Harnessing Knowledge and Reasoning for Human-Like Natural Language Generation: A Brief Review

Jiangjie Chen
Fudan University
jjchen19@fudan.edu.cn

Yanghua Xiao
Fudan University
shawyh@fudan.edu.cn

Abstract

The rapid development and application of natural language generation (NLG) techniques has revolutionized the field of automatic text production. However, these techniques are still limited in their ability to produce human-like text that is truly reasonable and informative. In this paper, we explore the importance of NLG being guided by knowledge, in order to convey human-like reasoning through language generation. We propose ten goals for intelligent NLG systems to pursue, and briefly review the achievement of NLG techniques guided by knowledge and reasoning. We also conclude by envisioning future directions and challenges in the pursuit of these goals.

1 Introduction

Language, as the vehicle of thought [128], is one of the most fundamental means for humans to reason and communicate with each other. Hence, the technology of natural language generation (NLG) has always been one of the main focuses throughout the history of AI research [126]. In the age of modern deep learning, NLG techniques have been widely deployed in real-life applications, including machine translation [130], news reporter [134], dialogue systems [115], automatic report generation [40], etc. Beyond that, NLG models also serve as a rather universal workhorse in other types of NLP tasks, such as structured prediction [85].

The achievement so far for NLG research has enabled the wide application of NLG techniques, but the dangers beneath them are still far from being resolved. The foundation of NLG models has shifted over the years, from rule-based models [62, 81] to statistical models [27], and now at pre-trained language models (PLMs) [92, 34, 93, 64, 11]. However, rule-based models are too rigid to generate natural language. Statistical models based on neural networks and training datasets suffer from limitations such as reporting bias, exposure bias, generalization issue, etc. Recent years, PLMs have shown their excellent capabilities of understanding and generating natural language. However, the research on PLMs does not necessarily solve the fundamental problems NLG, as the alleviation of these problems comes from exposing the models on much more data with self-supervised training. Moreover, multiple problems still occur, including model explainability [72], hallucination [95, 132], ethical risks [117, 148], logical inconsistency [60, 36], etc.

Towards these challenges, the research community has formulated an important perspective with symbolic knowledge, covering rules [55], commonsense knowledge [109], world knowledge [2, 120], etc. Since symbolic

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

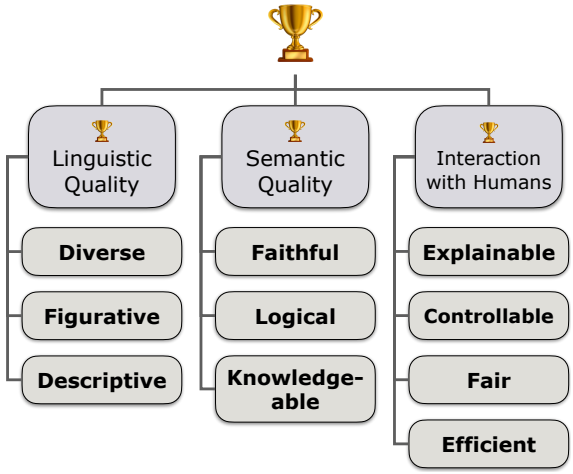

Figure 1: The taxonomy of the goals in NLG.

knowledge can be used to describe complicated concepts and their connections, it would be easier for the models to comprehend the textual world. Therefore, symbolic knowledge brings significant opportunities towards NLG models in these respects: 1) symbolic knowledge provides rich background materials for NLG models to generate from; 2) symbolic knowledge serves as regularization to NLG models for expressing constraints and desired properties related to a given task; and 3) the formal rules and structures that symbolic knowledge established can be used to improve the reasoning ability of NLG models.

In this paper, we stress the importance and necessity of NLG techniques to be guided by knowledge, so that human-like reasoning can be conveyed through language generation. To this end, we propose ten goals for intelligent NLG systems to pursue (§ 2), Next, we briefly review the achievement of NLG techniques with knowledge (§ 3), picturing the mutual enhancement of guiding NLG with knowledge and acquiring knowledge with NLG. Then, we summarize how to make rational usage of knowledge to approach human-level reasoning with NLG techniques (§ 4), showing reasoning-guided NLG and how NLG can be used to chain up reasoning. Finally, we conclude this paper and envision future directions and challenges in the pursuit of these goals (§ 5).

2 Holy Grails in NLG

To begin with, we list ten most-desired goals (with room for more) for machine-generated text, in order to build an AI system that can freely interact and communicate with humans with language. We categorize these goals into three classes based on the linguistic quality of generated text, conveying semantic information with generated text and the interaction with humans, as shown in Figure 1. These goals are still quite challenging for modern NLG methods. While the research about most of them is still in its infancy, some of these goals have been moderately explored by the community. Since symbolic knowledge offers control of the process and outcome of generated text, many of research endeavors prove that symbolic knowledge plays an important role in steering NLG techniques towards these goals.

Diverse Given a source input, an ideal NLG model needs to be able to generate multiple and diverse outputs, where each of them needs to be equally valid. For example Diversified NLG is a practical goal, especially in the context of real-life applications such as dialogue generation [129], machine translation [101], headline or query generation in e-commerce [102] and text paraphrasing [103]. Advanced NLG models should be able to generate diverse but informative text from the large sentence space, especially involving diverse but relevant background knowledge in the context.

Figurative The generative text should be figurative to build up emotional significance to the readers. Figurative text typically includes idioms, sarcasm, simile, metaphor, etc. All of them are great ingredients in interesting and creative writing tasks such as story generation or narratives. For example, the sentence that "he is drowning in a sea of grief" expresses strong negative emotions to the reader, where *grief*, with the metaphor of a *sea*, vividly overwhelms *him*. To master the ability to write figurative text is non-trivial, and recent studies show that even strong modern language models still struggle at this objective [47, 70, 22]. Nevertheless, research shows that such a problem can be alleviated with knowledge-enhanced models, which enrich the context and constituents of figurative text with acquired knowledge [18].

Descriptive Advanced NLG techniques need to be descriptive, that is, vivid and colorful as if something is being experienced by the readers. Descriptive texts are rather common in books and novels, fascinated by image-like texts. For example, to describe the sunset: "*the sunset filled the entire sky with the deep color of rubies, setting the clouds ablaze.*"[1] It is worth noting that, unlike previous desired properties that interact mostly with textual data, a system needs to integrate multi-modal knowledge and reasoning (e.g., image, video, etc.) to be descriptive with visual-specific features [138, 105, 106]. However, the exact definition of descriptiveness in the context of the machine-generated text as well as its automatic evaluation are still great challenges.

