Strategy of the Negative Sampling for Training Retrieval-Based Dialogue Systems

Aigul Nugmanova
ITMO University
Speech Technology Center
St. Petersburg, Russia
nugmanova@speechpro.com

Andrei Smirnov*
andrey.smirnov@juvo.ru

Galina Lavrentyeva
STC-innovations Ltd
ITMO University
St. Petersburg, Russia
lavrentyeva@speechpro.com

Irina Chernykh
STC-innovations Ltd
ITMO University
St. Petersburg, Russia
chernykh-i@speechpro.com

***Abstract*—The article describes the new approach for quality improvement of automated dialogue systems for customer support service. The analysis produced in the paper demonstrates the dependency of the quality of the retrieval-based dialogue system on the choice of negative responses required for training the system. The proposed approach implies choosing the negative samples according to the distribution of responses in the training set. In this implementation, the negative samples are randomly chosen from the original response distribution and from the "artificial" distribution of negative responses, such as uniform distribution or the distribution obtained by transformation of the original one. The results obtained for the implemented systems and reported in this paper confirm the significant improvement of automated dialogue systems quality in case of using the negative responses from transformed distribution.**

Keywords—dialogue systems, negative sampling, dual encoder, retrieval-based dialogue systems

I. Introduction

Automated dialogue systems in the customer support service recently became more popular area of research in the field of natural language processing [1]. Such systems can be used as a separate models or as a part of pervasive and mobile speech systems [2]. One approach to develop such systems uses the retrieval-based dialogue models. Such models can use the unlabeled data during the training and their responses are predictable because they use only responses from the training set [3, 4, 5, 6, 7, 8, 9]. For training these models it is important not only to customize the architecture but also to create appropriate training data. For example, the most recent research [4] shows the impressive improvement of the dialog system quality by usimg the weighting model for preparing training data [4].

In this paper we show how negative sampling strategy affects the performance of dialogue system. The main goal of the investigation the negative sampling methods is to form more effective training set. Random selection of negative samples allows adding a lot of identical examples to the training set if the original data contains repetitions. Our research shows how the training set can be prepared to be more diverse after simple transformations. The similar approach was described earlier in [10, 11, 12]. In [10] authors introduce negative sampling idea based on the concept of noise contrastive estimation (similar to generative adversarial networks), which implies, that a good model should differentiate fake signal. To achieve this goal several negative examples for every positive example are sampled from training data as noise examples and used to train the model. Authors use the noise distribution to choose negative samples by transforming the unigram distribution. We take it into account in our research and try to improve our systems by transforming the response distribution in order to choose more appropriate negative responses for retrieval-based dialogue systems training.

The dialogue systems investigated in this paper use neural network architecture performed in [5]. Neural network was used in two ways: to calculate the response probability for current question and for obtaining the text representation in order to find the nearest question.

Section 2 describes the architecture of the dialogue system. In section 3 the negative sampling is performed. Section 4 and section 5 contain the data description and definition of the evaluation metrics used during the experiments. Obtained results are reported in section 6. And section 7 concludes the produced investigation.

II. Architecture

Dialogue systems, considered in this paper, are based on Siamese network like Dual Encoder Model, presented in [5]. It is the retrieval-based model. The main idea of this approach is to find the best response for current context The context here includes user's question and the previous utterances of the dialogue in training set.

We use this architecture in two ways and with two kinds of encoders. Fist approach uses the dual encoder model to find a pairwise probability of context and response similar to [3, 5]. The second approach uses encoder model only to get sentences embeddings. In this case two types of neural encoder are considered: first is based on GRU cell and second uses Attention layer only.

A. Dual Encoder Model

Similar to [5] we use pair probability of context and response to find the best response.

The process of calculating the probability between current context and response can be described as follows:

- Context and response are divided into words sequences and initialized with the word embeddings. In this way

*Work done while the author was at Speech Technology Center.

978-1-5386-9151-9/19/$31.00 ©2019 IEEE

two matrices with dimensions: sequence length, word embeddings dimension size are obtained;

- These matrices are used as input layer of the encoder. As an output the encoder produces the representation for context and response sequences as illustrated in Fig. 1;
- For pairwise probability calculation the sigmoid function is applied to the product of context vector c with weight matrix M and response vector r.

