Evaluation of Synthetic Datasets Generation for Intent Classification Tasks in Portuguese

Robson T. Paula[1], **Décio G. Aguiar Neto**[2], **Davi Romero**[1], **Paulo T. Guerra**[1]

[1]Federal University of Ceará, Campus Quixadá, Ceará, Brazil

[2]Institute of Computation, University of Campinas, São Paulo, Brazil

robson@alu.ufc.br, aguiar@ic.unicamp.br,
{daviromero,paulodetarso}@ufc.br

Abstract. *A chatbot is an artificial intelligence based system aimed at chatting with users, commonly used as a virtual assistant to help people or answer questions. Intent classification is an essential task for chatbots where it aims to identify what the user wants in a certain dialogue. However, for many domains, little data are available to properly train those systems. In this work, we evaluate the performance of two methods to generate synthetic data for chatbots, one based on template questions and another based on neural text generation. We build four datasets that are used training chatbot components in the intent classification task. We intend to simulate the task of migrating a search-based portal to an interactive dialogue-based information service by using artificial datasets for initial model training. Our results show that template-based datasets are slightly superior to those neural-based generated in our application domain, however, neural-generated present good results and they are a viable option when one has limited access to domain experts to hand-code text templates.*

1. Introduction

A chatbot is an artificial intelligence (AI) based system aimed at chatting with users, commonly used as a virtual assistant to help people or answer questions [Al-Sinani and Al-Saidi 2019]. RASA [Bocklisch et al. 2017] is an open-source tool that allows the development of *chatbots* by people who are not experts in this area [Bocklisch et al. 2017]. The tool works by identifying the intentions contained in the user's messages, classifying them according to the intent defined by the developer, and returning a response based on the intent, which may be an already defined text or a custom action, such as access to database [Bocklisch et al. 2017].

Intent classification is essential for a *chatbot* since it must properly identifies what the user wants and, consequently, what the response to be returned by the system. Thus both a good database and a good model are needed. However, a good database for a given task is not always available, especially when the task involves an unexplored domain, making it difficult to train the model.

In this work, we investigated methods to the development of good synthetic datasets with questions about public services of the government of the state of Ceará. The generated data has labels to classify intentions that indicate what type of information the question requires. We proposed methodologies to generate synthetic datasets based on templates and neural generation (using neural networks). To evaluate this generated

data, we trained deep learning models to classify intents to be used these models for the development of *chatbots*.

1.0.1. Related works.

In [Amin-Nejad et al. 2020] the authors propose a methodology to guide the generation with structured patient information in a sequence-to-sequence manner. They propose an experiment with state-of-the-art Transformer models and demonstrate that their augmented dataset is capable of beating baseline models on the downstream classification task.

In [Ive et al. 2020] the authors present an approach to generate artificial medical documents. They propose an approach to discharge summaries from a large mental healthcare provider and discharge summaries from an intensive care unit. They apply several measures of text preservation and how much the model memorizes training data. They estimate the clinical validity of the generated text based on a human evaluation task. They found that using their artificial data as training data can lead to classification results that are comparable to the original results.

Finally, in [Bird et al. 2020] the authors present an approach to the training of deep learning chatbots for task classification called *Chatbot Interaction with Artificial Intelligence* (CI-AI) framework. CI-AI augments human-sourced data via artificial paraphrasing by the T5 model in order to generate a large set of training data for further language learning approaches. The authors show that seven state-of-the-art transformer models are improved when training data is artificially augmented with an average increase of classification accuracy by 4%.

2. Language Understanding and Dialogue Systems

2.1. Language Models based on Transformers

In [Vaswani et al. 2017], the authors proposed the Transformer architecture. This architecture brought a new and simple network architecture based only on attention mechanisms, thus dispensing recurrence and convolutions. They show that machine translation tasks using these models are superior in quality and require significantly less time to train since they can perform a better parallel computation.

2.1.1. BERT

Bidirectional Encoding Representations of Transformers (BERT) is an architecture proposed by [Devlin et al. 2019] used as a pre-trained base model and which can be used to create different state-of-the-art models in natural language processing tasks. This model, based on the use of the transformer encoder proposed by Vaswani [Vaswani et al. 2017]. The BERT model is pre-trained using two tasks, masked language model (MLM) and next sentence prediction (NSP) on an English Wikipedia text corpus and BookCorpus. The model follows a multi-task learning strategy (MTL), where its parameters are trained in a shared way between the two tasks simultaneously, through a shared loss.

