Automating Dataset Production Using Generative Text and Image Models

Christopher Thierauf, Mitchell Abrams, Matthias Scheutz
Tufts University, Tufts University, Tufts University
{christopher.thierauf, mitchell.abrams, matthias.scheutz}@tufts.edu

Abstract

Practical and ethical dataset collection remains a challenge blocking many empirical methods in natural language processing, resulting in a lack of benchmarks or data on which to test hypotheses. We propose a solution to some of these areas by presenting a pipeline to reduce the research burden of producing image and text datasets when datasets may not exist. Our approach, with accompanying software tools, involves (1) generating text with LLMs; (2) creating accompanying image vignettes with text–to–image transformers; and (3) low-cost human validation. Based on existing literature that has struggled with quantitative evaluation (due to difficulty of data collection), we present the creation of 3 relevant datasets, and conduct a user study that demonstrates this approach is able to aid researchers in obtaining previously-challenging datasets. We provide sample data generated with this technique, the source code used to produce it, and discuss applicability and limitations.

Keywords: Language Resource Design, Evaluation Methodologies, Information Extraction

1. Introduction

Like many technical and research fields, our understanding of language, linguistics, and language processing depends on data: for example, to perform benchmarking or to validate hypotheses. While the research community has produced a wide variety of datasets for these purposes, either from automatically scraping web forums (Forbes et al., 2020), news sources (Pradhan et al., 2012), or asking humans to provide input (Chen et al., 2021), it is practically and ethically challenging to collect data for many tasks. Further, these datasets will reflect the data as it was collected, and so will likely include biases that must be addressed.

As one example, consider the Winograd schema challenge (Levesque et al., 2012), which is a library of referentially ambiguous cases that need common-sense reasoning to resolve. While datasets for examining reference resolution are generally readily available, datasets for more specific reference phenomena are either sparse or non-existent. Several requirements are established to create such problems: stripping linguistic cues to avoid resolution via selectional restriction, permitting only binary reference candidates, and changing contexts so that the referent is interpreted in the opposite direction. As such, it is practically difficult to collect naturally occurring contextual and linguistic examples that satisfy these exact requirements.

Therefore, we aim to *carefully* and *artificially* construct an artificial dataset while still reflecting natural phenomena that one might encounter. Then, using low-cost human intelligence tasks, features of the artificial dataset can be validated and evaluated. With this, a full dataset, plus evaluated data, has been rapidly developed.

In this paper, we propose a novel pipeline for text-image vignette generation that can create datasets for specific NLP tasks that explore language and context. We will present our technical approach to this pipeline, demonstrate a small validation study on one of the visually grounded reference resolution datasets, and describe other applicable research questions.

2. Related Works

Generating synthetic data is not new. But current approaches are not sufficient for the class of problems that require specific scene contexts. He et al. (2022a) have acknowledged the challenge of finding task-specific data for certain NLP tasks with complex inputs that are naturally low-resource. They addressed this problem with a *generate*, *annotate*, and *learn* (GAL) framework that uses LLMs to generate synthetic unlabeled text. The application of this work is different from ours, however, in that it focuses on self learning; classifiers give pseudo-labels for the generated text for training. LLMs have been widely used in similar ways for general Data Augmentation (DA) purposes, generating text (Liu et al., 2020; Anaby-Tavor et al., 2020; Kedzie and McKeown, 2019; Yang et al., 2020; Mahajan et al., 2022) and text-image pairs (Xu et al., 2020) to enhance classifier performance for low-resource datasets or imbalanced classes.

Our pipeline is fundamentally distinct from this previous work because we generate contextual visual scenes that *relate* to language. Additionally, the synthetic data generated by our pipeline is for human-in-the-loop facing applications, rather than training NLP models (which we will not attempt). Unlike DA approaches, our pipeline does not use

