A Report on the FigLang 2022 Shared Task on Understanding Figurative Language

Arkadiy Saakyan[1] **Tuhin Chakrabarty**[1] **Debanjan Ghosh**[2] **Smaranda Muresan**[1]
[1]Department of Computer Science, Columbia University
[2]Educational Testing Service
a.saakyan@columbia.edu, tuhin.chakr@cs.columbia.edu, dghosh@ets.org, smara@cs.columbia.edu

Abstract

We present the results of the Shared Task on Understanding Figurative Language that we conducted as a part of the 3rd Workshop on Figurative Language Processing (FigLang 2022) at EMNLP 2022. The shared task is based on the FLUTE dataset (Chakrabarty et al., 2022), which consists of NLI pairs containing figurative language along with free text explanations for each NLI instance. The task challenged participants to build models that are able to not only predict the right label for a figurative NLI instance, but also generate a convincing free-text explanation. The participants were able to significantly improve upon provided baselines in both automatic and human evaluation settings. We further summarize the submitted systems and discuss the evaluation results.

1 Introduction

Figurative language such as metaphors, similes or sarcasm plays an important role in enriching human communication, allowing us to express complex ideas and emotions in an implicit way (Roberts and Kreuz, 1994; Fussell and Moss, 1998). However, understanding figurative language still remains a bottleneck for natural language processing (Shutova, 2011). In spite of the fact that Transformer-based language models (LMs) get larger (Brown et al., 2020; Raffel et al., 2020), they are still incapable of comprehending the physical world, cultural knowledge, or social context of figurative language (Bisk et al., 2020).

In recent years, there have been several benchmarks dedicated to figurative language understanding, which generally frame "understanding" as a recognizing textual entailment (a.k.a natural language inference (NLI)) task — deciding whether one sentence (premise) entails/contradicts another (hypothesis) (Chakrabarty et al., 2021; Stowe et al., 2022; Srivastava et al., 2022). However, similar to general NLI datasets, these benchmarks suffer from spurious correlations and annotation artifacts (McCoy et al., 2019; Poliak et al., 2018). These can allow large language models (LLMs) to achieve near human-level performance on in-domain test sets, yet turn brittle when evaluated against out-of-domain or adversarial examples (Glockner et al., 2018; Ribeiro et al., 2016, 2020). To tackle these problems, research in NLI has argued that it is not enough to correctly predict the entail/contradict labels, but also to explain the decision using natural language explanations that are comprehensible to an end-user assessing model's reliability (Camburu et al., 2018; Majumder et al., 2021; Wiegreffe et al., 2021), leading to novel datasets such as e-SNLI (Camburu et al., 2018).

In this paper, we report on the shared task that aim to test the ability of models to not only predict the right label, but also provide a free-text explanation to the instance. This task was conducted as part of the 3rd Workshop on Figurative Language Processing (FigLang 2022) at EMNLP 2022. Section 2 provides a description of the shared task, datasets, and evaluation metrics. Section 3 contains brief summaries of each of the participating systems whereas Section 4 reports a comparative analysis of the participating systems.

2 Datasets and Task Description

As stated earlier, this shared task is based on the FLUTE dataset that was released by Chakrabarty et al. (2022). FLUTE consists of pairs of premises (literal sentences) and hypotheses (figurative sentences), with the corresponding entailment or contradiction labels (NLI instances), along with explanations for each instance (Table 1). This dataset is based on four types of figurative language - idiom, metaphor, sarcasm, and simile. Note, given sarcasm is the opposite of the literal meaning, we would only have contradictions in the dataset, thus we also generate a literal hypothesis that entails the literal premise. Table 1 contains a few examples

Proceedings of the 3rd Workshop on Figurative Language Processing (FLP), pages 178 - 183
December 8, 2022

