# Employing Argumentation Knowledge Graphs for Neural Argument Generation

**Khalid Al-Khatib**[1] **Lukas Trautner**[2] **Henning Wachsmuth**[3] **Yufang Hou**[4] **Benno Stein**[5]
[1] Leipzig University, Germany, khalid.alkhatib@uni-leipzig.de
[2] University of Erlangen-Nuremberg, Germany, lukas.trautner@fau.de
[3] Paderborn University, Germany, henningw@upb.de
[4] IBM Research Europe, Ireland, yhou@ie.ibm.com
[5] Bauhaus-Universität Weimar, Germany, benno.stein@uni-weimar.de

## Abstract

Generating high-quality arguments, while being challenging, may benefit a wide range of downstream applications, such as writing assistants and argument search engines. Motivated by the effectiveness of utilizing knowledge graphs for supporting general text generation tasks, this paper investigates the usage of argumentation-related knowledge graphs to control the generation of arguments. In particular, we construct and populate three knowledge graphs, employing several compositions of them to encode various knowledge into texts of debate portals and relevant paragraphs from Wikipedia. Then, the texts with the encoded knowledge are used to fine-tune a pre-trained text generation model, GPT-2. We evaluate the newly created arguments manually and automatically, based on several dimensions important in argumentative contexts, including argumentativeness and plausibility. The results demonstrate the positive impact of encoding the graphs' knowledge into debate portal texts for generating arguments with superior quality than those generated without knowledge.

## 1 Introduction

Arguments are our means to build stances on controversial topics, to persuade others, or to negotiate. Automatic argument generation has the potential to effectively support such tasks: it may not only regenerate known arguments but also uncover new facets of a topic. Existing argument generation approaches work either in an end-to-end fashion (Hua and Wang, 2018) or they are controlled with respect to the argument's topic, aspects, or stance (Gretz et al., 2020; Schiller et al., 2021). In contrast, no approach integrates external knowledge into the generation process so far, even though knowledge graphs have been shown to be useful for supporting text generation models in other areas (Koncel-Kedziorski et al., 2019a; Ribeiro et al., 2020).

Previous research has proposed *argumentation knowledge graphs (AKGs)* that model supporting and attacking interactions between concepts (Al-Khatib et al., 2020). Such an AKG may assist argument generation models in different ways. For example, meaningful *prompts* on controversial topics can be constructed from an AKG with simple hand-defined rules, such as '*geoengineering* reduces *atmospheric greenhouse gas*' for generating an argument on 'geoengineering.' Alternatively, an AKG may be employed to *control* the generation, making arguments adhere to knowledge covered in the graph. We hypothesize this to be particularly beneficial for the quality of arguments in terms of factuality, the richness of evidence, and similar.

This paper concentrates on such controlled argument generation, investigating for the first time the ability to generate high-quality and content-rich arguments by integrating knowledge from AKGs into standard neural-based generation models. To this end, we exploit multiple manually and automatically created knowledge graphs, devoting particular attention to *causal* knowledge (Al-Khatib et al., 2020; Heindorf et al., 2020). Causality plays a major role in argumentation due to its frequent usage in real-life discussions; *argument from cause to effect* and *argument from consequences* are frequently used argumentation schemes (Feng and Hirst, 2011; Reisert et al., 2018).

To utilize AKGs for argument generation, we collect argumentative texts from diverse sources such as online debate portals. In these texts, we find arguments that contain instances of the knowledge covered in the graphs. We encode this knowledge as keyphrases in the arguments. Unlike Gretz et al. (2020) and Schiller et al. (2021), our keyphrases cover multiple aspects and stances related to the same topic. The resulting texts are used to fine-tune a transformer-based generation model, GPT-2 (Radford et al., 2019). The underlying hypothesis is

*Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 4744–4754
August 1–6, 2021. ©2021 Association for Computational Linguistics

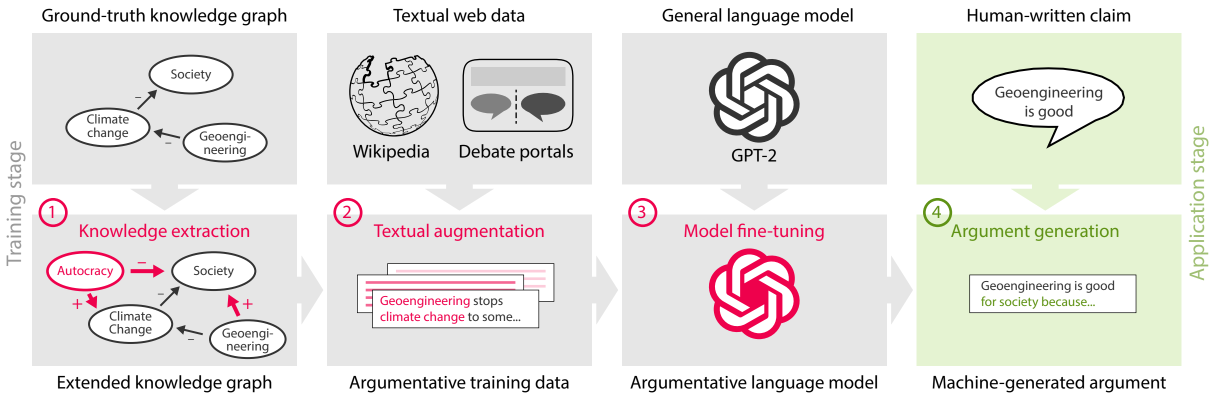

Figure 1: The main steps of our approach: (1) Given an argumentation knowledge graph, (possibly) extended by knowledge mined automatically, (2) texts are retrieved from the web to augment argument generation. (3) Pairs of text and knowledge are used to fine-tune GPT-2. (4) The model generates an argumentative text for a given prompt.

that GPT-2 will use the keyphrases to constrain the generation of arguments. During application, we provide the model with knowledge (as keyphrases) to obtain new arguments that further elaborate the knowledge. Figure 1 gives an overview of the main steps of our approach.

