The 35th Canadian Conference on Artificial Intelligence
DOI: 0

Canadian AI 2022
30 May–3 June 2022, Toronto

Dataset Augmentation Using Back-Translation to Improve Early Stage Dialog Systems

Marc Queudot[†,*], Louis Marceau[†,*], Raouf Belbahar[†], Éric Charton[†], Marie-Jean Meurs[‡,*]
[†] National Bank of Canada
[‡] Université du Québec à Montréal

Abstract

As dialog systems are increasingly used, a major challenge for building new ones is the lack of annotated training data. The necessary data collection and annotation efforts are laborious and time-consuming. A potential solution is to augment initial seed data by automatically paraphrasing existing samples. In this paper, we propose a novel data-efficient approach towards this goal. Our method can kick-start a dialog system with minimum human effort while delivering a performance strong enough to allow real-world usage. We ran experiments using Neural Machine Translation on two open corpora. On both of them, the proposed approach improved the generalization capabilities of the model. Our results suggest that paraphrase generation techniques could be used as-is to provide a boost in performance to dialog systems in an early phase.

Keywords: Dialog Systems, Paraphrase Generation, Chatbot

1. Introduction

Dialog systems have reached a level of maturity that allows widespread application, leading to their development in various domains to meet different needs [1, 2]. These systems communicate with users in natural language through text, speech or both. Some are designed mostly to recreate conversational experiences that mimic humans. Others try to accomplish tasks on behalf of their user, such as booking flights or restaurants. Others still, fill informational needs in a conversational experience. We focus our work on this latter category, composed of *question answering dialog systems*. In this context, the Natural Language Understanding (NLU) module is a core component, which attempts to produce a formal meaning representation of an incoming user message by trying to determine which intent, from a set of known intents, is expressed, and ultimately select a satisfactory response. An important challenge arises from the fact that users can express the same meaning in many different ways, especially when their culture, knowledge and previous experience differ.

In order to reach a satisfactory performance level for most users, a considerable volume of data needs to be gathered to meet the system training needs. The initial annotation effort becomes prohibitively expensive [3] if it is not mitigated. This is especially true for dialog systems that include hundreds of intents (*i.e.*, the categories of information needs or questions users may have). The main challenge when building a new dialog system is that, typically, only very little training data is available. Manually collecting annotated data to reach a sufficient performance level before exposing the system to real users is expensive and time-consuming, slowing down the deployment of the system and the expansion of its capabilities. We therefore aim to reduce the needed time and efforts by automatically augmenting the volume of training samples per intent. The solution we propose relies on generating artificial paraphrases based on existing observations during the design phase of a new dialog system. Paraphrasing can automatically increase the amount of training data to help the intent classification model to learn and generalize even better. However, for the generated data to be useful, we have to ensure that they are different enough from the initial data and remain realistic. In our preliminary work [4], we compared the use of Neural

[*] {marc.queudot, louis.marceau}@bnc.ca, meurs.marie-jean@uqam.ca

Machine Translation (NMT) and Transformers for paraphrase generation. The experiments reported in this paper evaluate different neural approaches in order to augment our training data and overcome some limitations of our preliminary work. The main improvement areas we identified were the exploration of different pivot languages to generate paraphrases and the cleaning of low quality instances.

The paper is organized as follows: Section 2 surveys related previous work in the domain of data augmentation. Section 3 presents in details the datasets used in our experiments. In Section 4, we lay out our methodology and experiments. Section 5 reports and discusses our results obtained on intent classification datasets. Finally, Section 6 concludes this paper, and explores possible future works.

2. Related Work

Data augmentation techniques are used in many domains of Machine Learning. For instance, in computer vision, applying cropping, rotating, or adding Gaussian noise to initial images to create new ones helps systems to generalize better [5–7]. For Natural Language Processing (NLP) tasks, text generation techniques like paraphrasing can be used in order to artificially augment a dataset. Paraphrase generation was first performed using a hand-crafted set of substitution rules to generate alternate sentences [8, 9]. Other approaches to paraphrase generation used bilingual corpora containing pairs of sentences written in two languages. Paraphrases can be obtained by translating from the source to the other language and then back to the original language. That technique, called *pivoting*, allows to generate lexical and grammatical changes in the original sentence, while retaining the same meaning. The same principle applied with neural-based models [10, 11] yielded improved performance. Pre-trained NMT models achieving good performance are made available under the MarianNMT [12] open-source framework. We will present our use of some of those models in Section 4.

