# Neural-Augmented Retrieval for Open-Domain Dialogue Systems

*Eric Frankel,[1] Rohan Mehrotra,[2] Raphael Ruban[2]*

[1]*Department of Statistics, Stanford University*
[2]*Department of Computer Science, Stanford University*

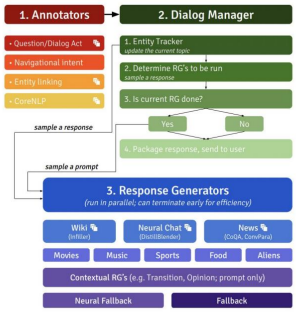
Stanford Computer Science

## 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 are used 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 **ColBERT-based retrieval** using a Faiss index to improve retrieval quality.
  - Applying alternative neural generation models for infilling, such as T5.
  - Benchmarking different **quantitative evaluation methods** for retrieved responses.
  - Integrated these models into ChirpyCardinal for end-to-end conversation

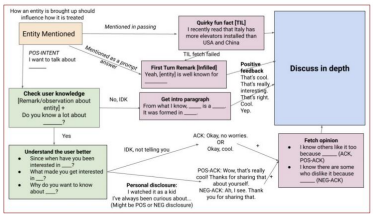
## ChirpyCardinal

- Open-source, end-to-end chatbot with foundation in multitopic, multithreaded response generators.




- This neural retrieval occurs in the Wiki response gen.

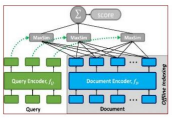
## Problem Setting



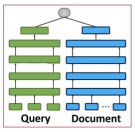

- Dataset: **May 2020 English Wikpedia Dump**
  - For GloVe search: filtered Wikipedia page corresponding to current entity.
  - For ColBERT: **21M+** passages of 180-tokens.
- Templates: **pre-created structures** ready for infilling given context.
  - Ex. Template: *I love how [[actor]] acted in [[film]], especially their <mask>.*
  - Ex. Infill: *I love how [Keanu Reaves] acted in [The Matrix], especially their ability to freeze time.*

## Task 1: Retrieval

- **ColBERT Retrieval**
  - **Data:** Faiss 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 (cont.)

- Quan. eval. method: avg. top-k retrieval relevance.

$$AS(k) = \frac{1}{20}\sum_{i=1}^{20} s_i, \quad s_i \in \{0, \ldots, k\}$$

| Retrieval Method | AS(5) |
|---|---|
| GloVe | 2.35 ± 0.11 |
| ColBERT | 2.91 ± 0.20 |

- Add. quan. method: "adapted MRR" (aMRR).

| Retrieval Method | aMRR@5 |
|---|---|
| GloVe | 0.231 |
| ColBERT | 0.532 |

- Ablation: use sentences retrieved by ColBERT for GloVe **(Augmented GloVe)**.

| Retrieval Method | AS(5) |
|---|---|
| GloVe | 2.35 ± 0.11 |
| Augmented GloVe | 2.56 ± 0.32 |
| ColBERT | 2.91 ± 0.20 |

- ColBERT's use of semantic information makes it superior retrieval method.

## Task 2: Infilling

- **Method:** Given retrieved template + context pairs, infill the mask in the template.
- **BART** and **T5** used for conditional generation.

| | | | |
|---|---|---|---|
| 4 | **User:** i love it when the batter is crunchy<br>**Bot:** I agree! The crunch and the texture is really good. It's a shame it's hard to find a good crispy batter these days, though. [ Neural ] Did you know that fish and chips were first made in England around the 16th century? I can't believe people have been eating fish and chips for so long! [ Food ] | Fish and chips<br>Fish and chips | Food gives a DistilBlender-generated acknowledgement, followed by a scripted prompt. |
| 5 | **User:** wow that's a long time<br>**Bot:** It really is. Anyway, I'm feeling hungry now! Thanks for recommending fish and chips! [ Food ] Speaking of Fish and chips, i was reading recently and found out that fish and chips is a common take-away food in the united kingdom and numerous other countries...do you wanna talk about the United Kingdom? [ Transition ] | None<br>United Kingdom | Transition generates a transition from *Fish and chips* to *United Kingdom* to continue the conversation smoothly. |
| 6 | **User:** sure<br>**Bot:** I just learned recently that violent pirates could have been executed in the united kingdom until 1998. [ Wiki ] | United Kingdom<br>United Kingdom | User accepts the suggested entity for discussion.<br>Wiki provides an template-infilled remark about the *United Kingdom*. |
| 7 | **User:** oh that's surprising<br>**Bot:** Isn't it? Come to think of it, I did read something the other day about United Kingdom's Culture and Geography, if you're interested in talking about any of that. [ Wiki ] | United Kingdom<br>United Kingdom | Wiki suggests aspects of the *United Kingdom* to discuss. |

## Task 2: Infilling

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

| Retrieval+Infilling Method | AS(5) |
|---|---|
| GloVe + BART | 2.42 ± 0.13 |
| GloVe + T5 | 2.23 ± 0.08 |
| ColBERT + BART | 3.21 ± 0.19 |
| ColBERT + T5 | 2.95 ± 0.24 |

- Results emphasize the importance of the accompanying context used during infilling.

## 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:** Refining the code to better leverage existing information in ChirpyCardinal will decrease latency and minimize bugs.
- **Latency Evaluation:** Further profiling of the latency of retrieval and infilling operations.

## References

[1] Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48, 2020.

[2] Ethan A. Chi, Chetanya Rastogi, Alexander Iyabor, Hari Sowrirajan, Avanika Narayan, and Ashwin Paranjape. Neural, neural everywhere: Controlled generation meets scaffolded, structured dialogue. 2021.

[3] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *International conference on learning representations*, 2017.

## 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.