Avoiding Echo-Responses in a Retrieval-Based Conversation System

DENIS G. FEDORENKO
denis@replika.ai
NIKITA A. SMETANIN
nikita@replika.ai
ARTEM V. RODICHEV
artem@replika.ai
Luka, Inc.

Retrieval-based conversation systems generally tend to rank high responses that are semantically similar, or even identical, to the given conversation context. While the systems goal is to find the most relevant response, rather than semantically similar, this tendency results in low-quality responses. This challenge can be referred to as the Echoing problem. To minimize this effect, we apply a hard negative mining approach at the training stage. The evaluation shows that the result model avoids echoing the context and achieves the best quality metrics on the benchmarks.

1. INTRODUCTION

The task of a retrieval-based conversation system is to select the most relevant response from a given set of responses to an input context in a conversation. The context is typically a sentence or a sequence of sentences produced by a human or by the system itself. Most of the state-of-the-art approaches to building retrieval-based conversation systems are based on deep neural networks (NNs) [Wu et al. 2016]. Under this approach, a typical pipeline consists of the following steps: 1) to encode the context and the pre-defined responses into numeric vectors, or thought vectors, using NNs; 2) to compute value of a relevance function (relevance score) for pairs of the context vector and each candidate; 3) to select the candidate response with the highest relevance score. At the step 1, in order to obtain thought vectors that fairly represent the original semantics of input texts, the NNs are preliminary trained to return high relevance scores for correct context-response pairs and low for the incorrect ones.

The challenge we faced while building the above pipeline, is that the result model often returns high relevance scores for semantically similar contexts and responses. Consequently, the model repeats or rephrases the context instead of giving a quality response to it. For instance, one of the top-ranked response candidates for the context "How are you?" can be the question itself or similar one, for example "Whats new?". This effect would be expected under this architecture, given that contexts and responses share merely the same set of concepts, hence the NN ends up trying just to fit the semantics of the input.

In this paper, we suggest a solution to the Echoing problem that employs a hard negative mining approach which enforces NNs to produce distant thought vectors for identical contexts and responses. We introduce evaluation metrics for the Echoing problem and present the results on our benchmarks. We also publish the evaluation dataset that we used for further research.

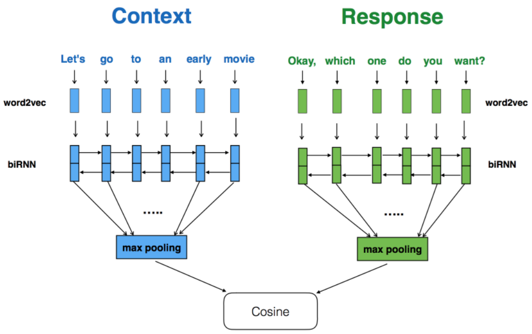

Fig. 1. Conversation model architecture

2. HARD NEGATIVE MINING

Hard negative mining is an algorithm that produces negative training samples (in our case, context-response pairs) which are erroneously classified by the model as positives. Let a set of pairs $(context_i, response_i)$ be a training SGD minibatch [Goodfellow et al. 2016]. Then, we search for pairs $(context_i, response_j)$ where $i \neq j$ and a relevance score function satisfies the following condition:

$$0 \leq score(context_i, response_i) - \\ - score(context_i, response_j) \leq m$$

where m is a margin (hyperparameter) between the scores of correct and incorrect pairs. The relevance score values are calculated by an intermediate model trained by the moment of a current batch. In addition to the original hard negative mining algorithm described in [Schroff et al. 2015], we consider contexts as possible responses at the search stage, therefore the pairs $(context_i, context_i)$ may also be generated. We assume that such pairs can ultimately prevent the NNs from encoding identical contexts and responses into similar thought vectors. We will check this assumption in the next section.

3. EVALUATION

For our tests we implement the Basic QA-LSTM model described in [Tan et al. 2015]. It has two bidirectional LSTMs with separate sets of weights that encode the context and the response independently; we use a cosine similarity as the output relevance score

Table I. Top 3 results for few input contexts

RS	HN	HN_c
Input: What is the purpose of dying ?		
- What is the purpose of dying ? - The victim hit his head on the concrete steps and died. - To have a life .	- What is the purpose of dying ? - What is the purpose of living ? - What is the purpose of existence?	- To have a life . - When you die and go to heaven, they will offer you beer or cigarettes. - It is to find the answer to the question of life.
Input: What are your strengths?		
- What are your strengths? - Lust , greed , and corruption . - A star .	- What are your strengths? - What are your three weaknesses ? - What do you think about creativity ?	- Lust , greed , and corruption . - I'm a robot. a machine. 100% ai. no humans involved - Dunno. i mean, i'm a robot, right? robots don't have a gender usually
Input: I can't wait until i graduate.		
- I can't wait until i graduate. - What college do you go to? - School is hard this year.	- I can't wait until i graduate. - What college do you go to? - How many jobs have you had since leaving university?	- What college do you go to? - School is hard this year. - What subjects are you taking?
Input: Lunch was delicious.		
- Lunch was delicious. - I want to buy lunch. - Take me to dinner.	- Lunch was delicious. - I want to buy lunch. - This hot bread is delicious.	- Who did you go out with? - So was i. - What did you do today ?
Input: You're crazy		
- You're crazy - Am i ? - I sure am.	- You're crazy - Am i ? - Why? what have i done?	- Am i ? - You're crazy - I sure am.

