Automated Template Paraphrasing for Conversational Assistants

Liane Vogel
Department of Computer Science
Technical University of Darmstadt

Lucie Flek
Conversational AI and
Social Analytics (CAISA) Lab
University of Marburg

Abstract

With synthetic data generation, the required amount of human-generated training data can be reduced significantly. In this work, we explore the usage of automatic paraphrasing models such as GPT-2 and CVAE to augment template phrases for task-oriented dialogue systems while preserving the slots. Additionally, we systematically analyze how far manually annotated training data can be reduced. We extrinsically evaluate the performance of a natural language understanding system on augmented data on various levels of data availability, reducing manually written templates by up to 75 percent while preserving the same level of accuracy. We further point out that the typical NLG quality metrics such as BLEU, utterance similarity, or utterance perplexity, are not suitable to assess the intrinsic quality of NLU paraphrases, and that public task-oriented NLU datasets such as ATIS and SNIPS have severe limitations.

1 Introduction

Task-oriented conversational assistants are designed to perform certain tasks to accomplish a user's goal, such as booking a table at a restaurant, or playing a specific song. Natural Language Understanding (NLU) systems are components of such assistants to perform the tasks of intent classification and slot filling (Tur and De Mori, 2011). Training data for NLU systems consists of user utterances such as "I want to book a table in an Italian restaurant", annotated with the intent (here *book a restaurant*) and slots (here *cuisine* = "Italian"). In order to increase wider availability of such systems, the manual effort required to conduct such annotations for every domain and language must be reduced. For this purpose, we explore automatic template phrase augmentation possibilities using paraphrasing techniques.

Our starting point are template phrases in natural language, such as "Play *[song]* by *[artist]*", containing placeholders for slot types. These template phrases can be specified by developers and are automatically populated from database entries. Our strategy is to directly generate a larger and more varied number of these templates rather than first generating and then paraphrasing example sentences. Comparing two different paraphrasing techniques, namely a Conditional Variational Autoencoder (CVAE) (Sohn et al., 2015) and the language model GPT-2 (Radford et al., 2019), we aim to analyze:

- Under which circumstances does paraphrasing template phrases increase the performance of an NLU system? What are the limitations of this approach?
- How far we can reduce the amount of manually annotated data without a significant loss in NLU performance?
- Which role does the quality of the generated paraphrases in terms of diversity, grammatical correctness and preservation of the intent play in downstream tasks?

Intent	Example utterance	Input	Generated
GetWeather	What's the weather going to be in <city><state>at <timeRange>	x	
GetWeather	Will it be getting <condition_description>on <timeRange>in <country>	x	
GetWeather	Is there any chance to change your forecast for <timeRange>in <country>		x
GetWeather	Will the temperature be going to be <condition_temperature>in <timeRange>		x
GetWeather	You can forecast the weather for <timerange>in <country>		x
PlayMusic	I would like to hear <track>	x	
PlayMusic	Please play a <music_item>off the <artist><music_item><album>	x	
PlayMusic	I would love to hear some <sort><music>		x
PlayMusic	Play some <artist>music that we think people will enjoy		x
PlayMusic	Play <music_item>by <artist>		x

Table 1: Example input and output template phrases finetuning GPT-2 on single template phrases

2 Related Work

Data augmentation for task-oriented dialogue systems has been explored in multiple directions (Yu et al., 2020; Louvan and Magnini, 2020; Kumar et al., 2020).
Witteveen et al. (2019) use a supervised approach to paraphrase sentences by fine-tuning GPT-2 on pairs of phrases that are separated with a delimiter. Similar to our approach, Malandrakis et al. (2019) explore variants of variational autoencoders to generate template phrases. However, they do not report results on publicly available benchmark datasets, and focus only on the task of intent classification, disregarding the often more challenging tasks of slot filling and slot preservation in paraphrasing.