With the introduction of Transformers [13], text-generation language models have been fine-tuned in order to output paraphrases in a seq-2-seq fashion [14]. In the context of dialog systems, our previous work [4] compared NMT methods and transformer-based models to generate paraphrases and improve the performance of a new dialog system in order to save cost and resources on corpus building and data acquisition. We expand on this work by adding more back-translation experiments, working exclusively on open corpora, and attempting to filter the paraphrases of lower quality. The quality and diversity of paraphrases are essential to the data augmentation method. A common method used to assess the quality of the generated sentences is human evaluation [15]. Other studies, used alignment-based metrics to examine the semantics preservation and syntactic conformity metrics [16]. In this work, the quality of paraphrases is indirectly evaluated by assessing the performance differential to a baseline trained on a non-expanded dataset.

3. Datasets

SCOPE and BANK, the open corpora used in our experiments, are described hereafter:
SCOPE. The CLINC150 dataset [17], contains requests from various domains annotated with 150 intents, as well as some out-of-scope data. To evaluate our systems on intent classification tasks, we exclude out-of-scope data and refer to the remaining dataset as SCOPE. We then build two smaller subsets of this dataset by randomly sampling either 5 or 10 examples per intent from the original dataset to create SCOPE-train-5 and SCOPE-train-10, respectively. While the original dataset contains many examples per intent (100 each, in the training set), these partitions have more realistic sizes when mimicking a seed corpus in the industry. A subset of the experiments were also run on a dataset including 50 original examples per intent, but none of the paraphrasing techniques we used made significant improvements on that task. While the exact number might differ depending on

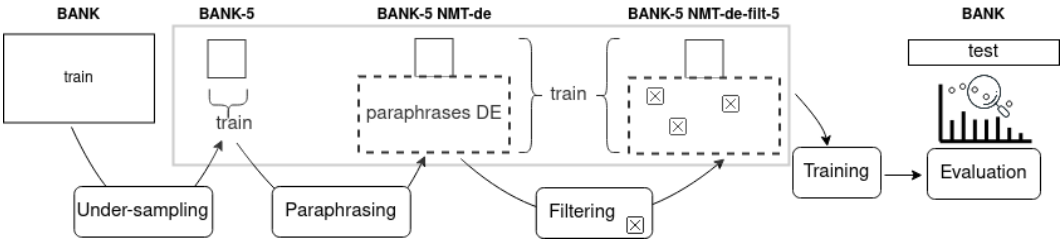
the classification problem, this phenomenon makes sense since artificial data are generally less useful than real data.

BANK. The second dataset is a domain-specific corpus annotated with 77 banking-related intents, and known as Banking77 [18]. It is composed of 13,083 customer service queries labeled with fine-grained intents. The mono-domain aspect is an additional challenge, since it makes inter-intent confusion more likely. We refer to this dataset as BANK. The dataset is split in training and test sets. The training set contains 10,003 examples and the test set is composed of 3,080 examples. We perform experiments on subsets of BANK by under-sampling to get fewer initial training examples per intent, *i.e.*, 5 and 10 each, in the same way as for SCOPE.

4. Methodology and Experiments

The goal of these experiments is to find an efficient way to improve the performance of dialog systems trained with datasets containing only few samples per intent, as it is common when building a new system from scratch. We compare several techniques to generate paraphrases from the artificially reduced datasets described in Section 3. We train one supervised classification model on each generated dataset with the same default pipeline from the Rasa [19] open-source framework. We evaluate those models on test sets that have not been altered, so they contain many more samples per intent than even our largest artificially augmented datasets. The data augmentation techniques are thus evaluated indirectly via their ability to improve the intent classification performance of resulting models. We used macro-averaged Precision, Recall and F1-score, as well as micro-averaged metrics (all equal in the multi-label context). This allows us to consider both per-class performance with no bias towards the majority classes and an instance-focused view that is closer to the performance users would experience in a real world setting. This experiments pipeline is illustrated in Figure 1. BANK-5 is a subset of the original BANK dataset created by under-sampling. Then, paraphrasing with the German pivot is used to augment BANK-5 and create BANK-5-NMT-de. In Section 4.2, we explain how filtering is applied on this dataset to then create BANK-5-NMT-de-filt-5. These different training sets allow to train models that are then compared with each other. The same process is applied to other pivot languages, with various under-sampling parameters and filtering thresholds.