BERT is presented in two versions: a base (BERT-base) and a large (BERT-large). BERT-base has 12 attention layers of 12 heads and a hidden layer with 768 neurons totaling 110M of parameters BERT-large has 24 attention layers with 16 heads each and one hidden state of 1024 totaling 340M of parameters.

Given the new state-of-the-art obtained by BERT, was proposed BERTimbau [Souza et al. 2020], a Portuguese version of BERT. This new model based on BERT was pre-trained on a large Brazilian Portuguese corpus named BrWaC (Brazilian Web as Corpus) using the same pre-training method as BERT.

2.1.2. T5

Text-to-Text Transfer Transformer (T5) is a training framework that aims to unify different natural language processing tasks, making the resulting model robust enough to handle a large set of tasks, defining them in the form of sequence-to-sequence problems. This model was developed using an incremental methodology where each step a NLP characteristic was evaluated, such as variable size architecture with respect to encoder and decoder layers and the attention mask that should be used in the model.

The T5 model were trained with unsupervised objectives providing mechanisms by which the model acquires general-purpose knowledge to apply to further tasks. And, similar to BERT, there is a T5 version pre-trained for Brazilian Portuguese data named PTT5 [Carmo et al. 2020].

2.2. Dialogue Systems and Chatbots

Dialogue Systems is a term used to denote systems that provide a conversational interface that allows users to interact with a computer using natural language. Chatbots describe dialogue systems that use text to interact with a user.

Natural language understanding is a key task for dialogue systems. Its main goal is to detect semantic information expressed in the current user utterance. Natural language understanding can be decomposed into different subtasks [Deriu et al. 2021]: (i) identification of domain (if multiple domains), (ii) identification of intents (that is, the question type, the dialogue act, etc.), and (iii) identification of the slots or concept detection. We focus in this work on identification of intents.

2.2.1. Intent Classification.

The NLU task of classifying an utterance into one of the pre-defined intents is called *intent classification* [Chen et al. 2017]. In an utterance such as, "`I want to change my last reservation.`", the intent classifier should identify the utterance intent as `update_reservation` rather than `cancel_reservation` or `new_reservation` intent. The intent classifier could also associate a confidence value to its output, as 0.9 to `update_reservation` and 0.1 to `cancel_reservation`.

Deep learning techniques have been successively applied in intent classification [Dauphin et al. 2013, Deng et al. 2012, Hashemi et al. 2016, Huang et al. 2013, Shen et al. 2014, Tur et al. 2012]. In particular, the RASA team introduces in

[Bunk et al. 2020] a flexible architecture for intent and entity modeling called Dual Intent and Entity Transformer (DIET). DIET outperforms state-of-the-art models even in a purely supervised setup without any pre-trained embedding.

2.3. Synthetic Data Generation Methods

Existing methods for synthetic text data generation can be summarized into two major categories [Peng et al. 2020]: (i) template-based methods that require domain experts to handcraft templates for each domain and a system fills in slot-values afterward; and (ii) statistical language models that learn to generate fluent responses via training from a labeled corpus.

2.3.1. Template-based Generation

As pointed in [Gatt and Krahmer 2018], when application domains are small and variation is expected to be minimal, sentence generation is a relatively straightforward task, and outputs can be specified using templates.

For example, the template `Turn [direction] onto [road] and continue for [distance] meters` is a meta-sentence with three placeholders, indicated by brackets, that can be filled with a direction, a thoroughfare indication, and the distance one must keep producing for example `Turn right onto 5th Avenue and continue for 200 meters`.

An advantage of templates is that they are easy to implement and they ensure a good quality of the output by avoiding the generation of grammatically incorrect text structures. The disadvantage of templates is that they might not scale well to applications that require considerable linguistic variation [Gatt and Krahmer 2018].

2.3.2. Neural-based Generation

Neural-based generation uses language models to generate natural, meaningful phrases and sentences. These models are used to generate natural sentences based on neural networks.

The use neural-based generation has the advantage of insert variability to the generated dataset and reduce the effort from the template designer. The designer just set a few examples in the template and, this generated example from a template is used to train the neural models.

3. Experimental Setup and Results

3.1. Datasets

We choose as a domain a collection of frequently asked questions (FAQ) about several public services of the government of the state of Ceará[1]. We intend to simulate part of the task of migrating a search-based portal to a dialog-based information service.

[1]Available at `https://cartadeservicos.ce.gov.br/`

We select six public services and nine classes of frequently asked questions related to them. Each class of questions is related to a possible intent that can be raised in a conversation in natural language. Table 1 shows the set of intents.