Faithful The generated text should be faithful to the input so that it correctly conveys and extends the input information without semantic violation. Otherwise, the credibility of an NLG system could be undermined when hallucinated texts are generated, which could be dangerous sometimes. There is a growing interest in enhancing and evaluating the (intrinsic or extrinsic) faithfulness of generated text in text summarization [17, 80, 63], dialogue systems [49], text simplification [33], etc. However, hallucinated texts are not always non-factual, because they may be faithful towards world knowledge [14]. Therefore, an NLG system should be faithful and can be verified by world knowledge [116, 20], except for applications such as fictional story generation.

Logical Logical reasoning is an essential part of human thinking and language; therefore, the generated text needs to be logically consistent and self-contained. Different from faithfulness which focuses on information consistency, logical consistency poses a challenge over the discourse of produced language of an NLG system [5, 25, 104, 107, 87]. However, the training objectives of current prevailing language modeling (e.g., masked language modeling and causal language modeling) prioritize recovering given text, which does not guarantee the logical reasoning ability to be effectively captured by models.

Knowledgeable An NLG model needs to be knowledgeable to generate text that is rich in knowledge [139]. When engaging in a conversation, current NLG models are known to be dependent on plain but safe responses. A knowledgeable NLG system, in contrast, should be able to actively initiate responses grounded with the background knowledge that is either retrieved [65] or inherent [75] in the system itself. Also, it is worth exploring whether current NLG models are using knowledge of their own or just relying on statistical patterns learned from (pre-)training [13].

Explainable An NLG system should be explainable for the trust of human users. Since language is the most natural tool for communication, the reasoning and decisions of a model should be explanatory through generated language [96]. Existing work usually focuses on generating natural language explanations [127] as a task, while how to develop universal explanation generators [137] is still challenging. More importantly, unlike natural language understanding models (NLU) [97, 100, 111], the explainability of NLG systems is severely underexplored.

[1] https://rescuewriting.org/featured/writing-descriptive-text/.

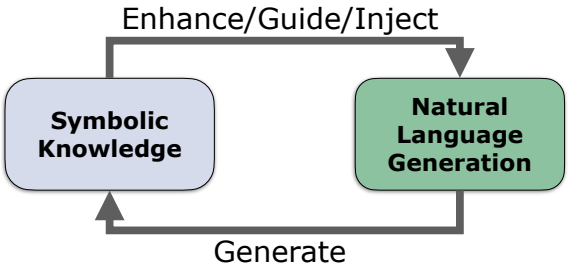

Figure 2: Relations between knowledge and NLG, where knowledge can enhance, guide and be injected into NLG systems, and NLG systems can be utilized as tools to generate new knowledge.

Controllable Controllable NLG aims to generate with desired attributes by users, which greatly broadens its applications. Current commonly used attributes focus on discriminative attributes such as sentiment and style, lexical constraints, and various properties of language such as lengths or complexity [50, 141, 37]. Moreover, there are still many interesting open questions for building an ideal NLG system, for example, more diverse applications of controlling factors, satisfying multiple types of constrained attributes (perhaps a mixture of them), and being controlled in a few-shot or even zero-shot manner. It is worth noting that recent studies in large language models [11] shed some light on these research problems.

Fair The generated text should be fair and contain no bias whatsoever. Methods to mitigate biases have been proposed w.r.t. gender, race, nationality, age, disability, religion, etc. [71, 131, 45]. Since text generation models have been widely applied in the age of the Internet, their outstanding performance could blind service providers so as to omit the negative social impacts that an unfair NLG system causes. However, since modern NLG systems are usually pre-trained on colossal unannotated corpus [92, 93, 64, 11], the biases within them are still rather difficult to eliminate.

Efficient An NLG system should be efficient to be deployed on high-demanding scenarios such as mobile devices or online industrial applications. Recent work on non-autoregressive generation (NAG) [41, 133], which generates tokens in parallel, has shown promising potential in efficiency. This enables NAG to achieve over $10\times$ speedup over the autoregressive counterpart (AG) that generates in a left-to-right manner [27]. However, NAG is still haunted by the multi-modality problem [145], its generality to tasks other than machine translation [89], and interaction with external knowledge [140].

3 Knowledge in Language Generation

In this section, we briefly review some representative work in knowledge-guided NLG algorithms and knowledge acquisition with NLG models, where NLG systems also show unique advantages in the latter. Figure 2 shows the relation between knowledge-guided NLG (§ 3.1) and knowledge acquisition with NLG (§ 3.2), where the two are mutually beneficial to each other.

3.1 Knowledge-guided NLG: Tricks of the Trade

NLG systems can be guided by multiple sources and types of knowledge to be logical, diverse, faithful, controllable, etc. In the following text, we list the knowledge sources that are commonly used, and showcase the common tricks of current knowledge-guided NLG algorithms, presenting a brief but as comprehensive review as possible.

Knowledge Sources According to Yu et al. [139], knowledge sources used in NLG systems can be categorized into internal and external knowledge, where: 1) internal knowledge is created within the given text, e.g., keywords, topics, linguistic features, etc., and 2) external knowledge exists outside the input of the NLG system, e.g., knowledge bases/graphs, grounded unstructured text. We add that knowledge mined from the corpus is also important to guide NLG systems, such as concepts, rules, and patterns [21]. Notably, NLG systems form an interesting dual learning loop where knowledge can be acquired by NLG systems (§ 3.2) and used in NLG systems [110, 16]. Next, we will detail some common practices of NLG models that involve knowledge.

Knowledge Incorporation in Model Architectures Modifying model architectures to incorporate knowledge is one of the most commonly used approaches in knowledge-guided NLG [136, 139]. Typical methods include using attention mechanism [3] (and its variants, e.g., copying mechanism, [42]) to attend additional knowledge sources, encoding knowledge with specific encoders (e.g., graph neural networks, [61]), adding specific layers in the neural network for knowledge storage (e.g., adapter layers, [44]), etc. In this line of work, a model trains to utilize the encoded knowledge for better NLG. However, such methods usually require the design of an ad hoc model architecture, which does not generalize well to new tasks, especially in the age of large PLMs [11].

