Focus on FoCus: Is FoCus focused on Context, Knowledge and Persona?

SeungYoon Lee[1†], **Jungseob Lee**[2†], **Chanjun Park**[2,3], **Sugyeong Eo**[2],
Hyeonseok Moon[2], **Jaehyung Seo**[2], **Jeongbae Park**[4*], **Heuiseok Lim**[2,4*]
[1]Chung-Ang University, [2]Korea University, [3]Upstage,
[4]Human Inspired Artificial Intelligence Research (HIAI)
dltmddbs100@cau.ac.kr, chanjun.park@upstage.ai
{omanma1928,bcj1210,djtnrud,glee889,seojae777,insmile,limhseok}@korea.ac.kr

Abstract

Rather than continuing the conversation based on personalized or implicit information, the existing conversation system generates dialogue by focusing only on the superficial content. To solve this problem, FoCus was recently released (Jang et al., 2022). FoCus is a persona-knowledge grounded dialogue generation dataset that leverages Wikipedia's knowledge and personal persona, focusing on the landmarks provided by Google, enabling user-centered conversation. However, a closer empirical study is needed since research in the field is still in its early stages. Therefore, we fling two research questions about FoCus. (i) *"Is the FoCus whether for conversation or question answering?"* to identify the structural problems of the dataset. (ii) *"Does the FoCus model do real knowledge blending?"* to closely demonstrate that the model acquires actual knowledge. As a result of the experiment, we present that the FoCus model could not correctly blend the knowledge according to the input dialogue and that the dataset design is unsuitable for the multi-turn conversation.

1 Introduction

Recent studies have widely considered knowledge grounded dialogue, user interest, and preference (Dinan et al., 2018; Zhou et al., 2018a; Zhang et al., 2019; Zhao et al., 2020; Zheng et al., 2020; Meng et al., 2020; Song et al., 2021; Majumder et al., 2021; Galetzka et al., 2021). Especially, Zhang et al. (2018) presented persona-chat dataset based on personalizing dialogue agents. Similarly, Rashkin et al. (2018) constructed conversation dataset with emotional labels according to the given situation. In different way, Dinan et al. (2018) and Zhou et al. (2018b) focused on generating dialogue based on knowledge retrieved from Wikipedia.

However, those existing dialogue datasets do not comprehensively consider the user persona of the given situation and knowledge of the grounded object. The biased dataset toward persona or knowledge quickly undermines the user's intention or purpose, and rendering high-quality answers is challenging. From this point of view, dialogue generation through the proper blending of persona and knowledge is a significant issue that should be considered in advanced research. To fulfill this purpose, Jang et al. (2022) has released FoCus and baseline models based on the grounded persona and knowledge.

To the best of our knowledge, FoCus is the first persona-knowledge grounded dialogue dataset that incorporates knowledge and customized persona. However, related research in the persona-knowledge is still insufficient, and Jang et al. (2022) only provides descriptions of the dataset and has not corroborated the weakness of the dataset and model architecture. Furthermore, no in-depth analysis has been conducted on whether the model engages in conversation based on the blended persona-knowledge.

To demonstrate these problems, we intend to conduct an in-depth analysis of the proposed model and dataset through an empirical study. In this paper, we execute probing tests by throwing research questions in terms of data-centric (Park et al., 2021; Seo et al., 2022) and model-centric (Park et al., 2020) in the Jang et al. (2022). First, we point out that FoCus is more of a question-answering than a multi-turn dialogue task, under the question *"Is the FoCus whether for conversation or question answering?"* Generally, a multi-turn dialogue dataset forms a context in which two or more speakers continue to conversation. Moreover, the memorable context significantly impacts the generation of the next utterance.