ID	**Intent**
servico	to know what a service is about
loc_presencial	to know whether a service requires presence *in loco*
loc_presencial_obj	to know whether a service is available in a given city
documentos	to know which documents are required by a service
doc_obj	to know whether a document is required by a service
doc_estado_civil	to know which documents are required by a service for a given marital status
doc_estado_civil_obj	to know whether a document is required by a service for a given marital status
hor_funcionamento	to know the opening hour of a given service provider
hor_funcionamento_obj	to know whether a service is available in a given time

Table 1. Intents of the public services chatbot domain.

First, we build five synthetic datasets to evaluate the effects of template-based and neural-based data for the initial training of chatbot systems (Table 2). Each dataset entry contains a text in natural language corresponding to a dialogue question, an intention descriptor, and a list of entities presented in it.

TBD1 and TBD2 datasets were generated by a set of hand-coded template questions. Each template question is a text associated with a single intention. Let `audiometria`, `mamografia`, and `radiografia` be terms related to an entity class called `servico` (service), the template `O que é [mamografia](servico)?` produces the grounded entries: `O que é audiometria?`, `O que é mamografia?`, and `O que é radiografia?`.

TBD2 dataset is built based on seven template questions for each intent with up to three placeholders for six different entities class: `servico` (service), `horario` (time), `dia` (date), `local` (local), `documento` (document), and `estado civil` (marital status). Each placeholder is replaced by terms related to its respective entity classes, such as `audiometria` and `mamografia` for `servico`, including some synonyms variation such as `radiografia` and `raio-x`.

Note that a template-based dataset will grow exponentially based on the number of placeholders within its template questions. Thus we limit in 600 the number of

Dataset	Description
TBD1	Template-based small dataset with 4844 entries
TBD2	Template-based large dataset with 5324 entries
TBTD	Template-based test dataset with 1254 entries
P5D1	PTT5-generated small dataset with 3391 entries
P5D2	PTT5-generated large dataset with 3727 entries

Table 2. Generated datasets.

Tokenizer	Intent Classifier
BERT	DIET
BERT Multilingual	DIET
BERTimbau	DIET
BERT Multilingual	BERT Multilingual
BERTimbau	BERTimbau

Table 3. NLU models used in the experiments.

grounded entries generated for each intent, randomizing the choice of terms to replace each placeholder.

In order to investigate how dataset variety influences the quality of the final result, we build the TBD1 dataset intentionally less diverse, with fewer synonyms for each entity and with placeholders happening only at the beginning or end of each template question.

TBTD is a test dataset generated using the same methodology of TBD2 but with no common entries between them. TBTD is meant to be used to test the overall performance of NLU models trained with the other synthetic datasets. We also built two datasets *TBD1small* and *TBD2small* to be used on the neural-based generation with 1453 and 1597 entries, respectively, with no intersection with TBD1 and TBD2.

We build P5D2 by training PTT5 to generate sentences in natural languages giving a pair intent-entities as input and comparing its output with the respective sentence associated in the *TBLSsmall* datasets. After training PTT5, we build P5D2 as the collection of sentences generated when we give the intent and entities presented in TBD2 as input to PTT5. The P5D2 dataset has BLEU of 56.6332, F1 of 0.7273, and exact match of 0.3565. The same process is applied to build P5D1 using *TBD1small* and TBD1 datasets. The P5D1 dataset has BLEU of 62.2126, F1 of 0.7670, and exact match of 0.4638.

3.2. Experimental Setup

We evaluate the use of synthetic data by training five different NLU models. Each model is a combination of a model tokenizer and an intent classifier as indicated in Table 3. Our intent is to emulate five different approaches that can be used in the development of a chatbot for Portuguese.

The first combination BERT+DIET[2] is one of the standard options when developing a chatbot using RASA. We use it as a baseline for our experiments. It is expected that this model does not perform well when dealing with Portuguese sentences. Thus we built two other models combining the DIET classifier with BERT Multilingual [Devlin et al. 2019] and BERTimbau [Souza et al. 2020], respectively.

We build two intent classifiers based on BERT multilingual and BERTimbau model architectures, simulating the case where a developer creates his own component. As before, our intention is also to approximate the intent classifier component to language models trained for Portuguese.

Based on these five models we design the following experiments: (a) each model will be fine-tuned with each one of TBD1, TBD2, P5D1, and P5D2 datasets; (b) we

[2]The RASA's BERT default tokenizer is loaded with `rasa/LaBSE` model weights available in https://huggingface.co/rasa/LaBSE.

