Towards Cost-effective Multi-style Conversations: A Pilot Study in Task-oriented Dialogue Generation

Tiziano Labruna[1,2], Bernardo Magnini[1]
[1]Fondazione Bruno Kessler, Via Sommarive 18, Trento, Italy
[2]Free University of Bozen-Bolzano, Piazza Università 1, Italy
{tlabruna, magnini}@fbk.eu

Abstract

Conversations exhibit significant variation when different styles are employed by participants, often leading to subpar performance when a dialogue model is exclusively trained on single-style datasets. We present a cost-effective methodology for generating datasets featuring multiple conversational styles, which can be used in the development of dialogue systems. The methodology only assumes the availability of a knowledge base for a certain conversational domain, and leverages the generative capabilities of large language models to produce dialogues in a particular style. In a pilot study focused on the generation component of task-oriented dialogues, we extended the well-known MULTIWOZ dataset to encompass multiple style variations, and generated a new multi-style dataset containing diverse styles while retaining core dialogue properties. Our findings highlight two key experimental outcomes: (i) novel, multi-style resources pose challenges for current single-style models, and (ii) multi-style resources enhance the dialogue model's resilience to stylistic variations.

Keywords: dialogue resources, conversational styles, natural language generation

1. Introduction

Task-oriented dialogue systems enable interactions with users to assist them in accomplishing specific tasks, such as booking a restaurant, purchasing a train ticket, or selecting music. Current task-oriented systems (McTear, 2020) are typically trained for a specific domain, whose content is usually stored in a knowledge base, like for instance a restaurant database for a specific area. During the training phase, the dialogue components, including Natural Language Understanding (NLU), Dialogue Manager (DM), and Natural Language Generation (NLG), learn from semantically annotated dialogues to build a model that can be employed during inference for conducting new conversations within the same domain.

However, most current dialogue systems are trained on a single conversational style and do not account for the diversity of styles encountered when interacting with different users. Previous studies have explored the ability of language models to render semantic content in various stylistic variations, assuming a definition of *style* based on psycholinguistic models of personality (Deborah, 1984). For instance, Oraby et al., 2018 created a parallel corpus in the restaurant domain, generating multiple outputs with varying styles using the Personage model, by means of stylistic parameters. These generators employ parameters based on the Big Five personality traits to match the perceived style of the user (Mairesse and Walker, 2010). Similarly, Harrison et al., 2019 present models for stylistic control of NLG, ensuring variation in personality while maintaining semantic accuracy. Recently,

MULTIWOZ: "What is the price range you are looking for?"

NEUTRALGPT: "Sure, we have several options available. Do you have any preferences for price range?"

FRIENDLYGPT: "Of course, man ! What' s your budget like?"

Figure 1: Different dialogue styles for the generation instruction REQUEST(PRICE=?).

Sun et al., 2022 introduce a scheme for generating personalized emotion-aware responses, characterizing the user's linguistic style and emotion polarity. In addition, Thomas et al., 2020 demonstrate that style can be measured in human-to-agent conversations, with people tending to align their style to the style of the agent.

Our work addresses stylistic changes in task-oriented dialogues, where a dialogue model trained for a specific style may struggle when used by users with different conversational styles. Stylistic changes share some features with domain changes in task-oriented dialogues, where the dialogue model needs to cope with frequent updates of domain knowledge (Labruna and Magnini, 2021, 2022). For instance, Labruna and Magnini, 2023 reported a significant decrease in dialogue system performance when new slot-values or instances are introduced to the conversational domain.

We investigate stylistic changes in task-oriented dialogues occurring when a dialogue model trained for a specific conversational style (e.g., formal) is

LREC-COLING 2024, pages 16473–16479
20-25 May, 2024.

employed in the same domain by users with a different conversational style (e.g., informal). Figure 1 provides an example of how the same NLG instruction (i.e., asking the user about the preferred price range for a restaurant) can result in different system responses. Similar to domain changes, addressing stylistic changes needs the ability to create training dialogues for a new dialogue style at a low cost. In this direction, we propose a cost-effective methodology for generating multiple styles conversations for the development of conversational agents. The methodology only assumes the availability of structured information about the domain, such as a knowledge base, and leverages the generative capabilities of large language models (LLMs) (Chen et al., 2019; Raffel et al., 2020) to produce multiple styles dialogues.