We evaluate the ability of our approach to generating new arguments for a variety of claim-like prompts: 400 generated arguments are manually assessed for their relevance to the prompt, argumentativeness, content richness, and plausibility. As a recent study indicates the adoption of bias from argumentative source data in word embeddings (Spliethöver and Wachsmuth, 2020), we also inspect potential social bias and abusive language in the generated arguments. Moreover, we evaluate the generated arguments automatically using recently developed argument mining techniques, in order to then examine correlations between manual and automatic evaluations. The results reveal an evident benefit of using the graphs' knowledge in generating controlled arguments that are rich in content and plausible. However, we also observe the presence of social bias in the outputs of GPT-2, suggesting the need for careful postproceeing step in argument generation.

Both the resources and the code developed in this paper will be made available.[1]

## 2 Knowledge Graphs for Argumentation

We use knowledge graphs (KGs) to plan the content of an argument to be generated and to control its *talking points*. A talking point is a specific aspect related to a given discussion topic. For instance, "health" is a talking point related to "smoking."

In this section, we describe the construction of three graphs related to argumentation: (1) a ground-truth argumentation knowledge graph, which is utilized based on Al-Khatib et al. (2020), (2) a generated argumentation knowledge graph, which is newly constructed from a set of argumentative texts, and (3) a causality graph, which is built upon Heindorf et al. (2020).

### 2.1 Ground-truth Knowledge Graph

Al-Khatib et al. (2020) propose a graph model that encodes the knowledge contained in arguments as relations (identified as the graph's edges) between concepts (identified as the graph's nodes). A concept is a noun phrase that represents an entity, an event, or an abstract idea. A relation represents the positive or negative effect that a concept has on another one. A relation is positive if *concept A promotes/causes/increases concept B*, and it is negative if *concept A suppresses/prevents/stops concept B*. A concept has two types of attributes: (1) groundings, which link concepts to the corresponding entries in a knowledge base such as Wikidata, (2) consequences, stating whether a concept is viewed as predominantly good or bad.

We slightly modify the outlined model to render the processing of the graph more amenable for our purposes. Instead of considering consequences as concept attributes, they are here modeled as an effect relation type: a good consequence is mapped to a positive effect, and a bad consequence to a negative effect. For example, "smoking is bad for health" is mapped to "smoking has a negative effect on health."

Accordingly, we populate the graph using the ar-

[1] https://github.com/webis-de/ACL-21

| Graph | #Nodes | #Edges | #Pos. | #Neg. |
|---|---|---|---|---|
| (A) Ground-truth | 4,607 | 9,100 | 4,904 | 4,196 |
| (B) Generated | 19,181 | 14,643 | 13,003 | 1,640 |
| (C) Causality | 74,356 | 179,701 | 179,701 | 0 |

Table 1: Counds of nodes and edges in the three graphs, the latter separated into positive and negative effect.

gumentation knowledge corpus of Al-Khatib et al. (2020), which comprises 16,429 *manual* annotations of 4,740 claims crawled from the online debate portal *debatepedia.org*. The population step results in the respective concept nodes (along with their groundings), which are connected by the two types of relations mentioned above.

We conduct a post-processing step to refine the graph including the removal of special characters and stop words at the beginning of the concepts, the changing of concepts from the plural to singular, and the decomposition of some concepts into two or more based on a set of conjunctions such as "and," "or," etc. For example, the concept of "depression and anxiety problems" will be decomposed into "depression" and "anxiety problems."

Table 1 (row A) shows statistics of this argumentation knowledge graph, which contains 4,607 nodes and 9,100 relations.

## 2.2 Generated Knowledge Graph

Since the ground-truth graph is limited in size, and since we aim for a higher coverage of knowledge from different controversial topics, we construct an additional new graph automatically.

**Data Source** The newly generated graph is derived from two resources: *args.me* and *kialo*.[2]

*Args.me* is the corpus underlying the argument search engine args.me (Ajjour et al., 2019). It comprises arguments from four online debate portals: *debate.org*, *debatewise.org*, *debatepedia.org* and *idebate.org*. We exclude *debate.org*, since it contains argumentative dialogues with frequent debate and user-meta information. In total, the corpus includes 30,748 arguments from the three considered debate portals.

Kialo is a debate portal in which argumentation is structured as trees. The platform comprises high-quality arguments as a result of the careful and substantial moderation. We crawled 1,640 discussions from *kialo.com*. From these, we obtained arguments by concatenating texts in the discussion levels of the tree (i.e., premises) with the texts in the tree roots (i.e., claims). Overall, we got 82,728 arguments from *Kialo*.