		F1			Acuracy			Recall		
Dataset	**NLU Model**	0%	70%	90%	0%	70%	90%	0%	70%	90%
TBD1	BERTimbau	0.988	0.965	0.849	0.988	0.977	0.818	0.988	0.943	0.818
	BERTimbau + DIET	0.949	0.903	0.874	0.949	0.901	0.874	0.949	0.901	0.874
	BERT multilingual	0.902	0.794	0.742	0.905	0.826	0.695	0.905	0.763	0.693
	BERT mult. + DIET	0.931	0.892	0.888	0.931	0.891	0.891	0.931	0.891	0.891
	BERT + DIET	0.881	0.876	0.846	0.883	0.872	0.849	0.881	0.872	0.849
TBD2	BERTimbau	0.993	0.993	0.885	0.997	0.997	0.869	0.998	0.998	0.885
	BERTimbau + DIET	0.999	0.997	0.997	0.999	0.998	0.998	0.998	0.998	0.998
	BERT multilingual	0.998	0.998	0.909	0.997	0.997	0.908	0.997	0.997	0.909
	BERT mult. + DIET	0.999	0.997	0.995	0.999	0.997	0.995	0.998	0.997	0.995
	BERT + DIET	0.997	0.997	0.996	0.998	0.997	0.996	0.997	0.997	0.997
P5D1	BERTimbau	0.998	0.897	0.696	0.998	0.946	0.639	0.998	0.864	0.639
	BERTimbau + DIET	0.861	0.839	0.817	0.873	0.850	0.834	0.873	0.850	0.845
	BERT multilingual	0.996	0.839	0.753	0.996	0.857	0.674	0.996	0.787	0.674
	BERT mult. + DIET	0.810	0.808	0.768	0.816	0.814	0.776	0.816	0.814	0.776
	BERT + DIET	0.750	0.729	0.724	0.758	0.736	0.715	0.759	0.736	0.725
P5D2	BERTimbau	0.971	0.954	0.853	0.972	0.966	0.813	0.972	0.938	0.813
	BERTimbau + DIET	0.916	0.910	0.899	0.915	0.912	0.767	0.921	0.912	0.904
	BERT multilingual	0.959	0.945	0.901	0.959	0.939	0.936	0.959	0.918	0.836
	BERT mult. + DIET	0.908	0.900	0.897	0.909	0.902	0.899	0.909	0.902	0.899
	BERT + DIET	0.840	0.793	0.784	0.847	0.803	0.801	0.846	0.803	0.795

Table 4. Result of the intent classification experiments.

measure accuracy, F1, and recall of predicting intents of the TBTD dataset; (c) these metrics will be evaluated with three confidence scenarios: minimum of 90%, minimum of 70% and no minimum required.

The confidence scenarios aim to emulate a common requirement of dialogue systems where a chatbot should proceed in a certain dialogue flow only when the confidence in the predicted intent exceeds a given threshold.

3.3. Intent Classification Results

Table 4 summarizes the results of our experiments. The overall result is that all generated datasets provide good training data. The experiments show that, with some exceptions, most of the trained models have an F1 score over 0.8 with these datasets.

The best result is obtained with TBD2 where all models show an F1 score of 0.99, except by BERTimbau and BERT Multilingual with 90% of confidence (0.885 and 0.909, respectively). This is an expected result since TBD2 is more structurally similar to the test dataset than the others.

The P5D1 dataset shows the lowest results among the training datasets. In the BERT+DIET experiment, for example, we obtained an F1 score of 0.72 with 90% of confidence (27% less than the TBD2 experiment). However, these models show F1 scores up to 0.99 when we relax the confidence restriction.

By comparing the intermediate results of TBD1 and P5D2, P5D2 seems to provide a better result than TBD1. Although for BERTimbau they show a variation of F1 score of 1% to 3%, for BERT multilingual the difference of F1 score is up to 15% in the experiment with 90% of confidence.

When we compare TBD1 to P5D1 and TBD2 to P5D2, it is possible to notice that the results of template-based datasets are slightly superior to those neural-based generated. This means that T5 was able to extract information from the training data and generate databases similar to those template-based generated, but with small errors in some texts, reducing the models' performance. The greater amount of entries in the template-based dataset could also be positively influencing the overall result.

Considering what was observed in these experiments, generating datasets from text templates seems a good choice when a real database is not available and you have an expert available for the task. One main concern is that it can lead to the generation of bases with many similar texts, limiting the model's ability to correctly classify texts that are very different from the training dataset On the other hand, datasets generated with T5 seem to be a good option when there is at least a small initial dataset to seed the neural-based generation, considering that this approach still achieves good general results.

