Cross-lingual Training for Retrieval-based Dialogue Systems*

Nikita Mamaev[1,2][0000−0002−4836−4044], Aigul Nugmanova[2][0000−0002−9167−5892], Mikhail Khovrichev[2][0000−0001−7436−9538], Anna Bulusheva[2][0000−0002−3602−5328], and Irina Chernykh[1,2][0000−0001−5774−6370]

[1] ITMO University, 49 Kronverksky Pr., Saint Petersburg, 197101, Russia
[2] Speech Technology Center, 4 Krasutskogo St., Saint Petersburg, 196084, Russia

{mamaev-n,nugmanova,khovrichev,bulusheva,chernykh-i}@speechpro.com

Abstract. In recent years, cross-lingual approaches have been successfully applied to a variety of tasks and have shown potential when available data is scarce. That potential can be effectively leveraged when building a retrieval-based dialogue system. In this paper we investigated different methods of cross-lingual training of retrieval-based dialogue systems. We compare several cross-lingual approaches, including adversarial pretraining on resource-rich language dataset and further fine-tuning of pretrained models with low-resource language dialogue data. This adversarial architecture extends Dual Encoder network with language discriminator. Other approaches are based on different training strategies, such as mixing data in different languages, adversarial learning and pretraining on big data. Experiments show that adversarial learning performs competitively, which is also true for the data mixing strategy.

Keywords: Cross-Lingual · Dialogue System · Retrieval-Based · Adversarial Learning.

1 Introduction

Retrieval-based dialogue systems have received a great amount of attention recently, mainly due to their predictability and more reasonable data requirements compared to their generative counterparts. However, retrieval-based models are also becoming larger, needing more data in order to be successfully trained. It is quite difficult to get a large dataset with human-human dialogues, sufficient for training a retrieval-based system. Especially when it is not English language because most languages have limited resources.

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

The lack of labeled data in most languages remains an open question and attracts more and more attention from researchers. For example, the work of Chen et al. [1] studies this issue; they attempt to compensate for the lack of labeled data in different languages. In the work of Chidambaram et al. [2] the investigated problem is to explore the cross-lingual approach for building retrieval-based cross-lingual dialogue systems by maximizing the representational similarity between sentence pairs drawn from parallel data.

In our work we focus on the possibility of pretraining models on large dataset and fine-tuning on the target case. One of the proposed approaches uses an adversarial component as language discriminator for transfer semantic relations from one language to another just as it was in the work of Chen et al. [1]. We suggest it simplifies the adaptation of model to target domain in dialogue systems too. Our experiments compare the proposed approach with several other training strategies, such as training on the blending bilingual data and consequent training with pretraining on `SOURCE` data and future fine-tuning on `TARGET`. Our Adversarial approach show good performance and in some cases surpasses all alternative approaches.

2 Architectures & methods

Retrieval-based dialogue systems. Over the last decade, various new architectures have been proposed in attempt to maximize performance on response retrieval datasets. Most of these use a recurrent network to construct either a flat [3], [4] or a hierarchical [5] context representation. For our experiments, we have chosen Dual Encoder for its simplicity. This model features an RNN encoder $\mathcal{E}$ with word embedding layer to produce context and response embeddings. The probability of an utterance being a suitable reply considering the current context is given by $\sigma(c^T M r)$, where c, r are the context and response embeddings, respectively, and M is a learnable parameter. We refer to the module that performs this computation as the matcher, $\mathcal{M}$. A diagram of Dual Encoder can be seen in Figure 1 (right).

Pretraining and transfer learning. Pretraining a deep neural model using data from a different domain is a common method for finding a good initialization point for successful further training, or fine-tuning, on target data. A large number of works from various fields prove the effectiveness of this procedure and explore the principle behind it. For example, the work of Erhan et al. [6] demonstrated that pretraining acts as a regularization mechanism, enabling better generalization in deep neural networks. Recent works in NLP also show the importance of unsupervised pretraining for language modeling [7] and machine translation [8].