Type	Premise (literal)	Hypothesis (figurative*)	Label	Explanation
Paraphrase + Sarcasm	My next door neighbors are *always arguing* in our shared hallway.	It's *so annoying* to have to hear my next door neighbors *argue all the time* in our shared hallway.	E	The sound of arguing neighbors can often be very disruptive and if it happens all the time in a common space like a shared hallway it is natural to find it annoying.
		It's *so pleasant* to have to hear my next door neighbors *argue all the time* in our shared hallway.	C	The sound of arguing neighbors can often be very disruptive and so someone considering it to be pleasant is not really accurate.
Simile	The assembly hall was now *hot and moist*, more so than usual.	In fact, the assembly hall was now *like a steam sauna*.	E	A sauna is a hot and moist environment, so the simile is saying that the hall is even hotter and more moist than usual.
	The assembly hall was now *cold and dry*, more so than usual.		C	A steam sauna is a small room or hut where people go to sweat in steam, so it would be hot and humid, not cold and dry.
Metaphor	He *mentally assimilated* the knowledge or beliefs of his tribe.	He *absorbed the knowledge* or beliefs of his tribe.	E	To absorb something is to take it in and make it part of yourself.
	He *utterly decimated* his tribe's most deeply held beliefs.		C	Absorbed typically means to take in or take up something, while "utterly decimated" means to destroy completely.
Idiom	Lady Southridge was wringing her hands, *trying hard and desperately to salvage* the bleak and miserable situation so that it somehow looks positive.	Lady southridge was wringing her hands, trying *to grasp at straws.*	E	To grasp at straws means to make a desperate attempt to salvage a bad situation, which is exactly what Lady Southridge is trying to do.
	Lady Southridge was wringing her hands, *doing absolutely nothing to overturn* the bleak and miserable situation so that it somehow looks positive.		C	To grasp at straws means to make a desperate attempt to salvage a bad situation, but the sentence describes not doing anything to change the situation

Table 1: FLUTE examples of figurative text (hypothesis) and their respective literal entailment (E) and contradiction (C) premises, along with the associated explanations. * For simile, metaphor, and idiom, figurative examples are the hypothesis whereas for sarcasm, we have both figurative and literal hypotheses.

	Entails	Contradicts	Total
Paraphrase	1339	-	1339
+ Sarcasm	-	2678	2678
Simile	750	750	1500
Metaphor	750	750	1500
Idiom	1000	1000	2000

Table 2: Dataset statistics showing distribution of Figurative Language across FLUTE.

taken fro the dataset. FLUTE contains 9,000 high quality <literal, figurative> sentence pairs with entail/contradict labels and the associated examples. Please refer to Table 2 for the dataset statistics.

2.1 Evaluation Setup

To evaluate the participant models, we built a test set by randomly selecting 750 instances (i.e., <premise, hypothesis> pairs with associated explanations) from the sarcasm dataset, and 250 examples each from simile, metaphor and idiom datasets, for a total of 1,500 instances. Below we describe several automatic metrics and human evaluations we consider to assess the models' ability to understand figurative language.

Automatic Metrics To judge the quality of the explanations we compute the average between BERTScore (Zhang et al., 2020) [1] and BLEURT (Sellam et al., 2020), which we refer to as *explanation score* (between 0 and 100). Instead of reporting only label accuracy, we report label accuracy at three thresholds of explanation score (0, 50, and 60). Accuracy@0 is equivalent to simply computing label accuracy, while Accuracy@50 counts as correct only the correctly predicted labels that achieve an explanation score greater than 50.

Human Evaluation We also measure the quality of the generated textual explanations via the MTurk platform. We recruit crowd workers with at least 98% HIT approval rate. We compute human judgement scores (H_{score}), identical to the e-ViL score in Kayser et al. (2021). We used instances that were used for evaluation in (Chakrabarty et al., 2022), and selected those on which all systems predicted correctly (a total of 150 samples, around 50 per figurative language type). We present five

[1] We use the DeBERTa-mnli version that has shown to have highest correlation with human judges (He et al., 2020).

textual explanations generated by the models and ask three workers the following question: *Given the two sentences, does the explanation justify the answer above?* We provide four options: *Yes* (1), *Weak Yes* ($\frac{2}{3}$), *Weak No* ($\frac{1}{3}$), and *No* (0). For each explanation, we average the scores by the three annotators and report the sample average in Table 4 as $\mathrm{H}_{\mathrm{score}}$.

3 Participants and Results

Training Phase The shared task started on July 10, 2022, when the training data and the auxiliary scripts were made available to all the registered participants. Participants were allowed to choose to partition the training data further to a validation set for tuning the hyper parameters. Likewise, they can also elect to use the training data to perform cross-validation.

Evaluation Phase In this phase, test instances for evaluation are released. We released the test data on August 15, 2022. Submissions were accepted until August 20, 2022. Out of all the submissions, five shared task system papers are accepted to the Workshop. Predictions are submitted to the Codalab site and evaluated against the gold labels of the test instances. We used Codalab for the shared task because it is easy to use, provided easy communication with the participants (e.g., allow mass-emailing to the participants), as well as tracks all the submissions updating the leader-board in real-time. We allowed up to five submissions per day for each participant team. We did setup our own GPU-based evaluation using a custom Docker architecture. The leader-board displayed the accuracy@60 scores on the descending order.

