FigurativeQA: A Test Benchmark for Figurativeness Comprehension for Question Answering

Geetanjali Rakshit and Jeffrey Flanigan
Computer Science and Engineering Department
UC Santa Cruz
{grakshit,jmflanig}@ucsc.edu

Abstract

Figurative language is widespread in human language (Lakoff and Johnson, 2008), posing potential challenges in NLP applications. In this paper, we investigate the effect of figurative language on the task of question answering (QA). We construct FigurativeQA, a test set of 400 yes-no questions with figurative and non-figurative contexts, extracted from product reviews and restaurant reviews. We demonstrate that a state-of-the-art RoBERTa QA model has considerably lower performance in question answering when the contexts are figurative rather than literal, indicating a gap in current models. We propose a general method for improving the performance of QA models by converting the figurative contexts into non-figurative by prompting GPT-3, and demonstrate its effectiveness. Our results indicate a need for building QA models infused with figurative language understanding capabilities.

1 Introduction

Understanding figurative language can be a challenging task for humans, let alone for machines (Zayed et al., 2020). Although native speakers may effortlessly infer the meaning of similes and metaphors, it may be particularly difficult for non-native speakers to understand. Effects of the presence of figurative language has been studied for various downstream NLP tasks such as machine translation (Dankers et al., 2022), recognizing textual entailment (Chakrabarty et al., 2021), and dialog models (Jhamtani et al., 2021), inter-alia.

To the best of our knowledge, there is no prior line of work investigating question answering (QA) on figurative text. Figurative language has a limited presence in existing question answering (QA) datasets in popular use such as SQuAD (Rajpurkar et al., 2018) and Natural Questions (Kwiatkowski et al., 2019), where the contexts are typically literal and factual, constructed from Wikipedia passages.[1]

While figurative language has a limited presence in many QA datasets, it does occur regularly in some domains, such as the reviews domain. User-written reviews, especially those with highly positive or highly negative ratings tend to use strong opinions and are more likely to contain figurative language (Mohammad et al., 2016). We show that it can be challenging for existing QA models to draw inferences from this kind of figurative text.

We propose a new task of answering questions from text that is figurative, and consequently, more challenging to answer. For this task, we present a test dataset, FigurativeQA, consisting of 400 questions and accompanying figurative contexts constructed from Amazon product reviews (Niculae and Danescu-Niculescu-Mizil, 2014) and Yelp restaurant reviews (Oraby et al., 2017). We leverage existing resources for figurative contexts (Niculae and Danescu-Niculescu-Mizil, 2014; Oraby et al., 2017) and manually construct question-answer pairs from these contexts. Further, we create non-figurative versions of this dataset, both automatically by prompting GPT-3 (Brown et al., 2020) as well as manually. We show that it is harder to answer questions from figurative context for current state-of-the-art models. In fig. 1, we show examples of figurative contexts from Amazon product reviews and Yelp restaurant reviews, a question answer pair for the contexts, along with automatically and manually constructed non-figurative versions of the context.

The contributions of this work are the following:

- FigurativeQA, a test set of 400 yes/no question-answer pairs with figurative and non-figurative contexts. For the 200 figurative contexts, we also provide manually created literal

[1]From a rough check of the SQuAD dev set, we observe that the questions themselves are also mostly non-figurative. We found two examples of figurative questions out of 5,928 answerable questions in the SQuAD dev set, one of them being *"Who is eligible to toss their name in the hat to be First Minister?"*.

Proceedings of the 3rd Workshop on Figurative Language Processing (FLP), pages 160 - 166

Figurative Context: *The album , like almost everything Krush has released , **slays** .*
Question: *Is the album good?*
Answer: *Yes*
Non-fig. version (manually created): *The album is really good, like most of Krush's work.*
Non-fig. version (from GPT-3): *The album is really good, like almost everything Krush has released.*

Figurative Context: *Although, the menu items doesnt **SCREAM** French cuisine. Most foods looks like you can get at any American place.*
Question: *Is the menu authentic french?*
Answer: *No*
Non-fig. version (manually created): *The menu items aren't typical of French cuisine. Rather, they are common at most American eateries.*
Non-fig. version (from GPT-3): *Although, the menu items doesn't look very French. Most foods look like you can get at any American place.*

Figure 1: Examples of figurative contexts from FigurativeQA. Example 1 is from Amazon product reviews and Example 2 from Yelp restaurant reviews.The figurative text fragments within the contexts are shown in bold and italics.

contexts for comparison.

