Large-scale Translation to Enable Response Selection in Low Resource Languages: A COVID-19 Chatbot Experiment

Lucas Almeida Aguiar[1], **Lívia Almada Cruz** [2], **Ticiana L. Coelho da Silva**[2,3], **Rafael Augusto Ferreira do Carmo**[3], **Matheus Henrique Esteves Paixao**[1]

[1]Centro de Ciências e Tecnologia – Universidade Estudal do Ceará (UECE)
Fortaleza – CE – Brasil

[2]Insight Data Science Lab – Universidade Federal do Ceará (UFC)
Quixadá – CE – Brasil

[3]Instituto Universidade Virtual – Universidade Federal do Ceará (UFC)
Fortaleza – CE – Brasil

lucas.aguiar@aluno.uece.br, livia.almada@ufc.br, ticianalc@insightlab.ufc.br,
carmorafael@virtual.ufc.br, matheus.paixao@uece.br

Abstract. *Natural Language Processing for Low Resource Languages is challenging. The lack of large-scale datasets affects the performance of data-hungry algorithms. To overcome this, we employ data augmentation to enlarge the training data for the task of response selection in multi-turn retrieval-based chatbots. We automatically translated a large-scale English dataset to Brazilian Portuguese (PT_BR) and used it to train a deep neural network. For a COVID-19 chatbot system, our results show that the combination of training with the translated dataset followed by a fine-tuning with the context-specific dataset provides the best results in terms of recall for all studied models. In addition, we make available the translated large-scale PT_BR dataset.*

1. Introdution

The task of response selection for *Low Resource Languages* (LRLS) is challenging due to the absence of a large corpus to guarantee the effectiveness of training neural networks in systems such as chatbots. One alternative for an LRL (for instance, Brazilian Portuguese) would be training neural networks using a corpus of a related language (for example, Spanish). However, this may not be ideal, considering that there may be many words and possibly even syntactic structures that cannot be shared between languages, even if they are highly related [Xia et al. 2019]. One can argue whether or not Brazilian Portuguese is an LRL, but we say that, at least for some tasks, there are fewer resources compared to English or other languages, as discussed by [Costa et al. 2020] and [Fischer et al. 2022].

Another alternative for LRLS is *data augmentation*. One of the strategies in this field is *back translation* [Sennrich et al. 2016], commonly used for training *Neural Machine Translation* (NMT). An NMT system is trained in the reverse translation direction (destination to source), and it is then used to translate monolingual (or resource-limited language - LRLS) data from the destination side back to the source language (in the reverse direction, hence the name *back translation*). The resulting sentence pairs constitute a synthetic parallel corpus that can be added to existing training data to learn a source-to-target model. Depending on the NLP (Natural Language Processing) task, there may be

different strategies for *data augmentation*. For text generation, for example, there can be data disturbance in building augmented samples, including corrupting the input text, the output text, or both [Bi et al. 2021].

We explore strategies to obtain LRLS training data (in our case, in Brazilian Portuguese) for training neural networks for the response selection task without changing the neural network's architecture. One solution to achieve this goal is to find a large public dataset in a resource-rich language, such as English, and automatically translate this dataset into Brazilian Portuguese. Then, train the response selection neural network model for the translated dataset. This alternative has been already investigated in different NLP applications, such as for question answering as [Carrino et al. 2020, Lee et al. 2019, Mozannar et al. 2019, von Essen and Hesslow 2020] and named entity recognition (NER) in natural language text dialogues [Coelho da Silva et al. 2020].

All in all, the contributions of this paper are: (1) the first UbuntuV2/PT-BR dataset, the first large text dialogue in Brazilian Portuguese available to download[1]; (2) performance evaluation of the architecture proposed by [Yoon et al. 2017] for multi-turn response selection in chatbots, comparing the efficiency of the response selection model trained with UbuntuV2/PT-BR dataset and a small text dialogue dataset in Brazilian Portuguese; (3) a response selection model was provided from fine-tuning using other datasets, although small, with improved performance and can be downloaded[2]. We highlight that our approach can use any neural network architecture for multi-turn response selection.

The remainder of this paper is organized as follows. Section 2 presents background concepts related to this work. Section 3 discusses the data used throughout this paper and the methods. Next, Section 4 discusses the results of our experimental evaluation. Section 5 presents our related works. Finally, in Section 6 we conclude the paper and discuss the future work.

