

Knowledge-Grounded Dialogue Act Transfer using Prompt-Based Learning for Controllable Open-Domain NLG

Alain Vazquez Risco and Angela Ramirez and Neha Pullabhotla
and Nan Qiang and Haoran Zhang and Marilyn Walker and M. Inés Torres
University of the Basque Country and University of California Santa Cruz
alain.vazquez@ehu.eus, aramir62@ucsc.edu, npullabh@ucsc.edu,
nqiang@ucsc.edu, hzhan264@ucsc.edu, mawalker@ucsc.edu, manes.torres@ehu.eus

Abstract

Open domain spoken dialogue systems need to controllably generate many different dialogue acts (DAs) to allow Natural Language Generation (NLG) to create interesting and engaging conversational interactions with users. We aim to create an NLG engine that can produce a variety of DAs that make substantive knowledge-grounded contributions to a conversation. Training such an NLG typically requires dialogue corpora that are labelled for DAs, which are expensive to produce and vulnerable to quality issues. Here, we present a prompt-based learning approach to transfer DAs from one domain, video games, to 7 new domains. For each novel domain, we first crawl WikiData to create Meaning Representations that systematically vary both the number of attributes and hops on the WikiData Knowledge Graph. The proposed method involves a self-training step to create prompt examples for each domain followed by an overgeneration and ranking step. The result is a novel, high-quality dataset, Wiki-Dialogue, of 71K knowledge-grounded utterances, covering 9 DAs and the Art, Movies, Music, Sports, TV, Animal, and Boardgames domains, whose combined DA and semantic accuracy is 89%. We assess the corpus quality using both automatic and human evaluations and find it high. The corpus is found to be safe, lexically rich, and large in vocabulary, when compared to similar datasets.

1 Introduction

Open domain spoken dialogue systems need to be able to controllably generate many different dialogue acts (DAs) in order to create interesting and engaging conversational interactions with users. For example, they should be able to ask questions of different types, inform the user of facts and express opinions, make recommendations and suggestions, and confirm what the user said. Moreover, using knowledge to ground DAs supports taking the initiative to drive the conversation forward, and

has been shown to help avoid hallucinations in generated outputs (Dziri et al., 2021; Gopalakrishnan et al., 2019; Chi et al., 2022).

We aim to create a Natural Language Generation (NLG) engine that can produce a variety of DAs and make substantive knowledge-grounded contributions to a conversation. A synthetic example of a conversation about music with a variety of DAs is shown in Table 1. This example is shown because it uses the nine different DAs from the ViGGO video games corpus (Juraska et al., 2019, 2021), with sets of triples from the Wikidata Knowledge Graph (KG) for knowledge grounding.

Dialogues like this require a semantically-controlled NLG that also controls the expression of DAs. In order to create such an NLG, training data consisting of dialogue corpora with utterances labeled with DAs, such as Multi-Woz (Budzianowski and Vulić, 2019), or a parallel corpus of domain-specific DA representations and reference utterances, such as ViGGO or RNNLG (Juraska et al., 2019; Wen et al., 2015), is typically needed. Such training data is typically collected via crowdsourcing, making it expensive to produce, and vulnerable to quality issues (Qian et al., 2021; Dušek et al., 2019). As a result, existing corpora primarily use the INFORM DA (Gardent et al., 2017; Novikova et al., 2017; Lebret et al., 2016; Nan et al., 2021; Parikh et al., 2020) or cover a limited number of topics (Wen et al., 2015; Budzianowski and Vulić, 2019; Juraska et al., 2019; Rastogi et al., 2020).

However, recent work using prompt-based learning (PBL) with LLMs shows that PBL can be used to create high quality dialogic datasets (Kim et al., 2023; Chen et al., 2023; Ramirez et al., 2023). For example, Ramirez et al. used PBL to enlarge the DA labelled ViGGO video games and RNN Laptop and TV corpora (Juraska et al., 2019; Wen et al., 2015). Here we extend their overgenerate-and-rank NLG method to transfer the DAs from the ViGGO corpus onto a set of 6 WikiData domains and a

Dialogue Act Representation	System Output
GREETING	System: Hi, it's so good to talk to you again! How are you doing today? User: I'm doing good, I want to talk about music
REQUEST-ATTRIBUTE (song genre.)	Sys: Sure, What kind of music do you like to listen to? User: I enjoy pop music
REQUEST (songs specifier catchy. songs genres pop.)	Sys: Do you know of any pop songs that are catchy? User: Yes, never gonna give you up
CONFIRM (Never Gonna Give You Up performer Rick Astley. Never Gonna Give You Up domain songs.)	Sys: Is it Rick Astley's song Never Gonna Give You Up that you're referring to? User: Yes
GIVE OPINION (Never Gonna Give You Up rating excellent. Never Gonna Give You Up performer Rick Astley. Never Gonna Give You Up publication year 1987. Never Gonna Give You Up domain songs.)	Sys: Rick Astley did a nice job with the single 'Never Gonna Give You Up' from 1987. It's quite good! User: That sounds great, I totally agree with you.
SUGGEST (Never Gonna Give You Up genre pop music. All I see genre pop music. All I see performer Kylie Minogue.)	Sys: Do you also like pop songs by Kylie Minogue, for example, All I See? User: No, I don't know about that song
INFORM (All I see rating excellent. All I see from album X. All I see performer Kylie Minogue. All I see domain songs.)	Sys: I must say, Kylie Minogue really nailed it with their performance of All I See from the album X. It's a fantastic track that I love listening to. User: I see. Are there any other songs that you can recommend
RECOMMEND (Kylie Minogue genre synth pop. Berlin genre synth pop. Berlin performer of Take My Breath Away.)	Sys: Speaking of synth-pop artists, have you heard the song Take My Breath Away by Berlin? User: Oh I love this artist and this song
REQUEST-EXPLANATION (songs rating excellent. songs performer Berlin.)	Sys: What do you think is it that makes the songs by Berlin so excellent? User: His songs always bring tears to my eyes

Table 1: Music conversation using a dialogue policy based on all the Dialogue Act (DA) types from ViGGO

database about Animals. Since there is no existing training data for these domains, we start with prompt examples from ViGGO, and add a self-training step. We show that self-training improves the results.

We first specify the constraints on different types of ViGGO DAs for each domain, then crawl Wiki-Data to create meaning representations (MRs) consisting of sets of KG triples for each DA type and domain. Example MRs can be seen on the left side of Table 1. We build on WebNLG’s approach for generating multihop KG Triples as the basis for some MRs, which then specify paths through the KG, e.g. talking about the spouse of a person’s sibling involves a path of two hops (Gardent et al., 2017). We systematically vary the complexity of the MRs in terms of number of attributes (up to 8) and number of hops through the KG (up to 3). To test generalizability, we then extend the method to create MRs for an Animals database whose relations and values are very different. To improve the quality of the corpus, we first overgenerate multiple outputs and then rank them, by automatically estimating DA and semantic accuracies and fluency at generation time. We then take the best rated outputs and repeat the procedure with a self-training step. The result is a novel, high-quality, synthetic dataset, Wiki-Dialogue, of knowledge-grounded DAs for the Art, Movies, Music, Sports, TV, Animal, and Boardgames domains, whose combined DA and semantic accuracy (PERFECT) is 89%. Our contributions include:

- Wiki-Dialogue: A new multi-domain dialogue

act and semantically-controlled corpus for the NLG community.

- A novel method that transfers DAs from one domain to another with prompt-based learning.
- A systematic analysis of methods for improving the quality of LLM generated corpora.

