ViGGO: A Video Game Corpus for Data-To-Text Generation in Open-Domain Conversation

Juraj Juraska, Kevin K. Bowden and **Marilyn Walker**
Natural Language and Dialogue Systems Lab
University of California, Santa Cruz
{jjuraska,kkbowden,mawalker}@ucsc.edu

Abstract

The uptake of deep learning in natural language generation (NLG) led to the release of both small and relatively large parallel corpora for training neural models. The existing data-to-text datasets are, however, aimed at task-oriented dialogue systems, and often thus limited in diversity and versatility. They are typically crowdsourced, with much of the noise left in them. Moreover, current neural NLG models do not take full advantage of large training data, and due to their strong generalizing properties produce sentences that look template-like regardless. We therefore present a new corpus of 7K samples, which (1) is clean despite being crowdsourced, (2) has utterances of 9 generalizable and conversational dialogue act types, making it more suitable for open-domain dialogue systems, and (3) explores the domain of video games, which is new to dialogue systems despite having excellent potential for supporting rich conversations.

1 Introduction

The recent adoption of deep learning methods in natural language generation (NLG) for dialogue systems resulted in an explosion of neural data-to-text generation models, which depend on large training data. These are typically trained on one of the few parallel corpora publicly available, in particular the E2E (Novikova et al., 2017) and the WebNLG (Gardent et al., 2017) datasets. Crowdsourcing large NLG datasets tends to be a costly and time-consuming process, making it impractical outside of task-oriented dialogue systems. At the same time, current neural NLG models struggle to replicate the high language diversity of the training sentences present in these large datasets, and instead they learn to produce the same generic type of sentences as with considerably less training data (Deriu and Cieliebak, 2018; Juraska and Walker, 2018; Dušek et al., 2019).

give_opinion(NAME [**SpellForce 3**], RATING [**poor**], GENRES [**real-time strategy, role-playing**], PLAYER_PERSPECTIVE [**bird view**])
I think that **SpellForce 3** is **one of the worst games** I've ever played. Trying to combine the **real-time strategy** and **role-playing** genres just doesn't work, and the **bird's eye view** makes it near impossible to play.
verify_attribute(NAME [**Little Big Adventure**], RATING [**average**], HAS_MULTIPLAYER [**no**], PLATFORMS [**PlayStation**])
I recall that you were **not that fond** of **Little Big Adventure**. Does **single-player** gaming on the **PlayStation** quickly get boring for you?

Table 1: Examples of MRs and corresponding reference utterances in the ViGGO dataset. The DA of the MRs is indicated in italics, and the slots in small caps. The slot mentions in the utterances are bolded.

Motivated by the rising interest in open-domain dialogue systems and conversational agents, we present ViGGO – a smaller but more comprehensive dataset in the video game domain, introducing several generalizable dialogue acts (DAs), making it more suitable for training versatile and more conversational NLG models.[1] The dataset provides almost 7K pairs of structured meaning representations (MRs) and crowdsourced reference utterances about more than 100 video games. Table 1 lists three examples.

Video games are a vast entertainment topic that can naturally be discussed in a casual conversation, similar to movies and music, yet in the dialogue systems community it does not enjoy popularity anywhere close to that of the latter two topics (Fazel-Zarandi et al., 2017; Li et al., 2017; Moghe et al., 2018; Shah et al., 2018; Khatri et al., 2018). Restaurants have served as the go-to topic in data-to-text NLG for decades, as they offer a sufficiently large set of various attributes and cor-

[1]The ViGGO corpus is available for download at: https://nlds.soe.ucsc.edu/viggo

responding values to talk about. While they certainly can be a topic of a casual conversation, the existing restaurant datasets (Stent et al., 2004; Gašić et al., 2008; Mairesse et al., 2010; Howcroft et al., 2013; Wen et al., 2015a; Nayak et al., 2017) are geared more toward a task-oriented dialogue where a system tries to narrow down a restaurant based on the user's preferences and ultimately give a recommendation. Our new video game dataset is designed to be more conversational, and to thus enable neural models to produce utterances more suitable for an open-domain dialogue system.