**Graph Construction** We followed the scheme of the manually generated argumentation knowledge graph described in the previous section, and identified concepts and relations in argumentative texts using the argument knowledge relation extraction approach of Al-Khatib et al. (2020). The approach comprised two main steps: (1) identifying whether a given text encodes an effect relation, and its type if any, and (2) finding the concepts of the identified relation. Specifically, for a given sentence, we extracted zero, one, or several argument knowledge relation instances in the format {*concept A, positive/negative effect, concept B*}.

We segmented all the arguments from the two sources into sentences and applied the argument knowledge relation extraction approach to all sentences, obtaining 11,537 and 17,688 relation instances from *args.me* and *Kialo*, respectively.

To improve the quality of the generated knowledge graph, we conducted the post-processing that we did for the manually generated argumentation knowledge graph. To reduce the observed noise and to exclude ill-formed concepts, we additionally filtered out concepts that are longer than seven words as well as those that comprise only one word, if it is not a noun. To increase the precision of the identified relation types, we extract the main verb of each sentence, and check the effect type of the verb using three lexicons: *+/-EffectWordNet* (Choi and Wiebe, 2014), *Connotation Frames* (Rashkin et al., 2015), and *ConnotationWordNet* (Kang et al., 2014). If the effect type of the knowledge relation instance obtained from this sentence contrasted with the effect type of its main verb (identified by any of the three lexicons), we excluded the instance obtained from this sentence.

Our new automatically-generated argumentation knowledge graph is built on top of these post-processed argument knowledge relation instances. Table 1 (row B) shows statistics of the new graph. It contains 19,181 nodes and 14,643 relations.

## 2.3 Causality Knowledge Graph

Recently, Heindorf et al. (2020) built a new causal knowledge graph which focuses on causal relations

[2] We also experimented with CMV, a discussion forum on the portal Reddit (i.e., a subreddit) which hosts argumentative discussions. However, due to the subreddit's dialogical nature and the use of informal language, the results were not convincing even when considering only the top-level posts.

between concepts. The construction of the KG was done by applying different information extraction techniques including bootstrapping, linguistic patterns, and sequence tagging on ClueWeb12 and Wikipedia. The corpus comes with two versions: a high-recall version with more than 11 million causal relations and a high-precision version with only around 200k relations. We make use of the high-precision version to build a new graph which is inline with the scheme of the two argumentation knowledge graphs described above. In particular, we map the *cause* relation to the *positive effect* relation since the former is a special case of the latter. We further exclude some noisy instances that contain the same concepts in a causal relation (e.g., *concept A* causes *concept A*). In total, the final graph comprises 74,356 nodes and 179,701 edges as shown in Table 1 (row C).

### 2.4 Graph Analysis

Table 2 shows examples of the knowledge in the graphs. To gain insights into the three graphs and their relationships, we analyzed the central concepts in each graph and the overlap between them.

**Graph Central Concepts** We use the centrality degree to get the most central nodes in each graph. For the graph constructed manually, we found the most central nodes to be controversial topics as well as some general concepts that affect our lives in general. A similar observation can be made for the second knowledge graph, but with an additional set of controversial topics. Most central concepts in the causality graph are related to health. Table 3 shows examples of the central concepts in the graphs.

**Graph Overlap.** We checked overlap between nodes among the three graphs. The ground-truth graph and the generated graphs have 1,424 overlapping nodes. Concretely, 908 nodes from the ground-truth KG match with those from the causality KG, and 2,326 from the generated KG match with those from the Causality KG. We note that the causality graph, albeit mostly covering general and health-related concepts, overlaps with the other two graphs in several controversial topics such as "*climate change*" and "*abortion*".

## 3 Neural Argument Generation

We now present our approach to integrate the argumentation knowledge graphs such as those described above into a neural text generation model.

| |
|---|
| Stability of a country bank system $\overset{\text{positive}}{\longmapsto}$ Economic stability |
| Raise oil price $\overset{\text{negative}}{\longmapsto}$ World oil industry |
| Legalizing marijuana $\overset{\text{positive}}{\longmapsto}$ Tourism industry |
| Online social vigilantism $\overset{\text{negative}}{\longmapsto}$ Insulting behavimy |
| Economic growth $\overset{\text{positive}}{\longmapsto}$ Global warming |
| Human parainfluenza viruse $\overset{\text{positive}}{\longmapsto}$ Viral pneumonium |

Table 2: Examples of the knowledge in the three constructed knowledge graphs.

| (A) Ground-truth | (B) Generated | (C) Causality |
|---|---|---|
| Global warming | Liquid democracy | Disease |
| Free speech | Unisex bathroom | Poverty |
| Public safety | Affirmative action | Violence |
| Public insurance | Religion | Confusion |
| Circumcision | Polygamy | Depression |
| Globalization | Capitalism | Obesity |

Table 3: Examples of the central concepts in the three constructed knowledge graphs.

### 3.1 Text Collection

To construct a dataset for fine-tuning a generation model, we first collect a set of argumentative texts which are likely aligned with the knowledge graphs we have constructed in Section 2.