The contributions of the paper are as follows: (i) we created multiple style variations of dialogues for the same task-oriented conversational domain; (ii) we demonstrate that the new resources pose challenges for a single-style dialogue generator; (iii) we show that multi-style resources enhance the dialogue model's robustness to stylistic variations. We publicly release all the generated resources[1].

2. Multi-style Dialogue Methodology

In this section we present the methodology employed to generate multiple styles dialogues for a given conversational, task-oriented, domain.

2.1. Domain Knowledge Base

In crafting multi-style dialogues, we begin from a task-oriented context. As outlined in previous literature (Budzianowski et al., 2018; Bordes et al., 2017; Mrkšić et al., 2017), a task-oriented dialogue between a system and a user unfolds as a sequence of turns, denoted as $t_1, t_2, ...t_n$. The primary objective of the dialogue system is to identify a set of relevant entities within a domain knowledge base (KB) that are aligned with the user's communicative goals.

A domain ontology O, serves as a blueprint for the KB, defining the relevant domain entities (e.g., RESTAURANT, MOVIE), each associated with a predefined set of slots (e.g., FOOD, PRICE, for the RESTAURANT entity), and corresponding values (e.g., EXPENSIVE, CHEAP for the PRICE slot). The KB is populated with instances of the domain entities based on the schema provided in the domain ontology through a set of slots-value pairs.

In this context, we distinguish between two types of slots, namely *Informable* slots (e.g., AREA), which the user employ to narrow down search, and *Requestable* slots (e.g., PHONENUMBER), which the user typically inquires about once an instance has been identified in the course of the dialogue. At each turn of the dialogue, both the user and the system may reference information in the KB. The user with the aim of locating entities that match his/her goals, while the system with the aim of suggesting entities that align with the user's objectives.

2.2. Dialogue Generation

Starting from a domain KB, the goal is to produce a corpus of task-oriented dialogues that are both stylistically marked and tightly integrated with the content of the KB. Particularly, system's responses must strictly rely on the KB content.

To create the dialogues, we prompt a LLM to generate a conversation between a user and a system. We structure the prompt as follows:

- Specification of the desired format and general dialogue characteristics (e.g., "a 7-8 turn conversation where the user seeks information and the system help achieving the goal").
- Indication of the conversational style (e.g., "maintain a friendly tone with an informal lexicon, as if they know each other").
- A list of instances from the KB that should dictate the content of the dialogue.

Given the potential size of KB, which may exceed the input capacity of the LLM, we partition the KB instances into K independent clusters, each intended to be used for the generation of a single dialogue. The size of each cluster is constrained by the LLM input capacity. We aim to maximize the similarity between instances within each cluster based on shared slot values (e.g., instances sharing FOOD=ITALIAN). This criterion selects instances likely to be used in dialogues where, for instance, the user is interested in restaurants serving Italian food. The rational is that in most cases the user starts by specifying a value for an informable slot, around which the dialogue revolves. For example, when the user asks for ITALIAN restaurants, the system may propose restaurants with different areas and price ranges, while sharing the same food type.

We use soft K-means clustering, where an instance is allowed to belong to more than one cluster, with the following parameters. K (with $K \geq 1$) denoting the number of clusters, corresponds to the number of dialogues to be generated. The objects O to be clustered are the instances in the initial knowledge base KB. The maximum elements per cluster, I (with $I \leq O$), is set to the capacity limit of the LLM. The similarity function $SIM(i_1, i_2)$ computes the number of shared slot values between two instances s_1 and s_2.

