Aspect-Controlled Neural Argument Generation

Benjamin Schiller and **Johannes Daxenberger** and **Iryna Gurevych**
Ubiquitous Knowledge Processing Lab
Department of Computer Science, Technical University of Darmstadt
http://www.ukp.tu-darmstadt.de/

Abstract

We rely on arguments in our daily lives to deliver our opinions and base them on evidence, making them more convincing in turn. However, finding and formulating arguments can be challenging. In this work, we present the *Arg-CTRL*—a language model for argument generation that can be controlled to generate sentence-level arguments for a given topic, stance, and aspect. We define argument aspect detection as a necessary method to allow this fine-granular control and crowdsource a dataset with 5,032 arguments annotated with aspects. Our evaluation shows that the Arg-CTRL is able to generate high-quality, aspect-specific arguments, applicable to automatic counter-argument generation. We publish the model weights and all datasets and code to train the Arg-CTRL.[1]

1 Introduction

Language models (Bengio et al., 2003) allow to generate text through learned distributions of a language and have been applied to a variety of areas like machine translation (Bahdanau et al., 2015), summarization (Paulus et al., 2018), or dialogue systems (Wen et al., 2017). A rather new field for these models is the task of producing text with argumentative content (Wang and Ling, 2016). We believe this technology can support humans in the challenging task of finding and formulating arguments. A politician might use this to prepare for a debate with a political opponent or for a press conference. It may be used to support students in writing argumentative essays or to enrich one-sided discussions with counter-arguments. In contrast to retrieval methods, generation allows to combine and stylistically adapt text (e.g. arguments) based on a given input (usually the beginning of a sentence). Current argument generation models, however, produce lengthy texts and allow the user little control over the aspect the argument should address (Hua et al., 2019; Hua and Wang, 2018). We show that argument generation can be enhanced by allowing for a fine-grained control and limiting the argument to a single but concise sentence.

Controllable language models like the *CTRL* (Keskar et al., 2019) allow to condition the model at training time to certain control codes. At inference, these can be used to direct the model's output with regard to content or style. We build upon this architecture to control argument generation based solely on a given topic, stance, and argument aspect. For instance, to enforce focus on the aspect of *cancer* for the topic of *nuclear energy*, we input a control code "*Nuclear Energy CON cancer*" that creates a contra argument discussing this aspect, for instance: "*Studies show that people living next to nuclear power plants have a higher risk of developing cancer.*".

To obtain control codes from training data, we pre-define a set of topics to retrieve documents for and rely on an existing stance detection model to classify whether a sentence argues in favor (*pro*) or against (*con*) the given topic (Stab et al., 2018a). Regarding argument aspect detection, however, past work has two drawbacks: it either uses simple rule-based extraction of verb- and noun-phrases (Fujii and Ishikawa, 2006) or the definition of aspects is based on target-concepts located within the same sentence (Gemechu and Reed, 2019). Aspects as we require and define them are not bound to any part-of-speech tag and (1) hold the core reason upon which the conclusion/evidence is built and (2) encode the stance towards a general but not necessarily explicitly mentioned topic the argument discusses. For instance:

> **Topic**: *Nuclear Energy*
> **Argument**: *Running nuclear reactors is costly as it involves long-time disposal of radioactive waste.*

The evidence of this argument is based upon the two underlined aspects. While these aspects encode

[1] https://github.com/UKPLab/controlled-argument-generation

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 380–396
June 6–11, 2021.

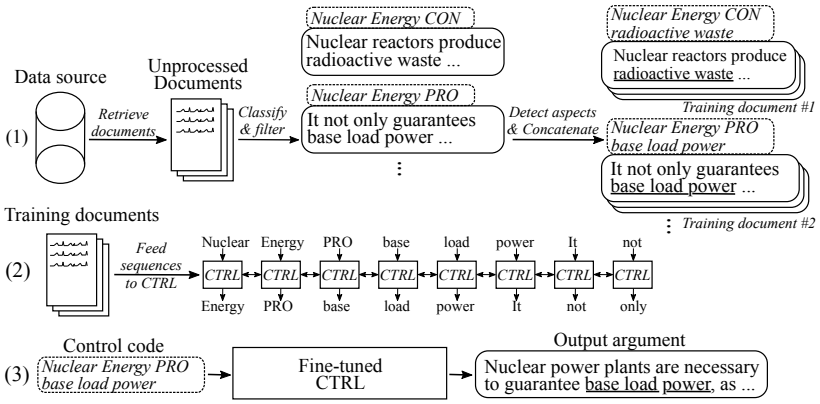

Figure 1: (1) Gather data from large data sources. Extract sentences, classify arguments, and detect aspects. Arguments sharing a topic & stance & aspect ($\hat{=}$ control code) are concatenated into training documents. (2) The model is fine-tuned on each document with the control code prepended to each input sequence. (3) At inference, the model only needs a control code to generate an argument that follows the given topic & stance & aspect.

a negative stance towards the topic of "Nuclear Energy", the topic itself is not mentioned explicitly in the argument.

Our final controlled argument generation pipeline (see Figure 1) works as follows: (1) We gather several million documents for eight different topics from two large data sources. All sentences are classified into pro-, con-, and non-arguments. We detect aspects of all arguments with a model trained on a novel dataset and concatenate arguments with the same topic, stance, and aspect into training documents. (2) We use the collected classified data to condition the Arg-CTRL on the topics, stances, and aspects of all gathered arguments. (3) At inference, passing the control code *[Topic] [Stance] [Aspect]* to the model will generate an argument that follows these commands.

Our evaluation shows that the Arg-CTRL is able to produce aspect-specific, high-quality arguments, applicable to automatic counter-argument generation. The contributions are as follows: (i) We adapt and fine-tune the CTRL for aspect-controlled neural argument generation. (ii) We show that detecting argument aspects and conditioning the generation model on them are necessary steps to control the model's training process and its perspective while generating. (iii) We propose several methods to analyze and evaluate the quality of (controllable) argument generation models. (iv) We develop a new scheme to annotate argument aspects and release a dataset with 5,032 samples.