Since our goal is to lead the text generation process towards arguments, we use texts from *args.me* and *kialo* (see Section 2). The two resources contain mostly argumentative texts, many of which cover concepts from the graphs. In addition, we use *Wikipedia* as we expect it to cover various facts for a large portion of concepts in the graphs. Specifically, we sample a set of articles from Wikipedia that address the concept groundings present in the ground-truth argumentation knowledge graph (altogether 2,050 articles). The articles are split into 81,872 paragraphs based on their structure.

### 3.2 Text-Knowledge Encoding

In each paragraph from all three sources described above, we identify all concepts found in the knowledge graphs using string matching. We add pairs of concepts that are connected in the graph to the beginning of the paragraph, encoding them with the type of effect relation between them as keyphrases separated by special tokens. We use 'positive' and 'negative' to represent the effect relations. For example, the paragraph

> *"Animal studies suggests marijuana causes physical dependence, and serious problems"*

will be transformed into:

*“<|startoftext|>’[’marijuana»positive»physical-dependence’, ’mariguana»positive»problems’] @ Animal studies suggests ...’<|endoftext|>”*

While this way of matching and encoding has limitations, it has shown good results in practice when used with pre-trained neural models (Witteveen and Andrews, 2019; Cachola et al., 2020).

### 3.3 Neural Language Model Fine-tuning

We use our text-knowledge encoding dataset to fine-tune the GPT-2 neural language model (Radford et al., 2019) for argument generation. Since GPT-2 cannot deal with graph structure as input directly, we fine-tune it on all paragraphs, including those with encoded relations as textual representations (i.e., keyphrases). We expect to thereby leverage the powerful generation capabilities of GPT-2 while biasing it to generate texts related to the encoded relations.

It is worth noting that, in training, we encode multiple relations at once and the generated arguments are paragraphs. The encoded relations are often related to different aspects of the same topic. This is different from previous studies (Gretz et al., 2020; Schiller et al., 2021) which only focus on generating an argumentative sentence based on a single topic or one aspect/stance of a topic. As a result, we expect that our fine-tuning strategy based on knowledge graphs can assist users to plan several “talking points” and generate the corresponding argument which covers the different aspects.

## 4 Experiments and Results

In this section, we report on the manual and automatic evaluation of our approach from Section 3 to employ the three argumentation knowledge graphs from Section 2 for neural argument generation:

A. The ground-truth graph

B. The generated graph

C. The causality graph

### 4.1 Experimental Set-up

We used the following experimental setup:

**Model Parameters** In all experiments, we fine-tuned the pre-trained GPT-2 model with 127M parameters using *gpt-2-simple library*.[3] For argument generation, we follow Gretz et al. (2020) in setting *top_k* to 40 and *temperature* to 0.7. Also, we set the *batch_size* to 2 and the *steps* to 1500. We specify the *length* of the generated arguments to be 100 (approximately, the mean number of words of the arguments in our data). As postprocessing, we removed non-ASCII characters and several improper symbols from the generated arguments. The fine-tuning took around 16 hours on a GPU Tesla T4.

**Argument Generation Models** For fine-tuning the generation model, there are various possible combinations of the three constructed graphs and the datasets. Based on initial tests of potentially promising combinations, we decided to address the following models in order to examine the impact of the graphs as well as the data:

1. *GPT-2.* As a baseline, we use the raw GPT-2 model without any fine-tuning or graph usage.
2. *ArgData.* This model is based on fine-tuning GPT-2 using the argumentative texts from *Kialo* and *args.me* in our constructed data. No knowledge from the graphs is used here.
3. *AB-ArgData.* Similar to the previous model, but the knowledge of the graphs A and B are encoded into the argumentative texts. Concretely, we combine A and B as follows: First, we compute the intersection of A and B. Then, we add the nodes and edges of A to the resulting intersection subgraph of B, including the nodes of this subgraph as well as their neighbors. Thereby, we reduce the usage of noisy knowledge, preferring knowledge with direct connections.[4]
4. *ABC-ArgData.* Just like the previous model, but we consider the knowledge of graph C in addition to A and B. We compose the graph above and C analog to above. The rationale is here to prefer argumentative knowledge over more general knowledge. The graph C is several orders of magnitude larger than A and B; considering the complete graph of C would thus likely eliminate the impact of A and B.
5. *ABC-FullData.* Analogous to the model before, but here we use the Wikipedia subset of our data in addition to the argumentative one.

In general, those models help investigate the impact of adding one type of information (data or

[3] https://github.com/minimaxir/gpt-2-simple

[4] In other words, we consider the complete graph A, since A is the one with highest precision, and we induce a subset of graph B that is related to A. Our inspections suggested that this subset has much higher precision than the complete graph B.

**Prompt:** *Multiculturalism is positive for tolerant society.*

**GPT-2:** no guarantee that the world of cultural evolution is going to be one of a kind. in a world where the majority of people are now tolerant, where many people still believe in evolution, we have to accept the world of cultural evolution as being a far more complicated...
**ArgData:** multiculturalism is a good way to go about making the world a more tolerant place. in the u.s., more than half americans think their country has more tolerance and diversity than other countries...
**AB-ArgData:** multiculturalism will allow for more tolerant societies. multiculturalism is already a force for good, helping to bring tolerance and diversity to the world. a multicultural society will bring such things as tolerance, kindness, and respect for everyone...
**ABC-ArgData:** multiculturalism will increase the diversity of the population. the european union eu was created to foster tolerance towards many cultures, but it is still intolerant towards many other cultures...
**ABC-FullData:** multiculturalism is an accepted part of a multicultural society. the majority of the population of a multicultural society are not religious, not socially or culturally dominant, and do not have political power...

