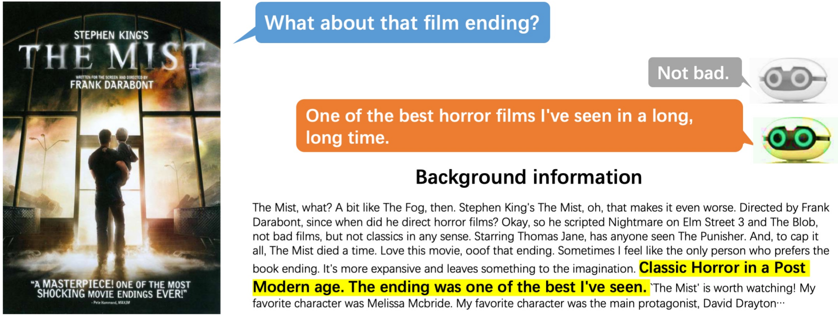
Improving Background Based Conversation with Context-aware Knowledge Pre-selection

Yangjun Zhang Pengjie Ren Maarten de Rijke

University of Amsterdam

Background Based Conversation (BBC)

- Aims to generate responses by referring to background information and considering the dialogue history at the same time



Extraction based methods

- Pros:
  - Better at locating the right background span than generation-based methods [Mogheet al., 2018]
- Cons:
  - Not suitable for BBCs:
    - BBCs do not have standard answers like those in RC tasks
    - Responses based on fixed extraction are directly copied from background sentences; neither fluent nor natural

Generation based methods

- Pros:
  - Response diversity and fluency improved; able to leverage background information
- Cons:
  - Selecting background knowledge by using decoder hidden states as query
  - Query not containing all information from context history since LSTM does not guarantee preserving information over many timesteps (Cho et al., 2014)

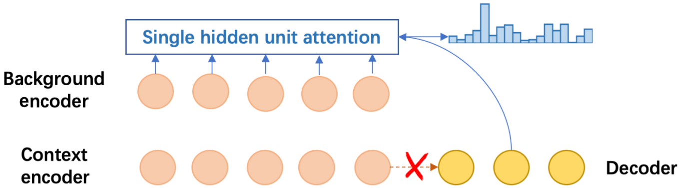


Figure: Previous generation based methods

Motivation

- The crucial role of context history in selecting appropriate background has not been fully explored by current methods
- Introducing **knowledge pre-selection** process to improve background knowledge selection by using the utterance history context as prior information

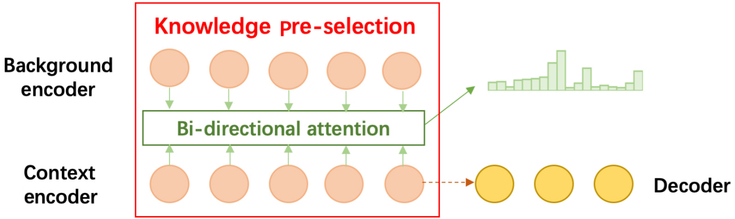


Figure: CaKe with knowledge pre-selection

Model overview

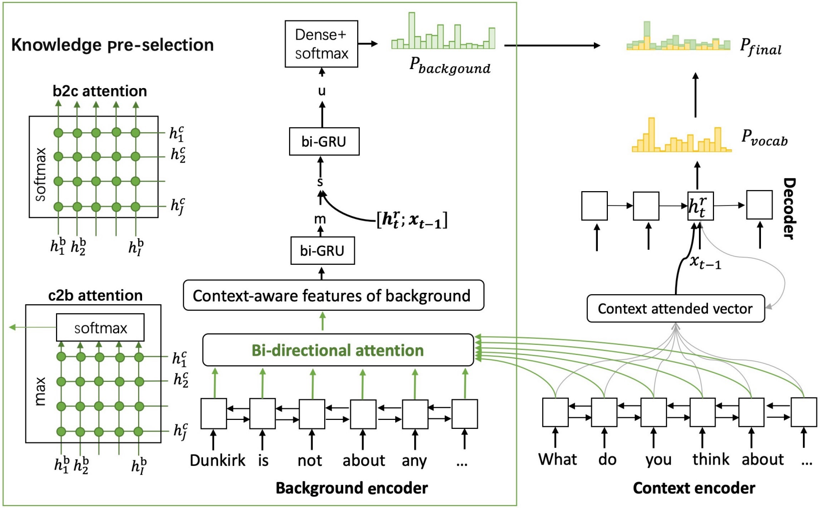
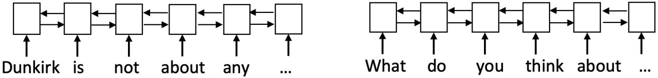