When clusters are formed, each cluster is appended to the generation prompt for a dialogue.

[1] https://github.com/mwozgpt/mwozgpt

The procedure facilitates the creation of coherent dialogue corpora by organizing instances into contextually relevant clusters, ensuring the diversity and richness of dialogues for the given domain.

2.3. Dialogue Annotation

Generated dialogues are then semantically annotated, in order to be used for training and evaluating purposes. Each utterance in the dialogue, both user's and system's turns, has to be annotated with a dialogue-act (Bunt, 2012) representing the communicative goal of the utterance (i.e., intents), and a list of slot-value pairs, which express the content of the dialogue-act. For instance, the user utterance *I'm looking for Italian food*, is annotated as INFORM(FOOD=ITALIAN). The annotation process involves two steps. Initially, we employ an LLM to generate annotations; however, modern LLMs may not consistently produce flawless annotations (Yu et al., 2023; Ashwin et al., 2023). As a result, the second crucial step involves manual correction to rectify inaccuracies or inconsistencies. Human revision is essential to ensure annotation precision and alignment with the dataset's intended purpose.

3. Multi-style Dialogue Resources

This section describes the multiple styles dialogue collections used in our pilot experiments. Each dialogue collection is based on the same KB, i.e., the portion of the MULTIWOZ KB referring to restaurant instances. The KB includes 13 intents, 12 slots (both informable and requestable) and 110 restaurant instances, and it is used for the style-oriented dialogue collections described in the rest of the Section. Examples are reported in Figure 1.

MULTIWOZ 2.4 (Ye et al., 2021) consists of task-oriented dialogues collected manually using the Wizard of Oz technique (Kelley, 1984). For our experiments, we focused on dialogues within the Restaurant domain, resulting in a training-set of 1,180 dialogues and a test-set of 131 dialogues. The style of the MULTIWOZ dialogues is predominantly formal, with the goal of gathering information from the user effectively. The style maintains a professional tone and usually follows a clear structure, with users asking questions and the system providing concise, informative responses.

NEUTRALGPT is designed to produce a neutral conversational style, similar to that found in MULTIWOZ. Dialogues have been automatically generated using as LLM the GPT-3.5-TURBO model, the most capable language model among the GPT-3.5 family[2], following the methodology described in Section 2. The generation was performed using the temperature set to 0.7, the top_p to 1, and the max_tokens to infinite.
We set the number of KB partitions K to 1311, which is the number of restaurant dialogues in MULTIWOZ, in order to create a comparable dataset. Regarding I, i.e., the number of instances for each partition, we decided to keep it to the value of 10: a smaller number would mean that we generate dialogues on a too limited view of the KB; a greater number was not manageable by GPT-3.5, as we observed from empirical tests. The training-set includes 1,180 dialogues that have not been manually corrected, while the test-set comprises 131 dialogues, which have undergone manual revision.

FRIENDLYGPT was also automatically generated using GPT-3.5-TURBO as LLM, with the same hyperparameters used for NEUTRALGPT. FRIENDLYGPT aims at simulating a conversational style between two friends who are excited to see each other, with one of them working for a restaurant reservation service and the other one looking for a restaurant to dine in. It is exclusively used for evaluating model performance and consists of 131 test-set dialogues, which were manually corrected.

MULTISTYLEGPT stands out as a unique case, differing from the aforementioned datasets as it does not adhere to one specific conversational style. Instead, it is a fusion of half of the MULTIWOZ and half of the FRIENDLYGPT datasets. This merging was done to maintain consistency in dataset dimensions for comparison with the others. The training set comprises 1,180 dialogues, evenly split between MULTIWOZ and FRIENDLYGPT, each contributing 590 dialogues. Similarly, the test set consists of 131 dialogues, with 75 randomly selected from MULTIWOZ and 76 from FRIENDLYGPT.