In the current work, we investigate the potential of pretraining a retrieval-based dialogue model using data in another language for improving performance in target language. Conceptually, we attempt to perform a sort of transfer learning by using weights of the model trained on the source dataset as a starting point for fine-tuning on the target dataset.

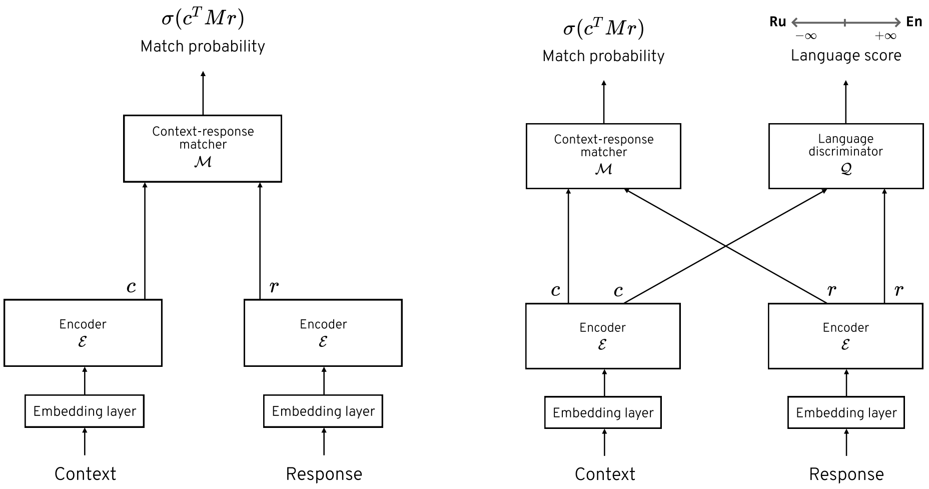

Fig. 1. Original Dual Encoder architecture (left) and architecture enabling adversarial learning of Dual Encoder (right).

Adversarial learning. To extend our base model to a scenario where the available data is bilingual, we add the language discriminator, $\mathcal{Q}$, which aims to identify the language of an input text. Specifically, the inputs of $\mathcal{Q}$ are embeddings given by $\mathcal{E}$ of pairs of contexts and responses in both languages. Similarly to ADAN [1], $\mathcal{Q}$ models an unbounded Lipschitz function with $K = 0.05$, and strives to minimize the Wasserstein distance between $P_{\mathcal{E}}^{src}$ and $P_{\mathcal{E}}^{tgt}$, which are the distributions of hidden features of $\mathcal{E}$ when the data is either in the source or the target language, respectively:

$$\begin{aligned} J_q(\theta_e) \equiv \quad & \max_{\theta_q} \mathop{\mathbb{E}}_{\mathcal{E}(x) \sim P_{\mathcal{E}}^{src}} [\mathcal{Q}(\mathcal{E}(x))] \\ & - \mathop{\mathbb{E}}_{\mathcal{E}(x') \sim P_{\mathcal{E}}^{tgt}} [\mathcal{Q}(\mathcal{E}(x'))] \end{aligned} \tag{1}$$

Response matching loss is binary cross-entropy, denoted as $L_m(\cdot, \cdot)$, between prediction from $\mathcal{M}$ and the ground truth label:

$$J_m(\theta_e) \equiv \min_{M} \mathop{\mathbb{E}}_{x,y} [L_m(\mathcal{M}(\mathcal{E}(x)), y)]$$

Adversarial loss is therefore given by:

$$J_e \equiv \min_{\theta_e} J_m(\theta_e) + \lambda J_q(\theta_e) \tag{2}$$

Table 1. Results (R@1 / R@2 / R@5). Pretrained models were fine-tuned on Russian in each case.