2 Related Work

Argument Aspect Detection Early work by Fujii and Ishikawa (2006) focuses mainly on Japanese and restricts aspects to noun- and verb-phrases, extracted via hand-crafted rules. Boltužić and Šnajder (2017) extract noun-phrases and aggregate them into *concepts* to analyze the microstructure of claims. Misra et al. (2015) introduce facets as low level issues, used to support or attack an argumentation. In that, facets are conceptually similar to aspects, but not explicitly phrased and instead seen as abstract concepts that define clusters of semantically similar text-spans of summaries. Bilu et al. (2019) define commonplace arguments that are valid in several situations for specified actions (e.g. "ban") and topics (e.g. "smoking"). These actions are similar to aspects, but limited in number and manually defined. Gemechu and Reed (2019) detect, amongst others, concepts and aspects in arguments with models trained on expert annotations. However, in their definition, aspects have to point to a target concept mentioned in the argument. In our definition, aspects refer to a general topic which is not necessarily part of the sentence and our annotation scheme is applicable by non-experts.

The concept of framing dimensions (Boydstun et al., 2014) is close to argument aspects. In the field of argument mining, Ajjour et al. (2019) recently applied frames to label argument clusters. Yet, their method does not allow to detect frames. Other works present methods to automatically label sentences of news articles and online discussions with frames (Hartmann et al., 2019; Naderi and Hirst, 2017). These methods are, however, limited to a small set of predefined frames that represent high-level concepts. Contrarily, we operate on a fine-grained span-level to detect aspects that are explicitly mentioned in arguments.

Argument Generation Early approaches rely on rules from argumentation theory and user preference models (Carenini and Moore, 2006; Zukerman et al., 1998). In a more recent work, Sato et al. (2015) construct rules to find arguments in a large data source, which are then filtered and ordered with a neural network based ranker. Baff et al. (2019) use a clustering and regression approach to assemble discourse units (major claims, pro and con statements) to argumentative texts. However, most of these approaches rely on hand-crafted features and do not generalize well. Moreover, they all require permanent access to large data sources and are not able to generate new arguments.

Recently, research on generating arguments with language models gained more attention. Hua and Wang (2019) use a sequence to sequence model (Sutskever et al., 2014) to generate argumentative text by attending to the input and keyphrases automatically extracted for the input from, for example, Wikipedia. Other work focuses on generating argumentative dialogue (Le et al., 2018) and counter-arguments (Hidey and McKeown, 2019; Hua et al., 2019) based on a given input sentence, or on generating summaries from a set of arguments (Wang and Ling, 2016). Contrarily, we train a language model that does not require a sentence-level input for generation and allows for direct control over the topic, stance, and aspect of the produced argument.

Xing et al. (2017) design a language model that attends to topic information to generate responses for chatbots. Dathathri et al. (2019) train two models that control the sentiment and topic of the output of pre-trained language models at inference. Gretz et al. (2020a) fine-tune GPT-2 on existing, labeled datasets to generate claims for given topics. However, the latter works do not explore generation for such a fine-grained and explicit control as proposed in this work. We show that argument generation requires the concept of argument aspects to shape the produced argument's perspective and to allow for diverse arguments for a topic of interest.

3 Argument Aspect Detection

Argument aspect detection is necessary for our argument generation pipeline, as it allows for a fine-grained control over the generation process. We create a new dataset, as existing approaches either rely on coarse-grained frames or cannot be applied by non-expert annonators in a scalable manner.

3.1 Dataset Creation

We base our new aspect detection dataset on the UKP Sentential Argument Mining Corpus (UKP-Corpus) by Stab et al. (2018b), as it already contains sentence-level arguments and two of the control codes we aim to use: topics and stance labels. More precisely, it contains 25,474 manually labelled sentences for eight controversial topics in English. Each sample consists of a topic and a sentence, labelled as either being supporting, attacking, or no argument towards the given topic. As we are only interested in arguments, we do not consider the non-argumentative sentences.

Step 1: Preliminary annotations To ensure the feasibility of creating a dataset for this task, two experts (a post-doctoral researcher and an undergraduate student with NLP background) independently annotate 800 random samples (from four topics, 200 per topic) taken from the UKP-Corpus. The annotations are binary and on token-level, where multiple spans of tokens could be selected as aspects. The resulting inter-annotator agreement of this study is Krippendorff's $\alpha_u = .38$. While this shows that the task is generally feasible, the agreement on exact token spans is rather low. Hence, in the following steps, we reduce the complexity of the annotation task.

Step 2: Annotation scheme Instead of free span-level annotations, we present annotators with a ranked list of aspect recommendations. To generate meaningful recommendations, we train a ranking model using the preliminary annotations (Step 1).

Step 2a: Data preparation for ranking To create training data for the ranker, we use a simple heuristic to calculate scores between 0 and 1 for all N-grams of a sentence by dividing the number of aspect tokens within an N-gram by its length N: $\frac{\#\ aspect\ tokens}{N} \in [0, 1]$. Our analysis reveals that 96% (783 of 814) of all aspects in the preliminary annotation dataset only contain one to four tokens. We thus decide to ignore all candidates with more than four tokens. No other limitations or filtering mechanisms are applied.

Step 2b: Training the ranker We use BERT (Devlin et al., 2019) and MT-DNN[2] (Liu et al., 2019) (base and large) to train a ranker. For training, we create five splits: (1) one in-topic split using a random subset from all four topics and (2) four

[2]BERT, fine-tuned on several NLP tasks via multi-task learning.

Topic	Five most frequent aspects (frequency)
Gun control	right (30), protect (18), background checks (17), gun violence (14), criminal (13)
Death penalty	cost (16), innocent (12), retribution (10), murder rate (9), deterrent (8)
Abortion	right (21), pain (10), choice (10), right to life (9), risk (9)
Marijuana legalization	dangerous (16), cost (13), risk (12), harm (10), black market (9)
General aspects	dangerous (in 8 of 8 topics), cost / life / risk / safety (in 7 of 8 topics)

Table 1: The five most frequent aspects for four exemplary topics and overall.

Setting	Rec@5	Rec@10	Rec@15	Rec@20
In-topic	0.7701	0.8468	0.8661	0.8925
Cross-topic	0.5951	0.7415	0.8164	0.8630

Table 2: In- and cross-topic Recall@k of the ranker used for aspect recommendations.

cross-topic splits using a leave-one-topic-out strategy. The cross-topic setup allows us to estimate the ranker's performance on unseen topics of the UKP-Corpus.

A single data sample is represented by an argument and an 1- to 4-gram of this argument, separated by the BERT architecture's [SEP] token. This technique expands the 800 original samples of the dataset to around 80,336. The model is trained for 5 epochs, with a learning rate of 5×10^{-5}, and a batch size of 8. We use the mean squared error as loss and take the recall@k to compare the models. The in- and cross-topic results of the best-performing model (MT-DNN$_{BASE}$) are reported in Table 2. All results are the average over runs with five different seeds (and over all four splits for the cross-topic experiments).