2 Related Work

Most knowledge-grounded dialogue corpora are based on crowdsourcing utterances matching different types of MRs. One motivation for generating corpora using LLMs is that crowdsourcing is expensive, and crowdsourced corpora can be noisy, requiring extensive filtering or additional annotation to ensure accuracy. For example, Dušek et al. states that up to 40% of the utterances in the E2E corpus either omitted information that was present in the MR or contained additional information (Dušek et al., 2019), while Qian et al. state that 70% of the dialogues in MultiWOZ contained dialogue state (semantic) annotation errors, which had to be corrected (Ye et al., 2022). Web-NLG (Gardent et al., 2017) also contained mismatches between the MRs and the crowdsourced utterances that have been corrected.¹ We show in Section 4 that there are fewer semantic errors in Wiki-Dialogue than in similar crowdsourced corpora.

The most similar corpus to Wiki-Dialogue is WebNLG, a multi-domain corpus that has been used for NLG challenges (Colin et al., 2016;

¹<https://gitlab.com/shimorina/webnlg-dataset>

Zhou and Lampouras, 2020; Ferreira et al., 2018). WebNLG is based on the DBPedia KG (Lehmann et al., 2015), and covers 19 domains, with utterances realizing the MRs collected via crowdsourcing. The English dataset contains about 17,000 triple sets and 45,000 crowdsourced texts. However, the only DA in WebNLG is INFORM, and dialogues consisting of only INFORM utterances are unnatural (See et al., 2019; Rastogi et al., 2020; Hedayatnia et al., 2020), while Wiki-Dialogue provides DA diversity with nine DAs (Juraska et al., 2019).

Previous work by Moon et al. (2019) created the OpenDialKG dataset by crowdsourcing human-human dialogues consisting of 15K utterances for the Music, Movies, Sports and Books domains. Each utterance corresponds to either a one or two-hop fact in the Freebase KG (Bast et al., 2014) and the KG paths are hand-annotated on each utterance. This corpus contains different types of DAs, but there are no DA annotations.

The Schema-Guided Dialogue (SGD) dataset consists of over 20k annotated multi-domain, task-oriented conversations between a human and a virtual assistant (Rastogi et al., 2020). These conversations target interactions with services, such as travel, spanning 20 domains. The dialogues are generated in two steps: first a simulator automatically generates a dialogue, given a task, in the form of a sequence of DAs whose semantic values are filled by queries to Freebase. The DAs used by the simulator are distinct for the system and the user, with 10 system DAs and 11 user DAs. Then crowdworkers are tasked with paraphrasing each dialogue act/MR combination in a dialogue to ensure naturalness of the utterance realizations.

The Topical Chat corpus consists of 235K utterances from 8 domains, but differs from other knowledge-grounded corpora in that the knowledge is represented by sentences which are automatically aligned with the dialogues. DAs were automatically labelled on this corpus using the 11 DAs from the ISO DA standard with an F1 of 0.54 (Hedayatnia et al., 2020; mez). The DA labels in Wiki-Dialogue are much less noisy: See Table 10.

Other knowledge-grounded NLG corpora with rich sets of DAs have focused on specific domains. The ViGGO Video Games corpus contains 9 DAs (Juraska et al., 2019), the RNNLG corpus encompasses 13 DAs for domains such as laptops, TVs, hotels, and restaurants (Wen et al., 2015), and MultiWOZ offers 34 task-oriented, domain-specific

DAs, for the restaurant, hotel, attraction, taxi, train, hospital and police domains (Eric et al., 2021).

Recent work by Wu et al. (2023) on controlling DAs in NLG for task-oriented dialogue introduces DiactTOD, a model based on learning latent DAs from pre-existing datasets, achieving state-of-the-art performance on MultiWOZ (Ye et al., 2022). However, the evaluation DiactTOD is based on a benchmark set and automatic evaluation metrics, which do not evaluate DA accuracy. We use a fine-tuned classifier that filters for DA accuracy as one step in the automatic ranking of possible responses. We also apply both automatic and human evaluation, showing that we achieve average DA accuracies of .98 for one-hop and .89 for multi-hop.

Other work on creating synthetic data for NLG has focused on creating whole dialogues or augmenting existing corpora (Kim et al., 2023; Xu et al., 2021; Chen et al., 2023). One of the main challenges with synthetic dialogue generation is producing high quality outputs without human evaluation. Here we show that the quality of synthetic dialogue data can be improved using a cycle of self-training, along with an overgenerate-and-rank step that uses a DA classifier and semantic accuracy estimates. We build on previous work by Ramirez et al. by using their definitional prompt style, ranking function, and DA tagger. We extend their approach by transferring the DAs used for controlled generation to 7 new domains and incorporating a self-training step required to bootstrap high quality generation outputs for completely novel domains.

3 Experimental Method

Figure 1 provides an overview of the experimental architecture. Our method consists of five steps:

- Specifying DA constraints;
- Creating DA specific MRs;
- Prompt Creation and LLM selection;
- Overgenerate and Rank from the MRs for one round using ViGGO examples in the prompts;
- Self-Training: Select new in-domain prompt examples and conduct a second round of overgenerate-and-rank.

3.1 Specifying Dialogue Act Constraints

We use the DAs from the ViGGO corpus to enable more highly varied dialogue policies. One possible policy is illustrated by the conversation shown in Table 1. The utility of controlling DAs and being

DA	Num Rels	ViGGO		Athletes		Wild Animals	
		Mandatory Rels	Hops	Mandatory Rels	Hops	Mandatory Rels	Hops
INFORM	3-8	NAME, GENRES	1	NAME, SPORT	1-3	NAME, COMMON_CLASS	1
CONFIRM	2-3	NAME	1	NAME, SPORT	1	NAME, COMMON_CLASS	1
GIVE_OPINION	3-5	NAME, RATING	1	NAME, RATING, SPORT	1-3	NAME, RATING, COMMON_CLASS	1
RECOMMEND	2-3	NAME	1	NAME, SPORT	1-2	NAME, COMMON_CLASS	1
REQUEST	1-2	SPECIFIER	1	SPECIFIER	1	SPECIFIER	1
REQUEST_ATTRIBUTE	1	—	1	—	1	—	1
REQUEST_EXPLANATION	2-3	RATING	1	NAME, POPULARITY	1-2	POPULARITY	1
SUGGEST	2-3	NAME	1	NAME	1-2	NAME	1
VERIFY_ATTRIBUTE	3-4	NAME, RATING	1	NAME, RATING	1-2	NAME, POPULARITY	1

Table 2: Semantic Constraints on Dialogue Acts for VideoGames (ViGGO), Athletes and Wild Animals.

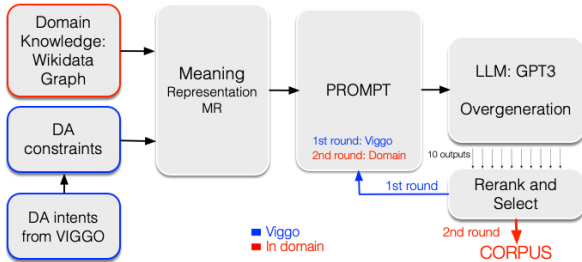


Figure 1: Experimental Architecture

able to produce different DAs is also illustrated by the utterances for the Art domain in Table 3, which demonstrate how the same MR can be realized very differently depending on the DA. A range of DAs allows a dialogue system to vary the dialogue policy in interesting ways (Juraska et al., 2021).

Attributes and Values

(NAME [Andromeda Chained to the Rocks], CREATOR [Rembrandt], GENRES [mythological painting, nude], INCEPTION [1630])

confirm

Is the painting you’re talking about **Andromeda Chained to the Rocks** by **Rembrandt**, a **mythological nude** piece from **1630**?

recommend

I recall you enjoy **mythological** paintings and **nudes**, so have you heard of **Rembrandt’s Andromeda Chained to the Rocks** from **1630**?

inform

Andromeda Chained to the Rocks, created by **Rembrandt** in **1630**, is a **mythological** painting featuring a **nude** figure.