	Ubuntu RU	Customer support
DE, ru	0.13 / 0.24 / 0.55	0.23 / 0.40 / 0.75
DE, en+ru	**0.30 / 0.48 / 0.80**	0.40 / 0.62 / 0.91
DE, pretr en	0.24 / 0.39 / 0.72	0.37 / 0.60 / 0.90
ADE, pretr en	0.27 / 0.42 / 0.75	**0.40 / 0.65 / 0.92**
ADE, pretr en+ru	0.17 / 0.32 / 0.66	0.27 / 0.49 / 0.82

We hypothesize that such training will allow $\mathcal{E}$ and $\mathcal{M}$ to learn language-independent knowledge to improve response retrieval performance on our target dataset compared to the baseline, the monolingual Dual Encoder.

3 Experiments

3.1 Data and Evaluation Metrics

Our experiments include target data of 2 types. First is Ubuntu Corpus [3], machine-translated to Russian, which is parallel to the original on a sentence level. We have extracted non-overlapping parts of both corpora for pretraining and fine-tuning. This corpus helps to study the problem of transfer knowledge between languages. In experiments using this corpus, we used the first 80 percent of the dialogues from the English corpus and 20 percent of the translated. The second type of data is customer service dialogs of a large Russian mobile network operator. The structure and domain of the data is very different from the Ubuntu corpus and these experiments reflect a more realistic setting.

The Ubuntu Corpus. The English data set is the Ubuntu Corpus which contains multi-turn dialogues collected from chat logs of the Ubuntu Forum. The data set consists of 1 million context-response pairs for training, 20 thousand pairs for validation, and 20 thousand pairs for testing. Positive responses are true responses from humans, and negative ones are randomly chosen sampled from other responses in the training set. The ratio of the positive and the negative is 1: 1 in training, and 1: 9 in validation and testing.

The Ubuntu Corpus, Machine Translated to Russian. Some of our experiments also used the Ubuntu corpus in Russian. The data was obtained using a machine translation of the full Ubuntu Corpus into Russian. Our goal was to keep data intersection to minimum. Therefore, in the Russian version, we transliterated all the English words, thus received zero intersection between dictionaries. In experiments using this corpus, we used the first 80 percent of the dialogues from the English corpus and 20 percent of the translated.

Russian Customer Support Dialogues. The Russian language corpus includes dialogues of customer support services of a large Russian mobile network operator. It consists of 200 thousand context-response pairs in training set and 20 thousand in validation and testing. Similar to Ubuntu corpus, negative examples were randomly selected from other parts of the training set. The ratio of the positive and the negative is 1: 1 in training, and 1: 9 in validation and testing. Dictionaries of current corpus and English Ubuntu intersect on 1 percent.

3.2 Evaluation Metrics.

For evaluation we use Recall@k (R@k) metric. The test dataset was prepared that for each context-response pair another 9 responses were selected from elsewhere in the test data. The 10 options for response were ranked, and the result was flagged as positive if the correct response was included in the top-k of ranked utterances. The percentage of positive results yields Recall@k, a conventional metric for evaluating retrieval-based models.

3.3 Experiments and Results

Our main results are shown in Table 1, which aggregates performance of the retrieval process for different pretraining and training methods. We conduct our experiments with random embeddings initialization (experiments with pretrained bilingual word embeddings (BWEs) didn't show any other correlations). For calculating R@k, $k \in \{1, 2, 5\}$ is being used. Whenever pretraining is conducted in `SOURCE` language, we save the best model with best recalls on `SOURCE` validation set, and the same is for `TARGET`. All results in the Table 1 have been evaluated with test set in `TARGET` language.

We conducted a series of experiments with classical Dual Encoder network to learn about generalization ability of this siamese RNN for the case of cross-lingual learning. First of all, we made straightforward training of DE with `TARGET` dataset. Following the idea of training with only low-resource data available, we consider `TARGET`-only training as baseline. Evaluation (Row 1) gives the lowest recalls (`TARGET` test set).

After that, we conduct two cross-lingual experiments. For that, we utilize `SOURCE` dataset, which volume is 4x larger than `TARGET` dataset. We exploit Ubuntu corpus to minimize domain divergence. The first cross-lingual setup implies `SOURCE` pretraining and subsequent fine-tuning with `TARGET` data. We observe recall increase during evaluation on the test set. The essence of the next experiment is blending languages during training. Without any pretraining , we train dual encoder on mixed bilingual data, so that for every 4 `SOURCE`-language documents there is one `TARGET`-language. Row (3) represents `TARGET`-test and rise of recalls compared with `TARGET`-only and pretraining methods.