Step 2c: Creating the annotation data For each of the four topics that are part of the preliminary annotation dataset, we use the in-topic model to predict aspects of 629 randomly chosen, unseen arguments from the UKP-Corpus. For the other four topics of the UKP-Corpus, we choose the best cross-topic model to predict aspects for the same amount of samples. To keep a recall of at least 80%, we choose the ten and fifteen highest-ranked aspect candidates for samples as predicted by the in-topic and cross-topic model, respectively. We remove aspect candidates that include punctuation, begin or end with stopwords, or contain digits.

Step 3: Annotation study We use Amazon Mechanical Turk to annotate each sample by eight different workers located in the US, paying $7.6 per hour (minimum wage is $7.25 per hour). Based on a subset of 232 samples, we compute an α_u of .67 between crowdworkers and experts (three doctoral researchers). Compared to the initial study, the new approach increases the inter-annotator agreement between experts by approx. 11 points (see App. A for further details on the annotation study). Based on this promising result, we create a dataset of 5,032 high-quality samples that are labelled with aspects, as well as with their original stance labels from the UKP-Corpus. We show the most frequent (lemmatized) aspects that appear in some topics in Table 1.

3.2 Evaluation

We create a cross-topic split with the data of two topics as test set (*gun control*, *school uniforms*), one topic as dev set (*death penalty*), and the remaining topics as train set and evaluate two models with it. First, we use the ranking approach described in Step 2a-2b to fine-tune MT-DNN$_{BASE}$ on the newly generated data ("Ranker"). At inference, we choose the top T aspects for each argument as candidates. We tune T on the dev set and find $T = 2$ to be the best choice. Second, we use BERT for sequence tagging (Wolf et al., 2020) and label all tokens of the samples with BIO tags. As previously done with the ranker, we experiment with BERT and MT-DNN weights and find BERT$_{LARGE}$ to be the best choice (trained for 5 epochs, with a learning rate of 1×10^{-5} and a batch size of 32). We flatten the predictions for all test samples and calculate the F_1, Precision, and Recall macro scores. All models are trained over five seeds and the averaged results are reported in Table 3.

BERT$_{LARGE}$ predicts classes B and I with an F_1 of .65 and .53, hence aspects with more than one token are less well identified. A difference is to be expected, as the class balance of B's to I's is 2,768 to 2,103. While the ranker performs worse based on the shown metrics, it has a slightly higher recall for class I. We assume this is due to the fact that it generally ranks aspects with more than one token on top, i.e. there will often be at least one or more I's in the prediction. In contrast to that, BERT$_{LARGE}$ focuses more on shorter aspects, which is also in accordance with the average aspect length of 1.8 tokens per aspect in the dataset.

Model	F_1 macro	Precision	Recall
Majority (baseline)	.3085	.2871	.3333
Ranker	.6522	.6685	.6474
BERT$_{BASE}$	.6980	.6927	**.7040**
BERT$_{LARGE}$	**.7100**	**.7240**	.6993

Table 3: Test set results of the models for aspect detection. Majority only predicts class O.

In total, BERT$_{LARGE}$ outperforms the ranker by almost 6 percentage points in F_1 macro.

4 Data Collection Pipeline

This section describes the data collection and preprocessing for the argument generation pipeline. We aim to train a model that is able to transfer argumentative information concisely within a single sentence. We define such an argument as the combination of a topic and a sentence holding evidence with a specific stance towards this topic (Stab et al., 2018b). Consequently, the following preprocessing steps ultimately target retrieval and classification of sentences. To evaluate different data sources, we use a dump from Common-Crawl[3] (CC) and Reddit comments[4] (REDDIT) to fine-tune two separate generation models. The CC dump is from July 2016 and contains 331M documents (3.6TB) after deduplication. The REDDIT dump contains 2.5B documents (1.6TB) from December 2012 to May 2019. We choose to compare these two sources, as REDDIT is focused around user discussions and CC contains mixed sources with potentially higher quality.

Document Retrieval We index REDDIT and CC with ElasticSearch[5] and, for both, gather up to 1.5M documents for each of the eight topics of the UKP-Corpus. To increase the search results, we add synonyms (see App. B) for most topics.

Argument and Stance Classification We split the sentences of all documents and remove duplicates. We notice that many sentences are not relevant with regard to the document's topic. To enforce topic-relevance, we decide to filter out all sentences that do not contain at least one token of the respective topic or its defined synonyms (see App. B). We use the ArgumenText API's[6] argument and stance classification models (Stab et al., 2018a) to classify all sentences into *argument* or *non-argument* (F_1 macro $= .7384$), and remaining arguments into *pro* or *con* with regard to the topic (F_1 macro $= .7661$).

Aspect Detection We detect aspects on all remaining arguments. To speed up the detection on millions of sentences, we use BERT$_{BASE}$ instead of BERT$_{LARGE}$ (see Table 3).

Training Document Generation We create the final training documents for the argument generation model by concatenating all arguments that have the same topic, stance, and aspect (i.e. the same control code). Further, we aggregate all arguments that include an aspect with the same stem into the same document (e.g. arguments with *cost* and *costs* as aspect). To cope with limited hardware resources, we restrict the total number of arguments for each topic and stance to 100,000 (i.e. 1.6M over all eight topics). Also, as some aspects dominate by means of quantity of related arguments and others appear only rarely, we empirically determine an upper and lower bound of 1,500 and 15 arguments for each document, which still allows us to retrieve the above defined amount of training arguments.

5 Model Training and Analysis

In the following, we describe the architecture and the training process of the Arg-CTRL and analyze its performance.