In total we have five participating teams alongside the organizing team of shared task. We describe the participating systems in the following section.

Team	**Acc@60**	H_{score}
TeamCoolDoge	**63.33 (1)**	*74.98 (2)*
rachneet	**63.33 (1)**	**75.28 (1)**
vund	*60.73 (2)*	71.82 (5)
yklal95	51.73 (3)	73.73 (4)
baseline	48.33 (4)	74.39 (3)

Table 3: Automatic (Accuracy@60) and Human evaluation results (H_{score}) by team with rank in parenthesis.

Baseline (Chakrabarty et al., 2022) The baseline is the system described in Chakrabarty et al. (2022). This system is trained to predict labels and rationales jointly using a T5-3B model (Raffel et al., 2020). Unlike other teams (Chakrabarty et al., 2022) verbalized inputs using natural language instruction: *Does the sentence **"P"** entail or contradict the sentence **"H"**? Please answer between **"Entails"** or **"Contradicts"** and explain your decision in a sentence.*

TeamCoolDoge (Gu et al., 2022b) present *DREAM-FLUTE* which first uses DREAM (Gu et al., 2022a) to generate an elaboration of the situation in the premise and hypothesis (separately), then uses this additional context for classification and explanation generation. They hypothesize that such additional, pertinent details could also improve a model's ability to judge whether it is an entailment or contradiction between the premise and hypothesis. This posit this could be especially helpful for the instances that use figurative language, where the underlying meaning might be opaque to the model and that further elaborating the context can make certain inferences more explicit. They take as input <*Premise*> <*Premise-elaboration-from-DREAM*> <*Hypothesis*> <*Hypothesis-elaboration-from-DREAM*> and fine-tune a T5-3B model to then jointly generate Label and Explanation. While the scene elaboration dimensions from DREAM can vary across the categories of *consequence, emotion, motivation, social norm* the winning submission is based on consequence elaboration dimension. It should be noted that the underlying model is similar to the baseline model (ablation without using DREAM), however the performance differs due to different hyperparameters.

Rachneet (Bigoulaeva et al., 2022) focus their efforts on the transfer of information from multiple related tasks for improved performance on FLUTE. They compare the effectiveness of *Sequential Fine Tuning* with that of *MultiTask Learning* in a context where one of the target tasks is dependent on the other. Their final submission which led to the highest Acc@60 on the FLUTE test set is a T5 (Raffel et al., 2020) based model where the label and rationales are predicted jointly. In particular their best submission is a sequentially fine-tuned model where they first finetune on eSNLI (Camburu et al., 2018) followed by IMPLI (Stowe et al., 2022) and

Team	idiom	metaphor	sarcasm	simile
TeamCoolDodge (AI2)	**74.85 (1)**	72.47 (3)	**75.71 (1)**	*77.33 (2)*
rachneet (UKP)	72.22 (3)	**77.27 (1)**	73.13 (4)	**79.11 (1)**
vund (UIT)	70.76 (4)	71.46 (4)	72.09 (5)	73.78 (3)
yklal95 (SBU)	70.76 (4)	*76.01 (2)*	73.64 (3)	73.78 (3)
debanjan (us)	*73.98 (2)*	*76.01 (2)*	*74.68 (2)*	71.11 (4)

Table 4: Human evaluation results (H_{score}) by team by figurative language type with rank in parenthesis.

finally FLUTE (Chakrabarty et al., 2022).

Vund (Phan et al., 2022) considered both the tasks: the NLI task, and the explanation generation task as two seq2seq tasks. They fine-tuned the two tasks separately as a simultaneous computation model. In addition, they also used the attribute about types of Figurative Language across the data as a predictor and treated it as seq2seq tasks. Therefore they have 3 component models based on fine-tuning pre-trained model T5 (Raffel et al., 2020) : NLI predictor, Type predictor, and Generator. Unlike other teams that predict label and rationale jointly here the team uses T5-large model in a pipeline fashion.

yklal (Lal and Bastan, 2022) propose a simple T5-large model fine-tuned on the FLUTE data, trained to generate the explanation before the label. The input format does not contain any task-specific keys and does not resemble any of the ones described in Raffel et al. (2020). The model uses a newline separator, which is a prominent part of how UnifiedQA (Khashabi et al., 2020) was built over T5.

4 Analysis

The best performing teams according to both human and automatic evaluation were TeamCoolDoge, rachneet, and vund (Table 3). For automatic metric we report Accuracy@60, i.e., accuracy score that counts as correct only the correctly predicted labels that achieve an explanation score greater than 60. We notice in Table 3 that TeamCoolDoge and rachneet have attain the highest score in case of accuracy score where team vund is slightly behind.