- We show that it is harder to answer questions from figurative contexts for models trained on QA data with non-figurative contexts, and that manually changing the figurative context to a meaning-preserving non-figurative version improves performance.
- We propose a method to use GPT-3 to automatically produce non-figurative contexts from figurative ones, and demonstrate that it improves our QA system on the FigurativeQA dataset.

The outline of the paper is as follows: after reviewing related work (§2), we introduce our new QA dataset for figurative language (§3). We next introduce a general method for converting figurative language to non-figurative language by prompting GPT-3 (§4), which we use to improve our baseline QA model. We then present our experimental results (§5), and finally conclude (§6).

2 Related Work

Handling of figurative language is of significance in many downstream NLP tasks such as machine translation (Mao et al., 2018; Dankers et al., 2022), recognizing textual entailment (Chakrabarty et al., 2021), sentiment analysis (Qadir et al., 2015), among others. Chakrabarty et al. (2021) investigate the robustness of state-of-the-art entailment models on figurative examples on test sets for similes, metaphors, and irony. Chakrabarty et al. (2022) test figurative language understanding in pre-trained language models by evaluating continuation of text in narratives, while (Liu et al., 2022) investigate non-literal reasoning capabilities of language models on a Winograd-style non-literal language understanding task.

The idea of converting metaphors to their literal counterparts has been previously explored for machine translation by Mao et al. (2018), where metaphors in English text are first identified and then converted to a literal version, by making use of word embeddings and WordNet, before doing machine translation into Chinese. In dialog systems, a similar approach was employed by Jhamtani et al. (2021), where idioms and metaphors in utterances are converted to literal versions using a dictionary lookup-based method. Our work is closest to Jhamtani et al. (2021), except that we explore the robustness of QA systems in a machine comprehension setup, instead of dialog models, to figurative language, which, to the best of our knowledge, is a first. Our automatic approach to creating rephrased non-figurative versions of figurative text is done using pre-trained language models, rather than rule-based methods which have been shown to be error-prone (Jhamtani et al., 2021).

Related QA datasets include the FriendsQA dataset (Yang and Choi, 2019), which is a dialog-based QA dataset constructed from dialogs from the TV series Friends. While it does contain metaphors and sarcasm, it may not be ideal to test figurative language understanding as it is unclear how much of the dataset is actually figurative. The dialogic nature of the dataset further contributes

to making it challenging. Another dataset that requires figurative language understanding is the RiddleSense dataset (Lin et al., 2021), which comprises of riddles, but unlike ours, it's modeled as an open domain QA task, rather than a machine comprehension task. Parde and Nielsen (2018) show that questions about novel metaphors from literature are judged to be deeper than non-metaphorical or non-conventional metaphors by humans, but their focus is on generating deep questions, rather than testing the robustness of QA models.

3 FigurativeQA Dataset

The figurative data in FigurativeQA comes from two sources: Amazon product reviews (Niculae and Danescu-Niculescu-Mizil, 2014), and Yelp restaurant reviews (Oraby et al., 2017). For comparison, we also extract non-figurative contexts from each of these sources to form the non-figurative split of FigurativeQA. We construct a question answering dataset on top of these contexts. For simplicity, we only work with yes-no questions. Fig 1 shows examples from the FigurativeQA dataset. The data statistics from each source (Amazon and Yelp) and each split (figurative and non-figurative) are summarized in Table 1.

We select figurative texts for annotation with question-answer pairs from Amazon product reviews using the following procedure. Niculae and Danescu-Niculescu-Mizil (2014) construct a dataset of figurative comparisons extracted using comparator patterns (such as *"like"*, *"as"*, or *"than"*) from Amazon product reviews, and then obtain 3 sets of figurativeness scores (on a scale of 1 to 4) on Amazon Mechanical Turk (with scores of 1–2 binned as literal and 3–4 as figurative). Of the 1260 comparisons in this dataset, we extract the sentences which have an average figurativeness score of greater than 3. This leaves us with 254 sentences, of which we manually pick 100 instances, and construct a yes-no question for each sentence.