2. Background

2.1. Conversational Systems

Conversational systems, commonly known as chatbots, are software systems capable of having a conversation with a person, either textually or vocally [Shum et al. 2018]. For simplicity, we focus on text-based chatbots in this paper. In general, chatbots need to perform two main tasks: i) understand the person's needs from their text messages and ii) display a text message in response to the person. These tasks can be called **input processing** and **response generation**, for brevity. Chatbots are complex systems in which various technologies and combinations can accomplish each of these tasks.

For the task of **response generation**, there are three main models proposed in the literature: **rule-based**, **generative** and **retrieval-based** models [Adamopoulou and Moussiades 2020]. Rule-based chatbots have a pre-defined (usually small) set of possible responses for a particular interaction scenario with a person. A specific rule will link the intent to a particular response based on the person's intent identified in the input processing. These are the most common chatbots found in commercial

[1] `https://shorturl.at/hsuBR`
[2] `https://shorturl.at/firu9`

applications nowadays. Generative chatbots do not have a pre-defined set of rules and responses. As the name suggests, they generate each response on-the-fly based on the inputs given by the person.

Retrieval-based chatbots search an index of possible responses based on the person's inputs. Instead of a small pre-defined set of responses for each particular scenario, retrieval-based models employ an extensive index (commonly composed of thousands of messages) that can serve answers for various inputs. The inputs are processed into a query, then executed into an information retrieval engine to select a suitable response within the index. When only the latest message by the person is used to compose the query, the chatbot is considered a **single-turn retrieval-based** chatbot [Wang et al. 2013]. Differently, a **multi-turn retrieval-based** model considers multiple messages from the person to query the index for the response. This paper focuses on the problem of response selection in multi-turn retrieval-based chatbots, as described next.

2.2. Response Selection in Multi-Turn Retrieval-Based Chatbots

Retrieval-based chatbots aim at leveraging existing dialogue data to build conversational systems. The core assumption is that a bot can use messages exchanged between people in the past to converse with a person in the future. Consider the PlantaoCOVID system, as described in Section 3.1, where nurses provide COVID-related consultations to patients through a website. Thousands of dialogues have been recorded in the system, where nurses have responded to various scenarios presented by the patients. Hence, by leveraging the responses given by the nurses to their patients in the past, a bot can be trained to mimic a nurse and provide consultations to patients in the future.

In a retrieval-based chatbot, the **index** is composed of the dataset of past dialogues, representing a pool of responses that will be queried to respond to a user. At any time in the conversation, the bot will respond to the user by considering the entire dialogue's context, i.e., all the messages exchanged by the user and bot, until the time to respond.

This characterizes a multi-turn retrieval-based chatbot, as previously discussed. Given an index of possible responses and the dialogue's context, the problem of response selection in retrieval-based chatbots automatically selects a response from the most appropriate index for the context at hand.

2.3. Terminology

The chatbot-related literature is new and diverse, resulting in a lack of default terminology within the community. To facilitate the understanding of our paper, Table 1 lists the essential concepts and definitions we employ in our research.

3. Data and Methods

3.1. PlantaoCOVID Dataset

During the COVID-19 pandemic, the Ceara State Government in Brazil launched PlantaoCOVID, a web-based system for online patient consultation. A patient uses the system through a text dialogue interface on the website. A nurse will respond to the patient and answer questions regarding several topics, such as the disease, primary care, testing locations, etc. Throughout its time online, many consultations were performed within

Concept	Definition
Dialogue	A textual conversation between two people. It is composed of multiple turns.
Turn	An interaction in the dialogue. It may be composed of multiple messages.
Message	A text message (including metadata) sent by a person in the dialogue.
Utterance	The textual content of a message.
Context	All the previous messages (from all turns) before a response.
Positive Response	A message was given by a person or bot that is appropriate for the context.
Negative Response	A message given by a person or bot that is not appropriate for the context.

Table 1. Terminology employed in the paper

PlantaoCOVID, resulting in several recorded dialogues between patients and nurses. A subset of the recorded consultations conducted within PlantaoCOVID was used to build a dataset for the problem of response selection in multi-turn retrieval-based chatbots.