Likewise, human evaluation results (Table 4) show relatively small difference between teams, indicating plausibility of explanations across systems and across different types of figurative language. These results support the high automatic evaluation scores the teams have achieved. Some discrepancies in human and automatic evaluation are present (e.g., the team TeamCoolDodge did not achieve the highest human score for metaphors and similes). This can be explained by high standard deviation in the human score (around 0.3, or one step increment in the answer), however, future work may explore spurious cues and lack of correlation in automatic metrics.

Across types of figurative language, explanations for similes and metaphors achieve higher human scores for the best submissions. This could be explained by the visual nature of comparisons drawing from commonsense property identification which can benefit from elaboration as used in the DREAM framework used by TeamCoolDoge.

5 Conclusion

This paper summarizes the results of the shared task on understanding figurative language organized as part of the 3rd Workshop on the Figurative Language Processing at EMNLP 2022 (FigLang 2022). This shared task aimed to not only predict the correct label for a figurative NLI instance but also generate a convincing explanation for the same. We provided basic description of each of the participating systems who submitted a shared task system paper (i.e. four qualifying submissions). All of the submitted systems by the participants attain higher accuracy than the baseline. We also conducted human evaluation via MTurk platform that shows the quality of explanations generated by the systems is comparable. Finally, to conclude, we hope the shared task will promote further exploration into figurative language understanding.

References

Irina Bigoulaeva, Rachneet Singh Sachdeva, Harish Tayyar Madabushi, Aline Villavicencio, and Iryna Gurevych. 2022. Effective cross-task transfer learning for explainable natural language inference with t5. In *Proceedings of the Third Workshop on Figurative Language Processing*. Association for Computational Linguistics.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online. Association for Computational Linguistics.

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.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Tuhin Chakrabarty, Debanjan Ghosh, Adam Poliak, and Smaranda Muresan. 2021. Figurative language in recognizing textual entailment. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3354–3361, Online. Association for Computational Linguistics.

Tuhin Chakrabarty, A. Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. Flute: Figurative language understanding through textual explanations.

Susan R Fussell and Mallie M Moss. 1998. Figurative language in emotional communication. *Social and cognitive approaches to interpersonal communication*, pages 113–141.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.

Yuling Gu, Bhavana Dalvi, and Peter Clark. 2022a. DREAM: Improving situational QA by first elaborating the situation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1115–1127, Seattle, United States. Association for Computational Linguistics.

Yuling Gu, Yao Fu, Valentina Pyatkin, Ian Magnusson, Bhavana Dalvi Mishra, and Peter Clark. 2022b. Just-dream-about-it: Figurative language understanding with dream-flute. In *Proceedings of the Third Workshop on Figurative Language Processing*. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. 2021. E-vil: A dataset and benchmark for natural language explanations in vision-language tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1244–1254.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online. Association for Computational Linguistics.

Yash Kumar Lal and Mohaddeseh Bastan. 2022. Sbu figures it out: Models explain figurative language. In *Proceedings of the Third Workshop on Figurative Language Processing*. Association for Computational Linguistics.

Bodhisattwa Prasad Majumder, Oana-Maria Camburu, Thomas Lukasiewicz, and Julian McAuley. 2021. Rationale-inspired natural language explanations with commonsense. *arXiv preprint arXiv:2106.13876*.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Khoa Thi Kim Phan, Duc-Vu Nguyen, and Ngan Luu-Thuy Nguyen. 2022. Nlp@uit at figlang-emnlp 2022: A divide-and-conquer system for shared task on understanding figurative language. In *Proceedings of the Third Workshop on Figurative Language Processing*. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. *arXiv preprint arXiv:1805.01042*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Richard M Roberts and Roger J Kreuz. 1994. Why do people use figurative language? *Psychological science*, 5(3):159–163.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Ekaterina V Shutova. 2011. Computational approaches to figurative language. Technical report, University of Cambridge, Computer Laboratory.

Aarohi Srivastava, Abhinav Rastogi, and Abhishek Rao. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. In *In preparation*.

Kevin Stowe, Prasetya Utama, and Iryna Gurevych. 2022. IMPLI: Investigating NLI models' performance on figurative language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5375–5388, Dublin, Ireland. Association for Computational Linguistics.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2021. Reframing human-ai collaboration for generating free-text explanations. *arXiv preprint arXiv:2112.08674*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.