Table 3: Sample dialogue acts (DAs) in the Art domain showing how the same attributes and values can be realized as different DAs.

The initial step of our method involves specifying semantic attribute constraints for each domain and entity and DA types. NUM RELS in Table 2 is the number of relations that can be included in a particular DA. As seen in Table 2, the CONFIRM DA (Row 2) should only include 2–3 relations, or it is difficult to understand, while the INFORM DA typically has more slots than other DAs.

We also specify mandatory and optional rela-

tions for each DA and entity type. MANDATORY RELS in Table 2 specifies relations that must be included for each DA type. The GIVE_OPINION DA (Row 3) requires an ordinal attribute similar to RATING in order to orient the polarity of the opinion. For domains such as Athletes and Animals, that do not have directly such an attribute, so the number of Wikipedia page views is used to create an ordinal and equivalent POPULARITY attribute, with values ranging from LOW to HIGH. Potential values for the SPECIFIER attribute needed by the REQUEST DA must be provided for each entity type, e.g. the specifier CATCHY used in the REQUEST DA in the conversation in Table 1. The ViGGO columns in Table 2 show the mandatory relations that were based on ViGGO’s 14 video game attributes. Each DA also has optional relations that define the attributes that can be added to the mandatory ones when creating the MRs. They are shown for each entity type and domain in the corpus GitHub.²

For each domain, there are two or more entity types, e.g. the entity types for Movies are Actor and Movie; for Music they are Song, Album and Musician; for Sports they are Athlete and Team; and for Animals they are Wild Animals, Cats and Dogs. Each domain’s specific slots are provided on GitHub with the corpus description. For the entity types that are media, like Movies, Songs or TV Shows, the constraints are identical to those shown for ViGGO in Table 2.

However, other domains require different semantic constraints as illustrated in the Athletes and Wild Animals columns of Table 2, e.g. for Athletes, the attribute of genre doesn’t apply, but the sport that the Athlete plays serves a similar function. Similarly, Animals doesn’t have a genre, but common_classes of animal, such as reptile, fish or mammal, are needed to specify the general type of the animal. For human entity types such as Actors, Musicians and Athletes the REQUEST_EXPLANATION DA (Row 7 of Table 2) must include the name

²<https://github.com/aramir62/Wiki-Dialogue>

slot: see the example conversation in Table 1. In addition, for both Athletes and Animals, the CONFIRM (Row 2), GIVE OPINION (Row 3), and RECOMMEND (Row 4) DAs require that the sport be mentioned for the athlete, and the common_class mentioned for the animal.

3.2 Generating KG Triples from WikiData

After specifying the DA attributes, we generate KG triples adhering to DA constraints using WikiData queries for Art, Movies, Music, TV Shows, Boardgames, and Sports. WikiData offers detailed knowledge across many domains and shares canonical IDs with Wikipedia. To test generalization, we used API Ninjas³ to create MRs for the Animal domain. For all entity types for each domain, we selected 60 entities: 30 popular and 30 lesser-known, based on Wikipedia page visits in the past 6 months. These entities become the starting nodes for all paths, and for multi-hop data, both incoming and outgoing relations are included.

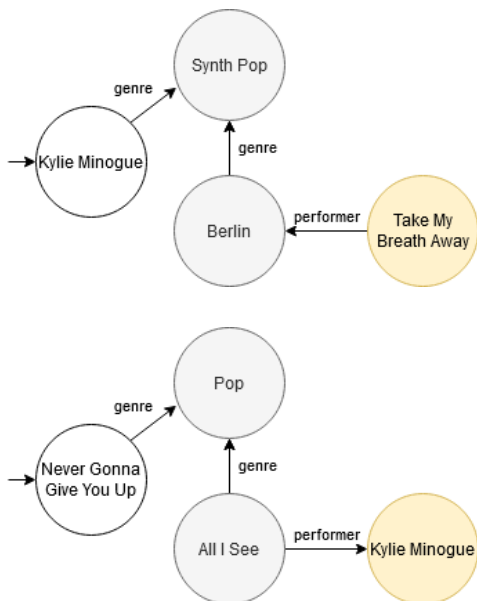


Figure 2: Subgraphs of Wikidata for the Recommend Dialogue Act and Suggest Dialogue Act in the music conversation in Table 1

For multi-hop paths, there are constraints on the maximum number of hops for each DA. This is necessary since some DAs such as CONFIRM should not realize multiple relations between entities, so it is constrained to being 1-hop. For other DAs, we performed a qualitative analysis that showed that hops larger than 3 often cause the model to fail to

³<https://api-ninjas.com/api/animals>

properly realize either the values or the relations. We thus restrict the number of hops to 3. Figure 2 shows two subgraphs of WikiData that correspond to 3-hops (paths of length 3). A 3-hop path derived from the second subgraph is below:

```
[Never Gonna Give You Up] - genre → [Pop]
- genre ← [All I See] - performer → [Kylie Minogue]
```

Before using this path as the MR for a prompt to an LLM, we convert it to a more text-like representation, namely the representation shown in the 6th Row of Table 1 for the SUGGEST DA.

We then utilize the MR generation package from ViGGO⁴ to generate the MRs. The mandatory relations are used first, and then optional relations are randomly sampled to vary the total number of relations in the MRs. In total we create ~71K MRs across the 7 domains, as summarized in the bottom row (Total) of Table 8.

3.3 Prompt Creation and LLM Selection

Domain	Dialogue Act	Example
Movies	give_opinion	description of give_opinion: An expression of opinion along with its justification. The response may consist of 1 or 2 sentences, but it must contain both an opinion and its justification. The justification will also be based on the given attributes. No extra information should be added in Data to Text for give_opinion>: Data: Apollo 13 review score excellent. Apollo 13 producer Brian Grazer. Apollo 13 domain movies. Data to Text for give_opinion: I think Apollo 13 is an excellent movie. Brian Grazer is a great producer and he did an outstanding job with this one.
Music	recommend	description of recommend: A question asking if your friend is familiar with a song you would recommend. In the response, bring up the given song (in a recommending way) because it has certain attributes that were discussed earlier in the conversation. Make sure you ask about the song , not the attributes. No extra information should be added in Data to Text for recommend: Data: Littlest Things part of Alright, Still. Alfie part of Alright, Still. Data to Text for recommend: Have you heard the album Alright, Still? It has two great tracks, Littlest Things and Alfie.

Table 4: Example of the Definitional Prompt for Give-Opinion and Recommend for Movies and Music.

Recent work on data-to-text NLG suggests that even though LLMs will have rarely, if at all, seen examples of data-to-text NLG in their training data (Brown et al., 2020; Raffel et al., 2020; Devlin et al., 2019), many LLMs do well on NLG for dialogue (Soltan et al., 2022; Ramirez et al., 2023). However, since we are transferring the DAs from the ViGGO corpus onto new domains, we start off with only ViGGO examples in the prompts. We

⁴<https://github.com/jjuraska/slug2slug>

piloted several prompt styles, and decided to use a Definitional prompt format (Gupta et al., 2022; Rastogi et al., 2020; Ramirez et al., 2023). Table 4 provides an example of a DA representation, used with a Definitional prompt, for two DAs across two domains. The section of the prompt labelled “Data” is a conversion of the KG path as described above.

After piloting our approach with ChatGPT-4, Vicuna 13B and GPT-3.5, we selected GPT-3.5 as the best performing LLM. For the model configuration, we employed gpt-3.5-turbo with a top-p value of 1, temperature set to 0.7, and a newline token as the stop token, following Ramirez et al. (2023). Then for all 71K MRs, we overgenerate 10 outputs, and then automatically rank them.

