Generating Factual Documents by Synthesizing Knowledge Sources

Shuyang Li, Jianmo Ni, Huanru Henry Mao, Julian McAuley
Computer Science and Engineering
University of California, San Diego
{shl008, jin018, hhmao, jmcauley}@ucsd.edu

Abstract

From youth, humans can read and process large amounts of information to write articles, book reports, and conduct deep conversation. Existing large-scale language models are yet incapable of such meaningful generation. We propose a knowledge-grounded document writing task for pre-training an encoder-decoder language model to enable such knowledge synthesis. We will pre-train a model on networks of knowledge-grounded documents from encyclopedias and news, leveraging high-quality source citations common in these fields. We present the datasets that we have collected thus far, methods for large-context knowledge grounded synthesis, and preliminary results indicating the applicability of our framework.

1 Problem Formulation

While the knowledge capacity of language models has been studied in context of fill-in-the-blank [3] and entity linking [2] tasks, we focus on tasks that involve synthesizing large amounts of information into long documents. We propose the unified **S**ource, **Q**uery, **T**arget framework for knowledge-grounded generation / synthesis tasks. In each task, we are given source document(s) consisting of factual texts (paragraphs, full articles etc.). These source documents are then synthesized along with a query to generate our target text. The query and target can take the form of question-answer pairs, dialog histories and utterances, or titles and article bodies.

2 Related Work

Large pre-trained language models [13] elicited much research into their capacity to store relational knowledge [12]. While their training data contain many factual statements, the lack of an explicit semantic model constraint can result in logical and factual inconsistencies in generated text [8]. Recent efforts have introduced common-sense and question-answering tasks as pre-training objectives in a multi-task framework [11, 14]. We propose a similar common framework for knowledge synthesis tasks, enabling us to pose a single pre-training objective that will aid a model in synthesis knowledge across a variety of generative domains.

Another approach involves leveraging an existing knowledge graph alongside the language model and guiding its generation [8]. Work on abstractive question-answering using the ELI5 sub-reddit has demonstrated the effectiveness of linearized knowledge graphs to summarize the knowledge from multiple long documents [4]. This approach was able to summarize a much larger context than previous extractive methods to select individual sentences from source documents, partially due to context size limitations [10]. Compared to these compressive approaches to knowledge extraction, we propose to learn a latent representation directly from full source documents via sparse encoder-decoder models with a higher context capacity. Additionally, we propose to use a persistent memory module [15] to learn factual encodings from our entire body of training data.

SoCal ML Symposium 2020, San Diego, California

Table 1: NPR news article length and citation statistics.

Split	Target	Source	Avg. Citations	Avg. Target Tokens	Avg. Source Tokens
Train	20,875	32,579	2.859	888	2,397
Valid	1,150	3,137	2.851	871	2,327
Test	1,161	3,064	2.814	892	2,378

3 Data

For preliminary study, we crawled a dataset of 23K target news articles from the NPR[1] website along with source articles linked from each target. Detailed dataset statistics are displayed in Table 1. We are in the process of crawling the articles and link structure for additional news sources, as well as leveraging the English Wikipedia hyperlink network [1] to generate a full wikipedia article given the other articles that it cites. We will evaluate our model on three sets of downstream tasks for knowledge synthesis: dialog modeling, abstractive question answering, and document synthesis. For dialog modeling, we investigate TopicalChat [6], a dataset of short dialogs grounded on "fun fact" snippets, as well as a novel interview response-generation task based on a large corpus of NPR interview transcripts. We will use ELI5 [5] to generate answers to questions involving 50+ articles on related topics. For document synthesis, we investigate the WikiSum task to generate the first paragraph of a Wikipedi article given its title and cited source articles [10].

4 Methods

We base our model on the Reformer architecture [9], an efficient transformer model that uses locality-sensitive hashing (LSH) to limit the attention computation, allowing for significantly longer contexts with log-linear attention complexity. As humans typically draw from their own accumulated world knowledge in daily life [7], we propose to maintain a persistent memory [15] to store widely applicable knowledge from each encountered document. The input to our model consists of the concatenation of the query string with the full text of all source documents. The decoder will auto-regressively predict the tokens of the target document. As baselines, we will evaluate unconditional pre-trained models such as GPT2 [13], as well as the Transformer-DMCA [10] which extracts full context sentences from each grounding document, and using linearized knowledge graphs as context [4].