5.1 Model and Training

Model The goal of a statistical language model is to learn the conditional probability of the next word given all (or a subset of) the previous ones (Bengio et al., 2003). That is, for a sequence of tokens $x = (x_1, ..., x_n)$, the model learns $p(x_i|x_{<i})$ where x_i is the i-th word of sequence x. For this work, we use the 1.63 billion-parameter *Conditional Transformer Language Model* (CTRL) by Keskar et al. (2019), which is built on a transformer-based sequence to sequence architecture (Vaswani et al., 2017). The CTRL has shown to produce high quality text, is general enough to be adapted for conditioning on the control codes we aim to use, and we do not need to pre-train the weights from scratch. Formally, the CTRL adds an extra condition to each sequence by prepending a control code c, hence learning $p(x_i|x_{<i}, c)$. The control code is represented by a single token and can then be used to direct the model output at inference. We extend the model from its previous limit of a single-token control code to accept multiple tokens. For

[3] https://commoncrawl.org

[4] https://files.pushshift.io/reddit/comments/

[5] https://www.elastic.co

[6] https://api.argumentsearch.com

cloning CON unrespectable . Cloning humans for reproductive purposes is unethical and unacceptable , but creating cloned embryos solely for research - which involves destroying them anyway - is downright criminal . (0.97) **cloning CON disfavored .** , cliques) to them . (0.36)
nuclear energy PRO safe . In addition , we must continue developing safer technologies like small modular reactors which will help us meet our nation 's need for reliable , emission-free sources of low-emission energy [...] . (0.96) **nuclear energy CON leak .** " We are concerned about the possibility of further releases of radioactivity due to possible melting or cracking of fuel rods at the No . (0.47)
marijuana legalization PRO safer : Legalizing cannabis will help reduce crime rates (especially violent crimes) and make society safer overall . (0.96) **marijuana legalization PRO benefits .** Decrease amount of police officers needed 6 . (0.37)

Table 4: Generated arguments of highest/lowest quality with Arg-CTRL$_{CC}$. Bold text shows the used control code. Quality score in brackets as predicted by the argument quality model. "[...]" signals shortened text.

decoding at inference, we use *penalized sampling* as proposed by Keskar et al. (2019). It defines a near-greedy sampling strategy that uses a penalty constant, effectively lowering the probability of previously generated tokens to prevent repetitions.

Training The CTRL was trained on 140GB of data from several large resources like Wikipedia, subreddits, and news data. We base our experiments on the pre-trained weights for a sequence length of 256 and fine-tune (see App. C for technical details) two models: Arg-CTRL$_{CC}$ (on the CC data) and Arg-CTRL$_{REDDIT}$ (on the REDDIT data). All training documents are sampled randomly for training. The respective control code is prepended to each sequence of 256 subwords of a document.

5.2 Analysis

Generation At inference, we gather multiple generated arguments from a control code input by splitting the generated output text into sentences with NLTK (Bird et al., 2009). We observe that for the first generated argument, the Arg-CTRL mostly outputs very short phrases, as it tries to incorporate the control code into a meaningful start of an argument. We prevent this by adding punctuation marks after each control code (e.g. a period or colon), signaling the model to start a new sentence. In this fashion, we generate *pro*- and *con*-arguments up to the pre-defined training split size[7] for each topic of the UKP-Corpus, resulting in 7,991 newly generated arguments. We do this with both models and use the generated arguments as a basis for the following analysis and evaluation methods. Examples of generated arguments can be found in Tables 4, 6, and 7 (as part of the evaluation, see Section 7).

Results With no other previous work on explicit control of argument generation (to the best of our knowledge), we decide to proof our concept of aspect-controlled neural argument generation by comparing both generation models to a retrieval approach as a strong upper bound. The retrieval approach returns all arguments from the classified training data (see Section 4) that match a given topic, stance, and aspect. Both the retrieval and generation approaches are evaluated against reference data from debate portals and compared via METEOR (Lavie and Agarwal, 2007) and ROUGE-L (Lin, 2004) metrics. The retrieval approach has an advantage in this setup, as the arguments are also of human origin and aspects are always explicitly stated within a belonging argument.

The reference data was crawled from two debate portals[8] and consists of pro- and con-paragraphs discussing the eight topics of the UKP-Corpus. As the paragraphs may include non-arguments, we filter these out by classifying all sentences with the ArgumenText API into arguments and non-arguments. This leaves us with 349 pro- and 355 con-arguments over all topics (see App. D for the topic-wise distribution). Next, we detect all aspects in these arguments. Arguments with the same topic, stance, and aspect are then grouped and used as reference for arguments from the (a) generated arguments and (b) retrieval approach arguments if these hold the same topic, stance, and aspect. The results reveal that both the average METEOR and ROUGE-L scores are only marginally lower than the retrieval scores (METEOR is 0.5/1.1 points lower for the Arg-CTRL$_{REDDIT}$/Arg-CTRL$_{CC}$, see Table 5). It not only shows the strength of the architecture, but also the success in generating sound aspect-specific arguments with our approach.

Overlap with Training Data We find arguments generated by the models to be genuine, i.e. demonstrating substantial differences to the training data. For each of the 7,991 generated arguments, we find the most similar argument in the training data based on the cosine similarity of their BERT embeddings

[7] Not counting non-arguments from the splits.

[8] procon.org and idebate.org

Model	METEOR	ROUGE-L
Retrieval (CC)	**17.85**	**14.72**
Arg-CTRL$_{CC}$	16.80	11.95
Retrieval (REDDIT)	**17.29**	**15.26**
Arg-CTRL$_{REDDIT}$	16.82	12.34

Table 5: Comparison of retrieval and generation approach with reference data from debate portals.

(CLS token). The average cosine similarity of the most similar pairs for both the Arg-CTRL$_{CC}$ and Arg-CTRL$_{REDDIT}$ is .92. However, this value is misleading, as even highly similar samples still show clear differences. This is also evident when looking at the average edit distances of 343 (Arg-CTRL$_{CC}$) and 163 (Arg-CTRL$_{REDDIT}$) for the pairs with highest similarity. Further comparison of these pairs for their longest common (string) overlap reveals only 9% (Arg-CTRL$_{CC}$) and 11% (Arg-CTRL$_{REDDIT}$) overlap on average, mostly consisting of stopwords. For illustration, we show two examples of highly similar pairs in Table 6.

Generated sentence: We do n't need more gun control laws when we already have enough restrictions on who can buy guns in this country . **Training sentence**: We have some of the strongest gun laws in the country , but guns do n't respect boundaries any more than criminals do . **Cosine similarity / edit distance / rel. overlap**: 95.59 / 88 / 8%
Generated sentence: The radioactivity of the spent fuel is a concern , as it can be used to make weapons and has been linked to cancer in humans . **Training sentence**: However , it does produce radioactive waste , which must be disposed of carefully as it can cause health problems and can be used to make nuclear weapons **Cosine similarity / edit distance / rel. overlap**: 92.40 / 99 / 17%

Table 6: Training data vs. generated arguments: examples of most similar arguments. Underlines mark the longest common overlap between generated and training sentences.

