Evaluating Humorous Response Generation to Playful Shopping Requests

Natalie Shapira*[1], Oren Kalinsky[2], Alex Libov[2], Chen Shani*[3], Sofia Tolmach[2]

[1] Bar-Ilan University `nd1234@gmail.com`
[2] Amazon Science `{orenk,alibov,sofiato}@amazon.com`
[3] The Hebrew University of Jerusalem `Chen.shani@mail.huji.ac.il`

Abstract. AI assistants are gradually becoming embedded in our lives, utilized for everyday tasks like shopping or music. In addition to the everyday utilization of AI assistants, many users engage them with playful shopping requests, gauging their ability to understand – or simply seeking amusement. However, these requests are often not being responded to in the same playful manner, causing dissatisfaction and even trust issues. In this work, we focus on equipping AI assistants with the ability to respond in a playful manner to irrational shopping requests. We first evaluate several neural generation models, which lead to unsuitable results – showing that this task is non-trivial. We devise a simple, yet effective, solution, that utilizes a knowledge graph to generate template-based responses grounded with commonsense. While the commonsense-aware solution is slightly less diverse than the generative models, it provides better responses to playful requests. This emphasizes the gap in commonsense exhibited by neural language models.

User: Buy me the moon!
Assistant:(Default) I'm sorry, you can't buy that.
(Desired) We only sell the moon when it is blue.

Fig. 1. Task illustrative example. The user playfully asks the AI assistant for a non-shoppable item. The goal is to provide a better response than the default.

1 Introduction

AI assistants such as Amazon's Alexa, Apple's Siri, Google Assistant, and Microsoft's Cortana are becoming increasingly popular. Users commonly expect the assistants to support a wide range of human capabilities, sometimes beyond their original intended tasks – such as carrying a conversation or responding to humor.

There has been a recent trend in the field of artificial intelligence toward addressing challenges related to social skills and commonsense [19, 8, 26, 21, 33, 6], including the challenging area of generative commonsense [10].

We focus here on the shopping scenario, where users playfully ask to purchase non-shoppable items (e.g., **"Buy me the moon"**). Recently, [22] analyzed Alexa

---

* Work was done during an internship at Amazon.

** Except for the first author, the rest of the authors follow the ABC of surnames.

traffic, showing this is a prevalent use-case. Moreover, they sketched a detection approach, leaving the problem of appropriate and scalable response generation as an open question. Responding to playful requests falls under the wider field of computational humor, considered an AI-complete problem [24].

Contemporary AI assistants rely on hand-curated responses (**"We only sell the moon when it is blue"**), which are hard to scale. Moreover, they provide a stopgap solution only for a handful of requests, missing many playful opportunities. For any detected playful request with no prepared response, a catch-all answer such as **"I'm sorry, you can't buy that"** can be applied. However, we strive to improve over this laconic, non-playful response.

To provide a scalable and suitable response mechanism, we first define the task of generating responses to playful shopping requests (see Section 3). We explore different methods for *automatically generating* satisfactory responses (Section 4). Motivated by neural-models' poor performance, we devise a simple knowledge graph and template based solution (Section 5). We hypothesize that incorporating commonsense will positively surprise users, as they do not expect it from a computer. Indeed, the commonsense-aware approach provides more suitable responses to playful requests. This is yet another example of the gap in commonsense that neural models exhibit, emphasizing its importance [20, 19, 21].

Our main contributions are: we 1) define a novel humor generation task of responding to playful shopping requests; 2) evaluate modern generative LMs on this task; and 3) devise a simple, yet effective, approach that leverages commonsense through templates and achieves state-of-the-art results.

We release our data to facilitate research in the challenging field of humor generation, and in particular on this novel task[4].

2 Related Work

In line with a recent review [1], we distinguish between two humor generation approaches, templates and neural networks.

Template-based systems often rely on external knowledge such as corpora or knowledge graphs [29, 15, 3, 25]. Their Achilles heel is low diversity, as they produce similar content with almost the same wording or repetitive humor mechanism.

Only a few humor-generation systems utilized neural-based solutions [12, 32, 18]. None of them tackled a question-answering scenario or involved a zero-shot setting with prompt engineering, as we present here. We note that although neural-based methods have achieved state-of-the-art results in many NLP tasks [5], generating *humorous* text is still in its infancy.

As for evaluation, a standardized methodology is still missing. The majority of studies used human judges to rate the generated texts on a 1-5 Likert scale. [28] proposed to measure the frequency of humorous outputs in the system's output (using a threshold on the numeric scores). [2] suggested a humor variation of the Turing test, in which an automatically generated pun is considered humorous if annotators are unable to tell it apart from human-generated puns.

