Using LLMs to Generate Training Data for Dialogue System NLUs

Bogdan Laszlo
University of Gothenburg
Master in Language Technology (MLT)
guslasbo@student.gu.se

Staffan Larsson **Asad Sayeed**
Dept. of Philosophy, Linguistics
and Theory of Science
University of Gothenburg
staffan.larsson@ling.gu.se
asad.sayeed@gu.se

Abstract

This paper explores using Large Language Models (LLMs) to generate dialogue datasets for training lightweight Natural Language Understanding (NLU) models for use in modular task-oriented dialogue systems. Employing a schema-guided framework and prompt engineering, we explore how synthetic dialogue data compares to MultiWoZ data on NLU tasks.

1 Introduction

LLMs are impressive in their capability to participate in open-domain dialogue, including understanding user utterances. At the same time there are problems with LLMs, such as producing misleading or false output ("hallucinations"), failure to adhere to instructions, sensitivity to small nuances in prompt design, costs and environmental impact (Rillig et al., 2023), and reliance on constant calls to proprietary LLMs in the cloud.

For many practical, domain-specific applications, a more lightweight controllable modular dialogue system may still be a viable alternative. However, it may often be desirable also in modular systems to make use of the advantages of LLMs. Using LLMs to generate training data for lightweight NLU models is one example of this. NLU models are designed to e.g. determine user intent, identify key entities and/or decipher sentiment.

Collecting datasets of human-human dialogue is labour-intensive, expensive, and may involve privacy concerns. Wizard-of-Oz (WoZ) data collection (Budzianowski et al., 2018) also requires manual effort for data cleaning and annotation. (Budzianowski, 2019).

Synthetic data generation offers a potentially viable and affordable solution for NLU training. However, synthetic datasets in general may exhibit biases in data distribution, may contain incomplete data and inconsistent annotations, and lack may diversity and nuance (Hao et al., 2024; Li et al., 2023).

In this paper, we investigate how NLU models trained on synthetic data compare to models trained on real-world data, when both are tested against real-world data.

2 Method and dataset

To generate synthetic dialogues, we use a schema-guided framework inspired by (Li et al., 2023) combined with strategic prompt engineering (Rastogi et al., 2020). The schema-guided approach involves defining a structured framework that outlines the possible states and transitions in a dialogue, ensuring that the generated dialogues are viable and aligned with specific conversational objectives.

3 Using LLMs to generate dialogues

As explored in Steindl et al. (2023) and Park et al. (2023), LLMs can produce dialogues that closely mimic human conversations. LLM dialogue generation can be fine-tuned for specific applications, such as asking relevant and context-specific questions (Horiuchi and Higashinaka, 2022), replicating complex dialogue patterns across various domains, Liu et al. (2023) and answer retrieval for a retrieval-based conversational character (Chen and Artstein, 2024).

4 Data and Models

Previous approaches to generating synthetic dialogue data have but encountered significant issues. These include models deviating from given templates (Steindl et al., 2023), generating contextually irrelevant responses (Liu et al., 2023), and facing scalability challenges (Rastogi et al., 2020).

The method proposed here tries to address these problems by enforcing strict dialogue schemas through prompt engineering, ensuring models adhere to templates. Additionally, the dialogue-

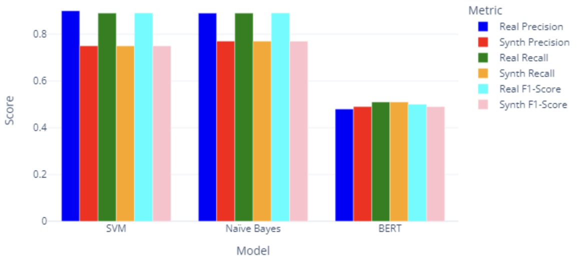

Figure 1: Experiment I — Domain Classification

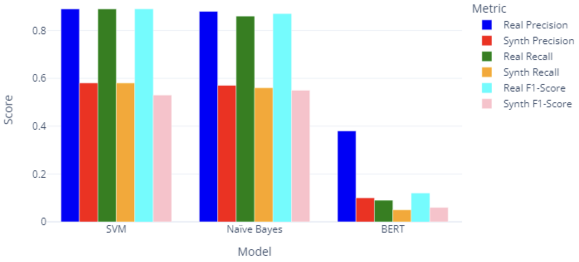

Figure 2: Experiment II — Intent Classification

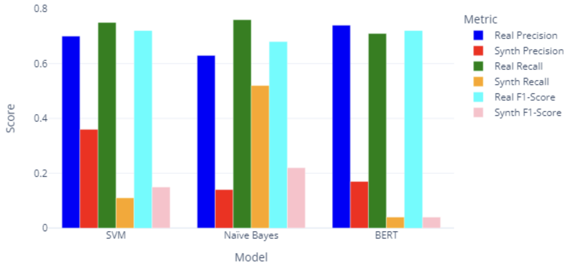

Figure 3: Experiment III — Slot labeling

generating model is exposed to the entire dialogue history in each iteration to prevent out-of-context utterances, in order to improve the coherence and relevance of synthetic dialogue.