3 Methodology

To simulate the industry development bootstrapping scenario on publicly available datasets, we automatically construct template phrases by replacing slot values in every utterance with generic slot tokens.
We start with a pre-trained GPT-2 model from the Huggingface Python library (Wolf et al., 2020), which we further fine-tune for the task of template phrase generation, treating the slot placeholders as tokens. We then sample from the fine-tuned model to obtain paraphrases. The fine-tuning is performed on unpaired template phrases, and only for a small number of epochs to avoid overfitting on the limited training data. The embedding layer of GPT-2, which is shared between the input and output layer, is kept fixed from the pre-training and is not trained during fine-tuning. We observe that not adapting the previously learned embeddings leads to more diverse paraphrases, as more of the knowledge acquired during pre-training can be incorporated. We train one model for each intent in a dataset separately. The training data is the set of template phrases from the same intent.
At inference time, the model receives only the Beginning-Of-Sentence token *[BOS]* as input. When sampling from the predicted distribution, it generates phrases that are similar to the training data. After generation, we automatically fill the generated templates with specific slot values from a database to obtain user utterances. For each slot type, we randomly pick a slot value to substitute the placeholder in the template phrase (e.g. date = "tomorrow evening"). A selection of input template phrases and generated phrases from GPT-2 are shown in Table 1.

4 Experiments

We conduct our experiments on the frequently used NLU benchmark datasets ATIS (Hemphill et al., 1990) and SNIPS (Coucke et al., 2018). From the datasets we automatically construct template phrases. These template phrases are used as training data for GPT-2 and a CVAE to generate additional template phrases. To compare the results of augmented and non-augmented data we use a bi-directional LSTM (Hochreiter and Schmidhuber, 1997) similar to Hakkani-Tür et al. (2016) that jointly performs slot filling and intent classification. In line with previous work on the used datasets, only slots where all tokens are correctly labeled are considered correct.

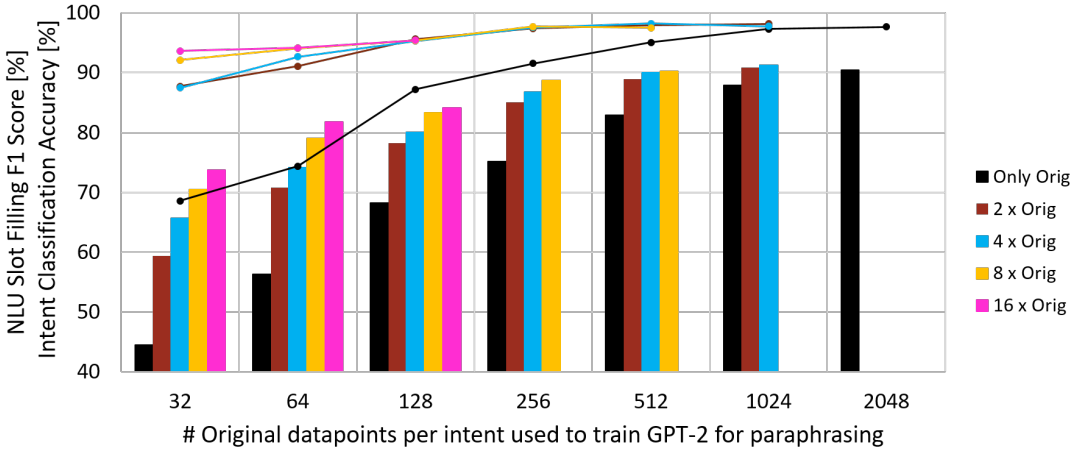

Figure 1: Scores for slot filling (bar chart) and intent classification (line chart) on the SNIPS test data augmented using GPT-2. The x-axis shows the number of original utterances per intent used to create templates to train paraphrasing models. The y-axis shows the performance of NLU models trained on each (augmented) split. Different colors represent different augmentation multipliers for the initial data.