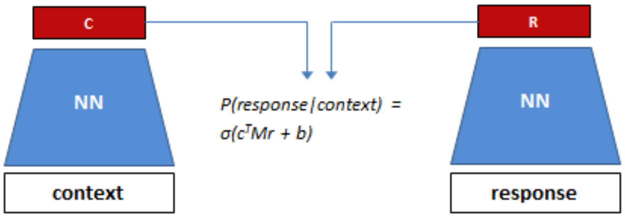

Fig. 1. Scheme of Dual Encoder model [3]

All the responses are then sorted by their probabilities. We presume that the highest-ranked response is the most appropriate response for current context.

The model used here is based on recurrent neural network with Gated Recurrent Unit (GRU) and hidden size 128. All models have the best result after 20000 iterations. For word embeddings we use N = 300 dimensional Word2Vec embeddings matrix pretrained on Russian dialogues corpus, included the target data and dialogues from popular websites. Words of the training set which are not included in the pretrained model are initialized by the average vector of word embeddings.

B. Embedding-Based Model

As an alternative approach, the architecture described in section 2.1 is used only for obtaining the contexts vectors. Here the output produced by the encoder is used to obtain the sentences representations.

In this approach we presume that the best response is in pair <nearest context, response> and we use this response as a correct answer. The similarity between current context vector and those from the training set is estimated using the cosine distance scoring. For searching the most similar context the representations of all context sequences in the training set are extracted. Further for current user question with previous dialog utterances, that all together contain the context of the dialog, the context representation vector is also obtained and the cosine distance is calculated.

In order to reduce the architecture of the network used for embedding extraction, the RNN layer was excluded from the original architecture leaving only the attention layer.

Our experiments show that the representations are more effective for employment in dialogue system if a linear combination of context and response representations is used for the search. It is called history vector and is expressed in (1).

$$history_i = context_i + c_r\, response_i \qquad (1)$$

where i is a number of pair in the training set, $context_i$ is a current context vector, $response_i$ is a current response vector, c_r is a response weight. In our experiments we use c_r as a free parameter and the best results are obtained for $c_r = 0.4$.

III. Negative Sampling

For training the systems, described in section 2, the negative sampling strategy is usually used. It helps to add the incorrect training examples into the training set. In this research we studied how the negative sampling strategy influences the quality of dialogue systems. We used several datasets prepared with the use of negative sampling methods described below.

For training the neural network with architecture described in Section 2 pairs <context, response> in each dialogue (where "context" is the concatenation of the current question and the previous utterances of the dialogue) are used as a training example of real (positive) responses. As negative samples N pairs <context, negative response> are used, where negative responses are incorrect answers selected from the training dialogues according to one of the techniques described below. We use a 1:5 ratio between positive and negative responses.

A popular approach to choose negative responses for concrete context is a random response selection from other dialogs. We suppose this approach is not optimal, because the most uninformative frequent utterances fall into subsamples more often than rare informative utterances. To overcome this problem we suggest to change the responses distribution and to choose responses for negative samples from the transformed distribution.

In our experiments 4 methods of negative response selection are considered:

- The real response distribution is taken into account. Responses for negative samples are selected randomly.
- The response distribution is transformed into the uniform distribution and responses for negative samples are selected from the obtained one.
- The responses are selected from the new transformed distribution obtained by raising the initial distribution to some power. It is important to note that negative degree helps to reduce the amount in frequent sentences of the base among negative samples.
- The latter approach also aims at bringing the response distribution closer to the uniform. But in this method, the responses distribution influences not only the choice of the negative answer but also the probability of the example entering the training set. For this purpose, the amount of occurrences in the dataset of dialogues for each answer from the current pair <context, answer> is calculated (N). The pair <context, answer> is added to the set of training data only with probability 1 / N.

To take into account the semantic similarity between phrases and to approximate the probability density for responses in the dataset by a continuous density function we apply a kernel-density estimation using Gaussian kernels. In our experiments we use the bandwidth value 0.4.

IV. Data

In this paper for training and evaluation of the proposed method the Russian language dataset with the human-human unstructured conversation without any labels was used. The dataset is a chat log of technical support of the web portal. It contains 25000 dialogues with an average of 4 turns.

TABLE I. Example of Dialogue Between a user and an Operator

English (translate):
Q1: Hello! How can I register in the web service? **A1:** Hello, my name is <name> and I will be glad to help you. Registration on the portal is available by a personal visit to the Service Center or on your own, which of the following ways would be more comfortable for you?
Q2: Thanks for the answer, I will come myself. **A2:** For registration, you need to contact the Service Center that is convenient for you. You can see the addresses by clicking on the <link>. You need to have a passport and SNILS.
Q3: OK. **A3:** Do you have any question about the portal?