3.4 Overgenerate-and-Rank

To create a high-quality data set, we use an overgenerate-and-rank method. The basis for using ranking is a direct translation of the probability of a generated output y , conditioned on a DA d , and an MR a , as in Equation 1. This requires a ranking function that selects outputs that maximize DA accuracy, semantic accuracy, and fluency by assigning a score to each utterance.

$$p(y|d, a) = p(d|y, a) * p(a|y) * p(y) \quad (1)$$

The term $p(d|y, a)$ requires a highly accurate DA classifier to use in automatic ranking. We utilize the ViGGO DA classifier, which achieves an average F1 over .97 for the ViGGO DAs.⁵

In order to estimate $p(a|y)$, semantic accuracy, at generation time in a domain-independent way, we use Beyond-BLEU (BBLEU) (Wieting et al., 2019), which was shown to perform better than other off-the-shelf measures of semantic accuracy such as BLEU, BERTScore and BLEURT (Papineni et al., 2002; Sellam et al., 2020; Zhang et al., 2019). Since these metrics require comparisons with reference utterances, which are not available at generation time, we define a referenceless version based on *pseudo-references*, S_{pseudo} , created from the input DAs (Juraska, 2022). For any set of KG triples, we create its S_{pseudo} by converting each triple to a simple sentence Ent1 relation Ent2, and then concatenating all the triples together. Because pseudo-references are available at generation time, we can use pseudo-Beyond-BLEU (pBBLEU) for ranking.

⁵<https://github.com/aramir62/da-nlg>

The term $p(y)$ requires an estimate of fluency. In general, NLG outputs from very large LLMs do not suffer from problems of fluency, but recent work suggests that the probability $P(S)$ of a generated output S according to an LLM is a good automatic measure of fluency (Kann et al., 2018; Suzgun et al., 2022). We thus adopt $P(S)$ to measure fluency, and use GPT-2 to calculate $P(S)$.

RF_{DA}: DAC | pBBLEU | P(S)

Table 5: Ranking function. DAC = probability of the correct DA using a classifier. pBBLEU = pseudo-Beyond-BLEU to measure semantic accuracy. $P(S)$ = LM probability to measure fluency. The | indicates stepwise evaluation.

For ranking, we adopt the ranking function RF_{DA} in Table 5. Ramirez et al. compared this ranking function with a ranking function that simply multiplies all the terms as in Equation 1. RF_{DA} filters first for DA correctness, and then for semantic accuracy, reflecting the importance of DA correctness. Interestingly, Ramirez et al. showed that the RF_{DA} ranking function also increased **semantic accuracy**, in addition to increasing DA accuracy.

3.5 Self Training

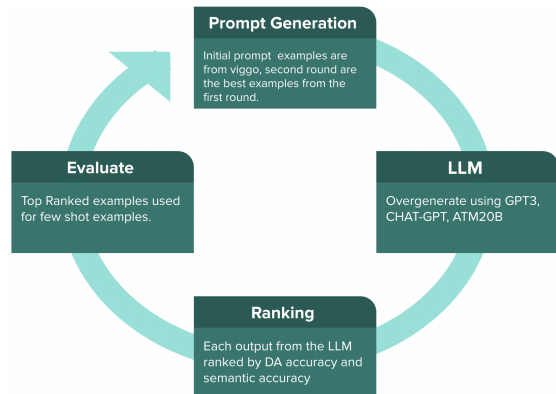


Figure 3: Self-Training Setup for In domain Prompts

One of the challenges with transferring the DA types to new domains is that we have no training data or prompt samples. We investigate a novel approach that uses ViGGO examples in the prompts followed by a round of self-training with silver-generated data. Figure 3 shows the self-training loop in more detail. We start with prompt generation using out of domain, ViGGO examples, then use an LLM to overgenerate examples using these prompts, and rank the outputs. We then select in-domain prompt examples from the top ranked

outputs for self-training. We show in Section 4 that self-training yields a significant improvement in the quality of the Wiki-Dialogue corpus.

The selection of the 10 examples for the in-domain prompts for each DA intent and domain is done manually. We select them following specific criteria in order to provide the LLM a wider knowledge of each DA intents’ realizations. For example, all the attributes of each DA intent are included in the prompt examples in a similar proportion. We also include examples with the different possible number of relations of each DA intent (Table 2). Finally, for scalar attributes with a reduced number of values like RATING, we ensure that all the values are present in a similar proportion.

Interestingly, this two-round self-training process worked successfully for every domain except for Animals, where we received error messages from the LLM complaining about being given data in the wrong domain when using the ViGGO examples. So for Animals, we constructed the 10 in-domain examples for the prompt of each DA intent by hand from a sample of MRs.

4 Results and Corpus Quality Evaluation

The Wiki-Dialogue corpus includes more than 71K utterances ($\sim 50K$ for one-hop and $\sim 21K$ for multi-hop). Table 8 presents a summary of the resulting corpus. Below, we show the benefits of our self-training and overgenerate-and-rank methods in Tables 6 and 7, respectively. We also carry out a comparison with other KG corpora (Table 9) and a human evaluation (Table 10).

4.1 With and Without Self-Training

Domain	BBLEU		DAC	
	N	ST	N	ST
Art	0.84	0.85	1.00	0.99
Music	0.81	0.82	0.98	0.97
Movies	0.77	0.78	0.96	0.97
Sports	0.84	0.86	0.99	0.97
TV	0.75	0.82	0.98	0.98
Boardgames	0.77	0.80	0.94	0.99

Table 6: No self-training = N (Out of Domain Prompts). After self-training = ST (In Domain Prompts). DAC is the DA accuracy using DA classifier. BBLEU is the Beyond BLEU score. Bolded text indicates a significant result (paired t-test, $p < 0.05$)

Table 6 compares the performance of generating outputs with the self-trained domain-specific examples and using ViGGO prompt examples for all the domains except animals (due to its problems with the generation with the ViGGO prompt examples).

We compare them using two metrics: BBLEU (Wieting et al., 2019), and DAC. After self-training (ST), the BBLEU scores show a significant improvement (paired t-test, $p < 0.05$), except for the Music domain. Overall for DAC, self-training improves some domains but not others. However, the DAC is consistently high across all domains, with values ranging from 0.97 to 0.99 after self-training.

4.2 Before and After Overgenerate-and-Rank

Domain	BBLEU		DAC	
	B	A	B	A
Art	0.80	0.85	0.91	0.99
Music	0.78	0.82	0.88	0.97
Movies	0.72	0.78	0.99	0.95
Sports	0.82	0.86	0.87	0.97
TV	0.68	0.82	0.89	0.98
Animals	0.61	0.75	0.86	0.90
Boardgames	0.78	0.80	0.72	0.99

Table 7: Before (B) and After (A) Overgenerate-and-Rank. DAC is the DA accuracy of the DA classifier using self-training examples. BBLEU is the Beyond BLEU score. Bolded text indicates a significant result (paired t-test $p < 0.05$).

Table 7 compares performance before and after applying the overgenerate-and-rank method. The results show that the method consistently improves the performance across all domains, e.g. in the Music domain, the BBLEU score increases from 0.78 to 0.82, and the DAC improves remarkably from 0.88 to 0.97, while in Sports, the BBLEU score increases from 0.82 to 0.86 and the DAC from 0.87 to 0.97 (paired t-test, $p < 0.05$). We see similar results for the rest of the domains. One reason for an increase in the BBLEU scores is that the overgenerate-and-rank method produces a diverse number of outputs, which increases the chances of a candidate output capturing all the attributes from an MR, while outputs that perform worse are dropped after ranking.

4.3 Automatic Evaluation

Table 8 presents a summary of the Wiki-Dialogue corpus, with results for each domain (and also the total) split into one-hop and multi-hop generation. The results for all the domains for BBLEU (> 0.75) and DAC (> 0.92) are very good.