Several dialogue datasets have been used for training NLU models. A prominent dataset is MultiWOZ, a multi-domain wizard-of-oz dataset (Budzianowski et al., 2018) that includes several annotations useful for training NLUs. We constructed a synthetic dataset that mirrors the structure and selected domains of MultiWoZ 2.2. We used 458 *train* and 500 *hotel* domain dialogues. The synthetic dataset was similarly constrained to approximately 516 *train* and 500 *hotel* domain dialogues. The creation of the synthetic dataset involved the following steps[1]: schema generation (using GPT-3.5), dialogue generation from schemas (using GPT-4), dialogue clean-up (to remove inconsistencies and errors introduced in generation), alignment of annotation alignment with MultiWoZ structure, and splitting the dataset into training, validation, and testing subsets with proportions of 80%, 5%, and 15%, respectively.

5 Experiment and results

We trained three different NLU models: Support Vector Machine(SVM), Naive Bayes, and BERT on both our synthetic dataset and MultiWoZ. Each model is evaluated on 3 tasks: domain classification, mulitclass intent classification, and slot multi-labelling.

On the domain classification task (Figure 1), the models trained on MultiWoZ perform better than those trained on the synthetic dataset, with the exception of the BERT model which performs poorly overall. However, models trained on the synthetic dataset perform better than a random baseline model. On the intent classification task (Figure 2), the MultiWoZ-trained models in general outperform the synthetic-trained ones.

In slot labeling (Figure 3), models detect the presence of slots without extracting their values. The synthetically-trained models have an unsatisfactory performance. Some slots were not labelled at all. SVM was not always able to beat the baseline model (which assigned the *book_train* category to every utterance). The Naïve Bayes model predicted the same label for every utterance.

6 Error analysis

Error analysis reveals some limitations of the method used, including inconsistent quality, where generated dialogues often lacked the complexity that characterizes natural dialogue. Another limitation is bias, causing repetitiveness with respect to phrasing and chosen topics and converging on a limited number of scenarios focusing primarily on Eurocentric settings (e.g. constantly referencing cities such as London and Cambridge). Furthermore, annotation quality is a concern, and the LLMs introduce hallucinated slots and intents that do not conform to the established MultiWOZ annotation guidelines.

7 Conclusion and future work

We conclude that synthetic data is useful for NLU training, and more so for low-granularity tasks, but not as useful as human data. This is in line with e.g. Chen and Artstein (2024).

[1] The source code for the dialogue generation framework is available at `https://github.com/Devix71/nlu_dialogue_dataset_generator`

Acknowledgments

This work was supported by Swedish Research Council grant 2014-39, Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

References

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*.

Paweł Budzianowski. 2019. The magic triangle of dialogue data collection. PolyAI Blog. [Online; accessed 2024-01-15].

Elizabeth Chen and Ron Artstein. 2024. Augmenting training data for a virtual character using gpt-3.5. In *The International FLAIRS Conference Proceedings*, volume 37.

Shuang Hao, Wenfeng Han, Tao Jiang, Yiping Li, Haonan Wu, Chunlin Zhong, Zhangjun Zhou, and He Tang. 2024. Synthetic data in ai: Challenges, applications, and ethical implications. *arXiv preprint arXiv:2401.01629*.

Sota Horiuchi and Ryuichiro Higashinaka. 2022. Learning to ask specific questions naturally in chat-oriented dialogue systems. In *Conversational AI for Natural Human-Centric Interaction: 12th International Workshop on Spoken Dialogue System Technology, IWSDS 2021, Singapore*, pages 263–276. Springer.

Bogdan Laszlo. 2024. Creating synthetic dialogue datasets for nlu training. an approach using large language models. Master's thesis, Master in Language Technology Programme, University of Gothenburg.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic data generation with large language models for text classification: Potential and limitations. *arXiv preprint arXiv:2310.07849*.

Mengjuan Liu, Chenyang Liu, Yunfan Yang, Jiang Liu, and Mohan Jing. 2023. Promoting open-domain dialogue generation through learning pattern information between contexts and responses. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 351–362. Springer.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8689–8696.

Matthias C Rillig, Marlene Ågerstrand, Mohan Bi, Kenneth A Gould, and Uli Sauerland. 2023. Risks and benefits of large language models for the environment. *Environmental Science & Technology*, 57(9):3464–3466.

Sebastian Steindl, Ulrich Schäfer, and Bernd Ludwig. 2023. Generating synthetic dialogues from prompts to improve task-oriented dialogue systems. In *German Conference on Artificial Intelligence (Künstliche Intelligenz)*, pages 207–214. Springer.