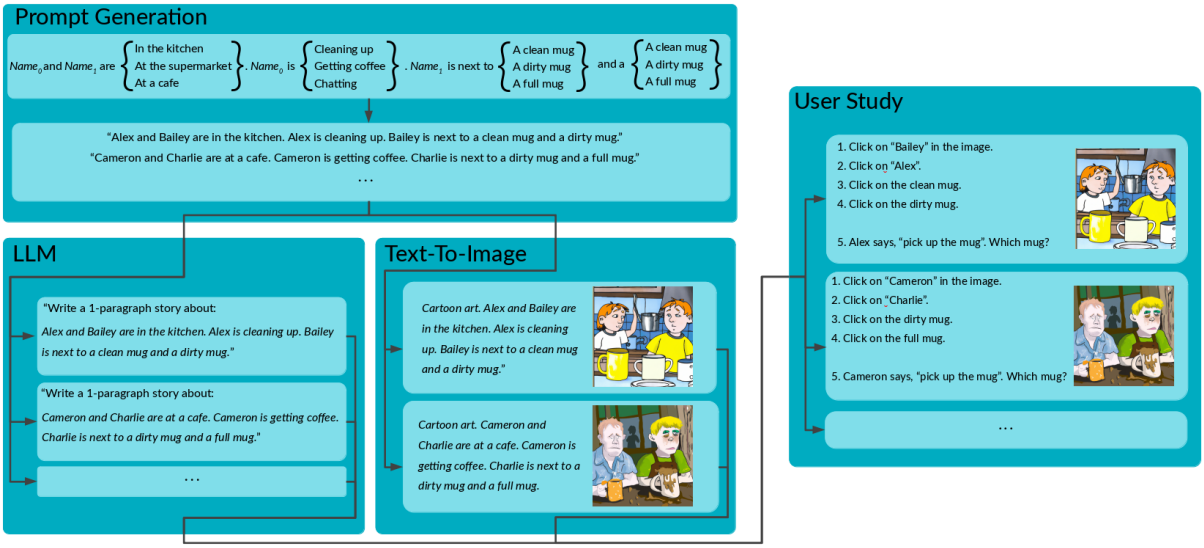

Figure 1: Provided approach deployed on motivating example.

a subset of an existing dataset for generating data (allowing for the explicit circumvention of dataset biases). Finally, the direction of our pipeline flows from prompt with specific constraints, to a generated dataset, to a final human validation step.

An exception is Chakrabarty et al. (2023), who present a similar pipeline flow through a collaboration of LLMs and diffusion-based text-to-image models to generate a dataset of visually metaphoric images from linguistic metaphors. Their generated images are evaluated by expert evaluators to rank the LLM-Diffusion Models and provide recommended feedback. This human evaluation is necessary as it helps to measure how well the image represents a metaphor. However, their pipeline diverges in a few distinct ways from ours. First, the original text for their pipeline (linguistic metaphors) are harvested from existing datasets. Our approach uses a more general, structured, and open-ended prompt construction template as a starting point. Second, expert human annotators provided feedback on how well the model represented the visual metaphor whereas our human-in-the-loop part of the pipeline is meant for general interpretation discoveries and ground truth annotations.

Our pipeline, therefore, is primarily focused on allowing researchers to have control over creating a *resource* with specific visual context that can be paired with language. This addresses an unmet need in various applications; in the reference resolution domain, one might want to study how people resolve ambiguous references and how pragmatic factors like social-normative constraints govern those interpretations. Thus, this task critically requires reference tasks where there are environmental ambiguities against a backdrop of a visual scene evoking a social-normative context. The multi-modal MS-COCO dataset (Lin et al., 2014) offers static images for referring expression grounding, and the VisDial v0.9 dataset for reference resolution in visual dialogue (Das et al., 2017). Yet these resources offer little environmental ambiguity and potentially insufficient examples with desired specifications; one can filter but not customize as needed[1]. Another promising path is creating reference scenes in virtual environments like the AI2Thor (Kolve et al., 2017) platform or the SIMMC 2.0 dataset (Kottur et al., 2021), built specifically for referential ambiguities in a virtual environment. But it can be time consuming to build such environments, especially with rendering a resource with extensive variation.

3. Technical Approach

Our proposed approach, summarized as Figure 1 and with source code provided[2], is to construct a generalized software toolkit that will:

1. Programmatically generate scene descriptions;
2. Add detail using an LLM;
3. Construct accompanying images using a generative image model; and
4. Ground items and get results in a user study.