---

[4] https://github.com/NatalieShapira/NonShoppableHumorGeneration.

We note that the evaluation effort focused on the quality of each generated output on its own, neglecting the systems' output diversity (ability to generate a variety of humorous texts). One notable example is the automated diversity, measured using the ratio of distinct uni- and bi-grams in the output [32]. More recently, [1] suggested measuring it in terms of syntactical and lexical features, as well as the joking mechanism.

Due to the highly subjective and individual nature of humor, previous research shows significant disagreement between different annotators [31, 15, 24, 4] - a finding that we reproduce here.

3 Problem Definition

We define a *non-shoppable item* as any object, entity, or concept that is impossible to purchase online via an AI assistant (e.g., love, brain, galaxy). We assume a detector exists and focus on *response generation*.

Given a *non-shoppable item*, we wish to automatically generate an *appropriate* response, such that: 1) the response expresses the assistant's understanding that the item is impossible to purchase, 2) it is sensible within the conversation's playful context, and 3) it has correct grammar. A *playful* response will also be humorous. The generic "I'm sorry, you can't buy that" is the *minimal* appropriate response, as it is very broad, laconic, and not playful (but satisfies all the requirements of an appropriate response).

While this definition refers to a single response, we extend the problem to also address diversity, which is the ability of the system to produce a variety of outputs for a given input.

4 Methods

In line with the division of [1], we explored two approaches towards automatically generating appropriate responses to non-shoppable requests: 1) off the shelf Generative Language Models (GLM), and 2) combining hand-crafted templates with commonsense knowledge graphs.

4.1 Generative Language Models (GLMs)

Large Language Models (LMs), neural-based, are widely used for text generation, where the input prompts affect their output drastically [11]. We employ T5-3B [17] and GPT-2 [16] in a zero-shot learning setup.

As input to these models, we curated general freestyle shopping-oriented response prompts (e.g., "It is impossible to buy <*non-shoppable-item*> because "). Additionally, we constructed commonsense-aware prompts based on the relations in ConceptNet, a commonsense knowledge graph [23]. This was done in an attempt to inject some commonsense-based wit into the neural LMs. For example, we constructed the prompt "You want to buy <non-shoppable-item>? Let me check if I have it next to ".[5]

---

[5] The full list is included in the code repository. T5 had 95 prompts, and GPT-2 had 89 (the prompts that were suffix-based are irrelevant to GPT-2 that attends to the prefix. Top-K=50, Top-P=0.95, Beam width=10, Max length GPT-2=50 T5-3B=20

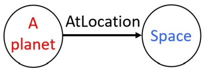
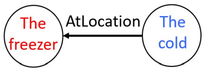
User: Assistant:	- I want to buy space. - Sorry, I'm all out. Maybe just a planet...?	- Do you sell the cold? - Just a second, I'm going to the freezer to get it for you.
Under the hood		
Case Template	AtLocation with arrow-start Sorry, I'm all out. Maybe just *<arrow-start>*...?	AtLocation with arrow-end Just a second, I'm going to *<arrow-end>* to get it for you.

Fig. 2. ConceptNet template example.

4.2 Knowledge-Graph Templates

We chose 30 relations from ConceptNet and manually created two templates – one for each edge direction. For example, consider the AtLocation relation: When the *non-shoppable-item* is in the start we produced *"Just a second, I'm going to <connected-item> to get it for you"*, while when the *non-shoppable-item* is at the end, the template is *"Sorry, I'm all out. Maybe just <connected-item>...?"* (See Figure 2). We also employed additional filtering steps removing concepts that are too similar to the non-shoppable item, or that are problematic (using age of acquisition score, readability score, and profanities filters).[6]

5 Evaluation

We first set out to create a dataset of non-shoppable items.[7] We used Amazon Mechanical Turk (MTurk) to generate a list of 100 non-shoppable items[8]. Specifically, we requested workers to be playful and write items that are impossible to buy via an AI assistant. All items were manually approved by our team. We denote this set as 100-NSI.

5.1 Response Sets

We used 100-NSI to automatically construct responses using the methods presented in Section 4.

1. **ConceptNet Templates**: For each 100-NSI item, we automatically generated all possible responses (4,198 in total). We then randomly chose 1,000 responses for further evaluation.
2. **T5**: For each 100-NSI item, we automatically generated up to 10 responses per prompt out of the 95 prompts (43,056 in total). We then randomly chose 10 random responses for each item, ending with 1,000 responses.
3. **GPT2**: For each 100-NSI item, we automatically generated up to 10 responses per prompt out of the 89 prompts (78,913 in total). We then randomly chose 10 random responses for each item, ending with 1,000 responses.
4. **Hand-Crafted**: Hand-curated responses created by a team of experts from a major commercial voice AI assistant. A total of 33 responses for 24 items.
5. **Random**: We constructed a baseline of arbitrary 2-20 words sentences that contain the non-shoppable items. Sentences extracted from C4 corpus [17].

