

## Introduction

This work considers the dialogue task on the Ubuntu Dialogue Corpus [1] with end-to-end architectures. Given a short conversation history between two users, the task is to generate what the next user will say (see Figure 1).

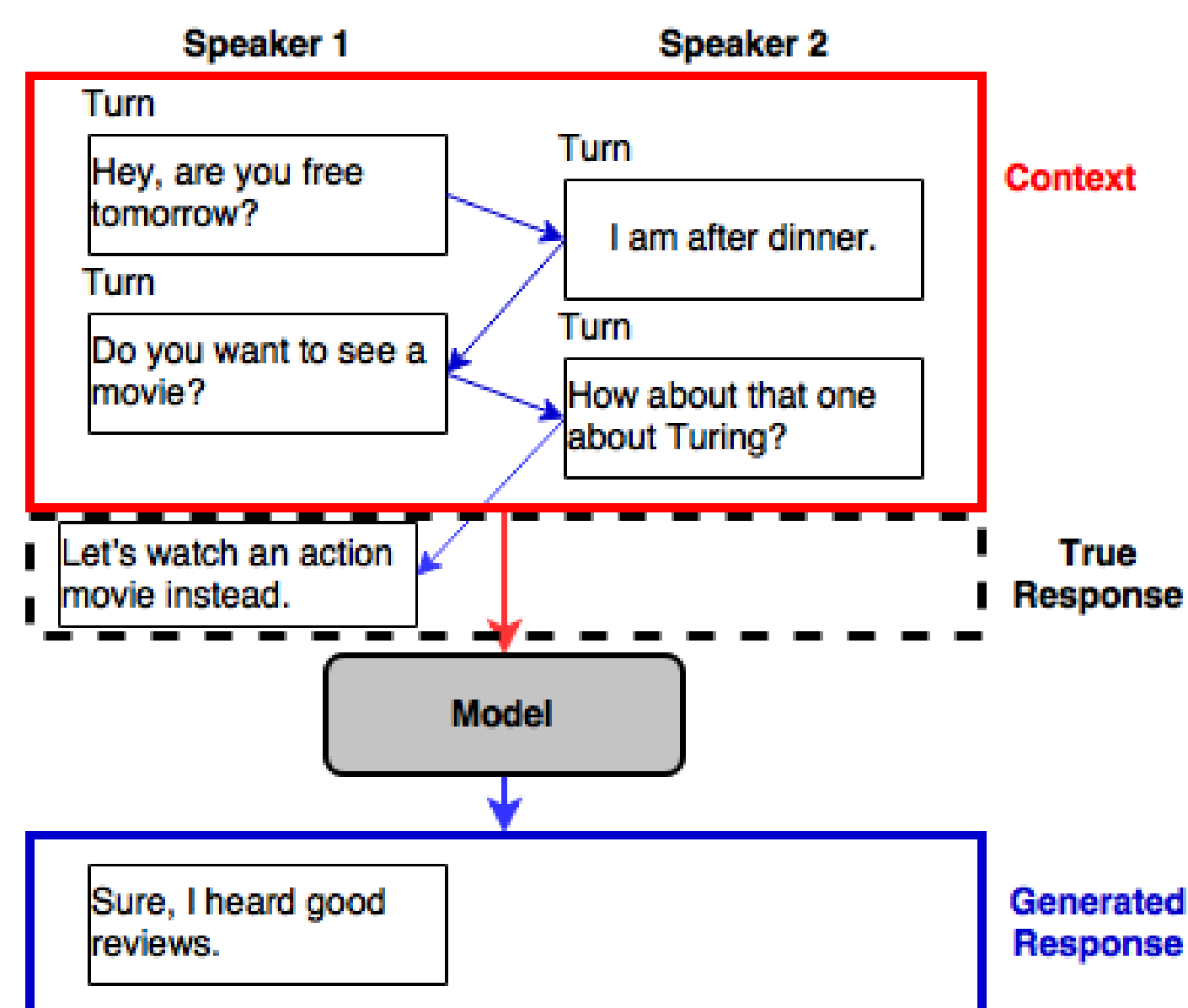


Figure 1: Description of the dialogue task.