Learning with Knowledge The most common method of knowledge-guided NLG is through supervised learning, where the guidance of knowledge can be well integrated into generation [139]. A simple way to do this is to append the acquired knowledge to the input as an additional context. This has been shown to be effective, especially when knowledge acquisition (e.g., retrieval) is jointly trained with generation [46, 65]. Another way is to guide learning objectives with knowledge. Studies in this direction are carried out by designing knowledge into one of the training targets [16] or pre-training [43], or by introducing additional knowledge-related objective functions to regularize training [51]. Also, reward design in a reinforcement learning (RL) framework is a flexible way of integrating knowledge into training objectives for NLG models. For an example of improving the consistency between source and target, [67] injects the entailment knowledge into NLG with an entailment model, and [53] designs a semantic cloze reward for faithful summarization.

Knowledge-Constrained Decoding Finally, we would like to emphasize knowledge incorporation during decoding. Knowledge injection through model architecture modification and new learning objectives, despite being effective, requires (re-)training. This usually causes inconvenience for deployment, especially with large PLMs. For this reason, constraining text decoding in inference time with knowledge has become a promising research topic [141, 37]. In this line of work, we briefly introduce the three most exemplary knowledge constraint types in NLG systems: 1) lexical knowledge constraints, such as keywords that must appear in the text, 2) discriminative constraints, such as the desired attributes of the generated text, and 3) structured knowledge constraints, such as the world knowledge that the generated text must be consistent with.

Imposing *lexical constraints* is challenging for NLG systems because the prevailing ones usually generate in an autoregressive manner, which is difficult to know when and where to keep the constraints. Popular solutions to this problem mainly focus on adding constraints during beam search [48, 88, 76] at the cost of computational complexity. Recent work [112, 135] also explores iterative lexically constrained decoding for non-autoregressive models, achieving significant speedups.

Discriminative constraints usually involve prior attribute models to control the generated text in terms of attributes such as topics (sport, finance, medical), sentiments (positive, negative), etc. Exemplary work is PPLM [30], which controls the attribute of generated text without changing the parameters of the backbone PLM (e.g., GPT-2, [92]). During sampling, PPLMs back-propagates the gradient from the attribute model to the hidden states of the PLM so that the generation can be steered by the attribute model. Under such a framework, it is easy to guide generation with prior knowledge. It is worth noting that work on text sampling is also a great framework to effectively incorporate the above-mentioned two types of constraints. In this framework, the desired

properties are designed as objective functions, guiding the sampling process, which is usually instantiated with Markov chain Monte Carlo methods [82, 143, 91]. Similar to PPLM, such methods ensure the generated text is constrained to the desired objectives without training.

Recent studies have also started to pay attention to the hallucination problem of NLG systems (represented by PLMs) from the perspective of constrained decoding with *structured knowledge constraints*, such as knowledge graphs. For example, [73] retrieves from knowledge graphs related to neighboring entities within source input. They are used to modify and constrain the distributions over the vocabulary during sampling, where tokens within knowledge graphs are rewarded. In this way, the generated text would be rich in semantic knowledge and faithful to world knowledge, making it an interesting and effective solution to the hallucination problem that is common in NLG.

3.2 Knowledge Acquisition with NLG

Knowledge acquisition based on NLG systems aims to generate knowledge from the input text. Compared with other paradigms of knowledge acquisition, automatic and generalizability to unseen knowledge are one of the most desired features of generative methods. In the following text, we first discuss current knowledge acquisition paradigms and show the advantages of generative methods in certain scenarios.

Paradigms of Knowledge Acquisition The knowledge acquisition methods can be divided into three categories: 1) crowd-sourcing, 2) extractive and 3) generative methods. The *crowd-sourcing* methods [83] invite some experts to label the knowledge in the corpus, which can acquire accurate knowledge but only have a small size due to costly annotation. The *extractive* methods [147] adopt language models to extract the knowledge explicitly mentioned in the input text automatically but ignore some symbolic and neural commonsense knowledge. Compared with them, generative knowledge acquisition methods enjoy the advantages of the ability to generate explicit and implicit knowledge, which we will soon discuss.

Meta Knowledge Generation Meta knowledge is knowledge about knowledge [31]. As a fundamental conceptual instrument in knowledge-based domains, meta-knowledge can greatly improve the performance of downstream tasks, such as text classification [23], question answering [38] and story generation [19]. However, meta-knowledge is hardly explicitly mentioned in the corpus, and thus it is difficult to directly extract meta-knowledge from the text. Alternatively, since NLG systems can generate information never mentioned in the text and has a strong generalization ability of unseen knowledge, recent work proposes to adopt generative methods to acquire meta knowledge, such as concepts [21, 66] and rules [24].

Generative Knowledge Retrieval Information retrieval aims to retrieve meaningful information from large Knowledge Bases (KB) given a textual input. Compared to extractive methods, generative information retrieval [15, 32, 98] can directly capture the relation between context and target information and reduce the memory footprint without negative data down-sampling.

Generative Knowledge Extraction Information extraction aims to obtain information from semi-structured and unstructured text, which suffers heterogeneous structures and domain-specific schemas [79]. The generative information extraction can end-to-end generate targeted structures directly. For example, [54] adopts autoregressive models to achieve the extraction of end-to-end relations. [78] proposes a sequence-to-structure generation method to directly extract events from the text. [52] formulates entity-based extraction as a template generation task to allow the generative framework to effectively capture cross-entity dependencies.

Knowledge Generation NLG models can also be used for the completion of knowledge bases, including commonsense knowledge graphs and rules. Instead of extracting semi-structured and unstructured text into knowledge, some work [9, 55] feeds large-scale language models with a massive corpus to obtain knowledge models, which can adapt their learned representations to knowledge generation and automatically construct KBs. Furthermore, the parameters of PLMs are shown to store vast amounts of linguistic knowledge [114]. A line of work further regards PLMs as knowledge bases and distills semantic knowledge from these models [86, 1].

4 Reasoning in Language Generation

We echo the argument that good reasoning should be right for the right reasons. Therefore, it is also crucial for an NLG model to make *rational* usage of knowledge to approach human-like reasoning skills. This section signifies the importance of reasoning in NLG, where we sketch two lines of research: 1) reasoning-guided NLG methods (§ 4.1) and 2) NLG for the purpose of reasoning (§ 4.2).