Table 8 also reports four more automatic metrics: vocabulary size, Canary% (Kim et al., 2022), MLTD (McCarthy and Jarvis, 2010) and Flesch-Kincaid (Kincaid et al., 1975). The vocabulary size is a common metric reported for NLG engines,

Domain	Counts		BBLEU		DAC		Vocab Size		Canary%		MLTD		Flesch-Kincaid	
	One	Multi	One	Multi	One	Multi	One	Multi	One	Multi	One	Multi	One	Multi
Art	6297	-	0.85	-	0.99	-	1953	-	0.11	-	56.15	-	9.16	-
Music	5342	3000	0.81	0.86	0.98	0.94	2047	2573	0.00	0.07	40.72	76.56	6.67	8.09
Sports	3473	3000	0.84	0.88	0.99	0.95	3025	3321	0.00	0.00	47.50	67.06	7.56	8.36
TV	7030	5956	0.80	0.85	0.97	1.00	2847	4640	0.00	0.02	45.99	38.10	8.07	9.07
Movies	7083	8295	0.78	0.78	0.97	0.99	3721	4053	0.00	0.00	48.68	36.40	7.37	7.02
Animals	19092	-	0.75	-	0.92	-	4248	-	0.74	-	45.91	-	7.31	-
Boardgames	1500	1500	0.75	0.85	1.00	0.99	504	913	0.40	0.00	54.89	77.23	6.64	7.10
TOTAL	49817	21751	0.78	0.83	0.96	0.98	12985	11051	0.31	0.01	47.36	45.43	7.60	7.94

Table 8: Automatic Evaluation Metrics. Counts are the number of unique MRs. DAC is the DA accuracy using a DA classifier. BBLEU is the Beyond BLEU score. Vocab size defines the number of unique tokens. Canary% is the percentage of sentences considered unsafe by the Canary model. MLTD is a measure of lexical richness. Flesch-Kincaid is a metric of readability. The data is split into one-hop (One) vs. multi-hop (Multi).

especially since neural training methods tend to reduce the size of the vocabulary from the original corpus (Juraska, 2022). The **Vocab Size** column shows that both one and multi-hop utterances use a large vocabulary.

The **Canary%** column is based on the use of the Canary model to analyze which utterances could be problematic in terms of ethical issues, rudeness, toxicity or bias, inspired by work on SODA (Kim et al., 2023). While Kim et al. filtered 5% of the SODA outputs based on Canary, we found that less than the 0.3% of the utterances are considered ethically inappropriate by the model. We did not filter these utterances because a manual check showed that Canary is very sensitive to certain entities, but the utterances are not actually ethically dangerous.

The MLTD and Flesch-Kincaid metrics estimate the lexical richness and readability of the corpus. For both metrics, the results for the Wiki-Dialogue corpus show no large differences across domains for one-hop and multi-hop. The Flesch-Kincaid values show that the Wiki-Dialogue outputs can be understood by the average American, so they are appropriate for a dialogue.

Table 9 shows a comparison of Wiki-Dialogue with other corpora based on KGs, namely WebNLG which is based on DBPedia and OpenDialKG, which is based on FreeBase (Han and Gardent, 2023; Moon et al., 2019). Wiki-Dialogue is larger than WebNLG but smaller than OpenDialKG (column **N**), even though WebNLG covers more domains, and OpenDialKG covers fewer domains. Wiki-Dialogue uses nine different DAs (column **DAs**), while WebNLG only has the **INFORM** DA, and OpenDialKG is not labelled for DAs. Wiki-Dialogue covers 7 domains (column **Dom**), while WebNLG covers 19 domains, and OpenDialKG covers 4 domains. The Music, Sports, and Movies domains are represented in all three datasets. This

suggests that future work could possibly benefit from using a combination of these corpora.

Corpus	Wiki-Dial	WebNLG	OpenDialKG
N	71568	47915	91829
DAs	9	1	?
Dom	7	19	4
Can%	0.22	0.15	0.03
Vocab	18359	6646	20574
MLTD	46.75	27.27	66.23
FK	7.69	8.93	3.71

Table 9: Comparison of Wiki-Dialogue with other corpora based on a Knowledge-Graph. **N** is the number of unique MRs. **DAs** is the number of Dialogue Act types. **Dom** is the number of domains. **FK** is Flesch-Kincaid.

The **Can%** column shows that all of the corpora are very safe (Kim et al., 2022), perhaps because they are all knowledge grounded. The **Vocab** and **MLTD** columns show that Wiki-Dialogue has a larger vocabulary and is more lexically diverse than WebNLG despite the fact that WebNLG covers more domains. Compared to OpenDialKG, Wiki-Dialogue has lower lexical diversity (column **MLTD**), which may be due to the fact that OpenDialKG is human-human. Both Wiki-Dialogue and WebNLG have a higher Flesch-Kincaid (column **FK**) reading level than OpenDialKG, probably because OpenDialKG restricts MRs to 1 and 2 hops, making utterances shorter on average.

4.4 Human Evaluation

Table 11 and Table 12 in the Appendix provide example realizations of every DA for all 7 domains. These examples show that the quality of the corpus is high: the realizations are natural and the DAs are correctly realized with high accuracy.

For human evaluation, we selected 100 examples from each domain for both one-hop and multi-hop yielding 1200 examples with 700 one-hop examples and 500 multi-hop examples. Five expert anno-

Domain	↓ HAL		↑ PERF		↑ DAC		↑ SAC	
	One	Multi	One	Multi	One	Multi	One	Multi
Art	0.01	-	0.98	-	1.00	-	0.98	-
Animals	0.11	-	0.82	-	0.89	-	0.93	-
BoardGames	0.00	0.18	0.97	0.76	1.00	0.81	0.97	0.95
Movies	0.01	0.06	0.91	0.82	0.98	0.82	0.94	1.00
Music	0.01	0.18	0.97	0.89	1.00	1.00	0.97	0.89
Sports	0.02	0.09	0.97	0.95	1.00	1.00	0.97	0.95
TV	0.00	0.10	0.93	0.73	0.98	0.80	0.95	0.92
TOTAL	0.02	0.10	0.94	0.83	0.98	0.89	0.96	0.94

Table 10: Human annotation results for HAL (Hallucinations), DAC (DA accuracy), and SAC (Semantic Accuracy). PERF (Perfect Semantic and DA accuracy) is calculated based on DAC and SAC.

tators were given a manual for DAs, and provided examples of hallucinations and utterances that were both perfect and not perfect. Each set of utterances were annotated for hallucinations (HAL), DA accuracy (DAC), and semantic accuracy (SAC). The 100 example-set for each domain and hop type was annotated by one annotator, and then 30 of these were re-annotated by a second expert to estimate inter-annotator agreement using Cohen’s Kappa. The average DAC Kappa is 0.94, and the SAC Kappa is 0.89 showing a very high level of agreement between the annotators.

We then calculated the percentage of utterances that had both perfect DAC and SAC (PERF). The results are in Table 10. Overall, Table 10 shows that the quality of the Wiki-Dialogue corpus is high, with perfect outputs that correctly realize both the specified DA and the set of KG triples in the MR ranging from 73% for TV multi-hop to 98% for Art, with an overall average over both hop types of 89% PERFECT outputs. We see that one-hop datasets have fewer hallucinations, better DA accuracy and more perfect utterances. While some values for hallucinations seem high, e.g. 18% for Music multi-hop, these values compare favorably to crowdsourced corpora such as MultiWOZ or E2E, as discussed in Section 2.

5 Conclusion and Future Work

This paper describes and provides a novel 71K utterance corpus called Wiki-Dialogue, covering 9 DAs and 7 KG based domains that are useful for both social conversation and task-oriented dialogue. The corpus includes both one and multi-hop sets of KG triples, and the MRs vary from a single triple for some REQUEST dialogue acts up to 8 triples for INFORM DAs.