To investigate effectiveness of adding language discriminator to DE architecture, we conduct two series of experiments, one with pretraining on `SOURCE`-only data and another one with blending `SOURCE` and `TARGET` languages. Following the design scheme of baselines, we prepare pretrained model and then fine-tune it with `TARGET` language, then evaluate with calculating recalls. Language discriminator is always trained on both `SOURCE` and `TARGET` languages, while encoder part of the system may be trained with `SOURCE`-only and `SOURCE-TARGET` joint mix, which led to two separate results. We compare ADE performance with original DE performance and observe that adversarial learning gives performance increase compared to train on `TARGET` language, thus it gives competitive results comparing to DE trained on the blended data.

3.4 Impact of Bilingual Word Embeddings

For the basic experiments we exploit initialization of random word embeddings by the embedding layer of the encoder. This allows us to track relative change of the evaluation metrics. For the experiment we use pre-trained cross-lingual word embeddings by Conneau et al. [9] with Russian-English align and joint Russian-English dictionary. Results show that exploiting Bilingual Word Embeddings (BWEs) is not always a winning practice, and there is no strict correlation between results with BWEs and without them. This uncertainty could be explained with the domain-specific nature of data we use in all the cases, and the fact pretrained BWEs are created for a general purpose.

3.5 Implementation details

For all our experiments on both languages, the encoder $\mathcal{E}$ is implemented as one-directional LSTM, with 300 hidden units. We set the 300-dimensional embedding layer to be trainable. Language discriminator is implemented with 2 linear layers and ReLU non-linearities. Dual Encoder and Language Discriminator are optimized using separate Adam optimizers [10]. We set learning rate equal for $\mathcal{Q}$ and $\mathcal{E}$ of 0.001. The weights of $\mathcal{Q}$ are clipped to $[-0.05, 0.05]$. We implement pretraining for both DE and ADE for 5 epochs and use early stopping technique exploiting harmonic mean of recalls as stopping criteria. For the further training, we run 13 epochs and also use early stopping. Both DE and ADE are implemented with PyTorch [11].

4 Discussion

According to the results of Table 1, it can be noted that DE model trained on blended data (DE, en+ru) and the model pretrained by the ADE method on English set (ADE, pretr en) outperform the DE model, which simply sequentially learns on English set and then fine-tune on Russian set (DE, pretr en). This effect can be explained by the fact that model (DE, pretr en) can have overfitting on target data. In this case, adding a mixture of data in (DE, en+ru) and discriminator (ADE, pretr en) serves as a kind of regularization.

In Table 2 we present the absolute values of the loss function on training data during the training model English and Russian Ubuntu with Random embeddings in order to study this issue in more detail. The results also confirm that

Table 2. Loss on the training data.

Method	Stage	RU loss	EN loss
DE, pretr en	pretrain	1.38	0.34
	finetune	0.18	0.67
DE, en+ru	finetune	0.27	0.31
ADE, pretr en	pretrain	0.85	0.41
	finetune	0.40	0.56

the absolute values of the loss function in model (DE pr(en)) on the TARGET language are lower than the others, and therefore the model in this case can be overfitted.

The results obtained with adversarial training on machine translation corpus make us think that the discriminator does not sufficiently bridges the gaps between languages. This comes from the fact that ADE performs no better than joint DE trained on the blend of English and Russian data. Therefore, it seems that the generalization ability of the encoder itself is higher when it gradually fed with bilingual data. Our main hypothesis was that adversarial component would make encoder to generate language-independent features. However, it can be seen that adversarial training suffice more when the target data is not directly matching source data. We speculate that the reason of this may lie into the hidden latent factors of the sentence representations. Along with language latent component, a representation contains semantic latent component. When we feed the discriminator with parallel data, it does not know a difference between these latent factors and strives to use all possible dissimilarities to predict whether the sentence is from SOURCE or TARGET. So, when the semantic factor becomes diminished, representations become meaningless — and correct response can hardly be matched to context. However, when datasets are not aligned perfectly, this effect does not occur, maybe due to correct consideration of latent factors by discriminator (or, at least, in a more balanced way). This side effect makes language discriminator, as we constructed it, not very effective in the terms of cross-lingual adaptation, but helps to bridge the gaps between different domains, which is also important.