Table 4: Examples of the arguments generated in response to the prompt by each of the evaluated approaches.

| Model | Args.me | Kialo | Wikipedia |
|---|---|---|---|
| AB-ArgData | 104,923 | 65,617 | – |
| ABC-ArgData | 367,697 | 204,651 | – |
| ABC-FullData | 367,697 | 204,651 | 943,070 |

Table 5: Number of relations (knowledge instances) for each of the graph models encoded in the argumentative texts from args.me and Kialo as well as in Wikipedia.

graph) on the quality of the generated arguments. Statistics of the knowledge encoded in the argumentative and full datasets are given in Table 5.

**Train-Test Data Split** We processed the data excluding all paragraphs related to five randomly-selected controversial topics: ‘Geoengineering’, ‘Renewable Energy’, ‘Illegal Immigration’, ‘Electoral College’, and ‘Multiculturalism’. The resulting paragraphs are used for training the models, while the five topics are used for generating prompts to test the models. Accordingly, the ArgData training set includes 112,658 arguments, and the FullData training set comprises 194,032 arguments and Wikipedia paragraphs.

**Model Prompts** We chose different knowledge instances related to the five selected topics and used them as prompts for the generation models. The knowledge includes the topic name (e.g., ‘Geoengineering’), edges from the graphs (e.g., ‘Geoengineering positive for climate change’), and graph paths (e.g., ‘geoengineering solutions are negative for atmospheric greenhouse gas, and atmospheric greenhouse gas are negative for earth’). For GPT-2 and ArgData, we represented the knowledge as coherent texts similar to the examples above. For the remaining models, we represented it in the same way that we encoded it in the data (e.g., ‘geoengineering»positive»unexpected consequences’).

## 4.2 Manual Evaluation

For evaluation, we generated 400 arguments using the prompts discussed above. Specifically, each model generated 16 arguments for each of the five test topics (80 arguments in total). Table 4 shows some examples of the generated arguments.

**Annotation Task** The evaluation was done by five workers hired on the freelancing platform, *Upwork*. The workers were writing experts, with a solid background in argumentation. They had at least 94% job success with more than 40 previous jobs on the platform. Each worker assessed the generated arguments from all models for two test topics, seeing all variants at the same time. Thus, each model was evaluated by two different workers. We paid each worker EUR 140 in total. The average time to complete the task was nine hours.

The assessment of the arguments given their prompts was conducted based on five dimensions:

- *Relevance.* Does the text comprise content relevant to the given knowledge?
- *Argumentativeness.* Does the text convey an explicit or implicit pro or con stance towards any topic?
- *Content Richness.* Does the text contain useful information and cover different aspects?
- *Plausibility.* Does the text comprise plausible content and does it not contrast with commonsense knowledge?
- *Bias.* Does the text include any social bias or abusive language?

The first four are adopted from Hua and Wang (2018) and Gretz et al. (2020). We added the last one in light of the observations of Spliethöver and Wachsmuth (2020). The first four dimensions were

| # | Model | Relevance | Argumentativeness | Content Richness | Plausibility | Bias |
|---|---|---|---|---|---|---|
| 1 | GPT-2 | 1.80 | 2.23 | 2.11 | 2.33 | 6% |
| 2 | ArgData | 1.91 | **2.50** | 2.10 | 2.20 | 13% |
| 3 | AB-ArgData | 2.00 | **2.50** | 2.14 | **2.34** | 6% |
| 4 | ABC-ArgData | **2.10** | 2.45 | **2.16** | 2.27 | 13% |
| 5 | ABC-FullData | 1.85 | 2.26 | 2.10 | 2.04 | 6% |

Table 6: Manual evaluation: Average scores between 1 (worst) and 3 (best) for the first four dimensions and proportion of generated arguments reported to have bias. The best values are marked bold.

scored from 1 to 3 (1 being worst), while the last one was answered with "yes" or "no".

We directed the workers to consider the length of the argument (100 words) in their assessments. We also asked them to keep in mind that the text should be self-contained; it should not be necessary to see the prompts to understand the text. As regards the argumentativeness dimension, we defined the scores to indicate 'no stance' (score 1), 'mixed stances' (2), and 'one stance' (3) of the generated argument. Unlike previous work, we omitted *fluency* as a dimension in our evaluation, since all the models are based on GPT-2, which is known to generate mostly fluent text. We manually checked a few samples, though, to confirm the reasonable fluency of the generated arguments.

**Results** Table 6 shows the resulting scores of all approaches in the manual evaluation. The inter-annotator agreement between the workers is 0.40 in terms of Fleiss' $\kappa$.

All models constructed with our data and graphs outperform the raw *GPT-2* model in most cases. For *relevance*, the model with the three graphs and the argumentative data, *ABC-ArgData*, performs best (2.10), followed by *AB-ArgData* (2.00). Such results clearly demonstrate the impact of the graphs in controlling the generated arguments. One exception is *ABC-FullData*, where it seems that using Wikipedia produces some shifts in topics in the generated arguments. Regarding *argumentativeness*, the models that were developed using the argumentative data achieve the highest score, leaving GPT-2 and ABC-FullData behind. As for *content richness*, ABC-ArgData reaches the highest scores, marginally higher than AB-ArgData and the other models. In general, all models show comparable performance for this dimension. For *plausibility*, the score of AB-ArgData is highest, closely followed by GPT-2, though. Despite failing on the other dimensions, GPT-2 apparently generates comparably plausible texts when having argumentation knowledge as prompts.