4.1 Reasoning-guided NLG

In contrast to general knowledge-guided NLG, reasoning-guided NLG systems make more rational and explainable usage of the knowledge (in wide forms). Since reasoning is highly correlated with knowledge (§ 3), we will discuss reasoning-guided NLG by highlighting the topics of graph reasoning and generative reasoning tasks in the following paragraphs.

Graph Reasoning Graph reasoning is one of the most commonly used techniques to guide NLP systems towards more multihop, controllable, and explainable reasoning. Therefore, we extend what has been discussed in § 3.1 in this paragraph for introducing graph reasoning-guided NLG. Graph reasoning implementations are usually built on retrieved subgraphs of external knowledge graphs [74] or internal graphs parsed from the input [121]. Most of them adopt graph embeddings [8] and graph neural networks [61, 119] to propagate information throughout the graph. These properties of graphs enable the multi-hop reasoning ability of NLG models for long-range text [57] and generating emerging concepts or topics guided by the graph [122, 142]. Due to the symbolic structure of graphs, heuristics and prior knowledge can be easily incorporated into graph construction, e.g., building graphs with various parsing tools [39] such as semantic role labeling or dependency parsing or guiding graph propagation with more information, such as popularity knowledge [102]. Also, such methods are explainable and easy to debug, since the weights on the graph nodes and edges greatly facilitate post hoc manual examination.

Reasoning Tasks One of the most direct ways to enable NLG models to reason is to design various reasoning tasks. During the solving of these tasks, researchers can develop and test their models w.r.t. corresponding reasoning skills, which makes it an important direction to guide AI models. Over the years, the community has accumulated many datasets of tasks that test various facets of machine reasoning in the form of text generation. Most of them are built from the point of view of human cognition, including tasks about logical reasoning [29], abductive reasoning [6], counterfactual reasoning [90], generative commonsense reasoning [69], social reasoning [99], physical reasoning [7], temporal reasoning [144], etc. [2] Not limited to these tasks, datasets on explanation generation [127] are also good sources to evaluate and improve the reasoning ability of NLG systems, including the explanations for natural language inference [12] commonsense reasoning [94], analogical reasoning [22], causal reasoning [35], multimodal reasoning [68], etc.

[2] Note that some of these reasoning tasks [99, 7, 144] take the form of question answering but can be solved in a generative manner [58].

4.2 Reasoning by NLG

Human language is a good vehicle for reasoning. Thus, the generation of language naturally resembles the way humans think and reason. With the success of PLMs, there is a growing interest in the AI community to use NLG models to generate a chain of reasoning for problem-solving.

Generative Reasoning Generative reasoning aims to generate intermediate reasons with NLG models for better problem solving. Such intermediate reasons take many forms, including deduction and abduction reasons [113], explanations [56], or decomposed subtasks of a complex one [59]. Some work [108, 4] even show that expanding the context of the input by generating more information would also help solve reasoning tasks. Moreover, [27] finds that transformer-based PLMs are effective soft reasoners on a toy deduction dataset, which consists of collections of text verbalized from artificial if-then rules and facts. Other studies [10, 5] corroborate this discovery and show that training generative models with artificial textual data that verbalize rule-based reasoning helps downstream logical reasoning tasks. We remark that generative reasoning provides a new perspective of breaking the black-box prediction of neural networks, which demonstrates the potential of achieving reasoning with NLG systems.

Reasoning with Large Language Models Entering the era of large language models (LLMs) such as GPT-3 [11, 84] and PaLM [26], there is a recent growing interest in exploring few-/zero-shot reasoning skills of these LLMs Since fine-tuning such tremendous language models is hardly possible, current work adopts prompt-based in-context learning methods [11, 77] to achieve few-/zero-shot learning with LLMs, where instructions and examples are demonstrated within the input prompts.

Similar to the above discussion of generative reasoning, this line of work aims to guide the language models to explicitly generate the intermediate thinking steps (or reasons) during reasoning. A representative work among them is the Chain-of-Thought prompting [125], where the intermediate thinking process is verbalized and integrated into the demonstrations. In this way, complex reasoning can be decomposed into multiple steps reflected by language, which is analogous to how humans solve complex tasks. Such methods achieve much better performance on a variety of reasoning tasks compared with normal prompting and even surpass fine-tuned methods in some cases. Moreover, the prompting strategy can be further refined [123, 146, 28], leading to generally better results. LLMs prompted with and asked to generate step-by-step reasoning chains also exhibit certain quantitative reasoning abilities such as solving math word problems [125], where the LLMs are not specifically trained on such tasks. However, reasoning abilities for LLMs are shown to be emergent [125, 124], i.e., effective only for really large language models (over 100 billion parameters). How to enable smaller language models with few-shot reasoning skills is still an open question.

5 Conclusion

In this work, we envision the ten most desired goals of an intelligent natural language generation system. In pursuit of these goals, the guidance of knowledge and reasoning plays a significant role in modern NLG models. We have revisited the achievements with knowledge in NLG w.r.t. knowledge-guided NLG and generative knowledge acquisition. We particularly highlight knowledge-constrained decoding for its wide application potential, where knowledge constraints can be incorporated into NLG models (especially large-scale ones) in a plug-and-play manner. We have also discussed current work on reasoning in NLG, which essentially makes rational usage of knowledge and datasets for various reasoning tasks. Moreover, NLG can verbalize intermediate thinking processes to facilitate complex reasoning, enabling few-/zero-shot reasoning abilities for large language models. However, current research is still far from realizing these goals, which we outline for future research. We hope

that this survey report can provide newcomers with a good entry point into the exciting area of knowledge-guided and reasoning-intensive NLG.

References

[1] Badr AlKhamissi, Millicent Li, Asli Celikyilmaz, Mona Diab, and Marjan Ghazvininejad. A review on language models as knowledge bases. arXiv preprint arXiv:2204.06031, 2022.

[2] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In Proceedings of the 6th International The Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference, ISWC'07/ASWC'07, page 722–735, Berlin, Heidelberg, 2007. Springer-Verlag.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun, editors, 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.

[4] Gregor Betz, Kyle Richardson, and Christian Voigt. Thinking aloud: Dynamic context generation improves zero-shot reasoning performance of gpt-2. arXiv preprint arXiv:2103.13033, 2021.

