Avoiding Echo-Responses in a Retrieval-Based Conversation System

Denis Fedorenko, Nikita Smetanin & Artem Rodichev
Replika AI Team @ Luka, Inc
{denis,nikita,artem}@replika.ai

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. Under this approach, a typical pipeline consists of the following steps: 1) to encode the context and the predefined 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.

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**: ***Echoing problem***. 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?.

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.

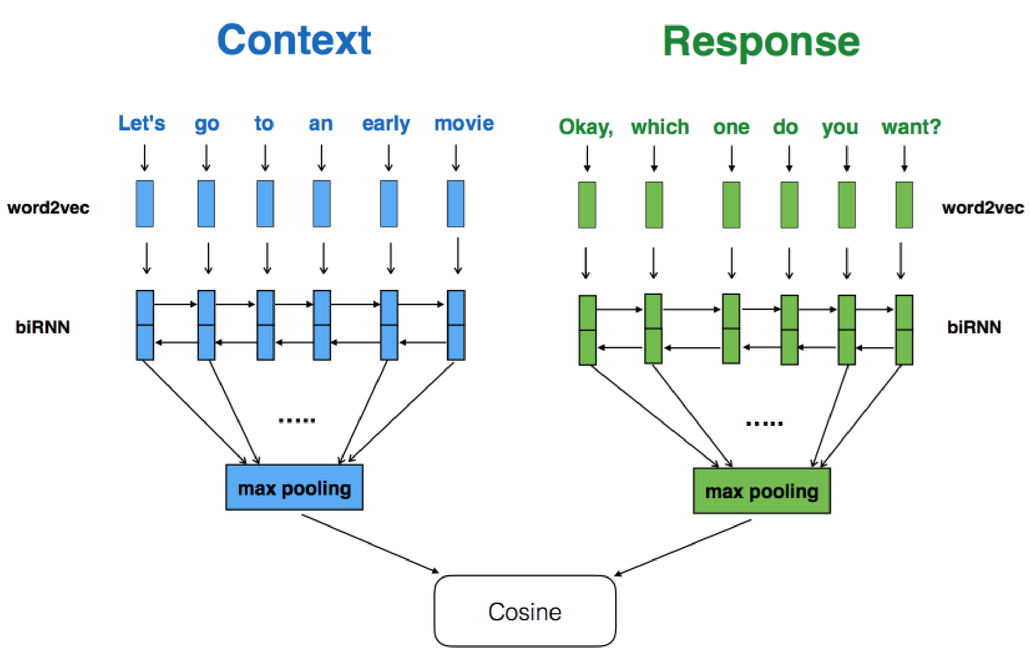

Figure 1: Conversation model architecture

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.

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 function. We represent the input words as a sequence of pre-trained 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.

We train the models on status-response pairs extracted from the Twitter data archive *https://archive.org/details/twitterstream*

We perform evaluation on our own dataset *https://raw.githubusercontent.com/lukalabs/replika-research/master/context-free-testset.tsv*. 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 1).

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.

Table 1: Evaluation dataset sample

- $rank_{context}$. 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

	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**

Table 2: Evaluation results. Metrics are averaged across all the test contexts

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 4 shows top responses for a few input contexts.

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

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

Selected references

1. 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
2. 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

Random Sampling (RS)	*Hard Negatives* (HN)	*Hard Negatives with user dialog contexts* (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 4: Top 3 results for few input contexts