Dialogue Response Generation

Research Project

Presented by Ariel Ramos

Table of Contents

01 Motivation

02 Data Preparation

03 Model Benchmark

04 Quantitative evaluation

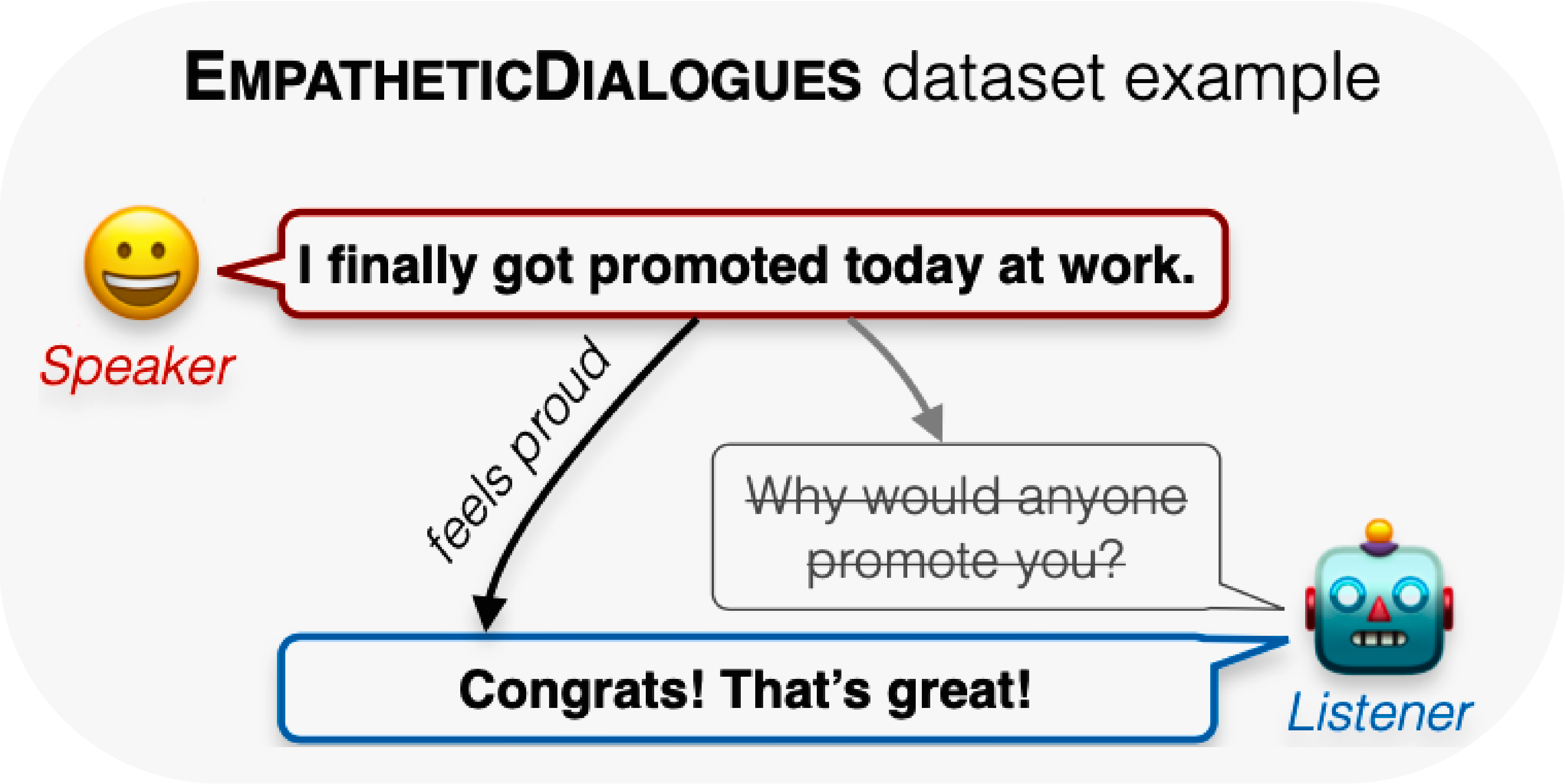
05 Qualitative evaluation

Motivation

- Relevance of responses to the user's social and emotional context.