6 Generation in Absence of Aspects

To show the necessity of having prior knowledge of aspects for our controlled argument generation approach, we create training data *without* prior knowledge of aspects, train a new generation model on it, and compare it to our previous models *with* prior knowledge of aspects. Equally to the original Arg-CTRL$_{CC}$'s procedure, we gather 100,000 sentences for each stance of a topic from the CC data. As we assume to have no knowledge about the aspects of the arguments, we randomly sample arguments from the CC source documents. We create training documents with numbers of arguments varying between 15 and 1,500 to mimic the data generation process of the original models and fine-tune a new generation model on them. After training, we generate the same number of arguments as for the other two models by using our default control code of *[Topic] [Stance] [Aspect]*. While the new model was only conditioned on topics and stances at training time, we make sure that all aspects used for generation appear in at least one argument of the model's training data.

We compare all models by verifying whether or not the aspect used for generation (including synonyms and their stems and lemmas) can be found in the generated arguments. For the original models conditioned on aspects, this is true in 79% of the cases for Arg-CTRL$_{REDDIT}$ and in 74% of the cases for Arg-CTRL$_{CC}$. For the model that was not conditioned on aspects, however, it is only true in 8% of the cases. It clearly shows the necessity to condition the model on aspects explicitly, implying the need for argument aspect detection, as the model is unable to learn generating aspect-related arguments otherwise. Moreover, without prior detection of aspects, we have no means for proper aggregation over aspects. We notice that for the model without prior knowledge of aspects, 79% of all aspects in the training data appear in only one argument. For these aspects, the model will likely not pick up a strong enough signal to learn them.

7 Evaluation

We evaluate the quality (intrinsic evaluation) of the Arg-CTRL and its performance on an exemplary task (extrinsic evaluation). As a basis, we use the 7,991 arguments generated in Section 5.

7.1 Intrinsic Evaluation

Human Evaluation We conduct an expert evaluation on a subset of generated arguments with two researchers (field of expertise is natural language processing) not involved in this paper. Two aspects are evaluated: *fluency* and *persuasiveness*. We consider a sentence as fluent if it is grammatically correct (Hua et al., 2019), i.e. contains neither semantic nor syntactic errors, and arrange this as a binary task. To reduce subjectivity for the persuasiveness evaluation, the experts do not annotate single arguments but instead compare pairs (Habernal and Gurevych, 2016) of generated and refer-

ence data arguments (see Section 5.2). The experts could either choose one argument as being more persuasive or both as being equally persuasive. In total, the experts compared 100 (randomly sorted and ordered) argument pairs for persuasiveness and fluency (50 from both the Arg-CTRL$_{\text{REDDIT}}$ and the Arg-CTRL$_{\text{CC}}$). A pair of arguments always had the same topic and stance. For fluency, only the annotations made for generated arguments were extracted and taken into account. Averaged results of both experts show that in 33% of the cases, the generated argument is either more convincing (29%) or as convincing (4%) as the reference argument. Moreover, 83% of generated arguments are fluent. The inter-annotator agreement (Cohen, 1960) between the two experts is Cohen's $\kappa = .30$ (percentage agreement: .62) for persuasiveness and $\kappa = .43$ (percentage agreement: .72) for fluency, which can be interpreted as "fair" and "moderate" agreement, respectively (Landis and Koch, 1977). As we compare to high-quality, curated data, the perceived persuasiveness of the generated arguments shows the potential of the work—further strengthened in the remainder of this section.

Argument Quality We introduce a novel method to evaluate generated arguments based on the argument quality detection approach proposed by Gretz et al. (2020b). They create an argument quality dataset that contains around 30,000 arguments over 71 topics. For each argument, annotators were asked whether or not they would recommend a friend to use the displayed argument in a speech. The quality scores for each argument result from a weighted average (WA) or MACE Probability function of all annotations and range between 0 (lowest quality) and 1.0 (highest quality). We use the *WA*-score as label, the same model (BERT$_{BASE}$) and hyperparameters as given in the original paper, and reproduce the reported correlations of .52 (Pearson) and .48 (Spearman) on the test dataset (averaged over five different seeds). The model predicts an average argument quality of .71 for the Arg-CTRL$_{\text{REDDIT}}$ arguments, .75 for the training arguments of the UKP-Corpus, and even .76 for the Arg-CTRL$_{\text{CC}}$ arguments. It shows that our model is able to produce arguments that are generally on a similar quality level as arguments made by humans. The difference of 5 points between the Arg-CTRL$_{\text{CC}}$ and the Arg-CTRL$_{\text{REDDIT}}$, however, shows that training on REDDIT data leads to a lower quality of generated arguments. Exemplary for three topics, we show the generated arguments with the highest and lowest argument quality in Table 4 (see App. E for the full table).

Human: Nuclear energy produces waste that stays radioactive for thousands of years and pollutes the environment.
Arg-CTRL$_{\text{CC}}$: (nuclear energy PRO waste .) The only thing that 's going to be left over is the spent fuel rods , which are not very radioactive and can be reprocessed [...]
Arg-CTRL$_{\text{CC}}$: (nuclear energy PRO radioactive .) The radioactivity of the waste is not a problem , because it will decay away in about 300 years or so [...]
Arg-CTRL$_{\text{CC}}$: (nuclear energy PRO environment .) The fact that the radioactive waste is not only stored in a safe manner , but also protected from any potential threat to human health or the environment for thousands of years [...]
Human: School uniforms are expensive and affect the pupil's individuality.
Arg-CTRL$_{\text{CC}}$: (school uniforms PRO expensive .) The cost of the uniform is not that much more than a regular shirt and shorts [...]
Arg-CTRL$_{\text{CC}}$: (school uniforms PRO individuality .) The uniform is a symbol of unity and identity , which helps to foster an environment where students can feel comfortable about expressing their own individual style without being judged [...]

Table 7: Generated counter-arguments with the Arg-CTRL$_{\text{CC}}$. Aspects in the initial argument are underlined and used for the counter-argument generation. Control code in brackets and "[...]" signals shortened text.

7.2 Extrinsic Evaluation: Counter-Arguments

Drafting counter-arguments is an important skill for debating, to provide constructive feedback, and to foster critical thinking. We lean onto the work of Wachsmuth et al. (2018) who describe a counter-argument as discussing the same aspect as an initial argument, but with a switched stance. Hence, given our defined control codes, our model is especially fit for counter-argument generation. Unlike current models for this task, we do not require a specific dataset with argument and counter-argument pairs (Hidey and McKeown, 2019; Hua et al., 2019). Also, in contrast to the model by Hua and Wang (2019) that implicitly integrates input-related "Keyphrases" into the process of counter-argument generation, our model is able to concentrate on every aspect of the input explicitly and with a separate argument, allowing for more transparency and interpretability over the process of counter-argument generation. We exemplary show how the combination of aspect detection and controlled argument generation can be successfully leveraged to tackle this task. For that, we manually

compose initial arguments for the topics *nuclear energy* and *school uniforms*. Then, we automatically detect their aspects and generate a counter-argument for each aspect by passing the topic, opposite stance of the original argument, and one of the aspects into the Arg-CTRL$_{CC}$. For both topics, the Arg-CTRL$_{CC}$ produces meaningful counter-arguments based on the detected aspects (see Table 7).

