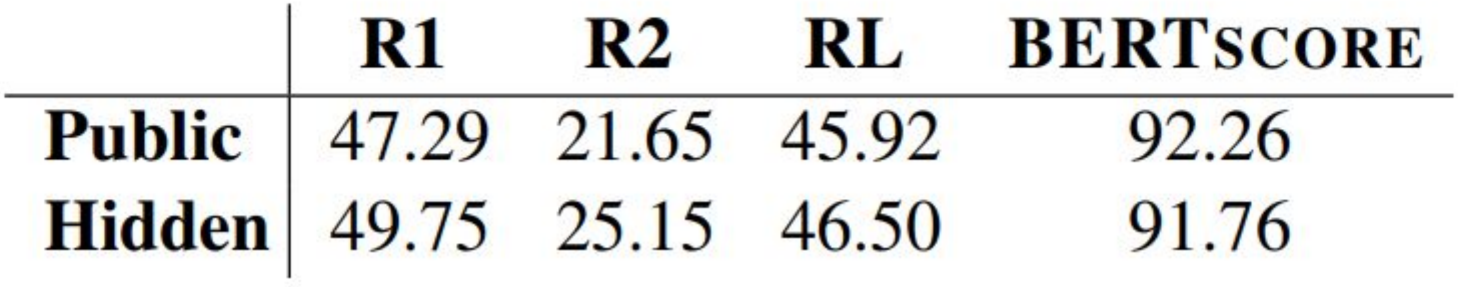


TARGET	*#Person1# tells Kate that Masha and Hero get divorced. Kate is surprised because she thought they are perfect couple.*
GENERATED	*#Person1# tells Kate Masha and Hero are getting divorced. Kate is surprised because she thought they are the perfect couple.*
TARGET	*#Person1# and Mike are discussing what kind of emotion should be expressed by Mike in this play. They have different understandings.*
GENERATED	*#Person1# thinks Mike is acting hurt and sad because that’s not how his character would act in this situation, but #Person2# thinks Jason and Laura had been together for 3 years so his reaction would be one of both anger and sadness.*

Table 1: Examples of a generated summary close to the target summary (above) and a less ideal generated summary (below)

Results

- Results very close to others on the leaderboard
- ROUGE scores on the hidden dataset were higher



	R1	R2	RL	BERTScore
Public	47.29	21.65	45.92	92.26
Hidden	49.75	25.15	46.50	91.76

DialogSum Challenge Website: https://cylnlp.github.io/dialogsum-challenge/

Conclusion

- Basic fine-tuned BART is able to achieve relatively successful dialogue summarization
- Compared to other submissions, we had good results on both the public testset and hidden testset
- Future work:
  - Intermediate task transfer learning on a different dataset or for more epochs
  - Directed to reported speech using better algorithm
  - Dataset augmentation with a different dataset

Thank you for your attention!

We encourage you to look at our paper to learn more!