Dataset Characteristics. Table 1 presents statistics of the four datasets. Notably, all the automatically generated datasets show both a higher number of turns and longer turn lengths compared to MULTIWOZ. FRIENDLYGPT emerges with the highest number of turns, while NEUTRALGPT exhibits the longest turn lengths. While the latter might be due to an inherent characteristic of GPT-3.5, which tends to be more verbose on average than a human worker from a Wizard of Oz setting, the former indicates a failure of the model to adhere to the 7-8 turns guideline specified in the prompt.
The average slots per message present a nuanced situation, with FRIENDLYGPT scoring the lowest value, probably implying a tendency towards casual conversation over domain-specific topics. Consequently, the total count of unique slot-values is also lowest for FRIENDLYGPT, while MULTIWOZ and

[2]https://platform.openai.com/docs/models/gpt-3-5

Dataset Characteristic	MULTIWOZ	NEUTRALGPT	FRIENDLYGPT	MULTISTYLEGPT
Avg. System turns per dialogue	4.39	6.05	7.28	5.32
Avg. Turn length	16.27	25.06	19.95	21.09
Avg. Slots per turn	2.56	3.02	1.96	2.73
Tot. Unique slot-values	443	439	400	474
Avg. Intents per turn	1.54	1.30	1.35	1.35
Avg. Utterances per turn	1.8	2.51	2.40	2.20
Type-Token Ratio	0.10	0.05	0.09	0.08

Table 1: Statistics of the dialogue collections used in the experiments.

NEUTRALGPT show comparable values.
The number of instances per turn mirrors the trend observed for slots per turn, while the number of utterances aligns with the length of turns. Lastly, the type-token ratio, indicative of lexical variation, is lowest for NEUTRALGPT, suggesting repetitive word usage, while is higher for MULTIWOZ, which is expected as a large number of diverse annotators contributed to collect the dataset. Interestingly, FRIENDLYGPT also exhibits a high type-token ratio, indicating a notable level of lexical diversity, compared to the other dataset generated by GPT-3.5.

4. Experiments

The goals of our experiments are threefold. Firstly, we aim to provide empirical evidence that a dialogue model trained on a single style struggles when exposed to a different dialogue style. Secondly, we aim to compare the automatically generated training-set, namely NEUTRALGPT, with MULTIWOZ, where dialogues are human-collected through Wizard of Oz. Thirdly, we intend to investigate whether a model trained on a multi-style dataset (combining different styles) is capable of achieving better performance compared to a one-style dataset. In our tests, we assess the ability of an NLG model to produce accurate system responses across various training and testing scenarios. For example, training the NLG model using the MULTIWOZ style and then testing it on the FRIENDLYGPT style allows us to determine the robustness of the model when exposed to style variations.

NLG Model. We employed RNNLG (Wen et al., 2015), a versatile NLG model based on RNNs. RNNLG integrates sentence planning and surface realization in a unified recurrent structure, employing a high-performing SC-LSTM generator with trainable semantic gates for various domains and ensuring competitive performance with limited data through data counterfeiting and discriminative training techniques.
Despite being a relatively old model, not achieving state-of-the-art performance compared to recent systems, RNNLG has proven its reliability and stability over the years. We selected this model because our focus is on comparing the results between datasets in relative terms rather than absolute performance.

Evaluation Metrics. As for the metrics employed to evaluate the generated responses, we focus on BLEU and BARTScore. BLEU (Lin and Och, 2004) is a widely used metric in NLP for assessing language generation tasks, including machine translation and summarization, providing a simple, language-independent measure known to correlate reasonably well with human judgment.
BARTScore (Yuan et al., 2021) is a metric designed for universal NLG evaluation, leveraging BART's (Lewis et al., 2020) generation probabilities to assess sentence quality. This is accomplished by comparing the log probabilities of each token in the generated text to the log probabilities of the corresponding tokens in the reference text.

Experimental Setting. We trained and tested RNNLG on the datasets presented in Section 3 evaluating the performance through the BLEU and BARTScore metrics. We used a learning rate of 0.1, with a learning rate decay of 0.5 and a learning rate divide of 3, ensuring a balance between model training speed and stability.