Even the most recent addition to the publicly available restaurant datasets for data-to-text NLG, the E2E dataset (Novikova et al., 2017), suffers from the lack of a conversational aspect. It has become popular, thanks to its unprecedented size and multiple reference utterances per MR, for training end-to-end neural models, yet it only provides a single DA type. In contrast with the E2E dataset, ViGGO presents utterances of 9 different DAs.

Other domains have been represented by task-oriented datasets with multiple DA types, for example the Hotel, Laptop, and TV datasets (Wen et al., 2015b, 2016). Nevertheless, the DAs in these datasets vary greatly in complexity, and their distribution is thus heavily skewed, typically with two or three similar DAs comprising almost the entire dataset. In our video game dataset, we omitted simple DAs, in particular those that do not require any slots, such as greetings or short prompts, and focused on a set of substantial DAs only.

The main contribution of our work is thus a new parallel data-to-text NLG corpus that (1) is more conversational, rather than information seeking or question answering, and thus more suitable for an open-domain dialogue system, (2) represents a new, unexplored domain which, however, has excellent potential for application in conversational agents, and (3) has high-quality, manually cleaned human-produced utterances.

2 The ViGGO Dataset

ViGGO features more than 100 different video game titles, whose attributes were harvested using free API access to two of the largest online video game databases: IGDB[2] and GiantBomb[3]. Using these attributes, we generated a set of 2,300 structured MRs. The human reference utterances for the generated MRs were then crowdsourced using vetted workers on the Amazon Mechanical Turk (MTurk) platform (Buhrmester et al., 2011), resulting in 6,900 MR-utterance pairs altogether. With the goal of creating a clean, high-quality dataset, we strived to obtain reference utterances with correct mentions of all slots in the corresponding MR through post-processing.

[2]https://www.igdb.com/

[3]https://www.giantbomb.com/

DA	Slot range	Mandatory slots	Additional common slots
inform	3-8	NAME, GENRES	RELEASE_YEAR, DEVELOPER, ESRB, GENRES, PLAYER_PERSPECTIVE, HAS_MULTIPLAYER, PLATFORMS, AVAILABLE_ON_STEAM, HAS_LINUX_RELEASE, HAS_MAC_RELEASE
confirm	2-3	NAME	
give_opinion	3-4	NAME, RATING	
recommend	2-3	NAME	
request	1-2	SPECIFIER	
request_attribute	1		
request_explanation	2-3	RATING	
suggest	2-3	NAME	
verify_attribute	3-4	NAME, RATING	

Table 2: Overview of mandatory and common possible slots for each DA in the ViGGO dataset. There is an additional slot, EXP_RELEASE_DATE, only possible in the *inform* and *confirm* DAs. Moreover, RATING is also possible in the *inform* DA, though not mandatory.

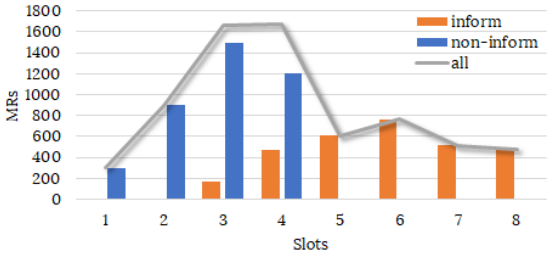

Figure 1: Distribution of the number of slots across all types of MRs, as well as the *inform* slot separately, and non-*inform* slots only.

2.1 Meaning Representations

The MRs in the ViGGO dataset range from 1 to 8 slot-value pairs, and the slots come from a set of 14 different video game attributes. Table 2 details how these slots may be distributed across the 9 different DAs. The *inform* DA, represented by 3,000 samples, is the most prevalent one, as the average number of slots it contains is significantly higher than that of all the other DAs. Figure 1 visualizes the MR length distribution across the entire dataset.

The slots can be classified into 5 general categories covering most types of information MRs typically convey in data-to-text generation scenar-

ios: *Boolean*, *Numeric*, *Scalar*, *Categorical*, and *List*. The first 4 categories are common in other NLG datasets, such as E2E, Laptop, TV, and Hotel, while the *List* slots are unique to ViGGO. *List* slots have values which may comprise multiple items from a discrete list of possible items.