As regards the last dimension, it seems that the output of all models sometimes conveys *bias*. However, this dimension appears to be very subjective, as only two workers reported biased arguments at all. Most of the reported arguments are about illegal immigration and multiculturalism. Examples include "the British are a big threat to the idea of multiculturalism" and "The latest attempt to bring the problem under control is the proposal to ban black people from entering the country."

### 4.3 Automatic Evaluation

In the automatic evaluation of arguments, we aimed to approximate dimensions from the manual evaluation. On one hand, this was to keep the focus on argumentation-related aspects. On the other hand, it allows for a rough comparison between the manual and the automatic evaluation results. Based on recent computational argumentation technologies, we assessed three dimensions as follows:

- *Relevance.* We computed the overlap between an argument's words and the prompt's words, after excluding stop words. To match the manual evaluation scores, we mapped full overlap to 3, partial overlap to 2, and no overlap to 1.
- *Argumentativeness.* We detected the stance of each argument using the approach of Stab et al. (2018), which has been shown to be effective in dealing with arguments from heterogeneous sources, topics, and domains. In particular, we checked the stance (pro or con) for each sentence, considering its topic. We scored the argument with 1 in case no stance is detected, 2 if two different stances are detected (pro and con), and 3 if only one stance is detected.
- *Content Richness.* As we consider an argument to be rich in content if it covers different aspects of a topic, we used the model of Schiller et al. (2021) for identifying aspects in arguments. We then mapped the number of detected aspects to scores heuristically: we

| Model | Relevance | Argumentativeness | Richness |
|---|---|---|---|
| GPT2 | 1.82 | 2.52 | 1.59 |
| ArgData | 2.26 | 2.70 | 1.94 |
| AB-ArgData | **2.36** | 2.79 | 2.02 |
| ABC-ArgData | 2.35 | **2.85** | **2.10** |
| ABC-FullData | 2.10 | 2.67 | 2.08 |

Table 7: The results of the automatic evaluation of the five models on the 400 generated arguments. The highest average score of each dimension is marked bold.

gave score 1 to arguments with maximum two aspects, score 2 for three to five aspects, and score 3 for more than five.

**Results** Table 7 presents the results of our automatic evaluation.

Again, all models perform better than *GPT-2*. In terms of *relevance*, *AB-ArgData* (2.36) and *ABC-ArgData* (2.35) are on par. Regarding *argumentativeness*, ABC-ArgData is the best with an average score of 2.85, and AB-ArgData follows with 2.79. Lastly, for *content richness*, ABC-ArgData again achieves the highest score (2.10), followed by ABC-FullData and AB-ArgData with 2.08 and 2.02, respectively. The results suggest that ABC-ArgData is the best model overall, followed by AB-ArgData. This emphasizes the impact of encoding the knowledge of the graphs into argumentative data for argument generation.

Comparing the scores of the automatic evaluation to the manual one, we observe rather comparable ranks of the models regarding the three dimensions considered.

### 4.4 Discussion

Inspecting the arguments generated by the models, we observe that their quality varies depending on the topic of the knowledge (e.g., nuclear energy) and their complexity (single or multiple-relations). We also find that the beginning of a generated argument often has higher quality than the end part. For example, some models start generating relations such as ‘x is positive for y’ instead of a text at the end of the arguments. The reason for this difference in quality could be the minimum length of arguments that we force the model to satisfy. Besides, the arguments have several problems, related to those that occur frequently with neural text generation models, such as duplication, contradicting statements, and topic shifting.

In general, we see that the quality of the automatically generated arguments still not on par with human written arguments. Nevertheless, the experiment results show that our approach for controlling the generated arguments using argumentation knowledge graphs improves the quality.

Still, our approach can be improved in several respects. First, argumentation knowledge graphs, especially those which are constructed automatically, might contain knowledge that is noisy, too specific, very abstract, or difficult to be interpreted without context. While we tried to limit such noise as much as possible (see Section 2.2), more sophisticated noise filtering and a ranking of knowledge based on its quality could be an essential improvement step. Besides, we used the simple method of string matching for finding the graphs' knowledge in the collected argumentative texts. Advanced methods utilizing semantic similarity could lead to more accurate matching. Moreover, although encoding the knowledge as keyphrases seems a reasonable method, different representations that consider the structure of the knowledge are worth investigating (see Section 3.2). Lastly, since our approach is meant as a proof of concept, we used the small GPT-2 model with the parameters adopted from Gretz et al. (2020). Using a larger model and exploring different sampling methods and parameter settings will probably result in a higher quality of the arguments generated.

## 5 Related Work

In this section, we outline related studies on argument generation, argumentation knowledge graphs, and graph-to-text generation.