---

[6] The full list of relations, templates, and filtering logic is included in the code repository

[7] The dataset of non-shoppable items and responses are included in the code repository

[8] Workers were paid 5 cents per generated non-shoppable item

Method	Better than Default	Diversity	Auto. Diversity (Unigram)	Auto. Diversity (Bigram)	Dispute
Hand-crafted	78.79%	3	57.83%	90.12%	60.61%
ConceptNet templates	**38.60%**	2	**16.47%**	37.34%	45.40%
T5-3B	15.30%	**2.5**	9.76%	27.59%	27.40%
GPT-2	9.20%	**2.5**	**16.47%**	**49.54%**	28.10%
Random	2.70%	3	32.17%	79.57%	14.40%

Table 1. Evaluation of the different response generation methods. The "Better than Default" column denotes the percentage of responses that were classified as better than the minimal appropriate response ("I'm sorry, you can't buy that"). Diversity classified according to [1]. Automated diversity is measured using the ratio of distinct unigrams and bigrams divided by the total number of words in the whole set, according to [32] . $N = 1000$ for all methods except for humans ($N = 33$). Results show that while the gap to human performance is still quite large, the commonsense-based templates outperform neural GLMs. However, it lacks diversity.

5.2 Annotation Task

We used MTurk to evaluate the quality of generated responses. The task placed the worker as a user initiating a non-shoppable request taken from 100-NSI (e.g., "I want to buy sleep") and the AI assistant's generated response (e.g., "Oh, you want to sleep? maybe you need to go to bed", generated by ConceptNet templates).

We then asked workers to evaluate the response (three workers per response). Due to the subjectivity of the task, we added an anchor – the rating was compared to the generic "I'm sorry, you can't buy that" response, which is the *minimal* appropriate response according to our definition (see Section 3).[9] Thus, we asked workers whether the responses they received is *a better experience than* "I'm sorry, you can't buy that" on the following Likert scale: really bad (-2), bad (-1), same experience (0), good (1), very good (2). The three ratings for each response were then averaged.

We note that this annotation is far from trivial due to: 1) its subjective nature, and 2) the difficulty in assessing generative output [27]. To account for these caveats we deployed several methods to raise the quality of annotation such as qualification tests and planted test questions with bad responses (See Section 6).

5.3 Results and Discussion

Results can be seen in Table 1. While all automatic generation methods are still far from the human baseline, they exceed the random baseline, proving they can generate appropriate responses. The template-based approach outperforms generative LM methods: 38.6% of ConceptNet template responses rated better than "I'm sorry, you can't buy that", utterly beating the generative models that reached only 15.3% and 9.2% (T5 and GPT-2, respectively). Interestingly, 14.4% of ConceptNet template responses were better scored than average human responses (compared to only 3.6% and 2.1% for T5 and GPT-2 respectively).

---

[9] Preliminary experiments showed that annotators tended to rank responses with a discourse issue as worse than the baseline response (-1/-2).

Diversity. To avoid distortion originated by a system having only a handful of responses, we also measure diversity. We use the measures proposed by [32] to compare the different methods' diversity (see Table 1). Results show that the GLM approach is more diverse than the templates one. Moreover, the gap to human-curated responses is still quite large.

A note about task's difficulty. We used a 5-point Likert scale. A response is classified as in dispute if at least one of the three annotators rated it as positive ($\{1, 2\}$) and one as negative ($\{-1, -2\}$). In line with previous works, we see high disagreement (24%). Interestingly, we found a highly significant positive correlation between the response's mean score (on the Likert scale) and disagreement ($R{=}0.48$, p-value$<$ 0.001). Meaning, bad responses are easy to agree upon, whereas good responses were more likely to be in dispute. While such an analysis was not done before, we hypothesize this is a general finding, since responses can often be objectively bad, while a good experience is often subjective and personal-taste dependent.

Table 2 presents the top and bottom-ranked samples from the user study. On one hand, there are cases where both ConceptNet templates and T5-3B generate great responses. Yet, some responses appear to be nonsense due to issues in the knowledge graph or poor performance of the models which were not tuned on this task. Finally, the table presents two cases of a strong dispute. It seems that in these cases, the AI assistant appears cynical or rude, and we expect this type of humor to be highly subjective.