2.2 Utterances

With neural language generation in mind, we crowdsourced 3 reference utterances for each MR so as to provide the models with the information about how the same content can be realized in multiple different ways. At the same time, this allows for a more reliable automatic evaluation by comparing the generated utterances with a set of different references each, covering a broader spectrum of correct ways of expressing the content given by the MR. The raw data, however, contains a significant amount of noise, as is inevitable when crowdsourcing. We therefore created and enforced a robust set of heuristics and regular expressions to account for typos, grammatical errors, undesirable abbreviations, unsolicited information, and missing or incorrect slot realizations.

2.3 Data Collection

The crowdsourcing of utterances on MTurk took place in three stages. After collecting one third of the utterances, we identified a pool of almost 30 workers who wrote the most diverse and natural-sounding sentences in the context of video games. We then filtered out all utterances of poor quality and had the qualified workers write new ones for the corresponding inputs. Finally, the remaining two thirds of utterances were completed by these workers exclusively.

For each DA we created a separate task in order to minimize the workers' confusion. The instructions contained several different examples, as well as counter-examples, and they situated the DA in the context of a hypothetical conversation. The video game attributes to be used were provided for the workers in the form of a table, with their order shuffled so as to avoid any kind of bias. Further details on the data collection and cleaning are included in the Appendix.

2.4 Train/Validation/Test Split

Despite the fact that the ViGGO dataset is not very large, we strived to make the test set reasonably challenging. To this end, we ensured that, after delexicalizing the NAME and the DEVELOPER slots, there were no common MRs between the train set and either of the validation or test set. We maintained a similar MR length and slot distribution across the three partitions. The distribution of DA types, on the other hand, is skewed slightly toward fewer *inform* DA instances and a higher proportion of the less prevalent DAs in the validation and test sets (see Figure 2). With the exact partition sizes indicated in the diagram, the final ratio of samples is approximately $7.5 : 1 : 1.5$.

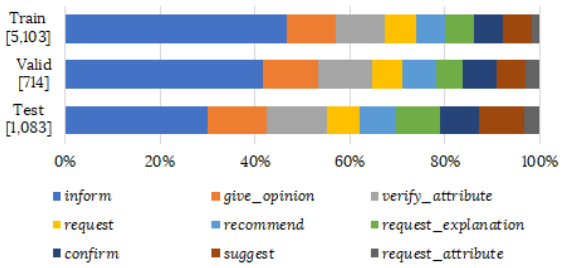

Figure 2: Distribution of the DAs across the train/validation/test split. For each partition the total count of DAs/MRs is indicated.

2.5 ViGGO vs. E2E

Our new dataset was constructed under different constraints than the E2E dataset. First, in ViGGO we did *not* allow any omissions of slot mentions, as those are not justifiable for data-to-text generation with no previous context, and it makes the evaluation ambiguous. Second, the MRs in ViGGO are grounded by *real* video game data, which can encourage richer and more natural-sounding reference utterances.

While ViGGO is only 13% the size of the E2E dataset, the lexical diversity of its utterances is 77% of that in the E2E dataset, as indicated by the "delexicalized vocabulary" column in Table 3. Part of the reason naturally is the presence of additional DAs in ViGGO, and therefore we also indicate the statistics in Table 3 for the *inform* samples only. The average *inform* utterance length in ViGGO turns out to be over 30% greater, in terms of both words and sentences per utterance.

Finally, we note that, unlike the E2E dataset, our test set does not place any specific emphasis on longer MRs. While the average number of slots per MR in the *inform* DAs are comparable to the E2E dataset, in general the video game MRs are significantly shorter. This is by design, as shorter, more focused responses are more conversational than consistently dense utterances.