Figure: Model overview

Encoders

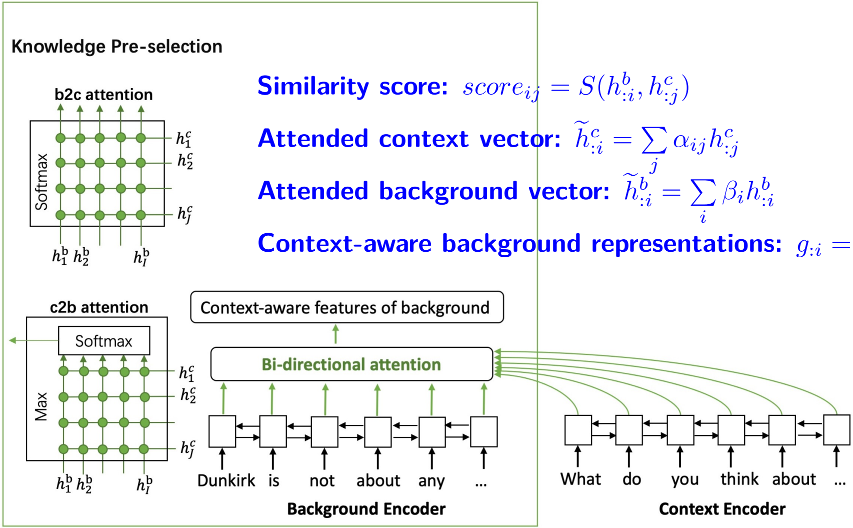
Background encoders: $h^b = (h_1^b, h_2^b, \ldots, h_i^b, \ldots, h_I^b)$

Context encoders: $h^c = (h_1^c, h_2^c, \ldots, h_j^c, \ldots, h_J^c)$



Background Encoder **Context Encoder**

Knowledge pre-selection



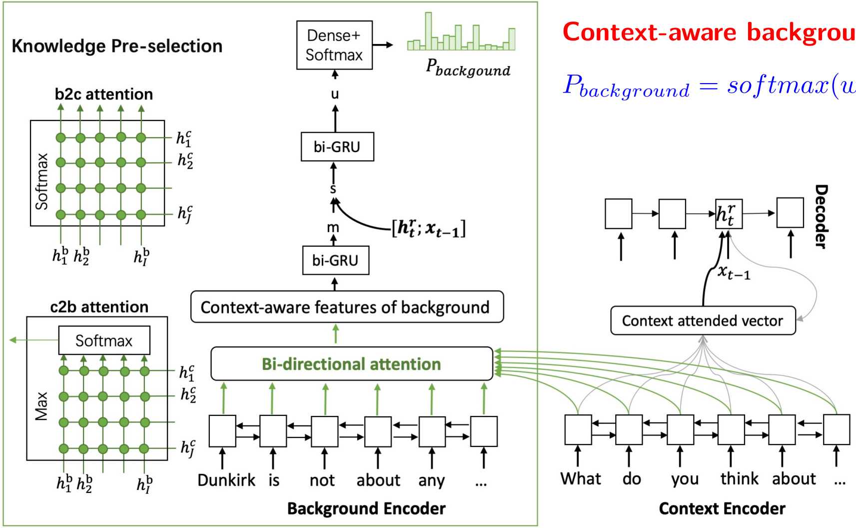
Similarity score: $score_{ij} = S(h^b_{:i}, h^c_{:j})$

Attended context vector: $\widetilde{h}^c_{:i} = \sum_j \alpha_{ij} h^c_{:j}$

Attended background vector: $\widetilde{h}^b_{:i} = \sum_i \beta_i h^b_{:i}$

Context-aware background representations: $g_{:i} = \eta(h^b_{:i}, \widetilde{h}^c_{:i}, \widetilde{h}^b_{:i})$