To conclude, evaluating responses to playful shopping requests is not a trivial task. However, when treated carefully, it is possible to achieve meaningful and insightful results. Our experiment shows that unlike some other natural language generation tasks, in which Generative LMs can outperform any other approach, here they still fall short. While template-based approaches show improved results, they are still far from human performance. Moreover, they are slightly less diverse compared to LMs. Further analysis shows that it is easier for annotators to agree in the bad-response region compared to the good-response region. The proposed method can be applied to a specific problem and may be relevant to other similar problems, but it does not provide a universal solution for humorous problems.

6 Ethical Considerations

We identify two ethical concerns – crowd worker wages and hurtful response generation.

Crowd worker wages. Payment for crowd workers was above the minimum wage in the US ($0.03 for a median work time of 7 seconds per hit; $15.43 per hour). Qualification-test included a warning about the rejection policy and candidates confirmed that they understood and agreed. The planted test questions were simple and clearly unambiguous, as five judges had 100% agreement and no concern someone else would disagree. For example, the sentence *"and leagues create custom wish Gamers also select participate. own to may their cups and teams they the"* is clearly broken and expected to be rated as a "very bad"

Experience	User	AI Assistant's Response	Method
Good	Order space	I have to warn you, it has zero pressure.	ConceptNet
	Do you sell a street	Would you like it to be one-way or two-way?	
	Buy me the moon	Sorry, I'm all out. Maybe just the stars...?	T5-3B
	Do you sell laugh?	We have no jokes left in our stock.	
Bad	Buy a smile	Are you motivated by heather told a joje?	ConceptNet
	Buy me a governor	Hmm... no, they wont help you, they could not a governor.	T5-3B
	Shop for budget	Do you have a great camera?	GPT-2
	Do you sell a farm?	I'll grab some iced tea for you on Friday.	
Dispute	Can I buy laziness?	Check with your local community service center.	GPT-2
	Can I buy a thought?	So I guess you're not really into art.	

Table 2. Examples of responses generated by the different methods, presenting the top-ranked and bottom ranked in the human evaluation, as well as a few with a strong dispute.

experience (we allowed for "bad" as well). In addition, we used a large safety margin (at least four errors) before auto-rejecting. Some edge cases were manually re-assessed by looking at the errors, error rate (total hits) and a sample of other hits. Overall, we rejected 5/75 workers (18.2% of the hits).

Hurtful response generation. T5 and GPT-2 can be potentially dangerous in manifesting undesirable biases and representational harm [30, 7, 9, 14]. ConceptNet, which is mostly generated by humans, may also reflect social biases [13]. User safety should be kept in mind before applying those methods in production, providing safety measures against responses proposing harmful actions. For example, the "sleep" request response: "Oh, you want to sleep? Maybe you need alcohol", or another example based on ConceptNet is the "happiness" request response: "Oh, you want happiness? Maybe you need to look better.".

7 Conclusions

In this work, we focus on equipping AI assistants with the ability to *respond* in a playful manner to irrational shopping questions. We compare between neural Generative LMs and a simple, yet effective, template-based solution, grounded with commonsense. Results show that generating appropriate responses is a non-trivial task – both in terms of the generation itself and in objectively measuring the output. Interestingly, the template-based approach outperforms state-of-the-art Generative LMs, providing further evidence on the importance of commonsense in constrained text generation. However, we note that the gap to human-expert responses is still quite large. Inspired with previous works showing that integrating real-world knowledge graphs can improve over the vanilla LM approach, an interesting line of research would be to constrain and guide LM generation by commonsense knowledge graphs.