	Instances	Unique MRs	Unique delex. MRs	Vocab	Delex. vocab	Avg. 3-gram freq.	Refs/ MR	Slots/ MR	W/ Ref	W/ Sent	Sents/ Ref
E2E	51,426	6,039	5,963	2,878	2,818	18.70	8.1	5.43	22.41	14.36	1.56
ViGGO$_{inf}$	3,000	1,000	997	1,378	1,102	8.33	3	5.81	30.62	15.01	2.04
ViGGO	6,900	2,253	2,066	2,427	2,178	6.91	3	4.18	25.01	15.04	1.66

Table 3: Dataset statistics comparing the ViGGO dataset, as well as its subset of *inform* DAs only (ViGGO$_{inf}$), with the E2E dataset. The average trigram frequency was calculated on trigrams that appear more than once.

	BLEU	METEOR	ROUGE	CIDEr	SER
Ao3	0.519	0.388	0.631	2.531	*2.55%*
Bo3	0.521	0.391	0.638	2.545	*2.48%*

Table 4: Baseline system performance on the ViGGO test set. Despite individual models (Bo3 – best of 3 experiments) often having better overall scores, we consider the Ao3 (average of 3) results the most objective.

	Naturalness		Coherence	
	Ref.	Gen. utt.	Ref.	Gen. utt.
E2E	4.48	4.67	4.57	4.77
ViGGO$_{inf}$	4.85	4.83	4.85	4.93
ViGGO	4.68	4.74	4.78	4.84

Table 5: Naturalness and coherence scores of our model's generated outputs compared to the reference utterances, as per the human evaluation. ViGGO$_{inf}$ corresponds to the subset of *inform* DAs only.

3 Baseline System Evaluation

The NLG model we use to establish a baseline for this dataset is a standard Transformer-based (Vaswani et al., 2017) sequence-to-sequence model. For decoding we employ beam search of width 10 ($\alpha = 1.0$). The generated candidates are then reranked according to the heuristically determined slot coverage score. Before training the model on the ViGGO dataset, we confirmed on the E2E dataset that it performed on par with, or even slightly better than, the strong baseline models from the E2E NLG Challenge[4], namely, TGEN (Dušek and Jurčíček, 2016) and SLUG2SLUG (Juraska et al., 2018).

Automatic Metrics We evaluate our model's performance on the ViGGO dataset using the following standard NLG metrics: BLEU (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007), ROUGE-L (Lin, 2004), and CIDEr (Vedantam et al., 2015). Additionally, with our heuristic slot error rate (SER) metric we approximate the percentage of failed slot realizations (i.e., missed, incorrect, or hallucinated) across the test set. The results are shown in Table 4.

Human Evaluation We let two expert annotators with no prior knowledge of the ViGGO dataset evaluate the outputs of our model. Their task was to rate 240 shuffled utterances (120 generated utterances and 120 human references) each on *naturalness* and *coherence* using a 5-point Likert scale. We define naturalness as a measure of how much one would expect to encounter an utterance in a conversation with a human, as opposed to sounding robotic, while coherence measures its grammaticality and fluency. Out of the 120 MRs in each partition, 40 were of the *inform* type, with the other 8 DAs represented by 10 samples each. In addition to that, we had the annotators rate a sample of 80 utterances from the E2E dataset (40 generated and 40 references) as a sort of a baseline for the human evaluation.

With both datasets, our model's outputs were highly rated on both naturalness and coherence (see Table 5). The scores for the ViGGO utterances were overall higher than those for the E2E ones, which we understand as an indication of the video game data being more fluent and conversational. At the same time, we observed that the utterances generated by our model tended to score higher than the reference utterances, though significantly more so for the E2E dataset. This is likely a consequence of the ViGGO dataset being cleaner and less noisy than the E2E dataset.

In an additional evaluation of ViGGO, we asked the annotators to classify the utterance samples into the 9 DA groups. For this task they were provided with a brief description of each DA type. The annotators identified the DA incorrectly in only 7% of the samples, which we interpret as a confirmation that our DAs are well-defined. Most of the mistakes can be ascribed to the inherent similarity of the *recommend* and the *suggest* DA, as

[4] http://www.macs.hw.ac.uk/InteractionLab/E2E/

well as to our model often generating *give_opinion* utterances that resemble the *inform* ones.