Table 1 demonstrates conversation examples translated to English language. The data were divided into 3 parts with ratio 80:10:10 for training, automatic evaluation and human evaluation.

Dialogues presented in the dataset contain a big amount of uninformative utterances. In this case the amount of uninformative responses in the training set will be much bigger than the amount of informative ones, which will lead to low performance of the final model. For example, the beginning of the most dialogues includes greetings and the ending of dialogues includes valedictions. Sometimes an operator can ask users to wait while he is looking for the information. Also, the examples of frequently responses are “Yes” or “No”.

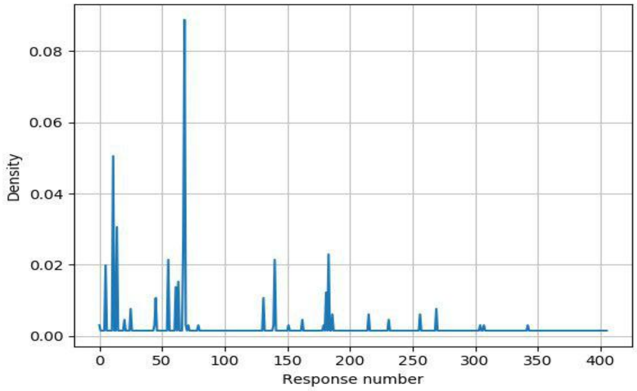

Fig. 2. The response distribution

Fig. 2 illustrates the responses distribution curve for the first 1000 dialogues in the dataset. This curve demonstrates that the set contains phrases with very high frequency. Most of them are uninformative. Reducing the influence of these responses is one of the aims to investigate negative sampling methods.

V. Evaluation Metric

To evaluate the quality of the proposed models the automatic and human evaluation methods are used. Both are based on the recall@k metric, similar to the evaluation of the retrieval-based systems in [4, 5, 6].

A. *Automatic evaluation*

Test set includes 2500 dialogues. For each pair <context, response> sampled from the test set m alternative negative responses are selected. These m+1 responses are then ranked according to its similarity to the context. The output is defined as the right answer if the original <context, response> pair appears in top-k among all m+1 candidates. In experiments we use m=9.

B. *Human evaluation*

We also apply a human evaluation in our research. For human evaluation 400 test questions were specially selected by the experts from the corresponding test set. These questions contained only targeted questions requiring a meaningful response. In the experiments the responses are chosen according to the ranking of the training dialogs and are selected according to the highest probability value. The responses obtained by several models trained on the data with negative sampling from different distribution are evaluated.

Our model selects three responses for each test question. For each of 1200 selected responses two assessors rate the consistency between context and response using a 4-point scale. The response is marked as: 0, if the response is incorrect; 1, if the response can be interpreted as correct by the user which is not an expert in the field; 2, if the response includes information of correct answer; 3, when it is a reference answer.

Also we take into account that the human marks can be changed over time between evaluations and therefore we fill in the test table by responses of different models and then shuffle it.

Based on the results of estimates two metrics are calculated: recall@3 for correct response (CR) and recall@3 for unsure response (UR). Recall@3 for correct response is equal 1 if in three responses selected by the model there is at least one with human mark above 1. Similarly, recall@3 for unsure response is declared to be 1 if there is at least one with human mark above 0 among three responses.

VI. Results

At first, we tested our models automatically with recall@k metrics. In the test set the alternative responses were chosen from the distribution of the training set. The model was trained on the training set with negative samples from initial and uniform distributions. Each model was then evaluated on the test sets with alternative responses from both of these distributions.

TABLE II. RECALL@3 FOR CORRECT RESPONSES (CR) AND UNSURE RESPONSES (UR) BASED ON HUMAN EVALUATION FOR MODELS TRAINED USING THE DATA WITH DIFFERENT DISTRIBUTIONS IN NEGATIVE SAMPLING (NS).

Approaches		**randomly NS (baseline)**	**NS from uniform distribution**	**NS from base distribution in -0.125 degree**	**NS from base distribution in -0.25 degree**	**NS from uniform distribution and filtered dialogues**
DE GRU	UR	0.40	0.45	0.49	0.46	0.44
	CR	0.24	0.23	0.22	0.22	0.20
DE emb GRU	UR	0.76	0.79	0.77	0.75	0.8
	CR	0.42	0.45	0.46	0.45	0.46
DE emb ATT	UR	0.7	0.7	0.71	0.72	0.71
	CR	0.40	0.44	0.41	0.47	0.48

The results presented in Table 3 confirm that models show poor quality on the test samples with transformed response distribution. This indicates that for automatic evaluation it is important that test and training responses are sampled from the same distribution. Otherwise the actual increase of the dialogue system quality with different negative sampling strategies cannot be estimated.

