INLG 2022 DialogSum Challenge: Dialogue Summarization using BART

Conrad Lundberg, Leyre Sánchez Viñuela, and Siena Biales
{conrad.lundberg, leyre.sanchez-vinuela, siena.biales}@student.uni-tuebingen.de
Seminar für Sprachwissenschaft, University of Tübingen

Introduction

- The DialogSum Challenge aims to summarize real-life scenario dialogues
- We investigate the capabilities of a fine-tuned BART model for this task
- We also explore other methods:
  - ○ Intermediate Task Transfer Learning
  - ○ Direct and Reported Speech
  - ○ Data Augmentation
- Evaluation is based on ROUGE scores, with BERTScore as secondary metric

Conclusions

- Basic fine-tuned BART model is able to achieve relatively successful dialogue summarization
- None of our other methods yielded improved results
- 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

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

Scores achieved on both the public and hidden test sets

- Some "good" summaries had low ROUGE scores
  - ○ Due to length discrepancies and 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.*

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

Data Augmentation

- Fine-tuned BART with merged SamSum and DialogSum datasets
- Results: lower ROUGE scores
- Possible reasons:
  - ○ Shorter length of SamSum dialogues and summaries
  - ○ Written dialogues (SamSum) vs. spoken conversations (DialogSum)

Intermediate Task Transfer Learning

- HellaSwag dataset
  - ○ Natural language inference dataset
  - ○ Multiple-choice questions
  - ○ Trained 1 epoch on 10% of the HellaSwag training split
  - ○ Result: lower ROUGE scores
  - ○ Discarded in final model
- XSum dataset
  - ○ News articles with 1-sentence summaries
  - ○ Trained 1 epoch on the XSum training split as intermediate task
  - ○ Result: lower ROUGE scores

Direct and Reported Speech

- Transform the dialogues to reported speech to reflect style of news articles
- More similar to the CNN/DailyMail that the model was originally fine-tuned on
- Fine-tune BART with the dialogues in their reported-speech form
- Result: lower ROUGE scores
- Possible reason:
  - ○ Poor quality of our rule-based direct-to-reported-speech algorithm