Qualitative Analysis Among all 9 DAs, the one posing the greatest challenge for our model was *give_opinion*, due to its high diversity of reference utterances. Despite the occasional incoherence, it learned to produce rich and sensible utterances, for instance "Little Nightmares is a pretty good game. Tarsier Studios is a talented developer and the side view perspective makes it easy to play.".

Since our baseline model does not implement any form of a copy mechanism, it fails on instances with out-of-vocabulary terms, such as the values of the SPECIFIER slot in the test set. These, in fact, account for almost half of the errors indicated by the SER metric in Table 4. Therefore, more robust models have good potential for improving on our scores.

4 Discussion

In Table 6 we demonstrate how the 9 DAs of the ViGGO dataset can support a natural multi-turn exchange on the topic of video games, as a part of a longer casual conversation on different topics. One caveat of using a language generator trained on this dataset in a dialogue system as-is is that multiple subsequent turns discussing the same video game would be repeating its full name. ViGGO was designed for grounded generation but without context, and therefore it is up to the dialogue manager to ensure that pronouns are substituted for the names whenever it would sound more natural in a dialogue. Alternately, the dataset can easily be augmented with automatically constructed samples which omit the NAME slot in the MR and replace the name with a pronoun in the reference utterance.

5 Conclusion

In this paper we presented a new parallel corpus for data-to-text NLG, which contains 9 dialogue acts, making it more conversational than other similar datasets. The crowdsourced utterances were thoroughly cleaned in order to obtain high-quality human references, which we hope will support the recent trend in research to train neural models on small but high-quality data, like humans can. This could possibly be achieved by transferring fundamental knowledge from larger available corpora, such as the E2E dataset, but perhaps by other, completely new, methods.

S:	. . . What about video games?
U:	Totally, I spend my entire evenings in front of my PS4.
S:	What is your favorite **genre** of video games? *request_attribute*(GENRES [])
U:	I guess I play a lot of shooter games.
S:	Have you played any **interesting** games lately? *request*(SPECIFIER [**interesting**])
U:	Tomb Raider is pretty good.
S:	Do you mean **Tomb Raider** from **2013**? *confirm*(NAME [**Tomb Raider**], RELEASE_YEAR [**2013**])
U:	Yes, that one with the epic E3 trailer.
S:	Do you also enjoy playing **first-person shooters**, such as **Half-Life 2**? *suggest*(NAME [**Half-Life 2**], GENRES [**shooter**], PLAYER_PERSPECTIVE [**first person**])
U:	I haven't played that one in a while. Would you recommend something else?
S:	Well, since you seem to be a fan of **M-rated PlayStation** games, I thought you would perhaps like **Max Payne 3**. Have you heard about that game? *recommend*(NAME [**Max Payne 3**], ESRB [**M (for Mature)**], PLATFORMS [**PlayStation**])
U:	No.
S:	**Max Payne 3** is a **multiplayer action-adventure shooter** from **2012**. *inform*(NAME [**Max Payne 3**], RELEASE_YEAR [**2012**], GENRES [**action-adventure; shooter**], MULTIPLAYER [**yes**])
U:	I see. So, you enjoyed it?
S:	Oh, I **love Max Payne 3**. When I see a game developed by **Rockstar Games**, I buy it without thinking. *give_opinion*(NAME [**Max Payne**], RATING [**excellent**], DEVELOPER [**Rockstar Games**])
U:	That sounds cool.
S:	Previously you told me that you **enjoyed Tomb Raider**. Is that the case with all games developed by **Crystal Dynamics**? *verify_attribute*(NAME [**Tomb Raider**], RATING [**good**], DEVELOPER [**Crystal Dynamics**])
U:	You bet.
S:	What is it that you **like** about **Crystal Dynamics** games? *request_explanation*(RATING [**good**], DEVELOPER [**Crystal Dynamics**])
U:	They usually have really good stories.
S:	. . .

Table 6: An example of a chit-chat about video games comprising utterances of DAs defined in ViGGO. "S" denotes the system and "U" the user turns.

References

Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. 2011. Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on psychological science*, 6(1):3–5.