## Ubuntu Corpus

- NLTK + Tweet tokenizers
- Named entities not replaced by generic placeholders makes the task harder than in original paper [1]
- Byte-Pair Encoding (*BPE*) [2] to reduce vocabulary size from 500,000 to 6,259.

## Model

- Baseline model: Hierarchical Recurrent Encoder Decoder (*HRED*) [3].

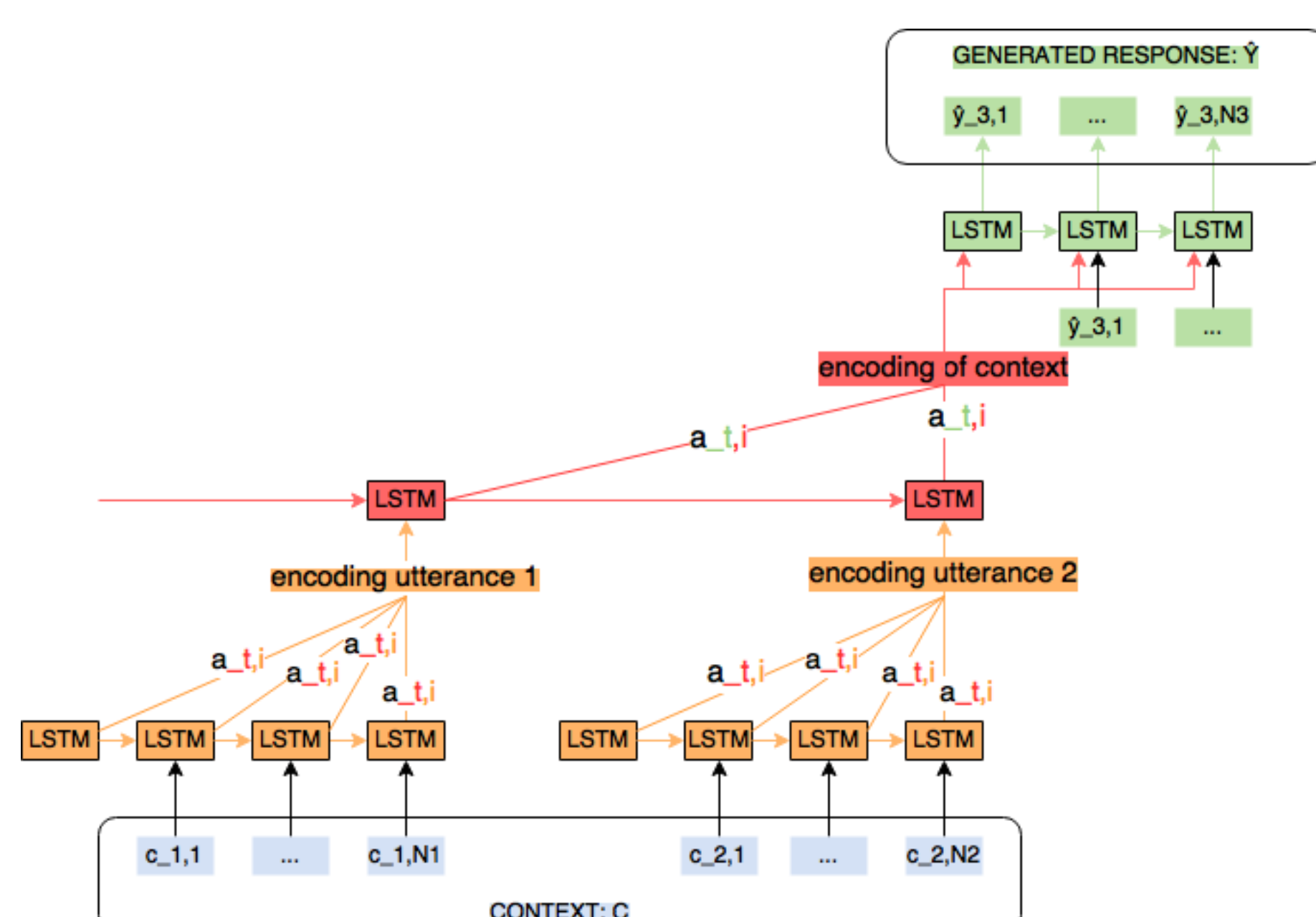


Figure 2: *HRED* model with attention between each layer

## Tasks

- 1 Prevent the repetition of past trigrams in the same response. Taking simple trick from [4]: when sampling, set the probability of tokens that will regenerate a past trigram to 0.0.
- 2 Enhance *HRED* by adding attention mechanisms between turn & dialogue encoders, and between the dialogue encoder & turn decoder (see Figure 2). Attention weights computed using simple matrix multiplication between hidden states:

$$e_{t,i} = h_t^{L2} \cdot W_{t,i} \cdot h_i^{L1}$$

$$\alpha_{t,i}^{L1} = \text{softmax}(e_{t,i}) = \frac{\exp(e_{t,i})}{\sum_j \exp(e_{t,j})}$$

$$c_t^{L1} = \sum_i \alpha_{t,i}^{L1} * h_i^{L1}$$

with upper layer *L2* paying attention over lower layer *L1*. At each time step *t* in *L2*, we compute  $c_t^{L1}$  instead of taking the last hidden state from *L1*.

- 3 Remove generic turns from dialogues in the training set: for each dialogue, flag each turn *t* as generic with probability

$$1 - \frac{1}{\text{unique\_previous\_turns}(t)}$$

Uniformly sample a turn flagged as generic and truncate the dialogue just before that. Resulting dialogue might still contain earlier turns which are generic. Reduces generic turns in the training set while not losing too much information.

## Metrics

- Perplexity:  $PPL = \exp(\text{negative log likelihood})$
- Recall@k (r@k): 1 true response & 9 false, correct when true response is ranked in top-k.
- Ubuntu pre-trained embeddings: measure cosine similarity between embeddings of generated and true responses:
  - Average embedding (AE): average of the word embeddings composing a response
  - Greedy matching (GM): match most similar word in true response to word in generated response, repeat in opposite direction & average cosine similarities
  - Vector extrema (VE): max value along each dimension of the word embeddings in a response

## Results

- Experiments with different RNN encoding sizes:

|    | PPL  | r@1   | r@2   | r@5   | AE   | GM   | VE   |
|----|------|-------|-------|-------|------|------|------|
| e1 | 51.4 | 21.98 | 35.21 | 66.20 | 0.62 | 0.58 | 0.37 |
| e2 | 51.7 | 22.51 | 35.72 | 66.40 | 0.63 | 0.58 | 0.37 |
| a1 | 14.3 | 21.15 | 35.29 | 66.26 | 0.57 | 0.56 | 0.34 |
| a2 | 13.1 | 19.07 | 32.10 | 63.28 | 0.55 | 0.57 | 0.33 |
| d2 | 13.0 | 18.82 | 32.08 | 63.55 | 0.60 | 0.56 | 0.35 |
| d3 | 12.6 | 19.73 | 33.53 | 64.89 | 0.40 | 0.57 | 0.27 |

Table 1: Results on test set. e1 & e2: task 1; a1 & a2: tasks 1 and 2; d2 & d3: tasks 1, 2 and 3.

- A human evaluation was performed on 200 randomly selected context/response pairs. Each pair was independently scored between 1 (very bad) and 5 (very good) by 3 NLU researchers.