However, we experimentally found that the previous conversation on the FoCus had little effect on

[*] Corresponding Authors

Proceedings of the 29th International Conference on Computational Linguistic, pages 1–8
October 12–17, 2022.

the next utterance. This is because each round consists of an independent set that is not involved in the context. Therefore, we demonstrate that the FoCus is more of a question-answering task rather than a consistent multi-turn conversation. To closely analyze this point, we conduct a case study according to the change in the conversation order and the inclusion of dialogue histories corresponding to the previous utterance.

Second, we ask, *"Does the FoCus model do real knowledge blending?"* and examine whether the baseline model presented in Jang et al. (2022) is blending properly with persona in selecting the appropriate knowledge for conversation. In this experiment, we proceed with various analyses using BM25 (Robertson et al., 1995), DPR (Karpukhin et al., 2020), STS (Reimers and Gurevych, 2019), and TF-IDF (Salton and Buckley, 1988) as a retrieval module. We analyze performance changes according to the knowledge selection of each search module. Based on this, we attempt to probe what problems the method of blending history and selected knowledge in an in-context approach causes in knowledge grounding.

2 FoCus

2.1 Dataset

FoCus presented by Jang et al. (2022) is a multi-turn dialogue dataset constructed on the landmark content provided by Google Landmarks Dataset v2 (GLDv2) (Weyand et al., 2020), enabling personalized conversation with relevant knowledge and various user personas. The purpose of this dataset is to take user utterances and generate responses leveraging landmark knowledge and an appropriate persona. FoCus comprises human-to-machine conversations, user persona, Wikipedia knowledge, and knowledge candidates. We report the detailed statistics and examples of the dataset in Appendix A.

2.2 Model

The baseline model in the Jang et al. (2022) consists of two steps: a retrieval module and a dialogue module. First, the TF-IDF score-based retrieving algorithm receives the user's utterance and chooses the top-5 knowledge to be transferred to the dialogue module. Next, the dialogue module takes input from the user's persona, selected knowledge and utterances through a learnable Transformer (Vaswani et al., 2017) and pre-trained language models such as GPT-2 (Radford et al., 2019), BART (Lewis et al., 2019). In this process, a subtask called knowledge grounding (KG) and persona grounding (PG) is performed. Based on the selected knowledge and utterance, KG determines the knowledge answer that matches the user utterance among the 10 knowledge candidates presented for each round. Instead of retrieving, PG adopts the persona answer that is consistent with the user utterance among the five persona candidates. The model receives the selected persona and knowledge vector, creating an utterance by blending them. An overview of the model is depicted in Appendix B.

3 Experiments

3.1 Experimental Design

We utilize the same train and validation set with FoCus (Jang et al., 2022) for objective verification. Experiments are implemented based on the baseline models released by the original research[1], the BART-base is adopted for our experiments. All hyper-parameters, including seeds, are run under the same settings, except for exceptional cases marked separately. We fine-tune the model using a single RTX-8000 GPU.

Research Question 01 In order to prove that FoCus is similar to QA composed of independent question-answering pair, we randomly shuffle the order of each round composed of user and model utterances, and then compare the generation score with the original order.

In this setting, the input utterance is mixed for each pair and cuts off the contextual flow according to the round. In addition, this data setting increases randomness because more history is considered during training as the history size increases, which means how far the model can consider past utterances. In accessing the model performance, we adopt the chrF++ (Popović, 2015) score, SacreBLEU (Post, 2018), and ROUGE (Lin, 2004) score for evaluation metrics and compare the average values. In this case, we also consider the history of the conversation during the evaluation.