Figure 1 shows that the performance of the NLU system increases for all used splits when adding data generated using GPT-2 to the SNIPS dataset. The results for the CVAE are similar, indicating that both approaches are suitable to generate useful paraphrases for improving the downstream tasks. We compute intrinsic metrics in order to assess diversity and grammatical correctness of the generated phrases. Especially language model perplexity scores (Chen et al., 1998) should in theory be lower for grammatically correct and coherent phrases. However, this is not well applicable on a task-oriented conversational assistant scenario relying on a multitude of slot-values. Table 2 shows several example utterances and their corresponding perplexity computed using a pre-trained GPT-2 model. BLEU scores (Papineni et al., 2002) did not capture enough structure of the phrases to evaluate for coherence.

Utterance	**Perplexity**	**Comment**
What will be the weather tomorrow in New York?	30.92	
What will be the weather on tomorrow in New York?	44.94	Wrong preposition
What will be the weather tomorrow in Cincinnati?	59.97	Uncommon Location
What will be the weather on Sunday in Cincinnati?	44.83	
What will be the best food tomorrow in New York?	41.78	Not GetWeather intent

Table 2: Comparison of perplexity scores for several utterances obtained using a pre-trained GPT-2 model. The selected examples show that perplexity is not always a reliable metric to evaluate grammatical correctness.

5 Conclusions and Future Work

In this paper, we investigate the paraphrasing of template phrases as a data augmentation method for task-oriented conversational assistants. Our results show that both used models, the CVAE and GPT-2, are suitable to generate useful paraphrases, improving the performance on downstream tasks. We further point out that we cannot properly assess the intrinsic quality of NLU paraphrases with traditional NLG quality metrics such as BLEU, utterance embedding similarity, or utterance perplexity, and show that these metrics do not correlate with downstream performance improvements. The main limitation for further improving the proposed approach is a lack of diversity in publicly available task-oriented NLU datasets. Both datasets used in this work, i.e. ATIS and SNIPS, contain only a very small number of intents that are very distinct from each other, compared to real-world applications.

References

Stanley F Chen, Douglas Beeferman, and Roni Rosenfeld. 1998. Evaluation metrics for language models.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*.

Dilek Hakkani-Tür, Gokhan Tur, Asli Celikyilmaz, Yun-Nung Chen, Jianfeng Gao, Li Deng, and Ye-Yi Wang. 2016. Multi-domain joint semantic frame parsing using bi-directional rnn-lstm. *Interspeech 2016*, pages 715–719.

Charles T Hemphill, John J Godfrey, and George R Doddington. 1990. The atis spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. *CoRR*, abs/2003.02245.

Samuel Louvan and Bernardo Magnini. 2020. Simple is better! lightweight data augmentation for low resource slot filling and intent classification. In Minh Le Nguyen, Mai Chi Luong, and Sanghoun Song, editors, *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation, PACLIC 2020, Hanoi, Vietnam, October 24-26, 2020*, pages 167–177. Association for Computational Linguistics.

Nikolaos Malandrakis, Minmin Shen, Anuj Goyal, Shuyang Gao, Abhishek Sethi, Angeliki Metallinou, and Amazon Alexa AI. 2019. Controlled text generation for data augmentation in intelligent artificial agents. *EMNLP-IJCNLP 2019*, page 90.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28:3483–3491.

Gokhan Tur and Renato De Mori. 2011. *Spoken language understanding: Systems for extracting semantic information from speech*. John Wiley & Sons.

Sam Witteveen, Red Dragon AI, and Martin Andrews. 2019. Paraphrasing with large language models. *EMNLP-IJCNLP 2019*, page 215.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Boya Yu, Konstantine Arkoudas, and Wael Hamza. 2020. Delexicalized paraphrase generation. In Ann Clifton and Courtney Napoles, editors, *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020 - Industry Track, Online, December 12, 2020*, pages 102–112. International Committee on Computational Linguistics.