Figure 1. Experiments pipeline (example of German pivot language)



4.1. Exploration of Additional Language Family as Pivot

One of the questions left unanswered in the previous work on NMT augmentation was whether the language family of the pivot has a significant impact on model performance. It is possible that languages that are quite different from English could generate paraphrases with more grammatical and lexical coverage. In order to test more languages than French (fr) and German (de), we added three new pivot languages: Russian (ru), Turkish (tr) and Chinese (zh). These languages were chosen because of their differences with the Romance and Germanic languages previously used, and the availability of pre-trained models with adequate performance on the NMT task itself [17].

4.2. **Post-processing Paraphrases by Model Confidence**

In the previous experiments, the set of paraphrases generated from each sample is kept constant. However, some sentences may be harder to translate in many different forms than others, leading to either poor quality of some translations, or a low diversity in the synthetic examples produced. Here, we designed a post-processing step applied to the generated paraphrases to try keeping only the best ones. In order to do that, we base a filter process on the model confidence score for each prediction (the `F0` field in Marian-NMT). These values are not meant to be interpreted as probabilities but they should be correlated with how relevant the generated samples are. A low confidence for a prediction could mean that other words or tokens could also have been used in the context, but could also be that the meaning of the produced sentence is unclear. Since the values of the prediction confidence can greatly vary, we pick relative values for a threshold of confidence under which the produced paraphrases will not be kept. This is opposed to simply picking absolute values. We first compute the p^{th} percentile (p being the percentile parameter) of model confidence and remove the examples where the confidence is under that amount. This step should ensure that we do not use translations of poor quality which could confuse the intent classification model. We experiment with different values of p, with higher values meaning that more lower quality paraphrases are dropped from the augmented training set (picking $p = 5$ would remove the bottom 5% of paraphrases in terms of model confidence).

5. **Results and Discussion**

Table 1 presents the results of our data augmentation experiments on the SCOPE and BANK datasets. We compare the intent classification performance of models trained on data augmented using different pivot languages. The naming of the systems is constructed in the following way: D-V-S M[-L]-[-`filt`-P] where:
D is the dataset short name (SCOPE or BANK); V is `paraph` if the dataset has been expanded with paraphrases and `train` otherwise; S is the number of samples per intent in the training set; M is the model used to generate the paraphrases (`NMT` for all models but baselines); L (optional) indicates the language of the pivot; and `filt`-P (optional) is the bottom percentile of the paraphrases used as cutoff for model confidence, prefixed with `filt-` (for filtering).

All paraphrase generation was done using a beam size (number of paraphrases per original sample) of $n = 5$ since we found in our prior work that this constituted a good trade-off between quality and quantity. The size of the augmented datasets are hence directly related to the number of original samples per intent. The results reported in Table 1 show a performance gain from the addition of synthetic data generated using NMT pivoting. While there exists some variance in the performance between experiments, we do not notice a trend that would favor the use of particular language as NMT translation pivot against other languages. In Table 1, we also report the results of our post-processing approach. This attempt at cleaning the dataset by removing samples where the model confidence was lower was not successful. It was actually slightly detrimental to the model performance in the case of SCOPE. Keeping examples slightly more varied that they would be after the cleaning step seems to help. As for BANK, the filtering step has more drastic consequences. One reason may be that samples produced with a low confidence end up helping the model to delineate the boundaries for each intent. This does not impact SCOPE systems much, but the intents from BANK all belong to the same semantic domain, so inter-intent confusion is naturally higher. It is likely that the Transformer model trained on both datasets for intent classification [20] is able to generalize well for a varied set of examples, but also to disregard some potentially redundant or low quality training examples, suppressing the need for further post-processing. Systems applying more filtering obtained results similar to `filt-5`.