We presume that the human evaluations show the difference between models better. Table 2 presents the CR and UR (such as in Section 5.2) metrics for three models: Dual Encoder with GRU cell (DE GRU), embeddings from Dual Encoder with GRU cell (DE emb GRU) and embeddings from Dual Encoder with Attention layer (DE emb ATT).

Evidently, any changes in the response distribution, which align it, leads to higher quality of dialogue systems based on embeddings from encoder in terms of human marks. Moreover when we use the dual encoder model to rank responses in the training set, we can use the degree of the response distribution as a free parameter and achieve improvement by selecting the more suitable degree value. For example we achieved the best quality using the degree=-0.125. Also table 2 demonstrates that filtering dialogues during the training can be effective for text representations, but it does not improve the dual encoder based model.

Comparisson of the experimental results for different models show that when the training set is not big enough the embeddings-based model works much better than the full model of the dual encoder (up to 2 times in our case with 20000 dialogues).

TABLE III. RECALL@1 VALUES FOR GRU DUAL ENCODER MODEL WITH NEGATIVE RESPONSES FROM THE ORIGINAL RESPONSE DISTRIBUTION AND UNIFORM DISTRIBUTION

Test set (alternative responses)	**Training set (negative samples)**	**Recall@1**
initial distribution	**initial**	**0.57**
	uniform	0.45
transformed distribution (**uniform**)	initial	0.61
	uniform	**0.69**

Also in Table 2 it is noticeable that the model that uses the GRU in the encoder shows better UR and CR estimations than the analogous architecture with only attention layer do.

VII. CONCLUSION

This paper reports the detailed analysis of the negative sampling strategy for training retrieval-based dialogue systems with several architectures. The conducted experiments confirm that using the proposed negative sampling strategy instead of the random sampling helps to achieve a relative improvement up to 20% in terms of the dialogue system quality by human evaluation. It is also shown that the embedding based model demonstrates twice better results than the full dual encoder model on our data.

Our future objective is to consider other methods of negative sampling with the use of additional information extracted from the data such as topic clustering or number of a turns in dialogues.

VIII. ACKNOWLEDGEMENTS

This work was financially supported by the Ministry of Education and Science of the Russian Federation, Contract 14.575.21.0178 (ID RFMEFI57518X0178)

REFERENCES

[1] Chakrabarti, Chayan, and George F. Luger. "Artificial conversations for customer service chatter bots: Architecture, algorithms, and evaluation metrics." Expert Systems with Applications 42.20 (2015): 6878-6897.

[2] McTear, Michael. "New directions in spoken dialogue technology for pervasive interfaces." *Robust and Adaptive Information Processing for Mobile Speech Interfaces* (2004): 57.

[3] Kadlec, Rudolf, Martin Schmid, and Jan Kleindienst. "Improved deep learning baselines for ubuntu corpus dialogs." arXiv preprint arXiv:1510.03753 (2015).

[4] Lison, Pierre, and Serge Bibauw. "Not All Dialogues are Created Equal: Instance Weighting for Neural Conversational Models." arXiv preprint arXiv:1704.08966 (2017).

[5] Lowe, Ryan, et al. "The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems." arXiv preprint arXiv:1506.08909 (2015).

[6] Ouchi, Hiroki, and Yuta Tsuboi. "Addressee and response selection for multi-party conversation." Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016.

[7] Wang S., Jiang J. Learning natural language inference with LSTM //arXiv preprint arXiv:1512.08849. – 2015.

[8] Inaba M., Takahashi K. Neural utterance ranking model for conversational dialogue systems //Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue. – 2016. – C. 393-403.

[9] Yan R., Song Y., Wu H. Learning to respond with deep neural networks for retrieval-based human-computer conversation system //Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. – ACM, 2016. – C. 55-64.

[10] Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems. 2013.

[11] Saeidi, Marzieh, et al. "The Effect of Negative Sampling Strategy on Capturing Semantic Similarity in Document Embeddings." Proceedings of the 2nd Workshop on Semantic Deep Learning (SemDeep-2). 2017.

[12] Zhekova, Desislava. "Instance Sampling for Multilingual Coreference Resolution." Proceedings of the Second Student Research Workshop associated with RANLP 2011. 2011.