The emergence of Large Language Models (LLMs) (e.g., ChatGPT (OpenAI, 2022), BERT (Devlin et al., 2019), BART (Lewis et al., 2020), LLaMA (Touvron et al., 2023), PaLM (Chowdhery et al., 2022)) enable the rapid production of text (and,

[1] One might desire a context where there are two mugs on a table—one clean and one dirty—with people chatting around the table. And with the same scene, the larger context can be modulated from a dining room to a cafe to a restaurants. It would be challenging to find this exact specification in existing image datasets

[2] https://github.com/hrilabtufts/data_gen_pipeline

if the appropriate textual scaffolding is provided, this text can be very high quality). Similarly, models which allow the production of images from textual inputs have seen rapid growth (e.g., DALL-E (Ramesh et al., 2021), Stable Diffusion (Rombach et al., 2022), IMAGEN (Saharia et al., 2022)). This, combined with platforms for human intelligence tasks[3], are what make our proposed pipeline viable. In this work, we specifically use ChatGPT, DALL-E, and Prolific, though certainly combination are feasible.

As an input to the system, we begin by constructing a series of problems which we are interested in exploring. In Figure 1, we imagine a case where there are ambiguous references to an object grounding problem, and the researchers would like empirical knowledge of what the correct referent is in these situations. As we will discuss in further detail, recent research (Abrams and Scheutz, 2022) has suggested that ambiguous references can be resolved (in part) by making use of social norms which may change based off of context: different settings, roles, and details about objects in the scene may change the appropriate referent. Thus, we generate a set of prompts representing these differing scenarios: while this could certainly be done by-hand, doing so would be time consuming, and so we generate these programmatically.

Each of these prompts contain the minimum set of necessary information to describe a scene: they contain basic information about people, objects, locations, etc. However, they are not immersive scenes that participants can imagine as a realistic setting. To produce this, we make use of an LLM to expand the prompt into a more verbose/descriptive format. Modern LLMs are capable of taking length constraints (e.g., "in one paragraph"), and so these descriptions can be kept fairly brief. For example, the consider the phrase:

"Alex and Bailey are in the kitchen. Alex is cleaning up. Bailey is next to a clean mug and a dirty mug."[4]

With the prompting to "describe this scene in one sentence", a more descriptive (yet still succinct) phrase is produced:

"In the kitchen, Alex takes charge of cleaning up while Bailey stands beside a clean mug and a dirty mug, contemplating their next move amidst the contrasting states of cleanliness."

This description retains the key information from the prompt, while remaining more suitable for presentation to participants as a scene. Importantly, it removes much of the monotony and repetition that these prompts may contain, encouraging participants to pay closer attention and resulting in higher quality responses.

Many of these LLMs, including ChatGPT, provide software interfaces (an Application Programming Interface, "api") that allow interaction with these systems from custom-built programs. Making use of these interfaces allows the programmatically-generated prompts to be directly converted into more descriptive formats, without need for time-consuming intervention from the researcher.

In addition to textual descriptions, images provide helpful accompanying data. Unfortunately, these models often struggle to create accurate images of human features. We work around this by providing them with a slightly modified prompt: by including the phrase "cartoon art", prompts are made in a less-precise style that retains the necessary points while minimizing distracting detail. This adjustment, as well as the image generation, is again automated to ensure minimal intervention from the researcher.

The approach concludes with a within-subjects user study to gather baseline information about each image. To accomplish this, participants select a point on the image in response to specific questions about the scene. By comparing these points across participants, we can determine the usefulness of an image (for example, if there is low consistency between the location of the "clean mug", this image prompt likely is failing to provide useful data). Additionally, this baseline data provides an indicator of which reference participants felt was the appropriate selection. Each image prompting set can then conclude with the question of interest to the researcher: in this case, "Alex says 'Pick up the mug'. Which mug?".

While surveys can be constructed by-hand, this is again time consuming for the researcher. However, we observe that the "Amazon Mechanical Turk" and the "Prolific" platforms, which are widely used for online human-intelligence tasks, both provide interfaces for constructing studies programmatically. With this final step, the full process of generating and constructing a full human-intelligence task can be largely automated, leaving the researcher with the final step of reviewing and approving for data collection.

4. Applications

To demonstrate the efficacy of this approach, we consider three unique applications[5].

Finding Contexts That Influence Reference Resolution. In Abrams and Scheutz (2022), the authors argue that shift in context (e.g., being in a taxicab vs. being in a friend's car) introduces different norms that modulate reference resolution (e.g., "take a seat" corresponding to the back or front

[3] e.g., Amazon Mechanical Turk or Prolific

[4] We prefer to use gender-neutral names to avoid confusion when images are generated.

[5] This footnote will be replaced with links to sample data after review.