We select examples for annotating with question-answer pairs from Yelp reviews using a similar procedure. Oraby et al. (2017) construct a dataset for NLG in the restaurant domain from Yelp reviews, which comes labeled with sentiment information (1-2 rating for negative, 3 for neutral and 4-5 for positive). Since positive or negative reviews are more likely to contain figurative language, from the set of positive and negative reviews, we extract instances using comparator patterns such as *"like"*, *"as"*, or *"than"*, similar to the procedure in Niculae and Danescu-Niculescu-Mizil (2014). We then manually choose 100 instances that contain rich, figurative language, and construct a yes-no question for each.

	avg. context length	**category**	**Yes**	**No**
fig.	9	**Amazon**	52	48
	16	**Yelp**	50	50
non-fig.	10	**Amazon**	50	50
	14	**Yelp**	49	51

Table 1: Number of yes-no questions from Amazon product reviews and Yelp restaurant reviews for figurative and non-figurative contexts, and average length of context (number of words)

The figurative contexts from FigurativeQA tend to contain more *similes*, since comparator patterns (*"like"*, *"as"*, or *"than"*) were used to extract the text. However, we observe that many of these examples also contain other kinds of figurative constructs such as metaphor, idiom, hyperbole, sarcasm, etc, because the nature of the reviews text is such that it is replete with figurative expressions.

For each context in FigurativeQA, we construct a yes-no question. For the figurative contexts, we make sure to pose a question such that answering it would require understanding of the figurative text present in the context. For the non-figurative contexts, we construct questions that look similar to the ones for the figurative contexts. Additionally, for the figurative contexts extracted from Amazon and Yelp, we manually create non-figurative counterparts that preserve the meaning and overall content.

4 Non-Figurative Contexts from GPT-3

We obtain non-figurative versions of the figurative contexts from FigurativeQA by prompting GPT-3 (da-vinci-002). Please refer to Appendix A for model details and the prompts we used.

As an intrinsic evaluation of the effectiveness of our prompting method, we manually evaluate the correctness of the non-figurative/literal contexts generated by prompting GPT-3. Table 2 shows the results from the manual evaluation of the GPT-3 outputs. We observe that GPT-3 performs fairly well in converting figurative language in FigurativeQA to literal, with nearly 81% and 75% of the outputs in Amazon and Yelp, respectively, judged

Amazon	**Context (original, figurative)**: *Second , the final confrontation was more **like a footnote in the story** than a real ending .* **[CORRECT] Non-fig. context (from GPT-3)**: *Second, the final confrontation was less significant than it should have been given its place in the story.*
	Context (original, figurative): *Her misty , vaguely dreamlike writing can pull a reader into the world of 1900s New Orleans and its society , but her heroine sometimes feels more **like a vessel** than a fully-realized person .* **Figurative Context**: *Not a storybook , by any means , this one is more **like a visit to the zoo** .* **[INCORRECT] Non-fig. context (from GPT-3)**: *This story is not like a fairytale, it's more like a visit to the zoo.*
	Context (original, figurative): *The books are **like potato chips** - you **can't eat just one** .* **[SAME] Non-fig. context (from GPT-3)**: *The books are like potato chips - you can't eat just one .*
Yelp	**Context (original, figurative)**: *i had the chicken fajitas , which came with a giant flour tortilla that was **as hot as hades** .* **[CORRECT] Non-fig. context (from GPT-3)**: *i had the chicken fajitas , which came with a really big and hot flour tortilla.*
	Context (original, figurative): *this is as authentic thai **as much as imitation crab is authentic crab** .* **[INCORRECT] Non-fig. context (from GPT-3)**: *this is as inauthentic thai as much as imitation crab is inauthentic crab.*
	Context (original, figurative): *the same thing with the steak and potatoes , it was almost as if they tried to **decorate the plate with salt** .* **[SAME] Non-fig. context (from GPT-3)**: *the same thing with the steak and potatoes , it was almost as if they tried to decorate the plate with salt .*

Figure 2: Examples of non-figurative contexts generated from GPT-3, for Amazon and Yelp. The figurative text fragments within the contexts are shown in bold and italics.