Systems	Micro Score (%)	Macro F1 (%)	Macro Precision (%)	Macro Recall (%)
SCOPE-train-5 (baseline)	54.7	52.9	54.9	54.7
SCOPE-paraph-5 NMT-de	62.0	59.8	61.3	62.0
SCOPE-paraph-5 NMT-fr	59.6	58.3	61.0	59.6
SCOPE-paraph-5 NMT-ru	60.8	59.2	61.7	60.8
SCOPE-paraph-5 NMT-tr	61.6	60.0	62.5	61.6
SCOPE-paraph-5 NMT-zh	61.2	59.5	62.1	61.2
SCOPE-train-10 (baseline)	69.2	68.1	69.5	69.2
SCOPE-paraph-10 NMT-de	73.2	72.4	73.7	73.2
SCOPE-paraph-10 NMT-fr	73.6	73.0	75.3	73.6
SCOPE-paraph-10 NMT-ru	71.3	70.0	71.5	71.3
SCOPE-paraph-10 NMT-tr	71.9	70.8	72.6	71.9
SCOPE-paraph-10 NMT-zh	72.1	71.0	72.8	72.1
BANK-train-5 (baseline)	45.6	44.8	46.7	45.6
BANK-paraph-5 NMT-de	52.6	51.9	55.0	52.6
BANK-paraph-5 NMT-fr	53.1	51.9	53.1	53.1
BANK-paraph-5 NMT-ru	52.2	51.6	54.3	52.2
BANK-paraph-5 NMT-tr	51.5	50.6	52.4	51.5
BANK-paraph-5 NMT-zh	49.6	48.9	51.2	49.6
BANK-train-10 (baseline)	56.8	56.3	58.7	56.8
BANK-paraph-10 NMT-de	68.5	68.3	70.0	68.5
BANK-paraph-10 NMT-fr	67.3	67.1	68.9	67.3
BANK-paraph-10 NMT-ru	68.8	68.6	70.1	68.8
BANK-paraph-10 NMT-tr	68.3	68.1	70.0	68.3
BANK-paraph-10 NMT-zh	66.3	66.0	68.6	66.3
SCOPE-train-5 (baseline)	54.7	52.9	54.9	54.7
SCOPE-paraph-5 NMT-de	62.0	59.8	61.3	62.0
SCOPE-paraph-5 NMT-de-filt-5	60.5	59.3	62.0	60.5
SCOPE-paraph-5 NMT-de-filt-10	59.8	58.9	61.9	59.8
SCOPE-paraph-5 NMT-de-filt-25	59.0	57.6	60.0	57.6
SCOPE-train-10 (baseline)	69.2	68.1	69.5	69.2
SCOPE-paraph-10 NMT-de	73.2	72.4	73.7	73.2
SCOPE-paraph-10 NMT-de-filt-5	71.0	70.0	71.4	71.0
BANK-train-5 (baseline)	45.6	44.8	46.7	45.6
BANK-paraph-5 NMT-de	52.6	51.9	55.0	52.6
BANK-paraph-5 NMT-de-filt-5	44.3	43.1	45.2	44.3

Table 1. Intent classification results on augmented datasets and after post-processing of NMT-generated paraphrases

6. Conclusion

In this work, we explored the benefit of automatically expanding intent classification datasets of small sizes to boost the performance of models trained on them. This process proved to make significant performance improvement on datasets where the initial number of samples per intent was low, which is usually the case when trying to build a new dialog system. Our findings were reproduced consistently on two open datasets of very different scopes (and thus different likelihoods of intent collisions). Using different pivot languages made no noticeable differences in the models trained on the generated data. The attempt at cleaning lower quality paraphrases was not beneficial either, and even detrimental to the performance when too many samples were removed. Our method can be quickly and efficiently applied to kick-start the development of a new dialog system. New data should still be collected from real interactions as the system is deployed, in the process of continuous improvement. The dataset paraphrase generation techniques offer an improved performance, which would allow the deployment of new dialog systems to occur sooner, which would then unlock the collection of real logs from users, thus compounding their usefulness in making that process more efficient. In the future, we plan on exploring paraphrasing models that are not based on NMT, such as Transformers or other language models fine-tuned

for paraphrasing. Preserving known translations and using entity linking [21] for specific keywords may also be an interesting avenue to avoid introducing noise and to collect relevant paraphrases including important words like product or organization names.