The development of the dataset was heavily inspired by the state-of-the-art UbuntuV2 dataset [Lowe et al. 2017]. The PlantaoCOVID dataset follows the same data structure as UbuntuV2, employing the same heuristics for processing the data and selecting the contexts and responses for each dialogue. We used data from two months (April/20 and May/20) to build the dataset since the same COVID protocol was followed during these two months in the state of Ceará.

The first message between patient and nurse is considered the first message in the dialogue. In PlantaoCOVID, the patient or the nurse can send multiple messages in a row before responding. In this case, even though there are multiple messages, the turn of the dialogue does not change. Hence, we concatenate subsequent messages in the same turn using the `__eou__` tag to indicate the end of the utterance. The end of the turn is marked by the `__eot__` tag. Dialogues with less than three turns do not characterize a conversation because there is no exchange of information between the patient and nurse. Thus, we discarded all the dialogues with less than three turns from our dataset.

We split the dataset into three: train, validation, and test. The training dataset accounts for 96% of the dialogues, while both validation and testing account for 2% of the dialogues each. We split the dataset chronologically. The first 96% of the dialogues compose the training dataset; the next 2% compose the validation dataset, and the last 2% compose the test dataset.

For each dialogue, we must select the context, positive response, and negative response(s). As previously explained, the context comprises all the turns before the response. The positive and negative responses are messages that are appropriate and not appropriate given the context, respectively. We cannot simply select the last turn as the positive response to train a model capable of interacting with patients at different moments in the dialogue. It is necessary to have contexts of different sizes and other moments in the dialogue. Hence, we randomly select one of the nurse's turns, following a uniform distribution for each dialogue. The selected turn is considered the positive response, and all the previous turns are considered the context.

The number of negative responses for a dialogue depends on the dataset. For the training dataset, each dialogue has only one negative response. For the validation and test dataset, each dialogue has nine negative responses. A negative response consists of

Dataset	#Dialogues	#Utterances	Turns per Dialogue				Words per Utterance			
			Avg	Median	Max	Min	Avg	Median	Max	Min
PlantaoCOVID	26,899	577,814	21.48	18	190	5	9.9	6	720	1
UbuntuV2	1,038,480	6,453,009	3.95	3	18	2	15.5	11	927	1
UbuntuV2/PT-BR	1,038,480	6,452,110	3.95	3	19	1	15.2	11	885	1

Table 2. Basic statistics of the datasets employed in this study

a randomly (with a uniform distribution) selected message from a nurse extracted from a different dialogue in the same dataset. For a dialogue in the training dataset, for example, one message sent by a nurse in a different dialogue of the training dataset will be selected as the negative response. For a dialogue in the validation dataset, nine messages sent by a nurse in different dialogues of the validation dataset will be selected as the negative responses. We repeat the same procedure for the test dataset.

The first line of Table 2 depicts the basic statistics for the complete PlantaoCOVID dataset (training, validation, and test combined). To the best of our knowledge, there are two alternative public datasets regarding Covid-19 dialogues: the CovidDialog dataset [Yang et al. 2020] contains 603 dialogues in English and 1,088 dialogues in Chinese and the ViraTrustData [Friedman et al. 2022] presents approximately 3,000 annotated dialogues. They account combined for 13% of the number of dialogues collected in the PlantaoCOVID dataset and 0.39% of the number of dialogues present in the UbuntuV2 [Lowe et al. 2017] dataset, which is presented in the following section.

3.2. UbuntuV2/PT-BR Dataset

Compared to other datasets used in the retrieval-based chatbots literature [Lowe et al. 2017, Wu et al. 2017, Zhang et al. 2018], the PlantaoCOVID dataset and the other publicly available datasets are an order of magnitude smaller. This may negatively affect the performance of response selection algorithms. In this scenario, techniques such as data augmentation may assist in boosting the performance of the response selection algorithms. We need a large response selection dataset in Brazilian Portuguese, which is the same language as the PlantaoCOVID dataset. However, such a dataset does not exist in the literature. To overcome this issue, we employed automatic translation.

The UbuntuV2 dataset [Lowe et al. 2017] is the de facto dataset used to evaluate response selection models and retrieval-based chatbots [Liu et al. 2016, Wu et al. 2017, Zhou et al. 2018]. The dataset was built using dialogues from Ubuntu's IRC network, a series of online chat rooms used by Ubuntu practitioners for technical support. The basic statistics of the UbuntuV2 dataset are displayed in Table 2. Due to its size and relevance in the retrieval-based chatbot literature, we chose the UbuntuV2 dataset to be translated to Brazilian Portuguese.