8 Conclusion

We apply the concept of controlled neural text generation to the domain of argument generation. Our Arg-CTRL is conditioned on topics, stances, and aspects and can reliably create arguments using these control codes. We show that arguments generated with our approach are genuine and of high argumentative and grammatical quality in general. Moreover, we show that our approach can be used to generate counter-arguments in a transparent and interpretable way. We fine-tune the Arg-CTRL on two different data sources and find that using mixed data from Common-Crawl results in a higher quality of generated arguments than using user discussions from Reddit-Comments. Further, we define argument aspect detection for controlled argument generation and introduce a novel annotation scheme to crowdsource argument aspect annotations, resulting in a high-quality dataset. We publish the model weights, data, and all code necessary to train the Arg-CTRL.

Ethics Statement

Models for argument and claim generation have been discussed in our related work and are widely available. Gretz et al. (2020a) suggest that, in order to allow for a fine-grained control over claim/argument generation, aspect selection needs to be handled carefully, which is what we have focused on in this work. The dangers of misuse of language models like the CTRL have been extensively discussed by its authors (Keskar et al., 2019). The ethical impact of these works has been weighed and deemed justifiable.

Argument generation—and natural language generation as a whole—is subject to dual use. The technology can be used to create arguments that cannot be distinguished from human-made arguments. While our intentions are to support society, to foster diversity in debates, and to encourage research on this important topic, we are aware of the possibility of harmful applications this model can be used for. For instance, the model could be used to generate only opposing (or supporting) arguments on one of the pretrained topics and aspects and, as such, bias a debate into a certain direction. Also, bots could use the generated arguments to spread them via social media. The same is true, however, for argument search engines, which can be used by malicious parties to retrieve (and then spread) potentially harmful information.

However, controllable argument generation can also be used to support finding and formulating (counter-)arguments for debates, for writing essays, to enrich one-sided discussions, and thus, to make discourse more diverse overall. For instance, anticipating opposing arguments is crucial for critical thinking, which is the foundation for any democratic society. The skill is extensively taught in school and university education. However, *confirmation bias* (or *myside bias*) (Stanovich et al., 2013), i.e. the tendency to ignore opposing arguments, is an ever-present issue. Technologies like ours could be used to mitigate this issue by, for instance, automatically providing topic- and aspect-specific counter-arguments for all arguments of a given text (this has been shown for single arguments in Section 7.2). We believe that working on and providing access to such models is of major importance and, overall, a benefit to society.

Open-sourcing such language models also encourages the work on counter-measures to detect malicious use: While many works have been published on the topic of automatic fake news detection in texts (Kaliyar et al., 2020; Reis et al., 2019; Hanselowski et al., 2018; Pérez-Rosas et al., 2018), the recent emergence of large-scale language models has also encouraged research to focus on detecting the creator of these texts (Varshney et al., 2020; Zellers et al., 2019). The former approaches are aimed at detecting fake news in general, i.e. independent of who (or what) composed a text, whereas the latter approaches are designed to recognize if a text was written by a human or generated by a language model. We encourage the work on both types of methods. Ideally, social networks and news platforms would indicate if a statement was automatically generated in addition to its factual correctness.

Further, we point out some limitations of the Arg-CTRL that mitigate the risks discussed before. One of these limitations is that it cannot be used

to generate arguments for unseen topics, which makes a widespread application (e.g. to produce fake news) rather unlikely (using an unseen topic as control code results in nonsensical repetitions of the input). The analysis in Section 6 of the paper shows that the model fails to produce aspect-specific sentences in 92% of the cases if it was not explicitly conditioned on them at training time. Even in case of success, the aspect has to exist in the training data. Also, the model is trained with balanced classes, i.e. both supporting and opposing arguments for each topic are seen with equal frequency to prevent possible bias into one or the other direction.

To further restrict malicious use, we release the training data for the Arg-CTRLs with an additional clause that forbids use for any other than research purposes. Also, all the training datasets for the Arg-CTRLs will be accessible only via access control (e-mail, name, and purpose of use). Lastly, this work has been reviewed by the ethics committee of the Technical University of Darmstadt that issued a positive vote.

Acknowledgements

We thank Tilman Beck and Nandan Thakur for their support in the human evaluation (Section 7.1). This work has been supported by the German Research Foundation within the project "Open Argument Mining" (GU 798/25-1), associated with the Priority Program "Robust Argumentation Machines (RATIO)" (SPP-1999).

References

Yamen Ajjour, Milad Alshomary, Henning Wachsmuth, and Benno Stein. 2019. Modeling frames in argumentation. 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 2922–2932, Hong Kong, China. 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 *12th International Natural Language Generation Conference (INLG 2019)*. ACL.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pages 1–15.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Yonatan Bilu, Ariel Gera, Daniel Hershcovich, Benjamin Sznajder, Dan Lahav, Guy Moshkowich, Anael Malet, Assaf Gavron, and Noam Slonim. 2019. Argument invention from first principles. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1013–1026, Florence, Italy. Association for Computational Linguistics.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*, 1st edition. O'Reilly Media, Inc.

Filip Boltužić and Jan Šnajder. 2017. Toward stance classification based on claim microstructures. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 74–80, Copenhagen, Denmark. Association for Computational Linguistics.

Amber E. Boydstun, Dallas Card, Justin Gross, Paul Resnick, and Noah A. Smith. 2014. Tracking the Development of Media Frames within and across Policy Issues. Carnegie Mellon University.

Giuseppe Carenini and Johanna D Moore. 2006. Generating and evaluating evaluative arguments. *Artificial Intelligence*, 170(11):925–952.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and Play Language Models: A Simple Approach to Controlled Text Generation. *arXiv, abs/1912.02164*.

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

Atsushi Fujii and Tetsuya Ishikawa. 2006. A system for summarizing and visualizing arguments in subjective documents: Toward supporting decision making. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, SST '06, pages 15–22, Stroudsburg, PA, 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. 2020a. The workweek is the best time to start a family – a study of gpt-2 based claim generation. *arXiv, abs/2010.06185*.