[5] Gregor Betz, Christian Voigt, and Kyle Richardson. Critical thinking for language models. In Proceedings of the 14th International Conference on Computational Semantics (IWCS), pages 63–75, Groningen, The Netherlands (online), June 2021. Association for Computational Linguistics.

[6] Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen tau Yih, and Yejin Choi. Abductive commonsense reasoning. In International Conference on Learning Representations, 2020.

[7] Yonatan Bisk, Rowan Zellers, Ronan Le bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language. Proceedings of the AAAI Conference on Artificial Intelligence, 34(05):7432–7439, Apr. 2020.

[8] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, Advances in Neural Information Processing Systems, volume 26. Curran Associates, Inc., 2013.

[9] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. COMET: Commonsense transformers for automatic knowledge graph construction. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4762–4779, Florence, Italy, July 2019. Association for Computational Linguistics.

[10] Kaj Bostrom, Xinyu Zhao, Swarat Chaudhuri, and Greg Durrett. Flexible generation of natural language deductions. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6266–6278, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[11] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark,

Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[12] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-snli: Natural language inference with natural language explanations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 31. Curran Associates, Inc., 2018.

[13] Boxi Cao, Hongyu Lin, Xianpei Han, Le Sun, Lingyong Yan, Meng Liao, Tong Xue, and Jin Xu. Knowledgeable or educated guess? revisiting language models as knowledge bases. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1860–1874, Online, August 2021. Association for Computational Linguistics.

[14] Meng Cao, Yue Dong, and Jackie Cheung. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3340–3354, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[15] Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. Autoregressive entity retrieval. In International Conference on Learning Representations, 2021.

[16] Ruisheng Cao, Su Zhu, Chenyu Yang, Chen Liu, Rao Ma, Yanbin Zhao, Lu Chen, and Kai Yu. Unsupervised dual paraphrasing for two-stage semantic parsing. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6806–6817, Online, July 2020. Association for Computational Linguistics.

[17] Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. Faithful to the original: Fact aware neural abstractive summarization. In Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

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

[19] Hong Chen, Yifei Huang, Hiroya Takamura, and Hideki Nakayama. Commonsense knowledge aware concept selection for diverse and informative visual storytelling. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 999–1008, 2021.

[20] Jiangjie Chen, Qiaoben Bao, Changzhi Sun, Xinbo Zhang, Jiaze Chen, Hao Zhou, Yanghua Xiao, and Lei Li. Loren: Logic-regularized reasoning for interpretable fact verification. Proceedings of the AAAI Conference on Artificial Intelligence, 36(10):10482–10491, Jun. 2022.

[21] Jiangjie Chen, Ao Wang, Haiyun Jiang, Suo Feng, Chenguang Li, and Yanghua Xiao. Ensuring readability and data-fidelity using head-modifier templates in deep type description generation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 2036–2046, Florence, Italy, July 2019. Association for Computational Linguistics.

[22] Jiangjie Chen, Rui Xu, Ziquan Fu, Wei Shi, Zhongqiao Li, Xinbo Zhang, Changzhi Sun, Lei Li, Yanghua Xiao, and Hao Zhou. E-KAR: A benchmark for rationalizing natural language analogical reasoning. In Findings of the Association for Computational Linguistics: ACL 2022, pages 3941–3955, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[23] Jindong Chen, Yizhou Hu, Jingping Liu, Yanghua Xiao, and Haiyun Jiang. Deep short text classification with knowledge powered attention. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 6252–6259, 2019.

[24] Lihan Chen, Sihang Jiang, Jingping Liu, Chao Wang, Sheng Zhang, Chenhao Xie, Jiaqing Liang, Yanghua Xiao, and Rui Song. Rule mining over knowledge graphs via reinforcement learning. Know.-Based Syst., 242(C), apr 2022.

[25] Wenhu Chen, Jianshu Chen, Yu Su, Zhiyu Chen, and William Yang Wang. Logical natural language generation from open-domain tables. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7929–7942, Online, July 2020. Association for Computational Linguistics.

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

[27] Peter Clark, Oyvind Tafjord, and Kyle Richardson. Transformers as soft reasoners over language. In Christian Bessiere, editor, Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20, pages 3882–3890. International Joint Conferences on Artificial Intelligence Organization, 7 2020. Main track.

[28] Antonia Creswell, Murray Shanahan, and Irina Higgins. Selection-inference: Exploiting large language models for interpretable logical reasoning. arXiv preprint arXiv:2205.09712, 2022.

[29] Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. Explaining answers with entailment trees. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 7358–7370, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[30] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. In International Conference on Learning Representations, 2020.

[31] Randall Davis and Bruce G. Buchanan. Meta-level knowledge: Overview and applications. In IJCAI, 1977.

[32] Nicola De Cao, Ledell Wu, Kashyap Popat, Mikel Artetxe, Naman Goyal, Mikhail Plekhanov, Luke Zettlemoyer, Nicola Cancedda, Sebastian Riedel, and Fabio Petroni. Multilingual Autoregressive Entity Linking. Transactions of the Association for Computational Linguistics, 10:274–290, 03 2022.

[33] Ashwin Devaraj, William Sheffield, Byron Wallace, and Junyi Jessy Li. Evaluating factuality in text simplification. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7331–7345, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[34] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, June 2019. Association for Computational Linguistics.

[35] Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. e-CARE: a new dataset for exploring explainable causal reasoning. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 432–446, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[36] Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and improving consistency in pretrained language models. Transactions of the Association for Computational Linguistics, 9:1012–1031, 2021.

[37] Cristina Garbacea and Qiaozhu Mei. Why is constrained neural language generation particularly challenging? arXiv preprint arXiv:2206.05395, 2022.

[38] François Gardères, Maryam Ziaeefard, Baptiste Abeloos, and Freddy Lecue. ConceptBert: Concept-aware representation for visual question answering. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 489–498, Online, November 2020. Association for Computational Linguistics.

[39] Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. AllenNLP: A deep semantic natural language processing platform. In Proceedings of Workshop for NLP Open Source Software (NLP-OSS), pages 1–6, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[40] Dimitra Gkatzia, Oliver Lemon, and Verena Rieser. Data-to-text generation improves decision-making under uncertainty. IEEE Computational Intelligence Magazine, 12:10–17, 2017.