	Amazon	Yelp
correct	81%	75%
incorrect	15%	19%
same	4 %	6%

Table 2: Evaluation of non-figurative outputs from GPT-3. **Correct** means the GPT-3 output is non-figurative and correctly preserves the meaning. **Same** means GPT-3 produced the exact same output as the input (no change). All other outputs are **incorrect**.

to be correct. In fig. 2, we show examples of non-figurative text generated from GPT-3.

5 Experiments and Results

As a baseline, we run RoBERTa-base (Liu et al., 2019) finetuned on the training set of BoolQ (Clark et al., 2019). The performance on FigurativeQA is summarized in Tables 3 and 4. We find that the RoBERTa QA model performs poorly on the figurative contexts compared to the non-figurative contexts, and that manually changing the figurative language to non-figurative language improves performance. This indicates that automatic conversion of figurative language to non-figurative language may improve performance.

	Amazon	Yelp
Fig (orig.)	83.43 ± 1.1	66.84 ± 2.61
Fig (man. non-fig)	93.5 ± 1.12	90 ± 1.44
Non-fig (orig.)	92 ± 1.42	89.6 ± 1.68

Table 3: Accuracy of RoBERTa-base fine-tuned on BoolQ, on the figurative split, manually created non-figurative version of the figurative split, and non-figurative split of FigurativeQA. (We reran experiments 1000 times with bootstrap resampling. The numbers reported are the mean and std-dev.)

To improve upon the baseline model, we pass the automatic non-figurative contexts from GPT-3 (§4) to our RoBERTa-base model. We find that this procedure improves the performance on figurative language split, and has no effect on the non-figurative language split, and improves overall

performance on FigurativeQA. As an additional comparison, we also prompt GPT-3 as a QA model and report its performance on FigurativeQA.[2]

	Amazon	Yelp
Baseline: Fig	83.43±1.1	66.84±2.61
Ours: Fig	86.51±1.13	73.5 ±1.66
Baseline: Non-fig	92±1.42	89.6±1.68
Ours: Non-fig	92.45±1.12	89.4±1.69
Baseline: Overall	87.71±0.89	78.21±.6
GPT-3: Overall	64.58±1.71	60.1±3.1
Ours: Overall	**89.5±3.18**	**81.46±1.19**

Table 4: QA performance on FigurativeQA. Our method uses GPT-3 prompting (zero-shot) to convert the figurative contexts to literal (We reran experiments for 1000 times with bootstrap resampling. The numbers reported are the mean and std-dev. The numbers in bold are the best results.)

6 Conclusion and Future Work

We show that current QA models do not perform so well when answering questions from figurative contexts as compared to non-figurative text. On manually created non-figurative versions of these contexts, we are able to show significant improvements. However, the manual annotation being an expensive step, we use an automatic method of prompting of GPT-3 and were still able to achieve performance gains. This highlights a need to build QA models that can handle figurative text. In the future, we would like to do a fine-grained analysis of QA performance on different kinds of figurative constructs, including similes, metaphors, irony, idioms, rhetorical questions, hyperbole, personification, etc.

Limitations

Our dataset contains the specific domains of Amazon and Yelp reviews, and is English-only. Results and conclusions may not generalize to other domains or languages.

References

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.

Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. 2022. It's not rocket science: Interpreting figurative language in narratives. *Transactions of the Association for Computational Linguistics*, 10:589–606.

Tuhin Chakrabarty, Debanjan Ghosh, Adam Poliak, and Smaranda Muresan. 2021. Figurative language in recognizing textual entailment. *arXiv preprint arXiv:2106.01195*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*.

