arXiv:1712.05626v1 [cs.CL] 15 Dec 2017

# Avoiding Echo-Responses in a Retrieval-Based Conversation System

**Denis Fedorenko**
Replika @ Luka, Inc
denis@replika.ai

**Nikita Smetanin**
Replika @ Luka, Inc
nikita@replika.ai

**Artem Rodichev**
Replika @ Luka, Inc
artem@replika.ai

## Abstract

Retrieval-based conversation systems generally tend to rank high responses that are semantically similar or even identical to the given conversation context. While the system's goal is to find the most appropriate response, rather than just 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 resulting model reduces echoing and achieves the best quality metrics on the benchmarks.

## 1 Introduction

The task of a retrieval-based conversation system is to select the most appropriate response from a set of responses given an input context of 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 retrieval-based conversation systems are based on deep neural networks (NNs) [8]. Under this approach, a typical pipeline consists of the following steps:

1. Encode the context and pre-defined candidate responses into numeric vectors, or thought vectors, using NNs;
2. Compute value of a matching function (matching score) for pairs consisting of the context vector and each candidate;
3. Select the candidate response with the highest matching score.

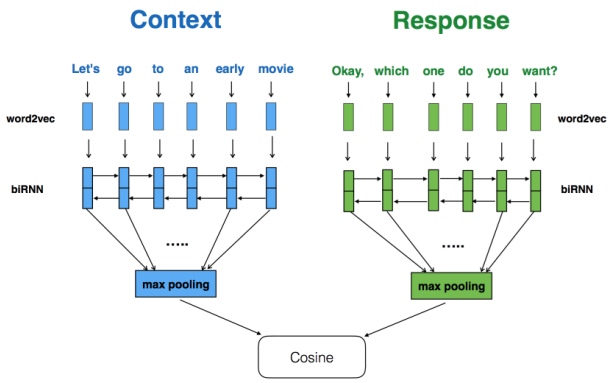


Figure 1: Conversation model architecture

At the step 1, in order to obtain thought vectors that fairly represent semantics of input contexts and responses, the conversation model is preliminarily trained to return high matching scores for correct context-response pairs and low for the incorrect ones.

The challenge we faced while building the above pipeline was that the resulting model often returned high matching scores for semantically similar contexts and responses. Consequently, the model frequently repeated or rephrased input contexts instead of giving quality responses. Consider the following examples:

A. Human: "What is the purpose of living?"
   Model: "What is the purpose of existence?"

B. Human: "What is the purpose of living?"
   Human: "It's a very philosophical question."

The effect in the conversation A contrasted by an appropriate human response in the conversation B is expected under this architecture. It is due to the fact that contexts and responses often contain the same concepts, hence during training the NNs simply ends up trying to fit the semantics of the input.

In this paper, we suggest a solution to the echoing problem based on a hard negative mining approach which enforces the conversation model to produce low matching scores for similar contexts and responses. We introduce evaluation metrics, our results and benchmarks for the echoing problem. We also publish the evaluation dataset for further research.

## 2 Hard Negative Mining

Suppose we have a dataset $D$, $|D| = N$ consisting of pairs $(context_i, response_i)$, $i \in \{1..N\}$. Our goal is to build a conversation model $M$:

$$M : (context, response) \to \mathbb{R}$$

that satisfies the following condition:

$$M(context_i, response_i) > M(context_i, response_j)$$

$\forall i, j \neq i$ and $response_j$ is not appropriate for $context_i$. In other words, the resulting model should return a higher matching score for correct responses than for incorrect ones.

To train this model, we also need incorrect context-response pairs as negative data in addition to the correct ones presented in $D$. Consider two approaches to obtain the negative pairs: random sampling and hard negative mining. Under the first approach, we randomly select $response_j$ from $D$ for each $context_i$. If $D$ is large and diverse enough, then a random $response_j$ is almost always inappropriate for a corresponding $context_i$.

In contrast to random sampling, hard negative mining imposes a special constraint on responses selected as negatives. Let $M_0$ be a conversation model trained on random pairs used as negative training data. Then, we search for a new set of negative pairs $(context_i, response_j)$, so that their matching score satisfies the following condition:

$$M_0(context_i, response_i) - M_0(context_i, response_j) \leq m$$

where $m$ is a margin (hyperparameter) between the scores of correct and incorrect pairs. The new set of pairs is used to train the next model $M_1$, which, in turn, generates the pairs for $M_2$, and so on. The process goes on until the model converge to some desired model $M_k$ [1].

The intuitive idea behind hard negative mining is to select only negatives that have relatively high matching scores and can be interpreted as errors of the conversation model. As a result, the model has faster convergence compared to random sampling [6].

Following this intuition, to solve the echoing problem we can consider contexts as possible responses, therefore the pairs $(context_i, context_i)$ may be selected as hard negatives. In the next section we demonstrate that this approach can ultimately prevent the conversation model from ranking high similar contexts and responses.