5 Conclusion and Future Work

In this work we investigated different methods of cross-lingual training of retrieval-based dialogue systems. These methods include pretraining on resource-rich language dataset and further fine-tuning of pretrained models with low-resource language dataset. Also, we experimented with blending data in different languages for improving training. We exploited LSTM-based Dual Encoder network as basic retrieval-based model. We also introduce adversarial cross-lingual architecture, ADE, which is based on Dual Encoder model and exploits language discriminator. We validate effectiveness of these methods conducting experiments on English-language Ubuntu Dialogue Corpus and two Russian-language corpora

— machine translation of the Ubuntu Corpus and Russian Customer Support Corpus. Experiments show effectiveness of mixing low- and rich-resource data and training improvement, compared to regular pretraining. Moreover, we observe competitive performance of ADE on cross-lingual tasks.

In future work we plan on adapting a more advanced retrieval-based model to an adversarial setting. We may also explore the impact of using a transformer model as the encoder of the proposed adversarial model.

References

1. Chen, X., Sun, Y., Athiwaratkun, B., Cardie, C., & Weinberger, K.: Adversarial deep averaging networks for cross-lingual sentiment classification. Transactions of the Association for Computational Linguistics, 6, 557-570 (2018)
2. Chidambaram, M., Yang, Y., Cer, D., Yuan, S., Sung, Y.-H., Strope, B., Kurzweil, R.: Learning cross-lingual sentence representations via a multi-task dual-encoder model. arXiv preprint arXiv:1810.12836 (2018)
3. Lowe, R., Pow, N., Serban, I., Pineau, J.: The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. arXiv preprint arXiv:1506.08909 (2015)
4. Kadlec, R., Martin, S., Kleindienst, J.: Improved Deep Learning Baselines for Ubuntu Corpus Dialogs. arXiv preprint arXiv:1510.03753 (2015)
5. Zhou, X., Dong, D., Wu, H., Zhao, S., Yu, D., Tian, H., Liu, X., Yan R.: Multi-view Response Selection for Human-Computer Conversation. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 372–381 (2016)
6. Erhan, D., Bengio, Y., Courville, A., Manzagol, P.-A., Vincent, P., Bengio, S.: Why Does Unsupervised Pre-training Help Deep Learning? Journal of Machine Learning Research, 11(Feb), 625-660 (2010)
7. Radford A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving Language Understanding by Generative Pre-Training. URL `https://www.semanticscholar.org/paper/Improving-Language-Understanding-by-Generative-Radford/cd18800a0fe0b668a1cc19f2ec95b5003d0a5035`. Last accessed Jan 17 2020 (2018)
8. Ramachandran, P., Liu, P., Le, Q.: Unsupervised Pretraining for Sequence to Sequence Learning. arXiv preprint arXiv:1611.02683 (2016)
9. Conneau, A., Lample, G., Ranzato, M. A., Denoyer, L., Jégou, H.: Word Translation Without Parallel Data. arXiv preprint arXiv:1710.04087 (2017)
10. Kingma, D. P., Ba, J.: Adam: A Method for Stochastic Optimization. arXiv preprint arXiv:1412.6980 (2014)
11. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., Devito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer A.: Automatic Differentiation in PyTorch. URL `https://pdfs.semanticscholar.org/b36a/5bb1707bb9c70025294b3a310138aae8327a.pdf?_ga=2.199060591.154285120.1579522748-1279907765.1576757633`. Last accessed Jan 17 2020 (2017)