Verna Dankers, Christopher G Lucas, and Ivan Titov. 2022. Can transformer be too compositional? analysing idiom processing in neural machine translation. *arXiv preprint arXiv:2205.15301*.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Taylor Berg-Kirkpatrick. 2021. Investigating robustness of dialog models to popular figurative language constructs. *arXiv preprint arXiv:2110.00687*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

George Lakoff and Mark Johnson. 2008. *Metaphors we live by*. University of Chicago press.

Bill Yuchen Lin, Ziyi Wu, Yichi Yang, Dong-Ho Lee, and Xiang Ren. 2021. Riddlesense: Reasoning about riddle questions featuring linguistic creativity and commonsense knowledge. *arXiv preprint arXiv:2101.00376*.

Emmy Liu, Chen Cui, Kenneth Zheng, and Graham Neubig. 2022. Testing the ability of language models to interpret figurative language. *arXiv preprint arXiv:2204.12632*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Rui Mao, Chenghua Lin, and Frank Guerin. 2018. Word embedding and wordnet based metaphor identification and interpretation. In *Proceedings of the 56th annual meeting of the association for computational linguistics*. Association for Computational Linguistics (ACL).

[2]Please refer to Appendix B for details about prompting GPT-3 as a QA system.

Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 23–33.

Vlad Niculae and Cristian Danescu-Niculescu-Mizil. 2014. Brighter than gold: Figurative language in user generated comparisons. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 2008–2018.

Shereen Oraby, Sheideh Homayon, and Marilyn Walker. 2017. Harvesting creative templates for generating stylistically varied restaurant reviews. *arXiv preprint arXiv:1709.05308*.

Natalie Parde and Rodney Nielsen. 2018. Automatically generating questions about novel metaphors in literature. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 264–273.

Ashequl Qadir, Ellen Riloff, and Marilyn A Walker. 2015. *Learning to recognize affective polarity in similes*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

Zhengzhe Yang and Jinho D Choi. 2019. Friendsqa: Open-domain question answering on tv show transcripts. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 188–197.

Omnia Zayed, John Philip McCrae, and Paul Buitelaar. 2020. Figure me out: a gold standard dataset for metaphor interpretation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5810–5819.

A Appendix A: Prompting GPT-3 for figurative text

We use the da-vinci-002 model with temperature set to 0.3 and max-length set to 100. We used a prompt with 5 examples, as shown in Fig. 3.

For the following inputs, if the text contains figurative language, convert it to a literal version. Otherwise, output the same text as the input.

Input: It's inevitable. Their love was built on sand and this is why their marriage has landed on the rocks.
Output: It's inevitable. Their love was unstable and this is why their marriage has failed.

Input: The weather forecast predicted a heatwave this week across most of the country.
Output: The weather forecast predicted a heatwave this week across most of the country.

Input: During the heatwave, the entire house was like a furnace.
Output: During the heatwave, the entire house was uncomfortably hot.

Input: The brisket is nothing to write home about.
Output: There is nothing particularly remarkable about the brisket.

Input: The fries were served cold.
Output: The fries were served cold.

Input: The lamb had a melt in the mouth texture.
Output: The lamb was soft and well-cooked.

Input: The adapter worked like a charm.
Output: The adapter worked perfectly.

Figure 3: GPT-3 prompt to generate non-figurative versions of the figurative contexts.

B Appendix B: Prompting GPT-3 for QA

We use the da-vinci-002 model with temperature set to 0.3 and max-length set to 100. We used a prompt with 2 examples, as shown in Fig. 4.

Based on the passage, answer the following question with a yes or a no.

Passage:
Windows Movie Maker (formerly known as Windows Live Movie Maker in Windows 7) is a discontinued video editing software by Microsoft. It is a part of Windows Essentials software suite and offers the ability to create and edit videos as well as to publish them on OneDrive, Facebook, Vimeo, YouTube, and Flickr.
Question: Is windows movie maker part of windows essentials?
Answer: yes

Passage:
Both Jersey and Bank of England notes are legal tender in Jersey and circulate together, alongside the Guernsey pound and Scottish banknotes. The Jersey notes are not legal tender in the United Kingdom but are legal currency, so creditors and traders may accept them if they so choose.
Question: Is jersey currency legal tender in the uk?
Answer: no

Figure 4: GPT-3 prompt to get yes-no answers.