[41] Jiatao Gu, James Bradbury, Caiming Xiong, Victor O.K. Li, and Richard Socher. Non-autoregressive neural machine translation. In International Conference on Learning Representations, 2018.

[42] Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. Incorporating copying mechanism in sequence-to-sequence learning. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1631–1640, Berlin, Germany, August 2016. Association for Computational Linguistics.

[43] Jian Guan, Fei Huang, Zhihao Zhao, Xiaoyan Zhu, and Minlie Huang. A knowledge-enhanced pretraining model for commonsense story generation. Transactions of the Association for Computational Linguistics, 8:93–108, 2020.

[44] Junliang Guo, Zhirui Zhang, Linli Xu, Hao-Ran Wei, Boxing Chen, and Enhong Chen. Incorporating bert into parallel sequence decoding with adapters. Advances in Neural Information Processing Systems, 33:10843–10854, 2020.

[45] Umang Gupta, Jwala Dhamala, Varun Kumar, Apurv Verma, Yada Pruksachatkun, Satyapriya Krishna, Rahul Gupta, Kai-Wei Chang, Greg Ver Steeg, and Aram Galstyan. Mitigating gender bias in distilled language models via counterfactual role reversal. In Findings of the Association for Computational Linguistics: ACL 2022, pages 658–678, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[46] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In Hal Daumé III and Aarti Singh, editors, Proceedings of the 37th International Conference on Machine Learning, volume 119 of Proceedings of Machine Learning Research, pages 3929–3938. PMLR, 13–18 Jul 2020.

[47] Qianyu He, Sijie Cheng, Zhixu Li, Rui Xie, and Yanghua Xiao. Can pre-trained language models interpret similes as smart as human? In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7875–7887, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[48] Chris Hokamp and Qun Liu. Lexically constrained decoding for sequence generation using grid beam search. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1535–1546, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[49] Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 7856–7870, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[50] Zhiting Hu and Li Erran Li. A causal lens for controllable text generation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, Advances in Neural Information Processing Systems, volume 34, pages 24941–24955. Curran Associates, Inc., 2021.

[51] Zhiting Hu, Zichao Yang, Russ R Salakhutdinov, LIANHUI Qin, Xiaodan Liang, Haoye Dong, and Eric P Xing. Deep generative models with learnable knowledge constraints. Advances in Neural Information Processing Systems, 31, 2018.

[52] Kung-Hsiang Huang, Sam Tang, and Nanyun Peng. Document-level entity-based extraction as template generation. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 5257–5269, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[53] Luyang Huang, Lingfei Wu, and Lu Wang. Knowledge graph-augmented abstractive summarization with semantic-driven cloze reward. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5094–5107, Online, July 2020. Association for Computational Linguistics.

[54] Pere-Lluís Huguet Cabot and Roberto Navigli. REBEL: Relation extraction by end-to-end language generation. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2370–2381, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[55] Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. In AAAI, 2021.

[56] Harsh Jhamtani and Peter Clark. Learning to explain: Datasets and models for identifying valid reasoning chains in multihop question-answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 137–150, Online, November 2020. Association for Computational Linguistics.

[57] Haozhe Ji, Pei Ke, Shaohan Huang, Furu Wei, Xiaoyan Zhu, and Minlie Huang. Language generation with multi-hop reasoning on commonsense knowledge graph. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 725–736, Online, November 2020. Association for Computational Linguistics.

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

[59] Tushar Khot, Daniel Khashabi, Kyle Richardson, Peter Clark, and Ashish Sabharwal. Text modular networks: Learning to decompose tasks in the language of existing models. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1264–1279, Online, June 2021. Association for Computational Linguistics.

[60] Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. Will I sound like me? improving persona consistency in dialogues through pragmatic self-consciousness. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 904–916, Online, November 2020. Association for Computational Linguistics.

[61] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907, 2016.

[62] Karen Kukich. Design of a knowledge-based report generator. In 21st Annual Meeting of the Association for Computational Linguistics, pages 145–150, Cambridge, Massachusetts, USA, June 1983. Association for Computational Linguistics.

[63] Faisal Ladhak, Esin Durmus, He He, Claire Cardie, and Kathleen McKeown. Faithful or extractive? on mitigating the faithfulness-abstractiveness trade-off in abstractive summarization. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1410–1421, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[64] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 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, July 2020. Association for Computational Linguistics.

[65] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems, volume 33, pages 9459–9474. Curran Associates, Inc., 2020.

[66] Chenguang Li, Jiaqing Liang, Yanghua Xiao, and Haiyun Jiang. Towards fine-grained concept generation. IEEE Transactions on Knowledge and Data Engineering, pages 1–1, 2021.

[67] Haoran Li, Junnan Zhu, Jiajun Zhang, and Chengqing Zong. Ensure the correctness of the summary: Incorporate entailment knowledge into abstractive sentence summarization. In Proceedings of the 27th International Conference on Computational Linguistics, pages 1430–1441, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[68] Qing Li, Qingyi Tao, Shafiq Joty, Jianfei Cai, and Jiebo Luo. Vqa-e: Explaining, elaborating, and enhancing your answers for visual questions. ECCV, 2018.

[69] Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. CommonGen: A constrained text generation challenge for generative commonsense reasoning.

In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1823–1840, Online, November 2020. Association for Computational Linguistics.

[70] Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. Testing the ability of language models to interpret figurative language. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4437–4452, Seattle, United States, July 2022. Association for Computational Linguistics.

[71] Haochen Liu, Jamell Dacon, Wenqi Fan, Hui Liu, Zitao Liu, and Jiliang Tang. Does gender matter? towards fairness in dialogue systems. In Proceedings of the 28th International Conference on Computational Linguistics, pages 4403–4416, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[72] Pengfei Liu, Jinlan Fu, Yang Xiao, Weizhe Yuan, Shuaichen Chang, Junqi Dai, Yixin Liu, Zihuiwen Ye, and Graham Neubig. ExplainaBoard: An explainable leaderboard for NLP. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations, pages 280–289, Online, August 2021. Association for Computational Linguistics.

[73] Ruibo Liu, Guoqing Zheng, Shashank Gupta, Radhika Gaonkar, Chongyang Gao, Soroush Vosoughi, Milad Shokouhi, and Ahmed Hassan Awadallah. Knowledge infused decoding. In International Conference on Learning Representations, 2022.