Our novel method involves a self-training step to create prompt examples for novel domains, fol-

lowed by an overgenerate-and-rank step, and we show that these two steps combined drastically improve the quality of the corpus. We assess the quality of the corpus with both automatic and human evaluation and find that the quality is high. We hand annotate for hallucinations and semantic errors and find their frequency to be lower than reported values for crowdsourced corpora such as E2E and MultiWOZ (Dušek et al., 2019, 2020; Qian et al., 2021). We also check the corpus for safety using Canary and find that only 0.22% of the utterances are flagged as needs_intervention, while a manual inspection of these suggests that there are no safety issues with them.

In a comparison with similar corpora such as WebNLG and OpenDialogKG, we observe that although WebNLG cover more domains Wiki-Dialogue is lexically richer and has a larger vocabulary. While WebNLG has only INFORM DAs, Wiki-Dialogue covers 9 DAs, providing a complementary and unique resource to the dialogue community.

Future work should explore how the Wiki-Dialogue corpus can be used to train an NLG for dialogue. In preliminary experiments, a subset of Wiki-Dialogue was used to fine-tune a 3B multi-domain NLG engine, which was tested in Athena, a real-time Amazon Alexa Prize system, with human users of Amazon “Let’s Chat” (Yue Fan and Wang, 2023). The deployment used a cross-domain universal dialogue policy based on the nine Wiki-Dialogue DAs. An example of this policy is shown in the music conversation in Figure 1.

Ethical Considerations. There are several potential risks with such an NLG. LLMs introduce the possibility of disinformation, often called hallucinations, whose control is an active area of research. One of the challenges is that it is very difficult to automatically identify them. Here we experiment with ranking functions for better control of hallucinations, hand-label hallucinations and characterize them. Another potential risk is that some of the DAs, like recommend and suggest, could be used in an application to persuade a user to buy something. **Acknowledgments.** This work has been partially funded by Spanish MCIU by the BEWORD project (grant number PID2021-126061OB-C42) and by the Basque Government under grant PRE 2020 1 0274.

References

- Hannah Bast, Florian Baurle, Björn Buchhold, and Elmar Haubmann. 2014. Easy access to the Freebase dataset. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 95–98.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Paweł Budzianowski and Ivan Vulić. 2019. Hello, It’s GPT-2 - How Can I Help You? Towards the Use of Pretrained Language Models for Task-Oriented Dialogue Systems. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 15–22.
- Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2023. **PLACES: Prompting language models for social conversation synthesis**. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 844–868, Dubrovnik, Croatia. Association for Computational Linguistics.
- Ethan A Chi, Ashwin Paranjape, Abigail See, Caleb Chiam, Trenton Chang, Kathleen Kenealy, Swee Kiat Lim, Amelia Hardy, Chetanya Rastogi, Haojun Li, et al. 2022. Neural generation meets real people: Building a social, informative open-domain dialogue agent. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 376–395.
- Emilie Colin, Claire Gardent, Yassine M’rabet, Shashi Narayan, and Laura Perez-Beltrachini. 2016. The WEBNLG challenge: Generating Text from DBpedia data. In *Proceedings of the 9th international natural language generation conference*, pages 163–167.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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.
- Ondřej Dušek, David M Howcroft, and Verena Rieser. 2019. Semantic noise matters for neural natural language generation. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 421–426.
- Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2020. Evaluating the state-of-the-art of end-to-end natural language generation: The E2E NLG challenge. *Computer Speech & Language*, 59:123–156.
- Nouha Dziri, Andrea Madotto, Osmar R Zaiane, and Avishek Joey Bose. 2021. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2197–2214.
- Mihail Eric, Nicole Chartier, Behnam Hedayatnia, Karthik Gopalakrishnan, Pankaj Rajan, Yang Liu, and Dilek Hakkani-Tur. 2021. **Multi-sentence knowledge selection in open-domain dialogue**. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 76–86, Aberdeen, Scotland, UK. Association for Computational Linguistics.
- Thiago Castro Ferreira, Diego Moussallem, Emiel Kraemer, and Sander Wubben. 2018. Enriching the WebNLG corpus. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 171–176.
- Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. Creating training corpora for nlg micro-planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 179–188.
- Karthik Gopalakrishnan, Behnam Hedayatnia, Qinqiang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, Dilek Hakkani-Tür, and Amazon Alexa AI. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations. *Proc. Interspeech 2019*, pages 1891–1895.
- Raghav Gupta, Harrison Lee, Jeffrey Zhao, Yuan Cao, Abhinav Rastogi, and Yonghui Wu. 2022. **Show, don’t tell: Demonstrations outperform descriptions for schema-guided task-oriented dialogue**. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4541–4549, Seattle, United States. Association for Computational Linguistics.
- Kelvin Han and Claire Gardent. 2023. **Generating and answering simple and complex questions from text and from knowledge graphs**. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 285–304, Nusa Dua, Bali. Association for Computational Linguistics.
- Behnam Hedayatnia, Karthik Gopalakrishnan, Seokhwan Kim, Yang Liu, Mihail Eric, and Dilek Hakkani-Tur. 2020. Policy-driven neural response generation for knowledge-grounded dialog systems. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 412–421.
- Juraj Juraska. 2022. *Diversifying Language Generated by Deep Learning Models in Dialogue Systems*. Ph.D. thesis, UC Santa Cruz.

- Juraj Juraska, Kevin Bowden, Lena Reed, Vrindavan Harrison, Wen Cui, Omkar Patil, Rishi Rajasekaran, Angela Ramirez, Cecilia Li, Eduardo Zamora, et al. 2021. Athena 2.0: Contextualized Dialogue Management for an Alexa Prize SocialBot. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 124–133.
- Juraj Juraska, Kevin K Bowden, and Marilyn Walker. 2019. ViGGO: A video game corpus for data-to-text generation in open-domain conversation. In *Proceedings of the 12th International Conference on Natural Language Generation*.
- Katharina Kann, Sascha Rothe, and Katja Filippova. 2018. Sentence-level fluency evaluation: References help, but can be spared! In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 313–323.
- Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Bras, Malihe Alikhani, Gunhee Kim, et al. 2023. SODA: Million-scale Dialogue Distillation with Social Commonsense Contextualization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12930–12949.
- Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. 2022. ProsocialDialog: A prosocial backbone for conversational agents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4005–4029, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated reliability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical Report Research Branch Report 8-75, Naval Technical Training, US Naval Air Station, Millington, TN.
- Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1213.
- Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. Dbpedia—a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2):167–195.
- Philip M McCarthy and Scott Jarvis. 2010. MTL, VOCD-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2):381–392.
- Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. OpenDialKG: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854, Florence, Italy. Association for Computational Linguistics.
- Linyong Nan, Dragomir Radev, Rui Zhang, Amrit Rau, Abhinand Sivaprasad, Chiachun Hsieh, Xiangru Tang, Aadit Vyas, Neha Verma, Pranav Krishna, et al. 2021. DART: Open-domain structured data record to text generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 432–447.
- Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2017. The E2E Dataset: New Challenges For End-to-End Generation. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 201–206.
- 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.
- Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqi, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. ToTTo: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186.
- Kun Qian, Ahmad Beirami, Zhouhan Lin, Ankita De, Alborz Geramifard, Zhou Yu, and Chinnadhurai Sankar. 2021. Annotation inconsistency and entity bias in MultiWOZ. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 326–337, Singapore and Online. Association for Computational Linguistics.
- 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*, 21:1–67.
- Angela Ramirez, Kartik Aggarwal, Juraj Juraska, Utkarsh Garg, and Marilyn Walker. 2023. Controllable generation of dialogue acts for dialogue systems via few-shot response generation and ranking. In *Meeting of the Special Interest Group on Discourse and Dialogue, SIGDIAL*.
- 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.

- Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. 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 1702–1723.
- Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. [BLEURT: Learning robust metrics for text generation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.
- Saleh Soltan, Shankar Ananthkrishnan, Jack FitzGerald, Rahul Gupta, Wael Hamza, Haidar Khan, Charith Peris, Stephen Rawls, Andy Rosenbaum, Anna Rumshisky, et al. 2022. AlexaTM 20b: Few-shot learning using a large-scale multilingual seq2seq model. arXiv preprint arXiv:2208.01448.
- Mirac Suzgun, Luke Melas-Kyriazi, and Dan Jurafsky. 2022. Prompt-and-rerank: A method for zero-shot and few-shot arbitrary textual style transfer with small language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2195–2222.
- Tsung-Hsien Wen, Milica Gasic, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721.
- John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019. [Beyond BLEU: training neural machine translation with semantic similarity](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4344–4355, Florence, Italy. Association for Computational Linguistics.
- Qingyang Wu, James Gung, Raphael Shu, and Yi Zhang. 2023. DiactTOD: Learning generalizable latent dialogue acts for controllable task-oriented dialogue systems. arXiv preprint arXiv:2308.00878.
- Xinnuo Xu, Guoyin Wang, Young-Bum Kim, and Sungjin Lee. 2021. [AugNLG: Few-shot natural language generation using self-trained data augmentation](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1183–1195, Online. Association for Computational Linguistics.
- Fanghua Ye, Jarana Manotumruksa, and Emine Yilmaz. 2022. MultiWOZ 2.4: A multi-domain task-oriented dialogue dataset with essential annotation corrections to improve state tracking evaluation. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 351–360.
- Wen Cui Winson Chen Vrindavan Harrison Angela Ramirez Saaket Agashe Xinyue Gabby Liu Neha Pulabhotla Nan Qiang Jeshwanth Bheemanpally Sugam Garg Marilyn Walker Yue Fan, Kevin K. Bowden and Xin Eric Wang. 2023. Athena 3.0: Personalized multimodal chatbot with neuro-symbolic dialogue generators. *Proceedings of the Alexa Prize*.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.
- Giulio Zhou and Gerasimos Lampouras. 2020. WebNLG Challenge 2020: Language agnostic delexicalisation for multilingual RDF-to-Text generation. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 186–191.

A Appendix

domain	MR	text
verify_attribute		
sports	Venus Williams rating okay. Venus Williams start time 1994. Venus Williams domain athlete.	You said you think Venus Williams is okay. Do you think she is an average athlete compared to other players who started in 1994?
movies	Avengers: Endgame review score good. Avengers: Endgame cast member Robert Redford, Taika Waititi. Avengers: Endgame domain movies.	I remember you thought Avengers: Endgame was good. Do you usually enjoy movies with Robert Redford and Taika Waititi?
music	Kings Of The Wild Frontier Good. Kings Of The Wild Frontier	I know you like Kings of the Wild Frontier. Do you like other songs from the Kings of the Wild Frontier album?
tv	The Crown review score average. The Crown genre biographical film, drama television series. The Crown domain tv.	I know you said The Crown had an average review score. Do you feel the same way about other biographical films and drama television series?
art	Raphael rating excellent. Raphael genres portrait, portrait painting. Raphael movements Italian Renaissance, High Renaissance.	You rated the portrait painting by Raphael as excellent. Would you say you generally have a strong liking for Italian Renaissance and High Renaissance portrait paintings?
animals	Carolina Dog good with other dogs friendly to other dogs. Carolina Dog good with children kid-friendly. Carolina Dog domain animals.	I know you said that Carolina Dogs are friendly to other dogs and kid-friendly. Are you generally interested in animals that are friendly to other dogs and kid-friendly?
board games	A Game of Thrones average rating good. A Game of Thrones genre fantasy. A Game of Thrones domain board game.	You said A Game of Thrones was good. Do you usually like fantasy board games?
recommend		
sports	Jake Paul sport boxing. Jake Paul height 1.83. Jake Paul domain athlete.	Have you heard of Jake Paul? He's a professional boxer who stands at 1.83 meters tall.
movies	Elvis review score good. Elvis producer Gail Berman, Catherine Martin, Baz Luhrmann. Elvis domain movies.	Have you seen Elvis, produced by Gail Berman, Catherine Martin, and Baz Luhrmann? It has a good review score.
music	Money For Nothing Rock Music Littlest Things Mark Ronson	Have you heard of the rock song Money for Nothing produced by Mark Ronson?
tv	Peaky Blinders review score good. Peaky Blinders creator David Caffrey. Peaky Blinders domain tv.	Have you seen Peaky Blinders from David Caffrey? It's got a really good review score.
art	The Haywain Triptych creator Hieronymus Bosch. The Haywain Triptych location Museo del Prado.	Have you seen The Haywain Triptych by Hieronymus Bosch at the Museo del Prado?
animals	Staffordshire Bull Terrier trainability very easy to train. Staffordshire Bull Terrier domain animals.	I think you'd really be interested in the Staffordshire Bull Terrier, which is very easy to train. Have you heard of it before?
board games	HeroQuest publisher Games Workshop. HeroQuest genre fantasy. HeroQuest domain board game.	Have you ever heard of HeroQuest, the fantasy board game from Games Workshop?
give_opinion		
sports	Ronaldo sport association football. Ronaldo rating good. Ronaldo domain athlete.	I think Ronaldo is a great footballer. He has consistently put in strong performances for his teams and is known for his skill and passion for the game.
movies	Scooby-Doo review score average. Scooby-Doo director Raja Gosnell. Scooby-Doo MPA film rating PG. Scooby-Doo domain movies.	Scooby-Doo is an average movie, but I think it's worth watching for the great direction of Raja Gosnell and the PG rating.
music	Y.M.C.A. rating excellent. Y.M.C.A. producer Jacques Morali. Y.M.C.A. record label Casablanca Records. Y.M.C.A. domain songs.	Y.M.C.A. by the Village People, produced by Jacques Morali and released by Casablanca Records, is an excellent song.
tv	Gunpowder review score average. Gunpowder screenwriter Ronan Bennett. Gunpowder domain tv.	Gunpowder was an average show, but I appreciate the work of the screenwriter, Ronan Bennett, for making the story interesting.
art	Henri de Toulouse-Lautrec rating good. Henri de Toulouse-Lautrec genres animal painting, portrait.	I think Henri de Toulouse-Lautrec is a good painting. The combination of animal painting and portraiture in this piece is quite engaging.
animals	Fin Whale popularity very popular. Fin Whale weight 1,800kg - 70,000kg 4,000lbs - 150,000lbs. Fin Whale domain animals.	I'm really impressed by the Fin Whale with its huge weight range of 1,800kg - 70,000kg (4,000lbs - 150,000lbs).
board games	Twilight Struggle average rating excellent. Twilight Struggle maximum number of players 2. Twilight Struggle domain board game.	I highly recommend Twilight Struggle. It's an excellent board game with a maximum number of players of 2, making it perfect for a competitive game night.
request_attribute		
sports	athletes sports teams.	What sports team does your favorite athlete play for?
movies	movies review score None. movies domain movies.	What rating would you give to movies that you have watched recently?
music	songs performer.	Who is your go-to artist for music?
tv	show screenwriter None. show domain tv.	What is your favorite TV show and its screenwriter?
art	paintings genres.	What types of painting genres resonate with you the most?
animals	animal good with other dogs. animal domain animals.	Is your favorite animal good with other dogs?
board games	board game genre. board game domain board game.	What board game genre do you find the most enjoyable?