(a) Heatmap produced when users are asked to select the speaker

(b) Heatmap produced when users are asked to select the referent

Figure 2: Converting generated images to useful data using a user study.

seat). However, owing to the difficulty of constructing text/image vignettes to evaluate these cases, only four competing contexts are evaluated. As a result, while the authors provide evidence to support their claim, they are unable to find a larger set of contexts and how they may modulate norms. Our approach can fill this gap by producing a set of possible contexts, and the accompanying text/image vignettes to explore them.

To demonstrate the feasibility of our approach, we constructed a small user study revolving around our motivating example. Ethical compliance was ensured by a human-subject IRB. We recruited 20 participants on Prolific (https://prolific.co). The participants were anonymized, received adequate payment for the study, and were free to abandon the study at any point. A consent form was presented at the beginning of each study with information on how their data would be used. We restricted our recruitment to people living in the United States and at the time of the study and native speakers of English.

The participants followed the experiment outlined in our motivating example: over the course of several variations of the "pick up the mug" instruction, participants saw a series of varying scenes where they were asked to identify each referent and to find the object being referred to in the scene. Participants select the speaker with little ambiguity (Figure 2a). When provided with an utterance and prompted to select the referent, there is again little ambiguity (Figure 2b).

However, this data is meaningless in isolation. It is for this reason that each participant was asked to first select unambiguous references: the data that "participant x thinks point y is the clean mug" allows us to resolve the selected points to the concrete referents, resulting in the final data points desired from the study. Had there been a low amount of agreement about the referent, it would have allowed us to explicitly identify and remove these prompts as not being constructive.

Discovering Frames. Our pipeline can facilitate the study of structured expectations across various contexts, a notion very central to frame analysis. Tannen (1993), in a classic sociolinguistic study, provides qualitative evidence that cultural backgrounds influence perception of linguistic *frames*—or, structures of expectation—and shows what these frames consist of. In the described experiment, a short film produced by the researchers was shown to participants who would later describe what happened in the film to another person who hasn't seen it. Evidence for such frames was realized in how participants linguistically altered objects and events from the actual story content. This, in turn, revealed expectations of the objects and events in the film. This approach allowed for valuable qualitative insights, but the high time cost of producing such a film limits the researcher's ability to explore a broader set of scenes. Images, alternatively, can also serve as the subject of narratives for frame analysis.

Our approach can generate textual-visual vignettes for specific narratives and test particular expectations around events similar to the study (e.g. *confrontations, theft, accident, personal encounters*). On a quantitative level, generating textual-visual vignettes can help to experimentally gather and validate semantic frames for FrameNet (Fillmore and Baker, 2009) and more appropriately, its extension into the multimodal domain (Belcavello et al., 2020).

Generating Scenes For Creative Visual Story Telling. Expectations can also be explored through the Creative Visual Story Telling domain, where textual narratives are created—either by a human or a language model—based on an images or series of images. Lukin et al. (2018), for instance, present a pipeline to train a computational model on three subtasks to achieve creative story telling: object identification, single-image inferencing, and multi-image narration. As Lukin et al., points out, high-quality real images are typically used for this domain (Plummer et al., 2015; Lin et al., 2014), yet this might not cover all domains. This paper particularly focuses on a low-resource domain of an "environment with odd surroundings taken from a camera mounted on a ground robot" (Lukin et al., 2018). Our pipeline complements this computational story telling pipeline, but generate images and captions to cover other low-resource domains—one's that don't necessarily require high-quality images. The narratives collected on this data can also be explored qualitatively through discourse analysis to discover expectations when telling a story through multi-image narration.

5. Conclusion

Our proposed approach makes it possible to create novel text-image resources at scale. We show that for specific applications related to visual grounding, frame analysis, and visual storytelling, one can generated a dataset prompted with specific constraints for human validation.

Limitations

While we are encouraged by the potential of this approach, there are some limitations stemming from these technologies that prevents it from being useful for some research questions.

Bias and related ethical concerns. LLMs and image generation models are frequently known to have problems with bias (for a recent exploration, see Salewski et al. (2023)), and will inherit problems from the underlying data. Further, the question of intellectual property attribution from model-generated data remains open at time of writing. The eventual answer to this question will have substantial implications on the feasability of this and similar approaches.