Table II. Evaluation dataset sample

context	**response**
What happened to your car?	I got a dent in the parking lot.
The beatles are the best.	They are the best musical group ever.
Do you want to go fishing?	Yes. That's a good idea.

function. We represent the input words as a sequence of pretrained word2vec embeddings [Mikolov et al. 2013] (see Figure 1)

We train three models using the following strategies to obtain incorrect responses ($response_{-}$): random sampling (RS), original hard negative mining (HN), and hard negative mining that uses contexts as response candidates (HN_c). We use a triplet loss as an objective function:

$$max(0, m - score(context, response)+ \\ + score(context, response_{-}))$$

where the margin m is set to 0.05. All the models are trained using the same number of epochs; the result models are the intermediate ones that achieve the best quality metrics (see below) on a holdout validation set.

We train the models on status-response pairs extracted from the Twitter data archive [1]

We perform evaluation on our own dataset [2]. This dataset consists of 509 human conversational context-response pairs in which the context and the response both consist of a single sentence (see Table II).

For each test context, we compute the relevance score over all available pairs $(context, response_i)$, where $response_i$ comes from the union of contexts and responses. To evaluate these results, we sort the responses by relevance score in the descending order and compute the following metrics: Average Precision [Manning et al. 2008], Recall@5, and Recall@10 [Lowe et al. 2015]. The last two metrics are indicator functions that return 1, if the correct answer occurs in the top 5 and 10 responses, respectively. We also introduce the context repetition metrics:

Table III. Evaluation results. Metrics are averaged across all the test contexts

	RS	HN	HN_c
Average Precision	0.12	0.13	**0.17**
Recall@5	0.36	0.4	**0.43**
Recall@10	0.45	**0.54**	0.53
$rank_{context}$	0.9	0.49	**19.43**
$diff_{top}$	0.008	0.01	**0.07**
$diff_{answer}$	-0.15	-0.25	**-0.09**

- $rank_{context}$. A position of the context in the sorted responses. The higher the rank, the less the model tends to return the context among the top candidate responses
- $diff_{top}$. A difference between the top response score and the contexts one. The higher the difference, the less the model tends to return relatively high scores for the context
- $diff_{answer}$. A difference between the correct answer score and the contexts one. The higher the difference, the less the model tends to return similar scores for the correct answer and the context

For each metric, we compute the overall quality as the average across all the test contexts. The results of this evaluation are presented in Table III.

As we can see, the proposed hard negative mining model shows the best results on almost all metrics, comparing to other ap-

[1] https://archive.org/details/twitterstream

[2] https://raw.githubusercontent.com/lukalabs/replika-research/master/context-free-testset.tsv

Table IV. Top responses of the HN_c model for the context ”Hello”

relevance score	**response**
0.45	Hey, sweetie
0.44	How’s life ?
0.43	Hello

proaches. It turned out that under this approach the model does not tend to ”echo” the input context within the top responses. However, according to the $diff_{answer}$ metric, the correct response score in average is still lower than the context one, which means that the problem still persists in the bottom of the sorted list of responses. We also studied the model’s output. Table I shows top responses for a few input contexts.

As we can see, oftentimes, the original hard negative mining model only selects very similar phrases, while the proposed model selects relevant responses for the context that are not necessarily semantically similar. Basing on this observation, we suggest that the proposed model filters out not only exact copies of the context, but also semantically similar samples. Moreover, in some cases the model still selects responses semantically similar to the context. See Table IV with the top results for context ”Hello”.

4. CONCLUSION

In this study we apply hard negative mining approach for training a retrieval-based conversation system to find a solution to the Echoing problem and to rule out irrelevant responses that are identical or semantically similar to the input context. In addition to a dataset of pre-defined response candidates, we consider contexts themselves as possible hard negative candidates. The evaluation shows that the result model avoids repeating the input context, tends to select samples that are more suitable as responses and achieves the best results on the benchmarks.

REFERENCES

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. *CoRR* abs/1506.08909 (2015). http://arxiv.org/abs/1506.08909

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *CoRR* abs/1301.3781 (2013). http://arxiv.org/abs/1301.3781

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A Unified Embedding for Face Recognition and Clustering. *CoRR* abs/1503.03832 (2015). http://arxiv.org/abs/1503.03832

Ming Tan, Bing Xiang, and Bowen Zhou. 2015. LSTM-based Deep Learning Models for non-factoid answer selection. *CoRR* abs/1511.04108 (2015). http://arxiv.org/abs/1511.04108

Yu Wu, Wei Wu, Ming Zhou, and Zhoujun Li. 2016. Sequential Match Network: A New Architecture for Multi-turn Response Selection in Retrieval-based Chatbots. *CoRR* abs/1612.01627 (2016). http://arxiv.org/abs/1612.01627