To create UbuntuV2/PT-BR, we used Google Translate, an automatic translation system provided by Google [Wu et al. 2016] that has already been previously used in the literature for similar purposes [von Essen and Hesslow 2020, Mozannar et al. 2019]. The GoogleTrans[3] and Deep Translator [4] libraries offer programmatic interfaces to Google

[3]https://pypi.org/project/googletrans/

[4]https://pypi.org/project/deep-translator/

Translate, which we then incorporated into a Python script. After acquiring the UbuntuV2 dataset from its official repository[5] used by the retrieval-based chatbots community[6], we executed the script, which works as follows.

For each dialogue in the three dataset splits (training, validation, test), we identified the turns and utterances using the `__eou__` and `__eot__` tags, respectively. Each turn was then submitted to the translation API. At the end of this step, there were both an automated and a manual process of fixing a few translation issues, such as known translation errors and broken tags. The entire UbuntuV2 dataset has been translated into the new UbuntuV2/PT-BR dataset at the end of this process. The third line of Table 2 depicts the basic statistics for the translated dataset.

3.3. Multi-turn response selection model

[Yoon et al. 2017] is a state-of-the-art paper that proposes a deep neural architecture in which the neural network is presented with two pieces of data, the dialogue, and a candidate response, and outputs a score value that represents the affinity of the given response for the given context. This process is repeated for each candidate response, then the response with the highest score is selected as the positive response. Also, these scores can be normalized and transformed into the probabilities of choosing each response candidate as the positive response depending on the approach.

This architecture is based on two main elements: Recurrent Dual Encoders (RDE) and Latent Topic Clustering (LTC). RDEs are recurrent neural networks (RNN) whose entries are a sequence of words and their internal hidden state. The base version of the proposed architecture uses one RDE block for encoding the dialogue and another for the candidate response, which is then used in a simple linear projection, which projects these vectors and uses the resulting score to calculate the pair's affinity. Hierarchical Recurrent Dual Encoder (HRDE) are extensions to the base RDE architecture, which tries to address the RNN's forgetting phenomenon [Yoon et al. 2017]. HRDE works by stacking two RNN blocks; the bottom one receives as input chunks of words from a sentence and outputs word-level codes which are in turn used as inputs for the top RNN block, that outputs chuck-level codes.

Complementary, Latent Topic Clustering is proposed by [Yoon et al. 2017] as an internal matrix of the form $\mathbb{R}^{m \times K}$, where K stands for the number of latent topics and m the dimension of topic representation, containing free model parameters that are learned and represent clusters/topics in the data. They act as an extra piece of information concatenated to a data representation vector so that a classifier can perform better by using it. Additional information on the block architecture can be seen in the original paper. Thus, the paper constructs four different approaches for response selection: one using only the RDE block, one using RDE and LTC, here named RDE-LTC, one using only the HRDE architecture, and the HRDE-LTC combining both HRDE and LTC blocks. In all approaches, a fixed embedding layer (in this case, the Glove embedding [Pennington et al. 2014]) transforms the words in the dialog into fixed-size vectors.

Note that we can use any neural network architecture for multi-turn response selection to solve our problem.

[5] `https://github.com/rkadlec/ubuntu-ranking-dataset-creator`

[6] `https://github.com/JasonForJoy/Leaderboards-for-Multi-Turn-Response-Selection`

Setting	1 in 2 (R@1)				1 in 10 (R@1)				1 in 10 (R@2)				1 in 10 (R@5)			
	RDE	RDE LTC	HRDE	HRDE LTC	RDE	RDE LTC	HRDE	HRDE LTC	RDE	RDE LTC	HRDE	HRDE LTC	RDE	RDE LTC	HRDE	HRDE LTC
Tr:Cov Te:Cov	0.716	0.733	0.834	0.822	0.304	0.310	0.453	0.417	0.484	0.453	0.630	0.609	0.781	0.791	**0.890**	0.876
Tr:Ubu Te:Cov	0.521	0.532	0.546	0.542	0.101	0.107	0.128	0.148	0.222	0.234	0.248	0.259	0.533	0.558	0.562	0.562
Tr:Ubu FT:Cov Te:Cov	0.795	0.784	**0.843**	0.830	0.414	0.400	**0.466**	0.494	0.572	0.587	0.658	0.650	0.855	0.857	**0.890**	0.880
Reject H_0?	No	No	No	No	Yes	Yes	No	Yes	Yes	Yes	No	No	Yes	No	No	Yes

Table 3. Experimental results for all models and settings considered in our study. Reject H_0? shows the result of the Wald test, at a 0.05 significance level, for each setting with hypothesis stating that Tr:Cov Te:Cov and Tr:Ubu FT:Cov Te:Cov performances are equal.