Modalities. Datasets produced by this approach are limited by the modalities available to various automation systems. Video, for example, is not yet a suitable approach: although impressive progress has been made in the last several years (Singer et al., 2022; Balaji et al., 2019; Wu et al., 2021; Hong et al., 2022), these videos are generally fairly short and low-precision (especially when multiple specific details are required). Even in the image modality, generating high volumes of high-quality images remains challenging, producing the need for text accompanyment. Thus, high-precision tasks (e.g., tasks exploring human timing or tasks requiring high accuracy images) are also not suitable as a result.

For this reason, we view the set of useful problems as datasets that can be described as text-only or text-image vignettes; and datasets that do not require high precision.

Text-image vignettes are uniquely positioned among other synthetically generated NLP data in that they can represent a relationship between language and visual context. This is particularly important for grounding or exophoric reference resolution problems where a visual referential scene and context can be modulated.

Ethics Statement

Our proposed approach has both positive and negative ethical implications, and so deployment must be done with care. First, as previously discussed, these approaches inherit ethical issues from LLMs and similar models. Data bias may creep into the produced question set, and therefore may have an impact on the produced data. In addition to being a research limitation, this has the potential to produce or reinforce biased or incorrect scientific results.

Beyond issues with data bias, the safety of the content produced cannot be guaranteed by researcher-provided prompts. Model safety and counter-bias approaches continue to be an area of active research, and cannot be considered a solved problem. For both of these issues, it remains necessary for researchers to carefully consider and review the data produced before presenting it to participants, or before presenting results to fellow researchers.

With careful deployment and continued development of model safety, however, these concerns can be substantially reduced. Doing so allows researchers to take advantage of positive ethical impacts of this approach: in particular, the ability to rapidly obtain human-validated results of niche content areas allows underrepresented areas of research to be more effectively explored.

6. Acknowledgments

This work was in part funded by ARO under contract W911NF2220252 and AFOSR grant FA9550-23-1-0425.

7. Bibliographical References

Mitchell Abrams and Matthias Scheutz. 2022. Social norms guide reference resolution. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1–11.

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2020. Do not have enough data? deep learning to the rescue! In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7383–7390.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multi-

ple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L_1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Anthony McEnery and others. 2004. *The EMILLE/CIIL Corpus*. EMILLE (Enabling Minority Language Engineering) Project. distributed via ELRA: ELRA-Id W0037, ISLRN 039-846-040-604-0.

Yogesh Balaji, Martin Renqiang Min, Bing Bai, Rama Chellappa, and Hans Peter Graf. 2019. Conditional gan with discriminative filter generation for text-to-video synthesis. In *IJCAI*, volume 1, page 2.

Frederico Belcavello, Marcelo Viridiano, Alexandre Diniz da Costa, Ely Edison da Silva Matos, and Tiago Timponi Torrent. 2020. Frame-based annotation of multimodal corpora: Tracking (a) synchronies in meaning construction. In *Proceedings of the International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet*, pages 23–30.

BSI. 1973a. *Natural Fibre Twines*, 3rd edition. British Standards Institution, London. BS 2570.

BSI. 1973b. Natural fibre twines. BS 2570, British Standards Institution, London. 3rd. edn.

A. Castor and L. E. Pollux. 1992. The use of user modelling to guide inference and learning. *Applied Intelligence*, 2(1):37–53.

Tuhin Chakrabarty, Arkadiy Saakyan, Olivia Winn, Artemis Panagopoulou, Yue Yang, Marianna Apidianaki, and Smaranda Muresan. 2023. I spy a metaphor: Large language models and diffusion models co-create visual metaphors. *arXiv preprint arXiv:2305.14724*.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Yixin Chen, Qing Li, Deqian Kong, Yik Lun Kei, Song-Chun Zhu, Tao Gao, Yixin Zhu, and Siyuan Huang. 2021. Yourefit: Embodied reference understanding with language and gesture. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1385–1395.

J.L. Chercheur. 1994. *Case-Based Reasoning*, 2nd edition. Morgan Kaufman Publishers, San Mateo, CA.

N. Chomsky. 1973. Conditions on transformations. In *A festschrift for Morris Halle*, New York. Holt, Rinehart & Winston.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

James W. Cooley and John W. Tukey. 1965. An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19(90):297–301.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Umberto Eco. 1990. *The Limits of Interpretation*. Indian University Press.