Jan Milan Deriu and Mark Cieliebak. 2018. Syntactic manipulation for generating more diverse and interesting texts. In *11th International Conference on Natural Language Generation (INLG 2018), Tilburg, The Netherlands, 05-08 November 2018*, pages 22–34. Association for Computational Linguistics.

Ondřej Dušek and Filip Jurčíček. 2016. Sequence-to-sequence generation for spoken dialogue via deep syntax trees and strings.

Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2019. Evaluating the state-of-the-art of end-to-end natural language generation: The e2e nlg challenge. *CoRR*.

Maryam Fazel-Zarandi, Shang-Wen Li, Jin Cao, Jared Casale, Peter Henderson, David Whitney, and Alborz Geramifard. 2017. Learning robust dialog policies in noisy environments. *NIPS 2017 Workshop on Conversational AI*.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. Creating training corpora for micro-planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada. Association for Computational Linguistics.

Milica Gašić, Simon Keizer, Francois Mairesse, Jost Schatzmann, Blaise Thomson, Kai Yu, and Steve Young. 2008. Training and evaluation of the HIS POMDP dialogue system in noise. In *Proceedings of the 9th SIGDIAL Workshop on Discourse and Dialogue*, pages 112–119. Association for Computational Linguistics.

David Howcroft, Crystal Nakatsu, and Michael White. 2013. Enhancing the expression of contrast in the SPaRKy restaurant corpus. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 30–39.

Juraj Juraska, Panagiotis Karagiannis, Kevin Bowden, and Marilyn Walker. 2018. A deep ensemble model with slot alignment for sequence-to-sequence natural language generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 152–162.

Juraj Juraska and Marilyn Walker. 2018. Characterizing variation in crowd-sourced data for training neural language generators to produce stylistically varied outputs. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 441–450.

Chandra Khatri, Behnam Hedayatnia, Anu Venkatesh, Jeff Nunn, Yi Pan, Qing Liu, Han Song, Anna Gottardi, Sanjeev Kwatra, Sanju Pancholi, et al. 2018. Advancing the state of the art in open domain dialog systems through the alexa prize. In *2018 Alexa Prize Proceedings*.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231. Association for Computational Linguistics.

Xiujun Li, Yun-Nung Chen, Lihong Li, Jianfeng Gao, and Asli Celikyilmaz. 2017. End-to-end task-completion neural dialogue systems. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 733–743.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

François Mairesse, Milica Gašić, Filip Jurčíček, Simon Keizer, Blaise Thomson, Kai Yu, and Steve Young. 2010. Phrase-based statistical language generation using graphical models and active learning. In *ACL*.

Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M Khapra. 2018. Towards exploiting background knowledge for building conversation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2322–2332.

Neha Nayak, Dilek Hakkani-Tür, Marilyn Walker, and Larry Heck. 2017. To plan or not to plan? discourse planning in slot-value informed sequence to sequence models for language generation. In *INTERSPEECH*.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2017. The E2E NLG shared task.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Pararth Shah, Dilek Hakkani-Tür, Gokhan Tür, Abhinav Rastogi, Ankur Bapna, Neha Nayak, and Larry Heck. 2018. Building a conversational agent overnight with dialogue self-play. *CoRR*.

Amanda Stent, Rashmi Prasad, and Marilyn Walker. 2004. Trainable sentence planning for complex information presentation in spoken dialog systems. In *Proceedings of the 42nd annual meeting on association for computational linguistics*, page 79. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International*

Conference on Neural Information Processing Systems, pages 6000–6010. Curran Associates Inc.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Tsung-Hsien Wen, Milica Gašić, Dongho Kim, Nikola Mrkšić, Pei hao Su, David Vandyke, and Steve Young. 2015a. Stochastic language generation in dialogue using recurrent neural networks with convolutional sentence reranking. In *SIGDIAL Conference*.

Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Lina M. Rojas-Barahona, Pei-Hao Su, David Vandyke, and Steve Young. 2016. Multi-domain neural network language generation for spoken dialogue systems. In *NAACL*.

Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015b. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *EMNLP*.

