Overview

- Progress in end-to-end deep neural dialogue agents is limited by their knowledge of world events and latency in responding to user inputs.
- Knowledge retrieval in ChirpyCardinal follows:
  - Wikipedia entities are identified from dialogue.
  - Entities + templates are passed to GloVe-based retrieval to return “knowledge statements” (KS).
  - KSs are used as input for template infilling.
- Challenges: GloVe-based retrieval has mixed performance, but large retrieval models w/ better performance based on LLMs have high latency.
- We explored:
  - Integrating CoBERT-based retrieval using a Falafel index to improve retrieval quality.
  - Applying alternative neural generation models for infilling, such as T5.
  - Benchmarking different quantitative evaluation methods for retrieving responses.
  - Integrated these models into ChirpyCardinal for end-to-end conversation

ChirpyCardinal

- Open-source, end-to-end chatbot with foundation in multilogic, multitreaded response generators.
- This neural retrieval occurs in the Wiki response gen.

Problem Setting

- Dataset: May 2020 English Wikipedia Dump
  - For GloVe search: filtered Wikipedia page corresponding to current entity.
  - For CoBERT: 2M+ passages of 180-tokens.
- Templates: pre-created structures ready for infilling given context.
  - Ex. Template: I love how [factor] played in [film], especially their <mask>.
  - Ex. Infill: I love how [President Reagan] acted in [The Mask], especially their ability to freeze time.

Task 1: Retrieval (cont.)

- CoBERT Retrieval
  - Data: Falafel index of BERT-embedded tokenized passages.
  - Method: Batch top-k retrieval queries corresponding to different templates.

- GloVe Retrieval
  - Data: filtered sentences from entity’s Wikipedia page.
  - Method: compute GloVe embeddings for each template, statement, then choose top-k pairs.

- Key difference in methods: better use of semantic context through BERT embeddings vs. GloVe.

Task 1: Retrieval

- Quan. eval. method: avg. top-k retrieval relevance.
  \[
  \text{AS(k)} = \frac{1}{k} \sum_{i=1}^{k} s_i \in [0, 1, \ldots, k] 
  \]

  - Retrieval Method
  - GloVe: 2.35 ± 0.11
  - CoBERT: 2.91 ± 0.20

- Add. quan. method: “adapted MRSE” (aMRSE)

  - Retrieval Method
  - aMRSE@5: 0.231
  - CoBERT: 0.932

- Ablation: use sentences retrieved by CoBERT for GloVe (Augmented GloVe)

  - Retrieval Method
  - Augmented GloVe: 2.06 ± 0.32
  - CoBERT: 2.91 ± 0.20

Task 2: Infilling

- We again use avg. top-k retrieval relevance for evaluating quality of infilled statements.

  - Retrieval-Infilling Method
  - GloVe + BART: 2.42 ± 0.13
  - GloVe + T5: 2.23 ± 0.08
  - CoBERT + BART: 3.21 ± 0.19
  - CoBERT + T5: 2.95 ± 0.24

Conclusion

- Existing neural retrieval used in ChirpyCardinal did not make full use of semantic context.
- Improved retrieval also benefits downstream infilling.
- Feasibly embedded within existing framework for end-to-end neural conversation.

Future Work

- Broader Quantitative Evaluation: Increase the number of people used for evaluating the quality of retrieved knowledge statements.
- Code Optimization: Refine the code to better leverage existing information in ChirpyCardinal while decreasing latency and minimize bugs.
- Latency Evaluation: Further profiling of the latency of retrieval and infilling operations.

References


Acknowledgements

- We would like to thank Ethan Chi for closely and attentively mentoring our project, which we learned a lot from. We would also like to thank Chris Manning for teaching CS 224N, as well as the rest of the course staff for running it so smoothly and making the class engaging and insightful.