**Argument Generation** Different approaches to the generation of arguments, or of components thereof, have been proposed in the last years. To create new claims, Bilu and Slonim (2016) recomposed predicates from existing claims with new topics. El Baff et al. (2019) composed complete arguments from given claims following specific rhetorical strategies based on the theoretical model of Wachsmuth et al. (2018). Unlike these approaches, we make use of neural language models.

Hidey and McKeown (2019) built a sequence-to-sequence model to rewrite claims into opposing claims, and Hua et al. (2019) presented a sophisticated approach that, given a stance on a controversial topic, combines retrieval with neural generation techniques to create full arguments with the opposite stance. Gretz et al. (2020) developed a transformer-based pipeline to generate coherent and plausible claims, whereas Schiller et al. (2021) proposed a language model that controls argument generation on a fine-grained level for a given topic, stance, and aspect. Lastly, Alshomary et al. (2021) generated belief-based claims, encoding the beliefs via conditional language models.

Most similar to our work are the studies of Gretz et al. (2020) and Schiller et al. (2021). Like us, the former also exploits the power of GPT-2, adding context to the model's training data. The latter is comparable in that it attempts to steer the generation towards aspect-specific arguments. To the best of our knowledge, however, our approach is the first to employ external knowledge from knowledge graphs for the task of argument generation.

**Argumentation Knowledge Graphs** Besides the argumentation knowledge graph of Al-Khatib et al. (2020), Toledo-Ronen et al. (2016) created an expert stance graph to support stance classification. Gemechu and Reed (2019) encoded the relations between segments of an argument into a graph and demonstrated the graph's effectiveness for argument mining. In our work, we utilize one of the available graphs, among others, using its knowledge to control the argument generation process.

Closely related to argumentation knowledge, causality graphs gained some attention recently. While general knowledge bases such as ConceptNet (Speer et al., 2017) contain causal knowledge, the causality graph of Heindorf et al. (2020) that we utilized is the largest source of causal knowledge, exceeding others by orders of magnitude.

**Graph-to-Text Generation** In the related area of neural graph-to-text generation, researchers have used various techniques (Song et al., 2018; Koncel-Kedziorski et al., 2019b; Schmitt et al., 2020). Within this area, the approaches most related to ours are those that exploit the usage of knowledge in graphs as input to sequence-to-sequence models (Moryossef et al., 2019) as well as those that make use of large pre-trained language models such as Liu et al. (2021), where the pretrained model BART is augmented by knowledge from a graph for generative commonsense reasoning.

Overall, our work concentrates on the context of argumentation, with an approach to encoding different types of argumentation knowledge into the pretrained model GPT-2 in order to allow for more controlled argument generation.

## 6 Conclusion

This paper tackles argument generation through the use of argumentation knowledge graphs. We have discussed how to take advantage of different manually and automatically created knowledge graphs to encode knowledge in argumentative texts, and how to utilize these texts to fine-tune GPT-2. Our approach is able to generate high-quality arguments for various inputs, including complex relational knowledge. Besides, we proposed a simple method for evaluating arguments automatically, with results correlating to those observed in the manual evaluation. In our future research, we plan to leverage more sources and evaluate other knowledge encoding methods. Moreover, we will study different directions to illuminate the possible social bias in argument generation methods.

## Ethics Statement

As this paper presents a computational method for generating arguments automatically, different ethical restrictions deserve discussion.

First, we have used only publicly available, non-personalized sources for our text collection. When crawling data from web platforms, we followed the platforms' policies, adhering to their usage rules.

Second, although we restricted the sources of our dataset and knowledge graphs to those trustworthy of having high quality, the generated arguments included some undesirable materials, such as abusive language and social bias. To account for these findings, we strongly suggest a postprocessing step to filter out such content when using respective data. Moreover, we explicitly checked for bias in the arguments we generated, as presented.

Arguments are a powerful means for changing people's stances and impact the attitude of communities. To prevent unethical use, such as generating arguments on controversial topics with specific stances and deploying them on social platforms, we will try to restrict the distribution of the data and code to researchers and academic institutions. This seems necessary since we are aware that there is no guarantee that the generated arguments are always factually correct.

## Acknowledgments

The first author is supported by the German Federal Ministry of Education and Research (BMBF, 01/S18026A-F) by funding the competence center for Big Data and AI (ScaDS.AI Dresden/Leipzig).

## References

Yamen Ajjour, Henning Wachsmuth, Johannes Kiesel, Martin Potthast, Matthias Hagen, and Benno Stein. 2019. Data acquisition for argument search: The args.me corpus. In *KI 2019: Advances in Artificial Intelligence - 42nd German Conference on AI, Kassel, Germany, September 23-26, 2019, Proceedings*, pages 48–59.

Khalid Al-Khatib, Yufang Hou, Henning Wachsmuth, Charles Jochim, Francesca Bonin, and Benno Stein. 2020. End-to-end argumentation knowledge graph construction. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 7367–7374. AAAI.

Milad Alshomary, Wei-Fan Chen, Timon Gurcke, and Henning Wachsmuth. 2021. Belief-based generation of argumentative claims. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 224–233, Online. Association for Computational Linguistics.

Yonatan Bilu and Noam Slonim. 2016. Claim synthesis via predicate recycling. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 525–530, Berlin, Germany. Association for Computational Linguistics.

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel Weld. 2020. TLDR: Extreme summarization of scientific documents. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4766–4777, Online. Association for Computational Linguistics.