A Appendix

A.1 Additional ViGGO Dataset Examples

In Table 7 we present one example of each DA in the ViGGO dataset, including the examples given in Table 1.

inform(NAME [**God of War**], RELEASE_YEAR [**2018**], DEVELOPER [**SIE Santa Monica Studio**], RATING [**excellent**], GENRES [**action-adventure, platformer, role-playing**], PLAYER_PERSPECTIVE [**third person**], HAS_MULTIPLAYER [**no**], PLATFORMS [**PlayStation**])
Developed by **SIE Santa Monica Studio** in **2018**, **God of War** is an **excellent single-player third person platformer** made exclusively for **PlayStation**. The **action-adventure** storyline involves **role-playing** as one of the dynamic characters.
confirm(NAME [**Hellblade: Senua's Sacrifice**], RELEASE_YEAR [**2017**], DEVELOPER [**Ninja Theory**])
Oh, do you mean the **2017** game from **Ninja Theory**, **Hellblade: Senua's Sacrifice**?
give_opinion(NAME [**SpellForce 3**], RATING [**poor**], GENRES [**real-time strategy, role-playing**], PLAYER_PERSPECTIVE [**bird view**])
I think that **SpellForce 3** is **one of the worst games** I've ever played. Trying to combine the **real-time strategy** and **role-playing** genres just doesn't work, and the **bird's eye view** makes it near impossible to play.
recommend(NAME [**Call of Duty: Advanced Warfare**], DEVELOPER [**Sledgehammer Games**], ESRB [**M (for Mature)**])
Speaking of **M rated** games developed by **Sledgehammer Games**, have you tried **Call of Duty: Advanced Warfare**?
request(DEVELOPER [**Guerrilla Games**], SPECIFIER [**overrated**])
What would you say is the most **overrated** game made by **Guerrilla Games**?
request_attribute(AVAILABLE_ON_STEAM [])
Do you prefer playing games that you can get on **Steam**?
request_explanation(RATING [**poor**], **HAS_MAC_RELEASE** [yes])
What is it about **Mac** games that you find **so disappointing**?
suggest(NAME [**Rocket League**], GENRES [**sport, vehicular combat**], PLAYER_PERSPECTIVE [**third person**])
Are you into **third person sport** games with **vehicular combat** like **Rocket League**?
verify_attribute(NAME [**Little Big Adventure**], RATING [**average**], HAS_MULTIPLAYER [**no**], PLATFORMS [**PlayStation**])
I recall that you were **not that fond** of **Little Big Adventure**. Does **single-player** gaming on the **PlayStation** quickly get boring for you?

Table 7: Examples of MRs and corresponding reference utterances in the ViGGO dataset. The DA of the MRs is indicated in italics, and the slots in small caps. The slot mentions in the utterances are bolded.

A.2 Slot Categories

In Section 2.1 we mentioned that the slots in the ViGGO dataset can be classified into 5 general categories. Here we provide more detailed descriptions of the categories:

1. *Boolean* – binary value, such as "yes"/"no" or "true"/"false" (e.g., HAS_MULTIPLAYER or AVAILABLE_ON_STEAM),
2. *Numeric* – value is a number or contains number(s) as the salient part (e.g., RELEASE_YEAR or EXP_RELEASE_DATE),
3. *Scalar* – values are on a distinct scale (e.g., RATING or ESRB),
4. *Categorical* – takes on virtually any value, typically coming from a certain category, such as names or types (e.g., NAME or DEVELOPER),
5. *List* – similar to categorical, where the value can, however, consist of multiple individual items (e.g., GENRES or PLAYER_PERSPECTIVE).

Note that in ViGGO the items in the value of a *List* slot are comma-separated, and therefore the individual items must not contain a comma. There are no restrictions as to whether the values are single-word or multi-word in any of the categories.

A.3 Data Collection

When generating the MRs for the *inform* DA, we fixed the slot ratios: the NAME and GENRES slots were mandatory in every MR, the PLAYER_PERSPECTIVE and RELEASE_YEAR were enforced in about half of the MRs, while the remaining slots are present in about 25% of the MRs. At the same time we imposed two constraints on the slot combinations: (1) whenever one of the Steam, Linux or Mac related boolean slots is present in an MR, the PLATFORMS slot must be included too, and (2) whenever either of the Linux or Mac slots was picked for an MR, the other one was automatically added too. These two constraints were introduced so as to encourage reference utterances with natural aggregations and contrast relations.