4. Results

We evaluate the data augmentation strategy to train a response selection model on LRLS data using the four different architectures proposed by [Yoon et al. 2017]. We performed rigorous experiments using three different learning approaches. Our base approach, dubbed **Tr:Cov;Te:Cov**, trains the four different NN architectures using the training data of the PlantaoCOVID dataset and then tests the adjusted model on the test set of the same dataset. **Tr:Ubu;Te:Cov**, trains the neural networks using the training data of the UbuntuV2/PT-BR dataset and tests the model using the test set of the PlantaoCOVID dataset. **Tr:Ubu;FT:Cov;Te:Cov**, trains the neural networks using the training data of the UbuntuV2/PT-BR dataset, performs a fine-tuning of the adjusted model using the training data of the PlantaoCOVID dataset, and finally tests the resulting model using the test set of the PlantaoCOVID dataset. By doing so, we aim to discover which model fits best for response selection in the context of the PlantaoCOVID task.

Table 3 summarizes the results obtained in our experiments. We used the standard *Recall at top-N*(**R@N**) metric found in the literature. This metric stands for the recall (true positive divided by the true positive plus false negative), i.e., the number of times the correct positive response is selected among the top N-selected responses. For example, R@1 is the fraction of times the correct positive answer is the selected answer, while R@5 stands for a fraction of times it is among the top 5 responses. **1 in 2** in the second column of Table 3 shows the recall metric when the model is presented with one correct positive response and one random negative response. At the same time, **1 in 10** presents the results for a positive response versus nine negative ones. As in the original paper [Yoon et al. 2017], we used a fixed embedding layer (Glove PT_BR [7]) to encode words in the dialogues.

We highlight in bold the best results. It is easy to perceive by inspecting Table 3 that **Tr:Ubu;Te:Cov** is the worst performing training strategy; its recorded recall values are consistently the lowest in every scenario. One can justify this result by the different contexts in which the model is trained and tested. The generalization provided by the massive UbuntuV2/PT-BR is not able, in these settings, to allow a flexible model that can perform well in a very different context. On the other side, **Tr:Cov;Te:Cov** consistently performs better than **Tr:Ubu;Te:Cov** in every setting by a large margin, showing that,

[7] http://nilc.icmc.usp.br/nilc/index.php/repositorio-de-word-embeddings-do-nilc

for this context, the usage of the small domain-specific dataset outperforms more general capabilities brought by training using the large UbuntuV2/PT-BR dataset.

The best approach in this experiment turned out to be **Tr:Ubu;FT:Cov;Te:Cov**, the one that combines the best features of both competing approaches. Using the automatically translated UbuntuV2/PT-BR as the main training dataset, followed by the fine-tuning process using the domain-specific PlantaoCOVID dataset. This approach numerically outperforms **Tr:Cov;Te:Cov** in most scenarios, with the same performance only in the R@5 metric for the HRDE model.

We performed the Wald test [Wasserman 2004, Chapter 10] for each algorithm (RDE, RDE LTC, HRDE, and HRDE LTC) on each of the performance metrics for the **Tr:Cov;Te:Cov** and **Tr:Ubu;FT:Cov;Te:Cov** approaches to assess if the obtained performances (in terms of recall) are statistically different. At a 0.05 significance level, the tests showed that the approaches were not significantly different in the easiest task: finding the correct positive response versus a single random negative response. In 7 out of 12 scenarios for the other tasks (R@1, R@2, and R@5), the tests provided evidence to reject H_0, i.e., one cannot state that the performances are expected to be equal. Additionally, in 4 out of the other 5 scenarios, the recall for **Tr:Ubu;FT:Cov;Te:Cov** was better than **Tr:Cov;Te:Cov**.