Research Question 02 We analyze the role of retrieval module in knowledge blending. In the existing baseline, Wikipedia knowledge is included in the conversation through TF-IDF. We additionally use BM25, DPR, and STS to select knowledge and measure grounding performance to check whether

[1] https://github.com/pkchat-focus/FoCus

Models	Generation					Average
	chrF++	BLEU	R-1	R-2	R-L	
BART + history = 2	0.2941	11.61	36.84	19.87	32.43	20.21
BART + history = 3	0.2983	12.01	37.19	20.31	32.78	20.52
BART + history = 4	0.2988	12.04	37.37	20.41	32.97	20.62
BART + shuffle + history = 2	0.2991	11.93	37.15	19.98	32.72	20.42
BART + shuffle + history = 3	0.2950	11.96	36.94	19.98	32.63	20.36
BART + shuffle + history = 4	0.2982	11.87	37.12	20.06	32.56	20.38

Table 1: Generation score of validation set under the same experimental environment as FoCus. History refers to how many past conversations are included in model training and evaluation. We randomly shuffle rounds of dialogue and compare them to their original order.

Retrieval	Grounding (Acc.)	
	Persona	Knowledge
TF-IDF	67.43	70.1
BM25	67.43	70.1
DPR	67.43	70.1
STS	67.43	70.1

Table 2: Knowledge and persona grounding performances from four different retrieval modules.

the selected knowledge is normally reflected according to different types of retrieval. We use accuracy as an evaluation metric for KG and PG tasks.

3.2 Is the FoCus whether for conversation or question answering? (Data-Centric)

The experimental results are shown in Table 1. If the dataset has the contextual multi-turn for the utterances flow, the order of previous utterances provides essential information for contextual understanding. However, when comparing the results of randomly shuffling the dialogue turns with the baseline results, there is no significant difference in the model's performance even if the turn of dialogue order is arbitrarily mixed.

As the history size increases, the average of the generation score increases from 20.21 to 20.62, as shown top of Table 1. However, when random shuffling is applied, the largest difference is insignificant at 0.24 (bottom of Table 1) compared to the same history size as random shuffling is not applied. Even when the history size is 2, which is the case of learning only the previous conversation, the score of the shuffled case is higher by 0.21. In general, a long input size leads to an increase in noise as well. Considering this, since we compare under the same conditions within the same history size, the difference in performance cannot be attributed to noise.

Generation scores in random order are similar to correct order, although FoCus has a multi-turn configuration. This result indicates that the dataset is more of a QA rather than a context-influenced conversation.

3.3 Does the FoCus model do real knowledge blending? (Model-Centric)

Table 2 shows the evaluation results on four different retrieval modules. First, as a result of quantitative analysis, even when four different modules are applied, PG accuracy is 67.43 and KG accuracy is 70.1, which is the same for all four modules. Second, we proceed with the qualitative evaluation results for knowledge extraction, and we are able to confirm that each module extracts different knowledge (The top-5 knowledge analysis results extracted by each module are described in Appendix C).

Combining and interpreting the two results, each module shows the same grounding accuracy despite selecting different knowledge. This is presumed to be caused by improper blending in the process of training the knowledge extracted by the model. The knowledge vector extracted from retrieval used in training is quite small compared to the size of persona and history vector to be concatenated. Therefore, it appears that there is relatively little effect on KG.

3.4 Additional Analysis

Ablation study for max-length As a result of comparing the generated sentences with the gold labels in the experimental process for the data-centric approach, we find that the max-length among the hyper-parameters during generation is presented too low. Since the generated sentence is forcibly

max-length	Generation				
	chrF++	BLEU	R-1	R-2	R-L
20 (baseline)	0.2821	10.74	34.84	18.55	30.6
50	0.3259	13.48	37.96	20.66	33.11
75	0.3259	**14.23**	**38.69**	**21.35**	**33.79**
100	**0.3322**	13.98	38.26	20.94	33.41

Table 3: Generation scores according to the extension of max-length.