The remaining 8 DAs, however, contain significantly fewer slots each (see Table 2). We therefore decided to have the MTurk workers select 5 unique slot combinations for each given video game before writing the corresponding utterances. Since for these DAs we collected less data, we tried to ensure in this way that we have a sufficient number of samples for those slot combinations that are most natural to be mentioned in each of the DAs. While fixing mandatory slots for each DA, we instructed the workers to choose 1 or 2 additional slots depending on the task. The data collection for MRs with only 1 additional slot and for those with 2 was performed separately, so as to prevent workers from taking the easy way out by always selecting just a single slot, given the option.

Leaving the slot selection to crowdworkers yields a frequency distribution of all slot combinations, which presumably indicates the suitability of different slots to be mentioned together in a sentence. This meta-information can be made use of in a system's dialogue manager to sample from the observed slot combination distributions instead of sampling randomly or hard-coding the combinations. Figure 3 shows the distributions of the 8 slot pairs most commonly mentioned together in different DAs. These account for 53% of the selections among the 6 DAs that can take 2 additional

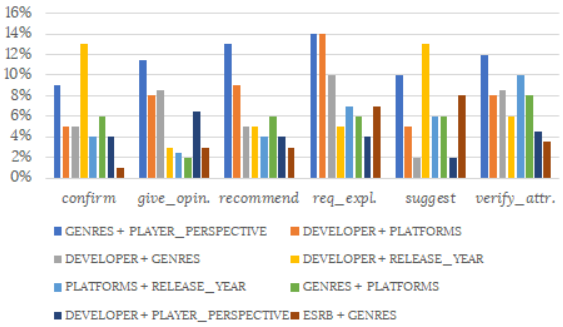

Figure 3: Distribution of the 8 most frequently selected slot combinations across different DAs.

slots besides the mandatory ones. We can observe some interesting trends in the distributions, such as that the DEVELOPER + RELEASE_YEAR combination was the most frequent one in the *confirm* DA, while fairly rare in most of the other DAs. This might be because this pair of a game's attributes is arguably the next best identifier of a game after its name.

A.4 Dataset Cleaning

A large proportion of the raw data collected contained typos and various errors, as is inevitable when crowdsourcing. We took the following three steps to clean the data.

First, we used regular expressions to enforce several standardization policies regarding special characters, punctuation, and the correction of undesired abbreviations/misspellings of standard domain-specific terms (e.g., we would change terms like "Play station" or "PS4" to the uniform "PlayStation"). At the same time, we removed or enforced hyphens uniformly in certain terms, for example, "single-player". Although phrases such as "first person" should correctly have a hyphen when used as adjective, the turkers used this rule very inconsistently. In order to avoid model outputs being penalized during the evaluation by the arbitrary choice of a hyphen presence or absence in the reference utterances, we decided to remove the hyphen in all such phrases regardless of the noun/adjective use.

Second, we developed an extensive set of heuristics to identify slot-related errors. This process revealed the vast majority of missing or incorrect slot mentions, which we subsequently fixed according to the corresponding MRs. Turkers would sometimes also inject a piece of information which was not present in the MR, some of

which is not even represented by any of the slots, e.g., plot or main characters. We remove this extraneous information from the utterances so as to avoid confusing the neural model. This step thus involved certain manual work and was thus performed jointly with the third step.

Finally, we further resolved the remaining typos, grammatical errors, and unsolicited information.

A.5 Model Parameters

Even though on the small datasets we work with we do not necessarily expect the Transformer model to perform better than recurrent neural networks, we chose this model for its significantly faster training, without sacrificing the performance. For our experiments a small 2-layer Transformer with 8 heads proved to be sufficient. The input tokens are encoded into embeddings of size 256, and the target sequences were truncated to 60 tokens. The model performed best with dropout values of 0.2. For training of the Transformer models we used the Adam optimizer with a custom learning rate schedule including a brief linear warm-up and a cosine decay.