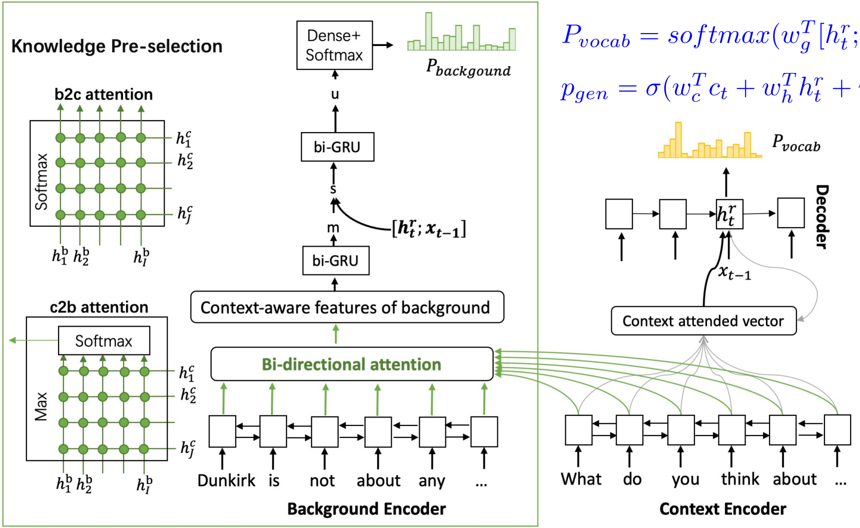
Knowledge pre-selection



Context-aware background distribution:

$$P_{background} = softmax(w_{p1}^T[g; m; s; u] + b_{bg})$$

Generator

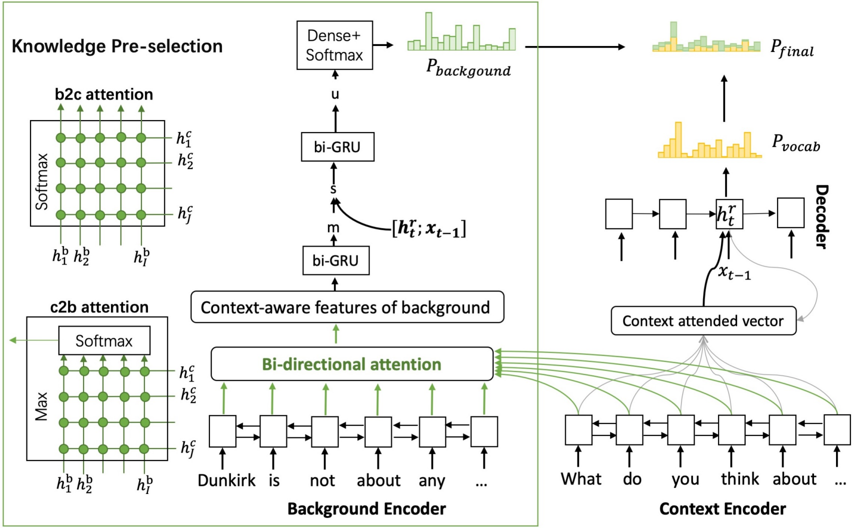


$$P_{vocab} = softmax(w_g^T[h_t^r; c_t] + b_v)$$

$$p_{gen} = \sigma(w_c^T c_t + w_h^T h_t^r + w_x^T x_t + b_{gen})$$

Final distribution

$$P_{final}(w) = p_{gen}P_{vocab}(w) + (1 - p_{gen})P_{background}(w)$$



Loss function

$$loss_t = -\log P(w_t^*)$$

$$loss = \frac{1}{T}\sum_{t=0}^{T} loss_t$$

$$L(\theta) = \sum_{n=0}^{N} loss$$

Experimental Setup

Datasets

- Holl-E dataset: contains background documents including review, plot, comment and fact table of 921 movies and 9071 conversations
  - Oracle background uses the actual resource part from the background documents
  - 256 words background is generated by truncating the background sentences

Baselines

- Sequence to Sequence (S2S)(Sutskever et al., 2014)
- Hierarchical Recurrent Encoder-decoder Architecture (HRED)(Serban et al., 2016)
- Sequence to Sequence with Attention (S2SA)(Bahdanau et al., 2015)
- Bi-Directional Attention Flow (BiDAF)(Seo et al., 2017):
- Get To The Point (GTTP)(See et al., 2017; Moghe et al., 2018)

Experimental Setup

Our methods

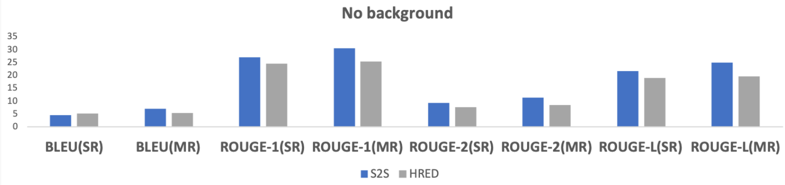
Apply knowledge pre-selection:

- 256-d hidden size GRU
- 45k vocabulary size
- 30 epochs

Evaluation

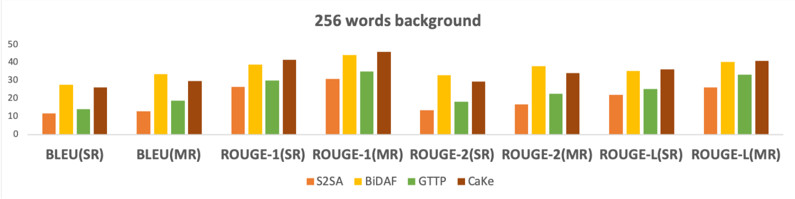
- The background knowledge and the corresponding conversations are restricted to a specific topic
- BLEU, ROUGE-1, ROUGE-2 and ROUGE-L as the automatic evaluation metrics

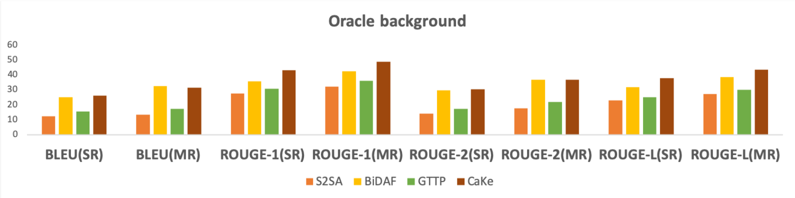
Overall Performance



- The models without background generate weak results

Overall Performance





- Slightly superior to BiDAF model; outperforms GTTP
- Performance reduces slightly when background becomes longer, but reduction is acceptable

Knowledge selection visualization

- Attention is very strong on several positions (b)
- Our pre-selection mechanism could help knowledge selection
- X: background word positions; Y: (a) b2c (b) c2b (c) final distribution (d) GTTP final distribution

- Background: I enjoyed it. Fun, August, action movie. It's so bad that it's good.
- GTTP: It was so bad that it's good.
- OURS: I agree, Fun, August, action movie.

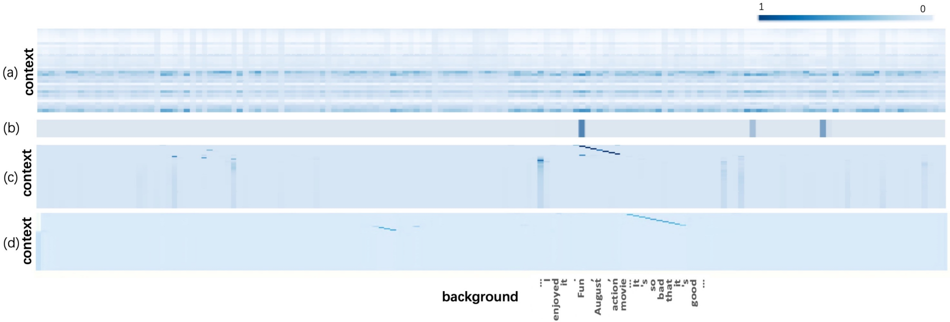


Figure: Knowledge selection visualization

Case study

- Context-aware Knowledge Pre-selection (CaKe) is able to generate more fluent responses than BiDAF and more informative responses than GTTP

Table: Case study

Background	The mist ... Classic Horror in a Post Modern age. The ending was one of the best I've seen ...
Context	Speaker 1: Which is your favorite character in this? Speaker 2: My favorite character was the main protagonist, David Drayton. Speaker 1: What about that ending?
Response	BiDAF: Classic horror in a post modern age. GTTP: They this how the mob mentality and religion turn people into monsters. **CaKe**: One of the best horror films I've seen in a long, long time.

Conclusion

1. We propose knowledge pre-selection process for the BBC task; explore selecting relevant knowledge by using context as prior query
2. Experiments show that CaKe outperforms the state-of-art
3. Limitation: Performance of our pre-selection process decreases when the background becomes longer

Future Work

1. Improve the selector and generator module, by methods such as multi-agent learning, transformer models and other attention mechanisms
2. Conduct human evaluations
3. Increase the diversity of CaKe results by incorporating mechanisms such as leveraging mutual information

Thank You

Source code

`https://github.com/repozhang/bbc-pre-selection`

Contact

- Yangjun Zhang
- y.zhang6@uva.nl

Thanks for support: Ahold Delhaize, the Association of Universities in the Netherlands (VSNU), the China Scholarship Council (CSC), the Innovation Center for Artificial Intelligence (ICAI), Huawei, Microsoft, Naver and Google.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *International Conference on Learning Representations*.

Kyunghyun Cho, Bart Van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. In *Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. 103–111.

Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M Khapra. 2018. Towards Exploiting Background Knowledge for Building Conversation Systems. In *2018 Conference on Empirical Methods in Natural Language Processing*. 2322–2332.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. In *The 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1073–1083.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *International Conference on Learning Representations*.

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*. 3104–3112.