5. Results and Discussion

The experiments presented in Table 2 aim to evaluate the impact of conversational style on the performance of an NLG model, namely RNNLG. The primary objective of these experiments is to evaluate the capability of RNNLG when trained on dialogues from the same domain but with distinct styles. Higher BLEU and BARTScore values indicate better quality in the output utterances of the dialogue system.
When we train RNNLG on MULTIWOZ and test it on the same style, the model achieves a BLEU score of 0.437 and a BARTScore of -4.388, indicating good performance. However, when we test the same model on NEUTRALGPT, the performance significantly drop, with a BLEU score of 0.184 and a

Training-set	Test-set	BLEU	BARTScore
MULTIWOZ	MULTIWOZ	0.437	-4.388
NEUTRALGPT	MULTIWOZ	0.088	-5.465
MULTISTYLEGPT	MULTIWOZ	0.340	-4.443
MULTIWOZ	NEUTRALGPT	0.184	-4.841
NEUTRALGPT	NEUTRALGPT	0.365	-4.422
MULTISTYLEGPT	NEUTRALGPT	0.343	-4.736
MULTIWOZ	FRIENDLYGPT	0.122	-4.991
NEUTRALGPT	FRIENDLYGPT	0.181	-4.616
MULTISTYLEGPT	FRIENDLYGPT	0.199	-4.665

Table 2: Results of the NLG experiments conducted by training RNNLG on different datasets and testing the models on both the MULTIWOZ and the ChatGPT generated test-sets.

BARTScore of -4.841. The same trend is observed when training on NEUTRALGPT and testing on MULTIWOZ, suggesting that the NLG model struggles with adapting to different dialogue styles.
Additionally, we conducted experiments using the multi-style dataset (MULTISTYLEGPT) and observed improvements in performance compared to the case where models are trained on MULTIWOZ and tested on NEUTRALGPT, and vice versa. This outcome was expected, since half of the MULTISTYLEGPT dataset comprises MULTIWOZ and the other half comprises NEUTRALGPT. Notably, the performance also improved compared to the test with FRIENDLYGPT, which was not seen during the training of any of the three datasets, thus contributing to consolidate our findings. The Pearson correlation coefficient between BLEU and BARTScore is 0.8195, indicating a strong positive correlation between the two metrics, again further consolidating the experimental results.
Overall, the results shed light on the challenges of dialogue model adaptation across various conversational styles. The findings suggest that single-style models may struggle to maintain optimal performance when exposed to diverse dialogue styles, as the performance of such systems is significantly influenced by the style of the training and test data. Conversely, training models on multi-style dialogues leads to improvements, underscoring the importance of incorporating diverse conversational styles in model training data for enhancing model robustness.

6. Conclusion

We have introduced a cost-effective methodology for generating datasets with diverse conversational styles to aid in the development of conversational agents. The methodology only assumes the availability of a conversational domain, such as a knowledge base, and leverages the generative capabilities of large language models. In a pilot study focusing on the generation aspect of task-oriented dialogues, we created variants of the well-known MULTIWOZ dataset featuring multiple conversational styles and demonstrated that models trained on one style struggle when exposed to a different style. Our experiments highlight two key findings: (i) the new resources present challenges for existing single-style models, and (ii) the inclusion of multi-style resources enhances the robustness of dialogue models to stylistic variations.

While GPT-3.5 models are undoubtedly powerful LLMs, they show several limitations when faced with dialogue and annotation generation. Major difficulties are related with ensuring that the generated output complies with the instructions of the desired format, number of turns, admissible intents and slots, and correct detection of slot-values in text. However, the model was used "as is", and a fine-tuning process could certainly help to overcome many of the above mentioned difficulties.
Finally, despite the promising results, there are several limitations that warrant consideration. Firstly, the generated styles, while distinct, may not fully capture the complexity and nuances of human conversational variation. Furthermore, the evaluation metrics employed, such as BLEU and BARTScore, have their own limitations and may not comprehensively capture the quality and appropriateness of the generated dialogues, especially when assessing stylistic variations. Future research should aim to address these limitations for a more comprehensive understanding of stylistic variations in task-oriented dialogues.