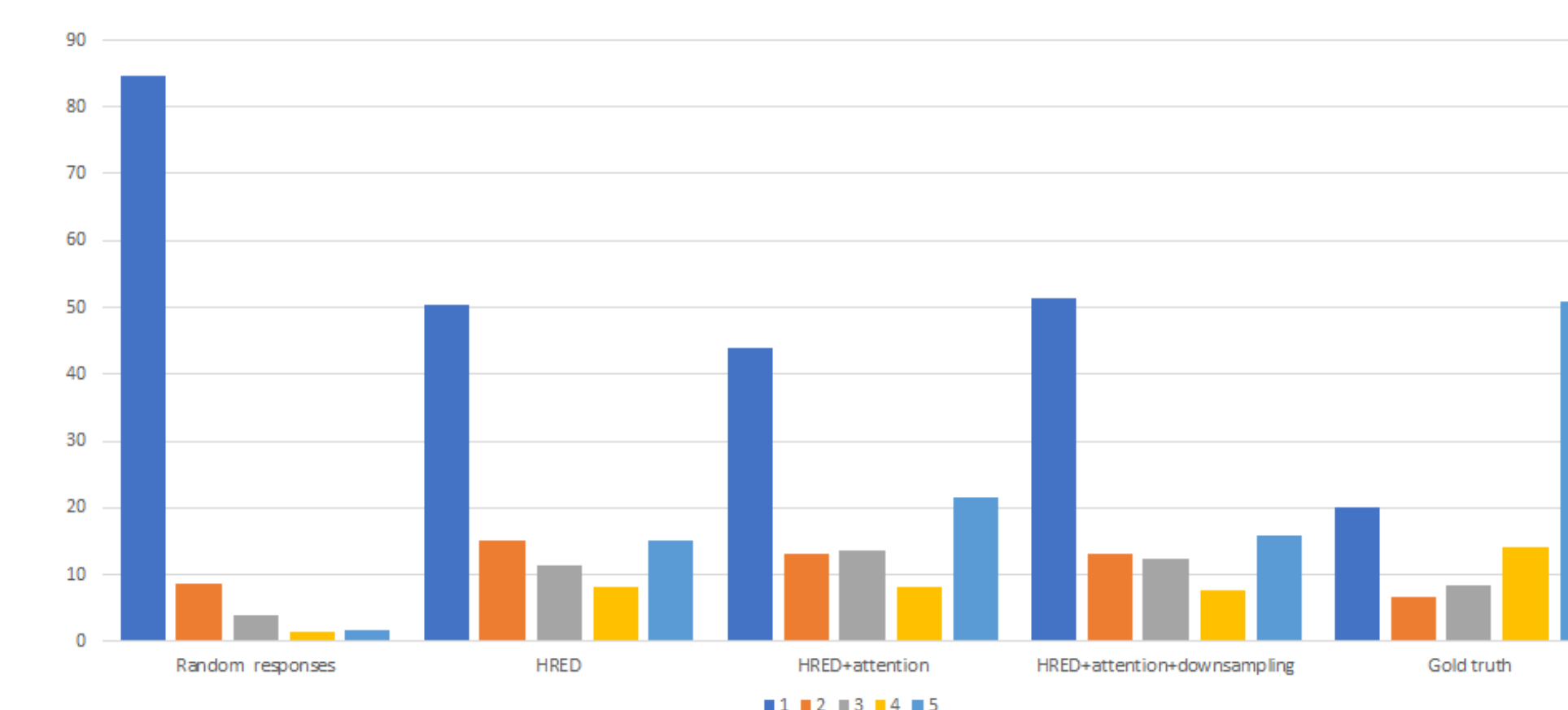


Figure 3: Human scores for each type of responses

- Sorted unique responses (x-axis) by number of occurrence to show the amount of repeated responses:

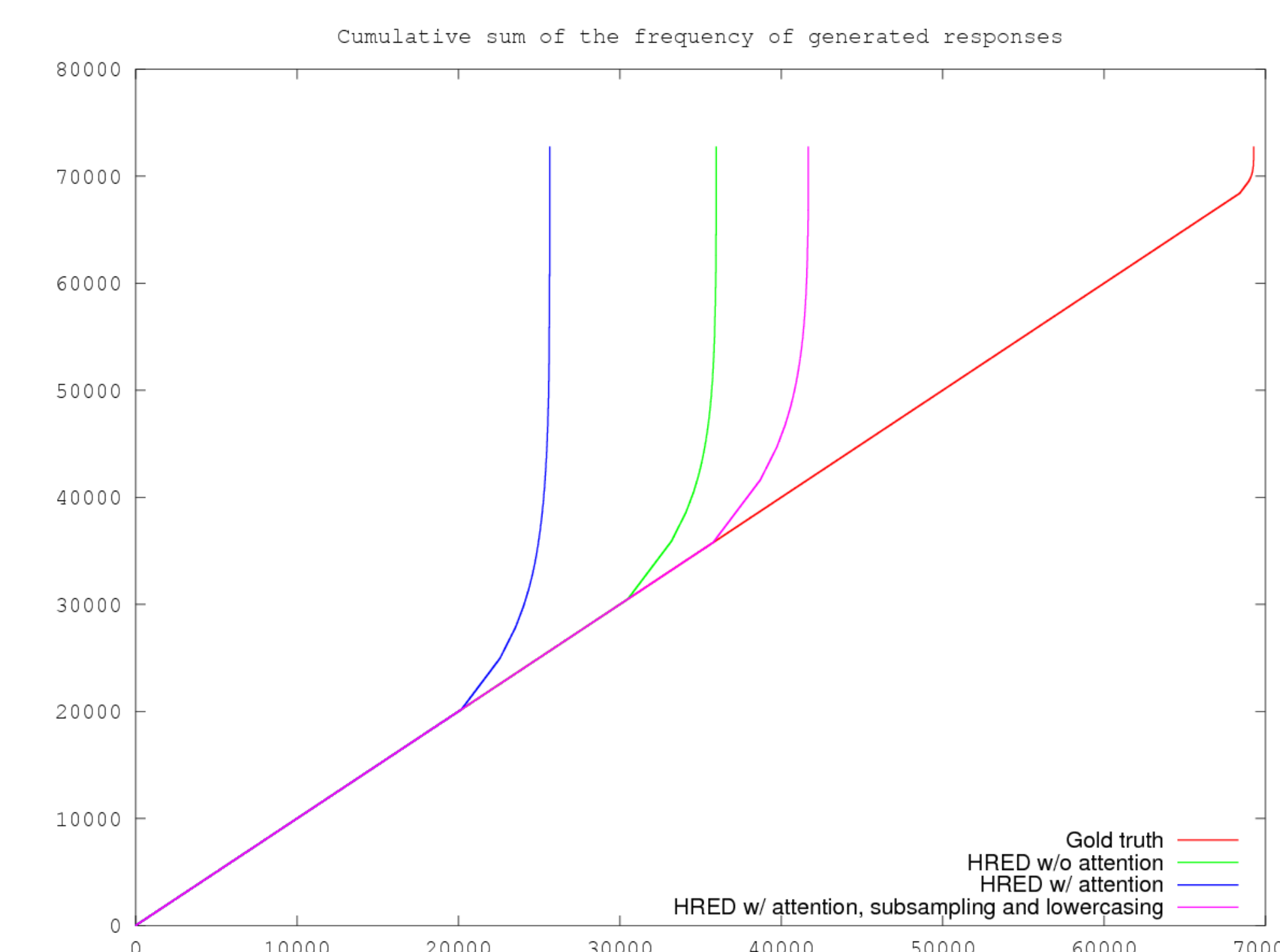


Figure 4: Cumulative sum of frequency of generated responses

## Conclusion

- Ubuntu corpus is noisy (dialogues artificially created from Ubuntu chat rooms) and assumes access to external information (e.g. Ubuntu's "paste-bin") which is not available, making the task even harder.
- HRED seems to be mostly ignoring encoded dialogue when generating responses. Adding attention mechanism didn't help.
- Generic responses are "frequent", short and thus easier to generate. While valid in a variety of contexts, more detailed responses are preferable.
- Evaluation of dialogue systems remains a challenge: different metrics tell different stories, human evaluation is still the only valid solution.

## References

- [1] Ryan Thomas Lowe, Nissan Pow, Iulian Vlad Serban, Laurent Charlin, Chia-Wei Liu, and Joelle Pineau. Training end-to-end dialogue systems with the ubuntu dialogue corpus, 2017.
- [2] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units, 2015.
- [3] Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models, 2015.
- [4] Romain Paulus, Caiming Xiong, and Richard Socher. A deep reinforced model for abstractive summarization, 2017.

## Acknowledgements

The authors gratefully acknowledge financial support for this work by Nuance Communications Inc. We would like to thank Iulian Serban for his input into these experiments and for interesting discussions.

## Contact Information

- Web: <http://cs.mcgill.ca/~nangel3>
- Email: [nicolas.angelard-gontier@mail.mcgill.ca](mailto:nicolas.angelard-gontier@mail.mcgill.ca)