5. Related Works

This section provides an overview of two main classes of works related to ours: the first presents models for response selection, including the papers that deal with the same issue of low resource datasets. At the same time, the second class approaches the works that deal with the low resource data problem but for different NLP applications.

Response Selection (with low resource data). The response selection proposals in the state-of-the-art chatbots consider different strategies: single-turn conversation, which only takes into account the last message to select a proper response for the current conversation [Wang et al. 2013], and multi-turn conversation, in which selection takes a message and utterances in its previous turns as input and selects a response that is natural and relevant to the whole context [Wu et al. 2017]. Several studies have been published for response selection. Some concatenate all utterances in context as a lengthy document to calculate the score corresponding to a candidate's response; most of these works use models based on recurrent networks, such as RNN and LSTM. For single-turn models, we have [Lowe et al. 2015, Wan et al. 2016, Wang and Jiang 2016]; others work using the multi-turn strategy [Wu et al. 2017, Zhou et al. 2018].

The paper [Lu et al. 2020] studies the effectiveness of three contextual language models, including BERT and two of its variations ($BERT_{WWM}$ and ROBERTa) as pre-trained models for the problem of selecting answers in chatbots. To increase the quality and quantity of training samples, the authors also proposed a Dialogue Augmentation method to provide enough conversations for training, randomly extracting coherent parts of the dialogues and combining them into positive and negative conversations of different time intervals. [Paul et al. 2019] attempts to provide easy implementation of a multilingual context-aware chatbot for any domain. The dataset was built using two different languages – Bangla and English. Bangla represents an example of a resource-poor language. [Paul et al. 2019] divides the response selection to a given query into two

steps: pre-processing and text classification. Pre-processing is needed to avoid variation of the exact words being considered different; this step recurs to stemming processing. [Paul et al. 2019] uses the Kolkata Bangla Academy Dictionary for Bangla Language to perform N-Gram stemming, which is a statistical approach and is entirely language independent because it only uses a word list to find the stem. In the text classification step, the authors investigated different algorithms to find the best possible relationship between a new query and an already existing dataset. [Li et al. 2022] design a multitasking model for retrieval-based dialogue systems. The multitask model introduces the target domain knowledge into the BERT model pre-trained on another domain. The model includes domain discrimination, a language model, and the prediction of the next sequence on the target domain. An adversarial network trained for a domain selector and the response selection model is used to adjust representations. [Shalyminov et al. 2021] fine-tune the transformer language model GPT-2 [Radford et al. 2019] pre-trained on very large data for dialogue generation task. For adaptation, they augment the input embedding in the dialogue with a speaker tag embedding and a turn identifier embedding. The input token is obtained by summing up these representations. They also add task-specific output layers, a language modeling head, and a next-sentence prediction head.

Other NLP applications with low resource data. Other NLP applications have tackled the same problem of the unavailability of large-scale datasets to train deep neural networks via translation. [Coelho da Silva et al. 2020] proposed Symptomatic – a NER model to identify COVID symptoms in textual dialogues in PT/BR. At the beginning of the pandemic, no model automatically captured the symptoms in a text in Brazilian Portuguese; [Coelho da Silva et al. 2020] used scispaCy – NER model for diseases in English, and through transfer learning, they trained Symptomatic. The training process's first step was translating the texts into Portuguese into English. Then, each input text (in English) passes through the scispaCy model, then the symptoms captured by the scispaCy model are translated from English to Portuguese. The training set for Symptomatic is composed of the original text and the symptoms captured by the scispaCy model in Portuguese. Symptomatic reached an F1 score of 85.66 on the PlantaoCOVID dataset, which is competitive compared to the English model of scispaCy, which has an F1 score of 85.02. The other two works [Lee et al. 2019], and [Carrino et al. 2020] use translated datasets for training neural models for QA (question answering). [Lee et al. 2019] examines the possibility of using translated resources for training QA systems. The work learns how to annotate small resources while taking advantage of the great resources developed for another language (English). [Carrino et al. 2020] proposes a method to automatically translate a QA dataset from English to Spanish, then apply it to the SQuAD dataset to generate a version in Spanish. The authors evaluate the dataset translated by training and fine-tuning two QA systems from a pre-trained and multilanguage BERT model. [Zhang et al. 2019] addresses the problem of the absence of large personal data for training personalized conversational robots. The work pre-trains an RNN-based sequence-to-sequence model on large-scale general data. The classical encoder-decoder model is changed by including a Learning to Start (LTS) component, proposed to predict the first token given the context vector. Then, the general response generation model is fine-tuned using small personal conversation data. The usage of automatic translation to provide datasets for low resource languages is well-known in the literature. For example, automatic translation of the GLUE benchmark, Stanford Natural Language Infer-

ence (SNLI) Corpus, and SciTail Dataset in Portuguese is provided by [GOMES 2020]. [Bonifacio et al. 2021] provides a multilingual translation of the MS MARCO passage ranking dataset.