Models	Generation					Average
	chrF++	BLEU	R-1	R-2	R-L	
BART + history = 2	0.3439	14.44	39.31	21.53	34.11	21.95
BART + history = 3	0.3565	13.86	39.3	21.26	33.49	21.65
BART + history = 4	0.3553	15.47	40.5	22.74	35.3	22.87
BART + history = 6	0.3563	15.17	40.32	22.34	34.9	22.62
BART + shuffle + history = 2	0.3604	14.73	39.39	21.67	33.79	21.99
BART + shuffle + history = 3	0.3455	15.24	39.85	22.36	35	22.56
BART + shuffle + history = 4	0.357	14.33	39.25	21.74	33.62	21.86
BART + shuffle + history = 6	0.3419	14.69	39.44	21.94	34.43	22.17

Table 4: Distribution of generation scores when history size is increased to 6 and max-length to 75. We adjust the history size and max-length to re-run the evaluation. All the other experimental settings are the same.

cut in the middle, it cannot contain as much contextual information as the expressive power of the model. This leads to negative effect on the point we attempt to experiment with, so we compare the parameters by giving them various values in a wider range.

We manually set the max-length to 50, 75, and 100, respectively, and re-evaluate. The experimental results are shown in Table 3. According to the Table 3, the adjustment of the max-length leads to a large improvement in the generation score. In particular, when max-length is set to 75, all scores except chrF++ are the highest. In other words, the existing max-length limits the performance of the model, which means that the model doesn't sufficiently capture the context of the conversation.

Expanding history size with max-length As the history size increases, which indicates the range of consideration of past utterances, the randomness also increases when the round is shuffled, making it more difficult to preserve the context information. Therefore, if the conversation has an element of connectivity, it can cause a significant drop in performance. To observe this more closely, we adjust the max-length to 75 under the same settings as in the previous experiments and conduct a case study by adding the case where the history size is 6. The results are shown in Table 4.

Similarly, the difference in average generation score between the two conditions is not significant in Table 4. Even if the history size is 2 or 3, the randomly mixed case has a higher score. In particular, in the case of shuffled history = 3, some scores are equal to or higher than that of unmixed history = 6. In addition, although the history size is increased to 6, there is a little performance difference between the shuffled case and the non-shuffled case.

This is because each round of the dataset is composed of an independent QA style, so even if the order of information is randomly reversed, it can be interpreted that the model concentrates only on the utterance of the user corresponding to the present and historical information is rarely used. This suggests that FoCus is difficult to be seen as a multi-turn dialogue dataset, and that more contextual information should be considered to construct close to practical dialogue.

4 Conclusion

In this work, we conduct an in-depth analysis of FoCus, which aims to blend knowledge and persona. Experimental results quantitatively demonstrate that the proposed content as a multi-turn dialogue is close to QA and that the model does not appropriately incorporate knowledge to persona. In the future, we plan to properly combine knowledge and persona based on the limitations we presented.

Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2021R1A6A1A03045425) and MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2022-2018-0-01405) supervised by the IITP(Institute for Information & Communications Technology Planning Evaluation).

References

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

Fabian Galetzka, Jewgeni Rose, David Schlangen, and Jens Lehmann. 2021. Space efficient context encoding for non-task-oriented dialogue generation with graph attention transformer. 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 7028–7041.

Yoonna Jang, Jungwoo Lim, Yuna Hur, Dongsuk Oh, Suhyune Son, Yeonsoo Lee, Donghoon Shin, Seungryong Kim, and Heuiseok Lim. 2022. Call for customized conversation: Customized conversation grounding persona and knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10803–10812.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Bodhisattwa Prasad Majumder, Taylor Berg-Kirkpatrick, Julian McAuley, and Harsh Jhamtani. 2021. Unsupervised enrichment of persona-grounded dialog with background stories. *arXiv preprint arXiv:2106.08364*.

Chuan Meng, Pengjie Ren, Zhumin Chen, Weiwei Sun, Zhaochun Ren, Zhaopeng Tu, and Maarten de Rijke. 2020. Dukenet: A dual knowledge interaction network for knowledge-grounded conversation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1151–1160.