References

[1] M. Queudot, É. Charton, and M.-J. Meurs. "Improving Access to Justice with Legal Chatbots". In: *Stats* 3.3 (2020).

[2] R. Bavaresco, D. Silveira, E. Reis, J. Barbosa, R. Righi, C. Costa, R. Antunes, M. Gomes, C. Gatti, M. Vanzin, et al. "Conversational agents in business: A systematic literature review and future research directions". In: *Computer Science Review* 36 (2020).

[3] T. Falke, M. Boese, D. Sorokin, C. Tirkaz, and P. Lehnen. "Leveraging user paraphrasing behavior in dialog systems to automatically collect annotations for long-tail utterances". In: *International Conference on Computational Linguistics: Industry Track*. 2020.

[4] L. Marceau, R. Belbahar, É. Charton, and M.-J. Meurs. "Quick Starting Dialog Systems with Paraphrase Generation". In: *arXiv preprint arXiv:2204.02546* (2022).

[5] P. Badimala, C. Mishra, R. K. Modam Venkataramana, S. Bukhari, and A. Dengel. "A Study of Various Text Augmentation Techniques for Relation Classification in Free Text". In: *International Conference on Pattern Recognition Applications and Methods*. 2019.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* 25 (2012).

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. "Imagenet: A large-scale hierarchical image database". In: *Computer Vision and Pattern Recognition*. 2009.

[8] K. R. McKeown. "Paraphrasing Using Given and New Information in a Question-Answer System". In: *Association for Computational Linguistics*. 1979.

[9] M. Meteer and V. Shaked. "Strategies for Effective Paraphrasing". In: *Conference on Computational Linguistics*. 1988.

[10] J. Mallinson, R. Sennrich, and M. Lapata. "Paraphrasing Revisited with Neural Machine Translation". In: *European Chapter of the Association for Computational Linguistics*. 2017.

[11] A. Sokolov and D. Filimonov. "Neural machine translation for paraphrase generation". In: *arXiv preprint arXiv:2006.14223* (2020).

[12] M. Junczys-Dowmunt, R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, U. Germann, A. F. Aji, N. Bogoychev, et al. "Marian: Fast neural machine translation in C++". In: *arXiv preprint arXiv:1804.00344* (2018).

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. "Attention is all you need". In: *Advances in Neural Information Processing Systems*. 2017.

[14] S. Witteveen and M. Andrews. "Paraphrasing with Large Language Models". In: *Workshop on Neural Generation and Translation*. 2019.

[15] M. Iyyer, J. Wieting, K. Gimpel, and L. Zettlemoyer. "Adversarial example generation with syntactically controlled paraphrase networks". In: *arXiv preprint arXiv:1804.06059* (2018).

[16] J. Sun, X. Ma, and N. Peng. "Aesop: Paraphrase generation with adaptive syntactic control". In: *Empirical Methods in Natural Language Processing*. 2021.

[17] S. Larson, A. Mahendran, J. J. Peper, C. Clarke, A. Lee, P. Hill, J. K. Kummerfeld, K. Leach, M. A. Laurenzano, L. Tang, et al. "An evaluation dataset for intent classification and out-of-scope prediction". In: *arXiv preprint arXiv:1909.02027* (2019).

[18] I. Casanueva, T. Temčinas, D. Gerz, M. Henderson, and I. Vulić. "Efficient Intent Detection with Dual Sentence Encoders". In: *Workshop on NLP for Conversational AI*. 2020.

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

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

[21] E. Charton, M.-J. Meurs, L. Jean-Louis, and M. Gagnon. "Improving entity linking using surface form refinement". In: *Int. Conf. on Language Resources and Evaluation*. 2014.