6. Conclusion

This paper tackles the problem of dealing with the LRLS dataset (in our case, in Brazilian Portuguese) for training a response selection neural network proposed in the literature. To solve this problem, we apply a simple data augmentation strategy that automatically translates a large public dataset from the resource-rich English language called UbuntuV2. We call the large translated Brazilian Portuguese dataset as UbuntuV2/PT-BR. Then, we train the response selector neural network using the translated dataset and fine-tune the model using a domain-specific dataset in Brazilian Portuguese called PlantaoCOVID. From the experimental results, we can see that by augmenting our training set and fine-tuning the model, our approach outperforms the model trained with the small dataset PlantaoCOVID and the one trained using only the UbuntuV2/PT-BR dataset.

For future works, we can explore other datasets and data augmentation techniques. Another possible future direction is investigating other neural networks for multi-turn response selection.

Acknowledgment

The research reported in this work was supported by the Cearence Foundation for Support of Research (FUNCAP) project "Big Data Platform to Accelerate the Digital Transformation of Ceará State" under the number 04772551/2020.

References

[Adamopoulou and Moussiades 2020] Adamopoulou, E. and Moussiades, L. (2020). An Overview of Chatbot Technology. In *IFIP Advances in Information and Communication Technology*, volume 584 IFIP, pages 373–383. Springer International Publishing.

[Bi et al. 2021] Bi, W., Li, H., and Huang, J. (2021). Data augmentation for text generation without any augmented data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2223–2237.

[Bonifacio et al. 2021] Bonifacio, L. H., Campiotti, I., Jeronymo, V., Lotufo, R., and Nogueira, R. (2021). mmarco: A multilingual version of the ms marco passage ranking dataset. *arXiv preprint arXiv:2108.13897*.

[Carrino et al. 2020] Carrino, C. P., Costa-jussà, M. R., and Fonollosa, J. A. (2020). Automatic spanish translation of squad dataset for multi-lingual question answering. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5515–5523.

[Coelho da Silva et al. 2020] Coelho da Silva, T. L., Ferreira, M. G. F., Magalhaes, R. P., de Macêdo, J. A. F., and da Silva Araújo, N. (2020). Rastreador de sintomas da covid19. *Simpósio Brasileiro de Banco de Dados*.

[Costa et al. 2020] Costa, F. A., Ferreira, T. C., Pagano, A., and Meira, W. (2020). Building the first english-brazilian portuguese corpus for automatic post-editing. In *Proceedings of the 28th international conference on computational linguistics*, pages 6063–6069.

[Fischer et al. 2022] Fischer, M., Haque, R., Stynes, P., and Pathak, P. (2022). Identifying fake news in brazilian portuguese. In *International Conference on Applications of Natural Language to Information Systems*, pages 111–118. Springer.

[Friedman et al. 2022] Friedman, R., Sedoc, J., Gretz, S., Toledo, A., Weeks, R., Bar-Zeev, N., Katz, Y., and Slonim, N. (2022). Viratrustdata: A trust-annotated corpus of human-chatbot conversations about covid-19 vaccines. *arXiv preprint arXiv:2205.12240*.

[GOMES 2020] GOMES, J. R. S. (2020). Plue: Portuguese language understanding evaluation. `https://github.com/ju-resplande/PLUE`.

[Lee et al. 2019] Lee, K., Yoon, K., Park, S., and Hwang, S. W. (2019). Semi-supervised training data generation for multilingual question answering. *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, pages 2758–2762.

[Li et al. 2022] Li, J., Tao, C., Hu, H., Xu, C., Chen, Y., and Jiang, D. (2022). Unsupervised cross-domain adaptation for response selection using self-supervised and adversarial training. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 562–570.