Yoonjung Choi and Janyce Wiebe. 2014. +/-EffectWordNet: Sense-level lexicon acquisition for opinion inference. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1181–1191, Doha, Qatar. Association for Computational Linguistics.

Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, Manfred Stede, and Benno Stein. 2019. Computational argumentation synthesis as a language modeling task. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 54–64, Tokyo, Japan. Association for Computational Linguistics.

Vanessa Wei Feng and Graeme Hirst. 2011. Classifying arguments by scheme. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 987–996, Portland, Oregon, USA. Association for Computational Linguistics.

Debela Gemechu and Chris Reed. 2019. Decompositional argument mining: A general purpose approach for argument graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 516–526, Florence, Italy. Association for Computational Linguistics.

Shai Gretz, Yonatan Bilu, Edo Cohen-Karlik, and Noam Slonim. 2020. The workweek is the best time to start a family – a study of GPT-2 based claim generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 528–544, Online. Association for Computational Linguistics.

Stefan Heindorf, Yan Scholten, Henning Wachsmuth, Axel-Cyrille Ngonga Ngomo, and Martin Potthast. 2020. Causenet: Towards a causality graph extracted from the web. In *CIKM*. ACM.

Christopher Hidey and Kathy McKeown. 2019. Fixed that for you: Generating contrastive claims with semantic edits. 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 1756–1767, Minneapolis, Minnesota. Association for Computational Linguistics.

Xinyu Hua, Zhe Hu, and Lu Wang. 2019. Argument generation with retrieval, planning, and realization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2661–2672, Florence, Italy. Association for Computational Linguistics.

Xinyu Hua and Lu Wang. 2018. Neural argument generation augmented with externally retrieved evidence. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 219–230, Melbourne, Australia. Association for Computational Linguistics.

Jun Seok Kang, Song Feng, Leman Akoglu, and Yejin Choi. 2014. Connotationwordnet: Learning connotation over the word+sense network. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 1544–1554. The Association for Computer Linguistics.

Rik Koncel-Kedziorski, Dhanush Bekal, Yi Luan, Mirella Lapata, and Hannaneh Hajishirzi. 2019a. Text Generation from Knowledge Graphs with Graph Transformers. 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 2284–2293. Association for Computational Linguistics.

Rik Koncel-Kedziorski, Dhanush Bekal, Yi Luan, Mirella Lapata, and Hannaneh Hajishirzi. 2019b. Text Generation from Knowledge Graphs with Graph Transformers. 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 2284–2293. Association for Computational Linguistics.

Ye Liu, Yao Wan, Lifang He, Hao Peng, and Philip S. Yu. 2021. KG-BART: Knowledge graph-augmented bart for generative commonsense reasoning. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence*. AAAI.

Amit Moryossef, Yoav Goldberg, and Ido Dagan. 2019. Step-by-step: Separating planning from realization in neural data-to-text generation. 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 2267–2277. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Hannah Rashkin, Sameer Singh, and Yejin Choi. 2015. Connotation frames: Typed relations of implied sentiment in predicate-argument structure. *CoRR*, abs/1506.02739.

Paul Reisert, Naoya Inoue, Tatsuki Kuribayashi, and Kentaro Inui. 2018. Feasible annotation scheme for capturing policy argument reasoning using argument templates. In *Proceedings of the 5th Workshop on Argument Mining*, pages 79–89, Brussels, Belgium. Association for Computational Linguistics.

Leonardo F. R. Ribeiro, Martin Schmitt, Hinrich Schütze, and Iryna Gurevych. 2020. Investigating pretrained language models for graph-to-text generation.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Aspect-controlled neural argument generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–396, Online. Association for Computational Linguistics.

Martin Schmitt, Sahand Sharifzadeh, Volker Tresp, and Hinrich Schütze. 2020. An unsupervised joint system for text generation from knowledge graphs and semantic parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7117–7130, Online. Association for Computational Linguistics.

Linfeng Song, Yue Zhang, Zhiguo Wang, and Daniel Gildea. 2018. A graph-to-sequence model for AMR-to-text generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1616–1626, Melbourne, Australia. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4444–4451. AAAI Press.

Maximilian Spliethöver and Henning Wachsmuth. 2020. Argument from old man's view: Assessing social bias in argumentation. In *Proceedings of the 7th Workshop on Argument Mining*, pages 76–87, Online. Association for Computational Linguistics.

Christian Stab, Johannes Daxenberger, Chris Stahlhut, Tristan Miller, Benjamin Schiller, Christopher Tauchmann, Steffen Eger, and Iryna Gurevych. 2018. ArgumenText: Searching for arguments in heterogeneous sources. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 21–25. Association for Computational Linguistics.

Orith Toledo-Ronen, Roy Bar-Haim, and Noam Slonim. 2016. Expert stance graphs for computational argumentation. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 119–123, Berlin, Germany. Association for Computational Linguistics.

Henning Wachsmuth, Manfred Stede, Roxanne El Baff, Khalid Al Khatib, Maria Skeppstedt, and Benno Stein. 2018. Argumentation synthesis following rhetorical strategies. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3753–3765. Association for Computational Linguistics.

Sam Witteveen and Martin Andrews. 2019. Paraphrasing with large language models. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 215–220, Hong Kong. Association for Computational Linguistics.