- Benchmark different response generation models:
    - GPT-2
    - DialoGPT
    - BART

- Evaluate responses quantitatively and qualitatively.


EMPATHETICDIALOGUES dataset example
Speaker
I finally got promoted today at work.
feels proud
~~Why would anyone promote you?~~
Congrats! That's great!
Listener

Data preparation

- Hugging Face Datasets:
  - daily_dialog
  - empathetic_dialogues

- Clean datasets
  - Removing end marks
  - Removing abbreviations
  - Assigning speaker ids (<sp1>, <sp2>)

- Build actual dialogues (for empathetic_dialogues) from row utterances.

Dataset\Split	**Train**	**Validation**	**Test**
daily_dialog	11.1k rows	1k rows	1k rows
empathetic_dialogues	76.7k rows	12k rows	10.9k rows

Model Benchmark

- GPT-2
  - transformers model pretrained on a very large corpus of English data in a self-supervised fashion.

- DialoGPT
  - Based on the GPT2 model architecture.
  - Powerful at response generation in open-domain dialogue systems.

- BART
  - seq2seq/machine translation architecture with a bidirectional encoder (like BERT) and a left-to-right decoder (like GPT).
  - Effective when fine tuned for text generation but also for comprehensions tasks.

OpenAI

Microsoft

Meta AI

Quantitative Evaluation

- Metrics
  - Sacreblue
  - Rouge
  - Bert score
  - chrf

- Hyperparameters:
  - Learning rate: 2e-5
  - Batch size: 8
  - Number of epochs: 6
  - top-p sampling = 0.8

- Generated responses (window=5):
  - daily_dialog: 1782
  - empathetic_dialogues: 174

daily_dialog

Model	SacreBleu	rouge1	Avg f1 Bert	chrf
GPT-2	0.965	0.118	0.862	**13.081**
DialoGPT	1.072	0.130	0.864	11.922
BART	**2.116**	**0.167**	**0.872**	12.071

empathetic dialogues

Model	SacreBleu	rouge1	Avg f1 Bert	chrf
GPT-2	0.376	0.106	0.842	**11.564**
DialoGPT	0.382	0.103	0.855	10.987
BART	**0.778**	**0.125**	**0.862**	10.616

Qualitative Evaluation

- 20 responses manually analysed.

empathetic dialogues

- Yesterday my car ran out of oil and broke down. I could not believe I would let such a thing happen!
- Oh no! Did it do any meaningful damage?
- The engine is totally shot, Its not designed to run without any. So now I have to go shopping for a new car.
- OuCH. New car is cheaper than replacing the engine?
- The new car is more expensive than a new engine. But I can't justify such a costly repair on my car.

Actual res: Yeah I get that. Plus I guess the value goes way down doing something like that.
GPT-2 res: Well, it works great. My car is completely restored!
DialoGPT res: That's what happens when you can't afford a new car,
BART res: I'm sure you can figure out a way to fix it.

- One of my favorite breakfast foods. She also made some thick sliced bacon. She treats me so well.
- That is great you have a sweet wife!
- For sure, No other woman I would rather have.
- It is so great to hear about a happy couple :).
- How can I complain when she gets up early to make me such a great meal!

Actual res: I bet she is a great cook!
GPT-2 res: Good job!
DialoGPT res: I know that feeling.
BART res: That sounds like a great wife!

Conclusions

- BART performed better in both quantitative and qualitative evaluations.

- Despite of being syntactically and semantically correct, some answers tend to have a lack of empathy towards the context given for a particular dialogue.

- Learning experience.
  - Coding
  - Research

- GitHub Repository containing notebooks and results.
  - https://github.com/arielramos97/Research_Project/