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2020b. A large-scale dataset for argument quality ranking: Construction and analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7805–7813.

Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599, Berlin, Germany. Association for Computational Linguistics.

Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M. Meyer, and Iryna Gurevych. 2018. A retrospective analysis of the fake news challenge stance-detection task. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1859–1874, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Mareike Hartmann, Tallulah Jansen, Isabelle Augenstein, and Anders Søgaard. 2019. Issue framing in online discussion fora. 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 1401–1407, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Xinyu Hua and Lu Wang. 2019. Sentence-level content planning and style specification for neural text 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 591–602, Hong Kong, China. Association for Computational Linguistics.

Rohit Kumar Kaliyar, Anurag Goswami, Pratik Narang, and Soumendu Sinha. 2020. Fndnet – a deep convolutional neural network for fake news detection. *Cogn. Syst. Res.*, 61(C):32–44.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv, abs/1909.05858*.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, page 228–231, USA. Association for Computational Linguistics.

Dieu Thu Le, Cam-Tu Nguyen, and Kim Anh Nguyen. 2018. Dave the debater: a retrieval-based and generative argumentative dialogue agent. In *Proceedings of the 5th Workshop on Argument Mining*, pages 121–130, Brussels, Belgium. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.

Amita Misra, Pranav Anand, Jean E. Fox Tree, and Marilyn Walker. 2015. Using summarization to discover argument facets in online idealogical dialog. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 430–440, Denver, Colorado. Association for Computational Linguistics.

Nona Naderi and Graeme Hirst. 2017. Classifying frames at the sentence level in news articles. In *Proceedings of the International Conference Recent*

Advances in Natural Language Processing, RANLP 2017, pages 536–542, Varna, Bulgaria. INCOMA Ltd.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

J. C. S. Reis, A. Correia, F. Murai, A. Veloso, and F. Benevenuto. 2019. Supervised learning for fake news detection. *IEEE Intelligent Systems*, 34(2):76–81.

Misa Sato, Kohsuke Yanai, Toshinori Miyoshi, Toshihiko Yanase, Makoto Iwayama, Qinghua Sun, and Yoshiki Niwa. 2015. End-to-end argument generation system in debating. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 109–114, Beijing, China. Association for Computational Linguistics and The Asian Federation of Natural Language Processing.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Christian Stab, Johannes Daxenberger, Chris Stahlhut, Tristan Miller, Benjamin Schiller, Christopher Tauchmann, Steffen Eger, and Iryna Gurevych. 2018a. 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, New Orleans, Louisiana. Association for Computational Linguistics.

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018b. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3664–3674. Association for Computational Linguistics.

Keith E. Stanovich, Richard F. West, and Maggie E. Toplak. 2013. Myside bias, rational thinking, and intelligence. *Current Directions in Psychological Science*, 22(4):259–264.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

L. R. Varshney, N. Shirish Keskar, and R. Socher. 2020. Limits of detecting text generated by large-scale language models. In *2020 Information Theory and Applications Workshop (ITA)*, pages 1–5.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018. Retrieval of the best counterargument without prior topic knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 241–251, Melbourne, Australia. Association for Computational Linguistics.

Lu Wang and Wang Ling. 2016. Neural network-based abstract generation for opinions and arguments. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 47–57, San Diego, California. Association for Computational Linguistics.

Tsung Hsien Wen, David Vandyke, Nikola Mrkšíc, Milica Gašíc, Lina M. Rojas-Barahona, Pei Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference*, 1:438–449.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Ingrid Zukerman, Richard McConachy, and Kevin B. Korb. 1998. Bayesian reasoning in an abductive mechanism for argument generation and analysis. In *Proceedings of the Fifteenth National/Tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence*, AAAI '98/IAAI '98,

page 833–838, USA. American Association for Artificial Intelligence.

A Argument Aspect Annotation Study

For the final crowdsourcing study, we use Amazon Mechanical Turk. Workers had to take a qualification test, have an acceptance rate of at least 95%, and location within the US. We paid $7.6 per hour (minimum wage is $7.25 per hour). Each data sample is annotated by eight crowdworkers. In case the ranker cut off the real aspect(s) from the list of candidates, crowdworkers could select any sequence up to four tokens from a second list.

Figure 2 shows the annotation guidelines for the Amazon Mechanical Turk study. Figure 3 shows one example of a HIT with two aspects selected. Selected aspects are highlighted in the sentence. We did not allow to choose overlapping aspects. If the aspect was not found in the first list provided by the learned ranker, crowdworkers could choose from as second list with the remaining 1-4-grams of the sentence (aspect candidates starting or ending with stopwords, as well as candidates with punctuation and numbers, were removed from the list). Additional checkboxes were added to choose from if the sentence contained no aspect or the aspect was not explicitly mentioned. Figure 4 shows a ranked list of aspect candidates for an example.

The structure of the final dataset is described in Section F. For reproducibility of results, we create fixed splits for in- and cross-topic experiments.

B Search Query and Topic Relevance Synonyms

Table 8 lists the ElasticSearch queries we used to retrieve the initial training documents from CC and REDDIT. Combinations of topics and data sources that are not listed in the table required no expansion of the query to gather enough documents for training. In Table 9, we show the synonyms used for filtering prior to the argument and stance classification step. We filtered out all sentences that did not contain tokens from the topic they belong to or any synonyms defined for this topic.

C Model Parameters and Details

All arguments of the training documents are tokenized with a BPE model (Sennrich et al., 2016) trained by the authors of the CTRL (Keskar et al., 2019). Both the Arg-CTRL$_{\text{CC}}$ and the Arg-CTRL$_{\text{REDDIT}}$ are fine-tuned on a Tesla V100 with 32 GB of Memory. We mainly keep the default hyperparameters but reduce the batch size to 4 and train both models for 1 epoch. Each model takes around five days to train on the 1.6M training sentences.

D Reference Data Statistics

Table 10 shows the sources and number of arguments for all topics of the reference dataset. The dataset is used to compare the argument generation models to a retrieval approach.

E Examples of Generated Arguments

For all eight topics, we show the generated argument with the highest and lowest argument quality score in tables 11 (Arg-CTRL$_{\text{CC}}$) and 12 (Arg-CTRL$_{\text{REDDIT}}$). Text in bold shows the given control code, text afterwards represents the generated argument. Numbers in brackets after the text show the quality score as predicted by the argument quality model.