Table 1: 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. |
| What do you think about Britney Spears? | Oh, she's a great singer. |
| White coffee, no sugar please. | Here you are. |
| I'm joining the army. | You're kidding. You might get killed. |

# 3 Experiments

For our experiments we implement a model similar to Basic QA-LSTM (see Figure 1) described in [7]. It has two bidirectional LSTMs of the size 1024 with separate sets of weights that encode context and response independently. We use a cosine similarity as the output matching function. We represent input words as a sequence of pre-trained word2vec embeddings of the size 256 [5]. Word sequences longer than 20 words are trimmed from the right, and the context encoder is fed with only one dialog step at a time.

## 3.1 Models

In order to study the impact of hard negative mining on the echoing problem, we train three models using the following strategies: random sampling ($RS$), a hard negative mining based on responses only ($HN$), and a hard negative mining based on both responses and contexts ($HN_c$).

## 3.2 Datasets

We train the models on 79M of tweet-reply pairs from a Twitter data archive [1].

We perform evaluation on our own dataset [2]. This dataset consists of 759 context-response pairs from human text conversations where context and response both consist of a single sentence (see Table 1). We split the dataset into validation and test subsets consisting of 250 and 509 pairs, respectively.

## 3.3 Training

The models are trained with the Adam optimizer [2] with the size of mini-batches set to 512. We train the models for one epoch (154821 mini-batches in total). Intermediate models that show best quality metrics on the validation set are selected as the resulting models.

We use a triplet loss [6] as an objective function:

$$max(0, m - M(context_i, response_i) + M(context_i, response_j))$$

where the margin $m$ is set to 0.05. For each pair $(context_i, response_i)$, $response_j$ is only selected within the current mini-batch using an intermediate model $M$ trained by the moment of this batch. We also apply an additional constraint to the hard negative responses:

$$M(context_i, response_i) > M(context_i, response_j)$$

This constraint results in a faster convergence of our models.

## 3.4 Metrics

For each $context$ from the test set, we compute matching score across all available pairs $(context, response_i)$, where $response_i$ comes from the union of contexts and responses. To evaluate these results, we sort responses by the matching score in a descending order and compute the following metrics: Average Precision [4], Recall@2, Recall@5, and Recall@10 [3]. The last three metrics are indicator functions that return 1, if the correct answer occurs in the top 2, 5 and 10 responses, respectively. We also introduce the context repetition metrics:

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

[2] https://github.com/lukalabs/replika-research/tree/master/context-free-dataset

Table 2: Evaluation results

| | $RS$ | $HN$ | $HN_c$ |
|---|---|---|---|
| Average Precision | 0.12 | 0.13 | **0.17** |
| Recall@2 | 0.18 | 0.23 | **0.29** |
| 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** |

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

For each metric, we compute the overall quality as the average across all the test contexts.

### 3.5 Results

The results of the evaluation are presented in Table 2. As we can see, the proposed hard negative mining model achieves the highest values in almost all metrics compared to other approaches. It turns out that under this approach the model does not tend to rank input contexts high and have them in the top candidate responses. Still, according to the $diff_{answer}$ metric, the score of a correct response is, on average, lower than the one of a context, which means that the problem still persists in the bottom of the sorted list of responses.

We also studied the model's output. Examples of top-ranked responses for different contexts are presented in Table 4. As we can see, oftentimes the $RS$ and $HN$ models select identical or very similar responses, while the $HN_c$ model selects appropriate responses that are not necessarily semantically similar. Based on this observation, we suggest that the proposed model filters out not only exact copies of the context, but also samples with a similar semantics. Moreover, in some cases the model selects semantically similar responses which are, at the same time, appropriate for a given context. See Table 3 with the top results for context “Hello".

## 4 Conclusion

In this study, we applied a hard negative mining approach to train a retrieval-based conversation system to find a solution to the echoing problem, that is to reduce inappropriate responses that are identical or too similar to the input context. In addition to a dataset of predefined response candidates, we consider contexts themselves as possible hard negative candidates. The evaluation shows that the resulting model avoids repeating the input context, tends to select samples that are more appropriate as responses and achieves the best results on a variety of benchmarks.

Table 3: Top responses of the $HN_c$ model for the context “Hello"

| **matching score** | **response** |
|---|---|
| 0.45 | Hey, sweetie |
| 0.44 | How’s life ? |
| 0.43 | Hello |

Table 4: Top 3 results for few input

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

## References


[1] Olivier Canévet and François Fleuret. Efficient sample mining for object detection. In *Proceedings of the 6th Asian Conference on Machine Learning (ACML)*, number EPFL-CONF-203847, 2014.

[2] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[3] Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *CoRR*, abs/1506.08909, 2015.

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

[5] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

[6] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *CoRR*, abs/1503.03832, 2015.

[7] Ming Tan, Bing Xiang, and Bowen Zhou. Lstm-based deep learning models for non-factoid answer selection. *CoRR*, abs/1511.04108, 2015.

[8] Yu Wu, Wei Wu, Ming Zhou, and Zhoujun Li. Sequential match network: A new architecture for multi-turn response selection in retrieval-based chatbots. *CoRR*, abs/1612.01627, 2016.