5 Preliminary Results

In preliminary experiments, we explore the applicability of extractive multi-document summarization models to knowledge-grounded dialog generation in TopicalChat. In our baseline model, we extracted 500 tokens of full source sentences and generated the next utterance (target) given a conversation history (query). We concatenated the query and sources as inputs to GPT2. A sample generation is provided in Table 5. We observe qualitatively that this model architecture applies in both summarization and dialog modeling settings, and is able to make specific references to source documents. However, conversations frequently span a breadth of information far beyond the context limits of this model, which suggests that models with larger context and the ability to remember world knowledge are necessary to achieve reasonable results in knowledge synthesis tasks.

Query	My article says that Pakistan has a jazz orchestra that is topping the charts. That's very interesting, I wonder what kind of a sound they have?
Sources	If I tried to insert a **Sonny Rollins** quote into a solo [...] Imitation is impossible, because he's so distinctive [...] Jazz has roots in West African cultural and musical expression [...]
Generated	It sounds "bluer" and not as good as **Sonny Rollins's**. They say that each style has its own charm. If you like classic jazz, you can't go wrong with **Sonny Rollins**.

Table 2: A sample generated utterance. **Bold** emphasizes references to sources.

[1] https://www.npr.org/

References

[1] P. Boldi, B. Codenotti, M. Santini, and S. Vigna. Ubicrawler: a scalable fully distributed web crawler. *Softw., Pract. Exper.*, 34(8):711–726, 2004.

[2] S. Broscheit. Investigating entity knowledge in BERT with simple neural end-to-end entity linking. In *CoNLL*, pages 677–685, 2019.

[3] J. Davison, J. Feldman, and A. M. Rush. Commonsense knowledge mining from pretrained models. In *EMNLP*, pages 1173–1178, 2019.

[4] A. Fan, C. Gardent, C. Braud, and A. Bordes. Using local knowledge graph construction to scale Seq2Seq models to multi-document inputs. In *EMNLP*, pages 4186–4196, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.

[5] A. Fan, Y. Jernite, E. Perez, D. Grangier, J. Weston, and M. Auli. ELI5: long form question answering. In *ACL*, pages 3558–3567, 2019.

[6] K. Gopalakrishnan, B. Hedayatnia, Q. Chen, A. Gottardi, S. Kwatra, A. Venkatesh, R. Gabriel, D. Hakkani-Tür, and A. A. AI. Topical-chat: Towards knowledge-grounded open-domain conversations. *Proc. Interspeech 2019*, pages 1891–1895, 2019.

[7] P. Hagoort, L. Hald, M. Bastiaansen, and K. M. Petersson. Integration of word meaning and world knowledge in language comprehension. *science*, 304(5669):438–441, 2004.

[8] R. L. L. IV, N. F. Liu, M. E. Peters, M. Gardner, and S. Singh. Barack's wife hillary: Using knowledge graphs for fact-aware language modeling. In *ACL*, pages 5962–5971, 2019.

[9] N. Kitaev, L. Kaiser, and A. Levskaya. Reformer: The efficient transformer. *CoRR*, abs/2001.04451, 2020.

[10] P. J. Liu, M. Saleh, E. Pot, B. Goodrich, R. Sepassi, L. Kaiser, and N. Shazeer. Generating wikipedia by summarizing long sequences. In *ICLR*, 2018.

[11] B. McCann, N. S. Keskar, C. Xiong, and R. Socher. The natural language decathlon: Multitask learning as question answering. *CoRR*, abs/1806.08730, 2018.

[12] F. Petroni, T. Rocktäschel, S. Riedel, P. Lewis, A. Bakhtin, Y. Wu, and A. Miller. Language models as knowledge bases? In *EMNLP*, pages 2463–2473, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.

[13] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.

[14] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683, 2019.

[15] S. Sukhbaatar, E. Grave, G. Lample, H. Jégou, and A. Joulin. Augmenting self-attention with persistent memory. *CoRR*, abs/1907.01470, 2019.