[Liu et al. 2016] Liu, C.-W., Lowe, R., Serban, I., Noseworthy, M., Charlin, L., and Pineau, J. (2016). How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *EMNLP*.

[Lowe et al. 2015] Lowe, R., Pow, N., Serban, I., and Pineau, J. (2015). The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909*.

[Lowe et al. 2017] Lowe, R., Pow, N., Serban, I. V., Charlin, L., Liu, C.-W., and Pineau, J. (2017). Training end-to-end dialogue systems with the ubuntu dialogue corpus. *Dialogue & Discourse*, 8(1):31–65.

[Lu et al. 2020] Lu, J., Ren, X., Ren, Y., Liu, A., and Xu, Z. (2020). Improving contextual language models for response retrieval in multi-turn conversation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1805–1808.

[Mozannar et al. 2019] Mozannar, H., Maamary, E., El Hajal, K., and Hajj, H. (2019). Neural Arabic Question Answering. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, number 1, pages 108–118, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Paul et al. 2019] Paul, A., Haque Latif, A., Amin Adnan, F., and Rahman, R. M. (2019). Focused domain contextual ai chatbot framework for resource poor languages. *Journal of Information and Telecommunication*, 3(2):248–269.

[Pennington et al. 2014] Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532–1543.

[Radford et al. 2019] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

[Sennrich et al. 2016] Sennrich, R., Haddow, B., and Birch, A. (2016). Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96.

[Shalyminov et al. 2021] Shalyminov, I., Sordoni, A., Atkinson, A., and Schulz, H. (2021). Grtr: Generative-retrieval transformers for data-efficient dialogue domain adaptation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2484–2492.

[Shum et al. 2018] Shum, H.-y., He, X.-d., and Li, D. (2018). From Eliza to XiaoIce: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, 19(1):10–26.

[von Essen and Hesslow 2020] von Essen, H. and Hesslow, D. (2020). Building a Swedish Question-Answering Model. *Proceedings of the Probability and Meaning Conference (PaM 2020)*, (PaM):117–127.

[Wan et al. 2016] Wan, S., Lan, Y., Xu, J., Guo, J., Pang, L., and Cheng, X. (2016). Matchsrnn: modeling the recursive matching structure with spatial rnn. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 2922–2928.

[Wang et al. 2013] Wang, H., Lu, Z., Li, H., and Chen, E. (2013). A dataset for research on short-text conversations. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 935–945.

[Wang and Jiang 2016] Wang, S. and Jiang, J. (2016). Learning natural language inference with lstm. In *Proceedings of NAACL-HLT*, pages 1442–1451.

[Wasserman 2004] Wasserman, L. (2004). *All of statistics: a concise course in statistical inference*, volume 26. Springer.

[Wu et al. 2016] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation.

[Wu et al. 2017] Wu, Y., Wu, W., Xing, C., Zhou, M., and Li, Z. (2017). Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 496–505.

[Xia et al. 2019] Xia, M., Kong, X., Anastasopoulos, A., and Neubig, G. (2019). Generalized data augmentation for low-resource translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5786–5796.

[Yang et al. 2020] Yang, W., Zeng, G., Tan, B., Ju, Z., Chakravorty, S., He, X., Chen, S., Yang, X., Wu, Q., Yu, Z., et al. (2020). On the generation of medical dialogues for covid-19. *arXiv preprint arXiv:2005.05442*.

[Yoon et al. 2017] Yoon, S., Shin, J., and Jung, K. (2017). Learning to rank question-answer pairs using hierarchical recurrent encoder with latent topic clustering. *arXiv preprint arXiv:1710.03430*.

[Zhang et al. 2019] Zhang, W.-N., Zhu, Q., Wang, Y., Zhao, Y., and Liu, T. (2019). Neural personalized response generation as domain adaptation. *World Wide Web*, 22(4):1427–1446.

[Zhang et al. 2018] Zhang, Z., Li, J., Zhu, P., Zhao, H., and Liu, G. (2018). Modeling multi-turn conversation with deep utterance aggregation. *arXiv preprint arXiv:1806.09102*.

[Zhou et al. 2018] Zhou, X., Li, L., Dong, D., Liu, Y., Chen, Y., Zhao, W. X., Yu, D., and Wu, H. (2018). Multi-turn response selection for chatbots with deep attention matching network. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1118–1127.