Chanjun Park, Jaehyung Seo, Seolhwa Lee, Chanhee Lee, Hyeonseok Moon, Sugyeong Eo, and Heui-Seok Lim. 2021. Bts: Back transcription for speech-to-text post-processor using text-to-speech-to-text. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 106–116.

Chanjun Park, Yeongwook Yang, Kinam Park, and Heuiseok Lim. 2020. Decoding strategies for improving low-resource machine translation. *Electronics*, 9(10):1562.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.

Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at trec-3. *Nist Special Publication Sp*, 109:109.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523.

Jaehyung Seo, Seounghoon Lee, Chanjun Park, Yoonna Jang, Hyeonseok Moon, Sugyeong Eo, Seonmin Koo, and Heui-Seok Lim. 2022. A dog is passing over the jet? a text-generation dataset for korean commonsense reasoning and evaluation. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2233–2249.

Haoyu Song, Yan Wang, Kaiyan Zhang, Wei-Nan Zhang, and Ting Liu. 2021. Bob: Bert over bert for training persona-based dialogue models from limited personalized data. *arXiv preprint arXiv:2106.06169*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. 2020. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2575–2584.

Houyu Zhang, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. 2019. Grounded conversation generation as guided traverses in commonsense knowledge graphs. *arXiv preprint arXiv:1911.02707*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020. Knowledge-grounded dialogue generation with pre-trained language models. *arXiv preprint arXiv:2010.08824*.

Chujie Zheng, Yunbo Cao, Daxin Jiang, and Minlie Huang. 2020. Difference-aware knowledge selection for knowledge-grounded conversation generation. *arXiv preprint arXiv:2009.09378*.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018a. Commonsense knowledge aware conversation generation with graph attention. In *IJCAI*, pages 4623–4629.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018b. A dataset for document grounded conversations. *arXiv preprint arXiv:1809.07358*.

A Statistics of FoCus

	Train	Validation
# Dialogues	11,562	1,445
# Average Rounds	6.00	6.00
Average. Len. Human's Utterance	40.94	40.89
Average. Len. Machine's Utterance	141.13	145.42
# Knowedge-Only Answer	35,580	4,501
# Persona-Knowledge Answer	33,792	4,169
# Landmarks	5,082	1,305

Table 5: Statistics of FoCus (Jang et al., 2022).

B Overview of baseline model architecture.