7. Acknowledgements

We acknowledge the support of the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU.

8. Bibliographical References

Julian Ashwin, Aditya Chhabra, and Vijayendra Rao. 2023. Using large language models for qualitative analysis can introduce serious bias. *arXiv preprint arXiv:2309.17147*.

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2017. Learning end-to-end goal-oriented dialog. In *ICLR*. OpenReview.net.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

H.C. Bunt. 2012. The semantics of feedback. In *Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogue (SEMDIAL 2012), Paris, France*, pages 118–127. University Paris-Diderot, Paris Sorbonne-Cite.

Qian Chen, Zhu Zhuo, and Wen Wang. 2019. Bert for joint intent classification and slot filling. *arXiv preprint arXiv:1902.10909*.

Tannen Deborah. 1984. Conversational style: Analyzing talk among friends. *Norwood, NJ, Ablex Pub*.

Vrindavan Harrison, Lena Reed, Shereen Oraby, and Marilyn Walker. 2019. Maximizing stylistic control and semantic accuracy in nlg: Personality variation and discourse contrast. *arXiv preprint arXiv:1907.09527*.

John F. Kelley. 1984. An iterative design methodology for user-friendly natural language office information applications. *ACM Trans. Inf. Syst.*, 2(1):26–41.

Tiziano Labruna and Bernardo Magnini. 2021. From cambridge to pisa: A journey into cross-lingual dialogue domain adaptation for conversational agents. *CLiC-it 2021*.

Tiziano Labruna and Bernardo Magnini. 2022. Fine-tuning bert for generative dialogue domain adaptation. In *Text, Speech, and Dialogue*, pages 490–501.

Tiziano Labruna and Bernardo Magnini. 2023. Addressing domain changes in task-oriented conversational agents through dialogue adaptation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 149–158.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin and Franz Josef Och. 2004. ORANGE: a method for evaluating automatic evaluation metrics for machine translation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 501–507, Geneva, Switzerland. COLING.

François Mairesse and Marilyn A Walker. 2010. Towards personality-based user adaptation: psychologically informed stylistic language generation. *User Modeling and User-Adapted Interaction*, 20:227–278.

Michaael McTear. 2020. *Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots*. Morgan and Claypool Publishers.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017. Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1777–1788. Association for Computational Linguistics.

Shereen Oraby, Lena Reed, Shubhangi Tandon, TS Sharath, Stephanie Lukin, and Marilyn Walker. 2018. Controlling personality-based stylistic variation with neural natural language generators. *arXiv preprint arXiv:1805.08352*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Teng Sun, Chun Wang, Xuemeng Song, Fuli Feng, and Liqiang Nie. 2022. Response generation by jointly modeling personalized linguistic styles and emotions. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(2):1–20.

Paul Thomas, Daniel McDuff, Mary Czerwinski, and Nick Craswell. 2020. Expressions of style in

information seeking conversation with an agent. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1171–1180.

Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Lina M. Rojas-Barahona, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Toward multi-domain language generation using recurrent neural networks. *NIPS Workshop on Machine Learning for Spoken Language Understanding and Interaction*.

Fanghua Ye, Jarana Manotumruksa, and Emine Yilmaz. 2021. Multiwoz 2.4: A multi-domain task-oriented dialogue dataset with essential annotation corrections to improve state tracking evaluation. *arXiv preprint arXiv:2104.00773*.

Danni Yu, Luyang Li, Hang Su, and Matteo Fuoli. 2023. Assessing the potential of ai-assisted pragmatic annotation: The case of apologies. arXiv:2305.08339. Version 3.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.