[74] Ye Liu, Yao Wan, Lifang He, Hao Peng, and Philip S. Yu. Kg-bart: Knowledge graph-augmented bart for generative commonsense reasoning. Proceedings of the AAAI Conference on Artificial Intelligence, 35(7):6418–6425, May 2021.

[75] Zihan Liu, Mostofa Patwary, Ryan Prenger, Shrimai Prabhumoye, Wei Ping, Mohammad Shoeybi, and Bryan Catanzaro. Multi-stage prompting for knowledgeable dialogue generation. In Findings of the Association for Computational Linguistics: ACL 2022, pages 1317–1337, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[76] Ximing Lu, Peter West, Rowan Zellers, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. NeuroLogic decoding: (un)supervised neural text generation with predicate logic constraints. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4288–4299, Online, June 2021. Association for Computational Linguistics.

[77] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8086–8098, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[78] Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. Text2Event: Controllable sequence-to-structure generation for end-to-end event extraction. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 2795–2806, Online, August 2021. Association for Computational Linguistics.

[79] Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. Unified structure generation for universal information extraction. In Proceedings of the 60th Annual Meeting

of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5755–5772, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[80] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1906–1919, Online, July 2020. Association for Computational Linguistics.

[81] Kathleen McKeown. Text generation. Cambridge University Press, 1992.

[82] Ning Miao, Hao Zhou, Lili Mou, Rui Yan, and Lei Li. Cgmh: Constrained sentence generation by metropolis-hastings sampling. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01):6834–6842, Jul. 2019.

[83] George A. Miller. Wordnet: A lexical database for english. Commun. ACM, 38(11):39–41, nov 1995.

[84] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, Advances in Neural Information Processing Systems, 2022.

[85] Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, RISHITA ANUBHAI, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. Structured prediction as translation between augmented natural languages. In International Conference on Learning Representations, 2021.

[86] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? 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), pages 2463–2473, Hong Kong, China, November 2019. Association for Computational Linguistics.

[87] Xinyu Pi, Wanjun Zhong, Yan Gao, Nan Duan, and Jian-Guang Lou. Logigan: Learning logical reasoning via adversarial pre-training. arXiv preprint arXiv:2205.08794, 2022.

[88] Matt Post and David Vilar. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1314–1324, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[89] Lihua Qian, Hao Zhou, Yu Bao, Mingxuan Wang, Lin Qiu, Weinan Zhang, Yong Yu, and Lei Li. Glancing transformer for non-autoregressive neural machine translation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1993–2003, Online, August 2021. Association for Computational Linguistics.

[90] Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. Counterfactual story reasoning and generation. 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), pages 5043–5053, Hong Kong, China, November 2019. Association for Computational Linguistics.

[91] Lianhui Qin, Sean Welleck, Daniel Khashabi, and Yejin Choi. Cold decoding: Energy-based constrained text generation with langevin dynamics. arXiv preprint arXiv:2202.11705, 2022.

[92] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. OpenAI Blog, 1(8):9, 2019.

[93] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140):1–67, 2020.

[94] Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. Explain yourself! leveraging language models for commonsense reasoning. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4932–4942, Florence, Italy, July 2019. Association for Computational Linguistics.

[95] Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. The curious case of hallucinations in neural machine translation. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1172–1183, Online, June 2021. Association for Computational Linguistics.

[96] Ehud Reiter. Natural language generation challenges for explainable ai. arXiv preprint arXiv:1911.08794, 2019.

[97] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "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, San Francisco, CA, USA, August 13-17, 2016, pages 1135–1144, 2016.

[98] Gaetano Rossiello, Nandana Mihindukulasooriya, Ibrahim Abdelaziz, Mihaela Bornea, Alfio Gliozzo, Tahira Naseem, and Pavan Kapanipathi. Generative relation linking for question answering over knowledge bases. In International Semantic Web Conference, pages 321–337. Springer, 2021.

[99] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Commonsense reasoning about social interactions. 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), pages 4463–4473, Hong Kong, China, November 2019. Association for Computational Linguistics.

[100] Karl Schulz, Leon Sixt, Federico Tombari, and Tim Landgraf. Restricting the flow: Information bottlenecks for attribution. In International Conference on Learning Representations, 2020.

[101] Tianxiao Shen, Myle Ott, Michael Auli, and Marc'Aurelio Ranzato. Mixture models for diverse machine translation: Tricks of the trade. In International conference on machine learning, pages 5719–5728. PMLR, 2019.

[102] Xinyao Shen, Jiangjie Chen, Jiaze Chen, Chun Zeng, and Yanghua Xiao. Diversified query generation guided by knowledge graph. In Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, WSDM '22, page 897–907, New York, NY, USA, 2022. Association for Computing Machinery.

[103] Xinyao Shen, Jiangjie Chen, and Yanghua Xiao. Diversified paraphrase generation with commonsense knowledge graph. In Lu Wang, Yansong Feng, Yu Hong, and Ruifang He, editors, Natural Language Processing and Chinese Computing, pages 353–364, Cham, 2021. Springer International Publishing.

[104] Weiyan Shi, Yu Li, Saurav Sahay, and Zhou Yu. Refine and imitate: Reducing repetition and inconsistency in persuasion dialogues via reinforcement learning and human demonstration. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 3478–3492, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[105] Zhan Shi, Hui Liu, and Xiaodan Zhu. Descriptive image captioning with salient retrieval priors. In Luiza Antonie and Pooya Moradian Zadeh, editors, Proceedings of the 34th Canadian Conference on Artificial Intelligence, Canadian AI 2021, online, May 2021. Canadian Artificial Intelligence Association, 2021.

[106] Zhan Shi, Hui Liu, and Xiaodan Zhu. Enhancing descriptive image captioning with natural language inference. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 269–277, Online, August 2021. Association for Computational Linguistics.

[107] Chang Shu, Yusen Zhang, Xiangyu Dong, Peng Shi, Tao Yu, and Rui Zhang. Logic-consistency text generation from semantic parses. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 4414–4426, Online, August 2021. Association for Computational Linguistics.

[108] Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Unsupervised commonsense question answering with self-talk. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4615–4629, Online, November 2020. Association for Computational Linguistics.

[109] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. Proceedings of the AAAI Conference on Artificial Intelligence, 31(1), Feb. 2017.