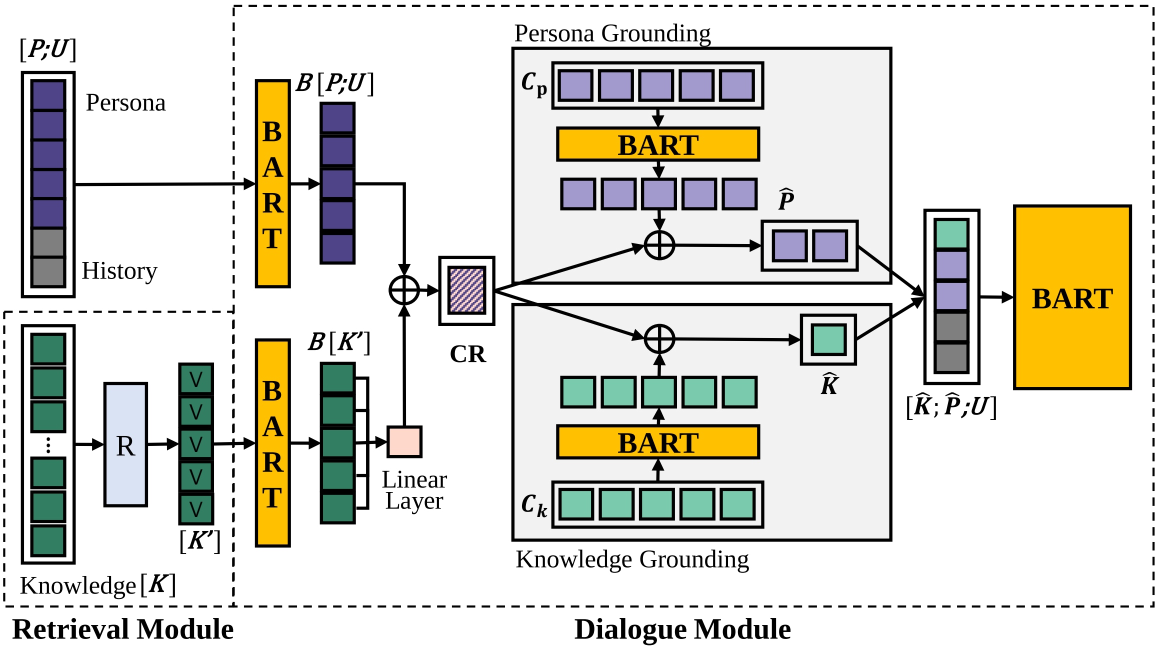

Figure 1: Architecture of proposed FoCus model using BART.

C Details of Knowledge Selection

This is an example of which knowledge is selected when the retrieval module is replaced with each of the TF-IDF, BM25, DPR, and STS modules. STS uses the `multi-qa-MiniLM-L6-dot-v1`[2] pre-trained model provided by Sentence-Transformer (Reimers and Gurevych, 2019) and uses the dot product for embedding of knowledge and user's utterance as a score.

User's Utterance: Where is this place?	
Knowledge	Model Selection
There were ten themed areas by the early 1980s. WaterWorld, ...	TF-IDF, STS
Six Flags purchased AstroWorld in 1975. The next year, Six Flags …	TF-IDF
WaterWorld opened in June 1983. The 10-acre 1.9 million-gallon water park …	TF-IDF
Peak attendance reached approximately 20,000 people on Saturdays. …	TF-IDF
The park had other seasonal attractions such as Alice Cooper's Brutal ...	TF-IDF
Roy Hofheinz acquired and developed 116 acres (47 ha) of land, …	BM25, DPR
While the original amusement park site was 57 acres, the Houston ...	BM25, STS
Six Flags AstroWorld, also known simply as AstroWorld, was a seasonally ..	BM25
AstroWorld was permanently closed by Six Flags after its final day of …	BM25
Thunder River was installed in 1980, has been described as the "first ...	BM25
An 8-foot (2.4) by 10-foot (3.0) 1967 model of Astroworld ...	DPR, STS
AstroWorld opened to the public with 50,000 guests visiting the first …	DPR
XLR-8 was installed in 1984. Looping Starship was installed in 1986. Ultra ...	DPR
Serial Thriller originally operated at AstroWorld starting in 1999. The ride …	DPR
In 2009, the former Astroworld site was still vacant. The land tract ...	STS
As of 2018, the HLSR owned the property at the former AstroWorld ...	STS
"Astrodomain" refers to an area of south Houston surrounding …	-
Hofheinz developed Astroworld just to the south of the Astrodome. …	-
During Astroworld's first twenty years, it entertained more than …	-
On September 12, 2005, Six Flags CEO Kieran Burke announced …	-
The final date of park operation was October 30, 2005. Following …	-
Other features included:	-
The Alpine Sleigh Ride, Astrowheel, and Mill Pond were among ...	-
Bamboo Shoot (formerly Ozarka Splash) was installed in 1969. Installed ...	-
The park's Southern Star Amphitheater opened in 1980 and hosted a …	-
Six Flags AstroWorld originated the "Fright Nights" special event …	-
Dan Dunn and Jeff Martin worked as a caricaturists at the park. Daniel ...	-
In 2018, former employees organized the AstroWorld 50th Anniversary ...	-

Table 6: Examples of Knowledge Selection in TF-IDF, BM25, DPR, and STS retrieval modules. The table is the result of the selected top-5 knowledge.

[2] https://huggingface.co/sentence-transformers/multi-qa-MiniLM-L6-dot-v1