References

1. Amin, M., Burghardt, M.: A survey on approaches to computational humor generation. In: Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature. pp. 29–41 (2020)
2. Binsted, K., Ritchie, G.: Computational rules for generating punning riddles (1997)
3. Dybala, P., Ptaszynski, M., Higuchi, S., Rzepka, R., Araki, K.: Humor prevails!-implementing a joke generator into a conversational system. In: Australasian Joint Conference on Artificial Intelligence. pp. 214–225. Springer (2008)
4. Goldberg, K., Roeder, T., Gupta, D., Perkins, C.: Eigentaste: A constant time collaborative filtering algorithm. information retrieval **4**(2), 133–151 (2001)
5. Goldberg, Y.: Neural network methods for natural language processing. Synthesis lectures on human language technologies **10**(1), 1–309 (2017)
6. Hessel, J., Marasović, A., Hwang, J.D., Lee, L., Da, J., Zellers, R., Mankoff, R., Choi, Y.: Do androids laugh at electric sheep? humor" understanding" benchmarks from the new yorker caption contest. arXiv preprint arXiv:2209.06293 (2022)
7. Kirk, H.R., Volpin, F., Iqbal, H., Benussi, E., Dreyer, F., Shtedritski, A., Asano, Y., et al.: Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. Advances in Neural Information Processing Systems **34** (2021)
8. Le, M., Boureau, Y.L., Nickel, M.: Revisiting the evaluation of theory of mind through question answering. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 5872–5877 (2019)
9. Liang, P.P., Wu, C., Morency, L.P., Salakhutdinov, R.: Towards understanding and mitigating social biases in language models. In: International Conference on Machine Learning. pp. 6565–6576. PMLR (2021)
10. Lin, B.Y., Zhou, W., Shen, M., Zhou, P., Bhagavatula, C., Choi, Y., Ren, X.: Commongen: A constrained text generation challenge for generative commonsense reasoning. arXiv preprint arXiv:1911.03705 (2019)
11. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G.: Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. arXiv preprint arXiv:2107.13586 (2021)
12. Luo, F., Li, S., Yang, P., Chang, B., Sui, Z., Sun, X., et al.: Pun-gan: Generative adversarial network for pun generation. arXiv preprint arXiv:1910.10950 (2019)
13. Mehrabi, N., Zhou, P., Morstatter, F., Pujara, J., Ren, X., Galstyan, A.: Lawyers are dishonest? quantifying representational harms in commonsense knowledge resources. arXiv preprint arXiv:2103.11320 (2021)
14. Nadeem, M., Bethke, A., Reddy, S.: Stereoset: Measuring stereotypical bias in pretrained language models. arXiv preprint arXiv:2004.09456 (2020)
15. Petrovic, S., Matthews, D.: Unsupervised joke generation from big data. In: ACL (2). pp. 228–232 (2013)
16. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
17. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:1910.10683 (2019)
18. Ren, H., Yang, Q.: Neural joke generation. Final Project Reports of Course CS224n (2017)

19. Sakaguchi, K., Bras, R.L., Bhagavatula, C., Choi, Y.: Winogrande: An adversarial winograd schema challenge at scale. Communications of the ACM **64**(9), 99–106 (2021)
20. Sap, M., LeBras, R., Fried, D., Choi, Y.: Neural theory-of-mind? on the limits of social intelligence in large lms. arXiv preprint arXiv:2210.13312 (2022)
21. Sap, M., Rashkin, H., Chen, D., LeBras, R., Choi, Y.: Socialiqa: Commonsense reasoning about social interactions. arXiv preprint arXiv:1904.09728 (2019)
22. Shani, C., Libov, A., Tolmach, S., Lewin-Eytan, L., Maarek, Y., Shahaf, D.: " alexa, what do you do for fun?" characterizing playful requests with virtual assistants. arXiv preprint arXiv:2105.05571 (2021)
23. Speer, R., Chin, J., Havasi, C.: Conceptnet 5.5: An open multilingual graph of general knowledge. In: Thirty-first AAAI conference on artificial intelligence (2017)
24. Stock, O., Strapparava, C.: Hahacronym: Humorous agents for humorous acronyms (2003)
25. Stock, O., Strapparava, C.: Hahacronym: A computational humor system. In: Proceedings of the ACL Interactive Poster and Demonstration Sessions. pp. 113–116 (2005)
26. Talmor, A., Yoran, O., Bras, R.L., Bhagavatula, C., Goldberg, Y., Choi, Y., Berant, J.: Commonsenseqa 2.0: Exposing the limits of ai through gamification. arXiv preprint arXiv:2201.05320 (2022)
27. Tevet, G., Habib, G., Shwartz, V., Berant, J.: Evaluating text gans as language models. arXiv preprint arXiv:1810.12686 (2018)
28. Valitutti, A.: How many jokes are really funny? In: Human-Machine Interaction in Translation: Proceedings of the 8th International NLPCS Workshop. vol. 41, p. 189. Samfundslitteratur (2011)
29. Valitutti, A., Doucet, A., Toivanen, J.M., Toivonen, H.: Computational generation and dissection of lexical replacement humor. Natural Language Engineering **22**(5), 727–749 (2016)
30. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., et al.: Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359 (2021)
31. Winters, T., Nys, V., Schreye, D.D.: Automatic joke generation: Learning humor from examples. In: International Conference on Distributed, Ambient, and Pervasive Interactions. pp. 360–377. Springer (2018)
32. Yu, Z., Tan, J., Wan, X.: A neural approach to pun generation. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1650–1660 (2018)
33. Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., Choi, Y.: Hellaswag: Can a machine really finish your sentence? arXiv preprint arXiv:1905.07830 (2019)