[110] Mingming Sun, Xu Li, and Ping Li. Logician and orator: Learning from the duality between language and knowledge in open domain. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2119–2130, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[111] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17, page 3319–3328. JMLR.org, 2017.

[112] Raymond Hendy Susanto, Shamil Chollampatt, and Liling Tan. Lexically constrained neural machine translation with Levenshtein transformer. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 3536–3543, Online, July 2020. Association for Computational Linguistics.

[113] Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 3621–3634, Online, August 2021. Association for Computational Linguistics.

[114] Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. What do you learn from context? probing for sentence structure in contextualized word representations. In International Conference on Learning Representations, 2019.

[115] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. Lamda: Language models for dialog applications. arXiv preprint arXiv:2201.08239, 2022.

[116] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 809–819, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[117] Lindsey Vanderlyn, Gianna Weber, Michael Neumann, Dirk Väth, Sarina Meyer, and Ngoc Thang Vu. "it seemed like an annoying woman": On the perception and ethical considerations of affective language in text-based conversational agents. In Proceedings of the 25th Conference on Computational Natural Language Learning, pages 44–57, Online, November 2021. Association for Computational Linguistics.

[118] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008, 2017.

[119] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net, 2018.

[120] Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. Communications of the ACM, 57(10):78–85, 2014.

[121] Danqing Wang, Pengfei Liu, Yining Zheng, Xipeng Qiu, and Xuanjing Huang. Heterogeneous graph neural networks for extractive document summarization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6209–6219, Online, July 2020. Association for Computational Linguistics.

[122] Qingyun Wang, Lifu Huang, Zhiying Jiang, Kevin Knight, Heng Ji, Mohit Bansal, and Yi Luan. PaperRobot: Incremental draft generation of scientific ideas. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1980–1991, Florence, Italy, July 2019. Association for Computational Linguistics.

[123] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171, 2022.

[124] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. Transactions on Machine Learning Research, 2022. Survey Certification.

[125] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, Advances in Neural Information Processing Systems, 2022.

[126] Joseph Weizenbaum. Eliza—a computer program for the study of natural language communication between man and machine. Communications of the ACM, 9(1):36–45, 1966.

[127] Sarah Wiegreffe and Ana Marasovic. Teach me to explain: A review of datasets for explainable natural language processing. In J. Vanschoren and S. Yeung, editors, Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks, volume 1, 2021.

[128] Ludwig Wittgenstein. The blue and brown books, volume 34. Blackwell Oxford, 1958.

[129] Sixing Wu, Ying Li, Dawei Zhang, Yang Zhou, and Zhonghai Wu. Diverse and informative dialogue generation with context-specific commonsense knowledge awareness. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5811–5820, Online, July 2020. Association for Computational Linguistics.

[130] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144, 2016.

[131] Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. Demoting racial bias in hate speech detection. In Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media, pages 7–14, Online, July 2020. Association for Computational Linguistics.

[132] Yijun Xiao and William Yang Wang. On hallucination and predictive uncertainty in conditional language generation. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 2734–2744, Online, April 2021. Association for Computational Linguistics.

[133] Yisheng Xiao, Lijun Wu, Junliang Guo, Juntao Li, Min Zhang, Tao Qin, and Tie-yan Liu. A survey on non-autoregressive generation for neural machine translation and beyond. arXiv preprint arXiv:2204.09269, 2022.

[134] Runxin Xu, Jun Cao, Mingxuan Wang, Jiaze Chen, Hao Zhou, Ying Zeng, Yuping Wang, Li Chen, Xiang Yin, Xijin Zhang, Songcheng Jiang, Yuxuan Wang, and Lei Li. Xiaomingbot: A Multilingual Robot News Reporter. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 1–8, Online, July 2020. Association for Computational Linguistics.

[135] Weijia Xu and Marine Carpuat. EDITOR: An Edit-Based Transformer with Repositioning for Neural Machine Translation with Soft Lexical Constraints. Transactions of the Association for Computational Linguistics, 9:311–328, 03 2021.

[136] Jian Yang, Gang Xiao, Yulong Shen, Wei Jiang, Xinyu Hu, Ying Zhang, and Jinghui Peng. A survey of knowledge enhanced pre-trained models. arXiv preprint arXiv:2110.00269, 2021.

[137] Xi Ye and Greg Durrett. The unreliability of explanations in few-shot in-context learning. arXiv preprint arXiv:2205.03401, 2022.

[138] Xuwang Yin and Vicente Ordonez. Obj2Text: Generating visually descriptive language from object layouts. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 177–187, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[139] Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. A survey of knowledge-enhanced text generation. ACM Computing Surveys (CSUR), 2022.

[140] Chun Zeng, Jiangjie Chen, Tianyi Zhuang, Rui Xu, Hao Yang, Qin Ying, Shimin Tao, and Yanghua Xiao. Neighbors are not strangers: Improving non-autoregressive translation under low-frequency lexical constraints. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5777–5790, Seattle, United States, July 2022. Association for Computational Linguistics.

[141] Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. A survey of controllable text generation using transformer-based pre-trained language models. arXiv preprint arXiv:2201.05337, 2022.

[142] Houyu Zhang, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. Grounded conversation generation as guided traverses in commonsense knowledge graphs. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 2031–2043, Online, July 2020. Association for Computational Linguistics.

[143] Maosen Zhang, Nan Jiang, Lei Li, and Yexiang Xue. Language generation via combinatorial constraint satisfaction: A tree search enhanced Monte-Carlo approach. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1286–1298, Online, November 2020. Association for Computational Linguistics.

[144] Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. 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), pages 3363–3369, Hong Kong, China, November 2019. Association for Computational Linguistics.

[145] Chunting Zhou, Jiatao Gu, and Graham Neubig. Understanding knowledge distillation in non-autoregressive machine translation. In International Conference on Learning Representations, 2020.

[146] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Olivier Bousquet, Quoc Le, and Ed Chi. Least-to-most prompting enables complex reasoning in large language models. arXiv preprint arXiv:2205.10625, 2022.

[147] Shaowen Zhou, Bowen Yu, Aixin Sun, Cheng Long, Jingyang Li, and Jian Sun. A survey on neural open information extraction: Current status and future directions. arXiv preprint arXiv:2205.11725, 2022.

[148] Caleb Ziems, Jane Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. The moral integrity corpus: A benchmark for ethical dialogue systems. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3755–3773, Dublin, Ireland, May 2022. Association for Computational Linguistics.