Charles J Fillmore and Collin Baker. 2009. A frames approach to semantic analysis.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Xuanli He, Islam Nassar, Jamie Kiros, Gholamreza Haffari, and Mohammad Norouzi. 2022a. Generate, annotate, and learn: Nlp with synthetic text. *Transactions of the Association for Computational Linguistics*, 10:826–842.

Xuanli He, Islam Nassar, Jamie Kiros, Gholamreza Haffari, and Mohammad Norouzi. 2022b. Generate, Annotate, and Learn: NLP with Synthetic Text. *Transactions of the Association for Computational Linguistics*, 10:826–842.

Paul Gerhard Hoel. 1971a. *Elementary Statistics*, 3rd edition. Wiley series in probability and mathematical statistics. Wiley, New York, Chichester. ISBN 0 471 40300.

Paul Gerhard Hoel. 1971b. *Elementary Statistics*, 3rd edition, Wiley series in probability and mathematical statistics, pages 19–33. Wiley, New York, Chichester. ISBN 0 471 40300.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. 2022. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*.

Otto Jespersen. 1922. *Language: Its Nature, Development, and Origin*. Allen and Unwin.

Chris Kedzie and Kathleen McKeown. 2019. A good sample is hard to find: Noise injection sampling and self-training for neural language generation models. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 584–593, Tokyo, Japan. Association for Computational Linguistics.

Khalid Choukri and Niklas Paullson. 2004. *The OrienTel Moroccan MCA (Modern Colloquial Arabic) database*. distributed via ELRA: ELRA-Id ELRA-S0183, ISLRN 613-578-868-832-2.

Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, et al. 2017. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*.

Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 2021. Simmc 2.0: A task-oriented dialog dataset for immersive multimodal conversations. *arXiv preprint arXiv:2104.08667*.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and Larry Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*. European Conference on Computer Vision.

Ruibo Liu, Guangxuan Xu, and Soroush Vosoughi. 2020. Enhanced offensive language detection through data augmentation. *arXiv preprint arXiv:2012.02954*.

Stephanie Lukin, Reginald Hobbs, and Clare Voss. 2018. A pipeline for creative visual storytelling. In *Proceedings of the First Workshop on Storytelling*, pages 20–32, New Orleans, Louisiana. Association for Computational Linguistics.

Khyati Mahajan, Soham Parikh, Quaizar Vohra, Mitul Tiwari, and Samira Shaikh. 2022. Improving dialogue act recognition with augmented data. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 471–479.

OpenAI. 2022. ChatGPT: Optimizing Language Models for Dialogue. https://openai.com/blog/chatgpt/.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Joint conference on EMNLP and CoNLL-shared task*, pages 1–40.

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.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695.

Roventini, Adriana and Marinelli, Rita and Bertagna, Francesca. 2016. *ItalWordNet v.2*. ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council, in Pisa, ISLRN 532-206-426-067-2. PID http://hdl.handle.net/20.500.11752/ILC-62. Note: You don't really need both an ISLRN and another PID, but it can't hurt.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494.

Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. 2023. In-context impersonation reveals large language models' strengths and biases. *arXiv preprint arXiv:2305.14930*.

Charles Joseph Singer, E. J. Holmyard, and A. R. Hall, editors. 1954–58. *A history of technology*. Oxford University Press, London. 5 vol.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. 2022. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*.

Speecon Consortium. 2011. *Catalan Speecon database*. SpeeCon. Speecon Project, distributed via ELRA: ELRA-Id S0327, Speecon resources, 1.0, ISLRN 935-211-147-357-5.

Jannik Strötgen and Michael Gertz. 2012. Temporal tagging on different domains: Challenges, strategies, and gold standards. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3746–3753, Istanbul, Turkey. European Language Resource Association (ELRA).

S. Superman, B. Batman, C. Catwoman, and S. Spiderman. 2000. *Superheroes experiences with books*, 20th edition. The Phantom Editors Associates, Gotham City.

Deborah Tannen. 1993. What's in a frame? surface evidence for underlying expectations. *Framing in discourse*, 14:56.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. 2021. Godiva: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806*.

Nan Xu, Wenji Mao, Penghui Wei, and Daniel Zeng. 2020. Mda: Multimodal data augmentation framework for boosting performance on sentiment/emotion classification tasks. *IEEE Intelligent Systems*, 36(6):3–12.

Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji-Ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. 2020. Generative data augmentation for commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1008–1025, Online. Association for Computational Linguistics.