Table 11: Examples of transferred DAs for `verify_attribute`, `recommend`, `give_opinion`, and `request_attribute` to all 7 domains.

domain	MR	text
suggest		
sports	Wilt Chamberlain Nba Most Valuable Player Award, Nba All-Star Game Kobe Bryant Most Valuable Player Award	Have you ever heard of Wilt Chamberlain, who won the NBA Most Valuable Player Award and the NBA All-Star Game Kobe Bryant Most Valuable Player Award?
movies	Ant-Man and the Wasp cast member Dax Griffin, Hannah John-Kamen. Ant-Man and the Wasp publication date 2018. Ant-Man and the Wasp domain movies.	Did you ever watch Ant-Man and the Wasp, starring Dax Griffin and Hannah John-Kamen, which was released in 2018?
music	Waka Waka performer Freshlyground, Shakira. Waka Waka domain songs.	Do you like the music of Shakira and Freshlyground, like Waka Waka?
tv	Breaking Bad character Tuco Salamanca, Mike Ehrmantraut. Breaking Bad domain tv.	Do you know any characters from Breaking Bad, such as Tuco Salamanca and Mike Ehrmantraut?
art	The Bridge at Narni creator Jean-Baptiste Camille Corot. The Bridge at Narni genres landscape art.	Have you explored landscape art, such as The Bridge at Narni by Jean-Baptiste Camille Corot?
animals	Pygora Goat weight 75-95 lbs. for males, 64-75 lbs. for females.. Pygora Goat lifespan 12-15 years. Pygora Goat domain animals.	Have you heard about the Pygora Goat? It typically weighs 75-95 lbs. for males and 64-75 lbs. for females, and has a lifespan of 12-15 years.
board games	Enchanted Forest maximum number of players 6. Enchanted Forest instance of board game. Enchanted Forest domain board game.	Have you ever played a board game for up to 6 players, like Enchanted Forest?
request		
sports	athletes specifier boring.	Who do you think is the most boring athlete to watch?
movies	movies specifier realistic. movies director Cary Joji Fukunaga. movies domain movies.	Have you seen any realistic movies directed by Cary Joji Fukunaga?
music	songs specifier beautiful.	What's the most beautiful song you can think of?'
tv	show specifier overrated. show screenwriter Arif Aliyev. show domain tv.	Do you think any of Arif Aliyev's TV shows are overrated?
art	paintings specifier bizarre. paintings location Museo Nacional Centro de Arte Reina Sofia.	Can you think of a bizarre painting from the Museo Nacional Centro de Arte Reina Sofia?
animals	animals specifier cheerful. animals habitat Tropical bushland, grassland and savannas.	Do you know of any cheerful animals that live in Tropical bushland, grassland and savannas?
board games	board game specifier frustrating. board game maximum number of players 5. board game domain board game.	What's the most frustrating board game you can play with up to 5 players?
request_explanation		
sports	athletes rating good. athletes team positions center.	What is it about athletes who play the center position that you find so good?
movies	movies review score good. movies screenwriter Steven Knight. movies domain movies.	What do you think makes the movies written by Steven Knight so good?
music	songs rating good. songs from album Katharine McPhee.	What do you like about the songs from Katharine McPhee's album that make them good?
tv	show review score average. show awards received None. show creator Jonathan Frakes, Hanelle Culpepper. show domain tv.	Given the average review score and lack of awards, what do you think Jonathan Frakes and Hanelle Culpepper did right in their TV shows?
art	painters rating good. painters movements Impressionism, post-impressionism.	What is it about Impressionism and post-impressionism paintings that you find so good?
animals	animal coat length short coat. animal protectiveness very alert. animal name Airedale Terrier. animal good with other dogs friendly to other dogs. animal domain animals.	What do you like about the Airedale Terrier that has a short coat, is very alert, and is friendly to other dogs?
board games	board game genre science fiction. board game average rating excellent. board game domain board game.	What makes science fiction board games your favorite type of game?
confirm		
sports	Dwayne Johnson sport professional wrestling. Dwayne Johnson sports teams Calgary Stampeders, Miami Hurricanes football. Dwayne Johnson domain athletes.	Are you talking about Dwayne Johnson, the professional wrestler who previously played for the Calgary Stampeders and the Miami Hurricanes football team?
movies	Thor: Love and Thunder cast member Jacob Yee, Gemma Dart. Thor: Love and Thunder domain movies.	Are you referring to the movie Thor: Love and Thunder, starring Jacob Yee and Gemma Dart?
music	Break My Soul genres house music. Break My Soul domain songs.	Are you referring to the house music song Break My Soul?
tv	Obi-Wan Kenobi character Darth Vader, Obi-Wan Kenobi. Obi-Wan Kenobi screenwriter Hossein Amini. Obi-Wan Kenobi domain tv.	Are you referring to the Obi-Wan Kenobi TV show, written by Hossein Amini, featuring Darth Vader and Obi-Wan Kenobi?
art	The Wayfarer creator Hieronymus Bosch. The Wayfarer genres genre art. The Wayfarer movements Early Netherlandish painting.	Are you referring to The Wayfarer, the Early Netherlandish genre art painting by Hieronymus Bosch?
animals	Carolina Parakeet common class Bird. Carolina Parakeet predators Birds of prey, raccoons, skunks, squirrels, and snakes. Carolina Parakeet domain animals.	Do you mean the Bird known as the Carolina Parakeet, which is preyed upon by Birds of prey, raccoons, skunks, squirrels, and snakes?
board games	Monopoly publisher Hasbro. Monopoly country of origin United States of America. Monopoly domain board game.	Are you talking about the Hasbro board game Monopoly, from the United States of America?
inform		
sports	Magic Johnson sport basketball. Magic Johnson stats leader 1992 NBA All-Star Game. Magic Johnson domain athletes.	Magic Johnson is a basketball athlete who was the leading stats leader for the 1992 NBA All-Star Game.
movies	Captain Phillips 2013 'Film Based On Books', 'Survival Film' 'Faysal Ahmed As Nour Najee' Excellent 'Louis Mahoney', 'Catherine Keener'	Captain Phillips is a 2013 survival film based on books, starring Faysal Ahmed as Nour Najee, Louis Mahoney, and Catherine Keener. It has an excellent review score.
music	As It Was record label Columbia Records. Mireille Mathieu record label Columbia Records. Mireille Mathieu genre pop music. As It Was genre synth-pop. As It Was performer Harry Styles.	Harry Styles' As It Was is a synth-pop album released on Columbia Records. Mireille Mathieu's record label is also Columbia Records and her genre is pop music. (3 hop)
tv	Peaky Blinders genre drama television series, costume drama. Peaky Blinders creator David Caffrey. Peaky Blinders screenwriter Steven Knight. Peaky Blinders review score good. Peaky Blinders character Arthur Shelby, Jessie Eden. Peaky Blinders domain tv.	Peaky Blinders is a good costume drama television series created by David Caffrey and written by Steven Knight. It stars Arthur Shelby and Jessie Eden.
art	The Coronation of Napoleon creator Jacques-Louis David. The Coronation of Napoleon materials canvas, oil paint. The Coronation of Napoleon genres history painting.	The Coronation of Napoleon, a history painting by Jacques-Louis David, is made of oil paint on canvas.
animals	Australian Stumpy Tail Cattle Dog shedding average shedding. Australian Stumpy Tail Cattle Dog domain animals.	The Australian Stumpy Tail Cattle Dog has average shedding.
board games	Enchanted Forest genre fantasy. Enchanted Forest maximum number of players 6. Enchanted Forest domain board game.	Enchanted Forest is a fantasy board game for up to 6 players.

Table 12: Examples of transferred DAs for suggest, request, request_explanation, confirm and inform to all 7 domains