4. Conclusions

In this work, we evaluate the performance of methods to generate synthetic data for chatbots in the domain of public services' FAQ provided by the government of the state of Ceará. We evaluate two methodologies to generate synthetic datasets: one based on template generation and the other based on neural generation using transformer networks. These datasets were used to train chatbot components in the intent classification task.

We generated two datasets, TBD1 and TBD2, from a set of hand-coded templates and two neural-generated datasets, P5D1 and P5D2, by training PTT5 to generate sentences in natural languages. We then train and evaluate five different models in the task of intent classification.

The best result is obtained with TBD2 while the P5D1 has the lowest. This was an expected result since TBD2 is indeed the most structurally similar to the test dataset and P5D1 besides being the smallest among them also presents small grammatical text errors due to neural-based generation. By comparing the intermediate results of TBD1 and P5D2, P5D2 seems to provide a better result than TBD1.

When we compare TBD1 to P5D1 and TBD2 to P5D2, it is possible to notice that the results of template-based datasets are slightly superior to those neural-based generated, however, datasets generated with T5 seem to be a good option when one has limited access to domain experts.

For future work, we intend to refine the quality of the generated datasets and expand the generation methods to include generation based on structured knowledge, such as an ontology-based generation approach.

Acknowledgments. This work is partially supported by the FUNCAP projects 04772314/2020.

References

Al-Sinani, A. H. and Al-Saidi, B. S. (2019). A survey of chatbot creation tools for non-coder. *Journal of Student Research*.

Amin-Nejad, A., Ive, J., and Velupillai, S. (2020). Exploring transformer text generation for medical dataset augmentation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4699–4708.

Bird, J. J., Ekárt, A., and Faria, D. R. (2020). Chatbot interaction with artificial intelligence: Human data augmentation with T5 and language transformer ensemble for text classification. *arXiv preprint arXiv:2010.05990*.

Bocklisch, T., Faulkner, J., Pawlowski, N., and Nichol, A. (2017). Rasa: Open source language understanding and dialogue management. *arXiv preprint arXiv:1712.05181*.

Bunk, T., Varshneya, D., Vlasov, V., and Nichol, A. (2020). Diet: Lightweight language understanding for dialogue systems. *arXiv preprint arXiv:2004.09936*.

Carmo, D., Piau, M., Campiotti, I., Nogueira, R., and Lotufo, R. (2020). PTT5: Pretraining and validating the T5 model on brazilian portuguese data. *arXiv preprint arXiv:2008.09144*.

Chen, H., Liu, X., Yin, D., and Tang, J. (2017). A survey on dialogue systems: Recent advances and new frontiers. *Acm Sigkdd Explorations Newsletter*, 19(2):25–35.

Dauphin, Y. N., Tur, G., Hakkani-Tur, D., and Heck, L. (2013). Zero-shot learning for semantic utterance classification. *arXiv preprint arXiv:1401.0509*.

Deng, L., Tur, G., He, X., and Hakkani-Tur, D. (2012). Use of kernel deep convex networks and end-to-end learning for spoken language understanding. In *2012 IEEE Spoken Language Technology Workshop (SLT)*, pages 210–215. IEEE.

Deriu, J., Rodrigo, A., Otegi, A., Echegoyen, G., Rosset, S., Agirre, E., and Cieliebak, M. (2021). Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54(1):755–810.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Gatt, A. and Krahmer, E. (2018). Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *J. Artif. Int. Res.*, 61(1):65–170.

Hashemi, H. B., Asiaee, A., and Kraft, R. (2016). Query intent detection using convolutional neural networks. In *International Conference on Web Search and Data Mining, Workshop on Query Understanding*.

Huang, P.-S., He, X., Gao, J., Deng, L., Acero, A., and Heck, L. (2013). Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 2333–2338.

Ive, J., Viani, N., Kam, J., Yin, L., Verma, S., Puntis, S., Cardinal, R. N., Roberts, A., Stewart, R., and Velupillai, S. (2020). Generation and evaluation of artificial mental health records for natural language processing. *NPJ Digital Medicine*, 3(1):1–9.

Peng, B., Zhu, C., Li, C., Li, X., Li, J., Zeng, M., and Gao, J. (2020). Few-shot natural language generation for task-oriented dialog. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 172–182, Online. Association for Computational Linguistics.

Shen, Y., He, X., Gao, J., Deng, L., and Mesnil, G. (2014). Learning semantic representations using convolutional neural networks for web search. In *Proceedings of the 23rd international conference on world wide web*, pages 373–374.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Tur, G., Deng, L., Hakkani-Tür, D., and He, X. (2012). Towards deeper understanding: Deep convex networks for semantic utterance classification. In *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5045–5048. IEEE.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, Long Beach, Califórnia.