F Argument Aspect Detection Dataset

The argument aspect detection dataset contains a total of 5,032 samples in JSONL-format, i.e. each dataset sample has a separate line and can be parsed as JSON. A sample contains the keys:

- **hash**: Unique identifier.
- **aspect_pos**: List of string tuples "(begin,length)", marking the character position and length of each aspect within the argument.
- **aspect_pos_string**: The aspects as a list of strings.
- **stance**: Original stance label of the argument towards the topic, taken from the UKP-Corpus (Stab et al., 2018b). Either "Argument_for" or "Argument_against".
- **topic**: The topic of the argument.
- **sentence**: The argument.

For reproducibility, we define a fixed cross-topic split with the data of two topics as test set (*gun control*, *school uniforms*), the data of one topic as development set (*death penalty*), and the data of the remaining five topics as train set. We also create a fixed in-topic split by randomly taking 3,532 samples of all topics for training, 500 for development, and 1,000 for testing.

Instructions

In the following, you will find a set of sentences. The sentences argue for or against a given topic (e.g. school uniforms, nuclear energy, etc.). Your task is to identify the main reasons why they argue in such a way - we refer to these reasons as *aspects*. Aspects answer the question why we should support or oppose a specific topic.
The motivation behind this task is to learn grouping sentences of the same topic (e.g. nuclear energy) by the aspects (e.g. costs, safety) they argue about.

Note: In some cases, the topic might be incorrect or only implicitly related, as it was tagged automatically. The topics are just listed to provide additional context and an incorrectly assigned topic should not worry you.

▾Guidelines:

- An aspect resembles the core reasoning of a sentence.
- For each sentence ask yourself: "Which word sequences (e.g. "radioactive waste") are relevant to determine whether to support/oppose the given topic (e.g. Nuclear energy)?" The shortest possible answer in the sentence is usually the aspect.
- A sentence can have (a) no aspect, (b) one aspect, or (c) several aspects, and one aspect can have several words.
- Each chosen aspect should be a valid aspect on its own.
- Aspects should be as short as possible.

▾ Examples:

- **TOPIC: Nuclear energy**
  - Sentence: It is pretty expensive to build and run nuclear power plants .
  - Aspect(s): [expensive].
  - Explanation: In this sentence, the relevant factor to determine whether to support/oppose nuclear energy is that it is "expensive". Please do not select adverbs like "pretty", "very", etc., as they do not add further information to the reasoning behind the aspect.
- **TOPIC: Nuclear energy**
  - Sentence: Compared to coal-fired power , nuclear energy is clean but the reactors can easily be targeted by terrorist attacks .
  - Aspect(s): [clean, terrorist attacks].
  - Explanation: In this sentence, the relevant factors to determine whether to support/oppose nuclear energy is that it is "clean" and a possible target of "terrorist attacks". A sentence can have several aspects and they can be opposing and supporting within the same sentence.
- **TOPIC: Marijuana legalization**
  - Sentence: Legalization will inevitably lead to a decrease in medical opioid users .
  - Aspect(s): [medical opioid users].
  - Explanation: In this sentence, the relevant factor to determine whether to support/oppose marijuana legalization is the issue of people who have to rely on opioids instead. Thus, the aspect is "medical opioid users". The words "decrease in" are not necessary to understand the reason for supporting marijuana legalization. Whenever the aspect you chose seems quite long, try to remove words and ask yourself if it would still count as a valid aspect (there are exceptions of course, when aspects are indeed long).

▾ How to use the HIT:

- To select an aspect, click in the first input field below a sentence and select as many aspects from the list as you see fit.
- You can also use the keyboard (arrow keys for selection, tab to jump to the next field, backspace to remove aspects).
- If the aspect is not in the first list, please click on *Aspect not in the first list, show more aspect candidates.* and select the missing aspects from the second list of remaining candidates. You can add aspects from both lists.
- If you are in doubt about two similar aspects in both lists, take the one you find in the first list.
- If the sentence contains no aspect or it is not in either of the two lists, hit the repsective checkbox and solve the math equation (digits only).
- If you select multiple aspects, they must not overlap (you will see a warning message if they do).

After processing all sentences, press the submit button to finish this HIT. If you have a remark or something is unclear, please inform us through the feedback box below each HIT.

Note: some hits include easy-to-answer quality control questions. These will ask you to select specific words as aspect(s). Please make sure to answer these questions correctly.

Figure 2: Guidelines for the final annotation study.

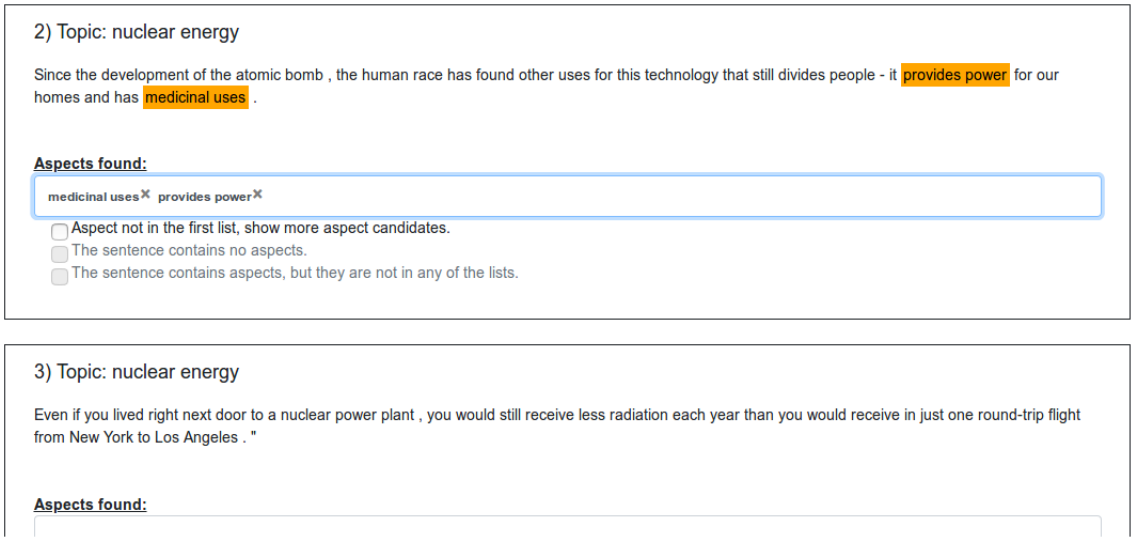

Figure 3: Example sentence of a HIT with two aspects selected.

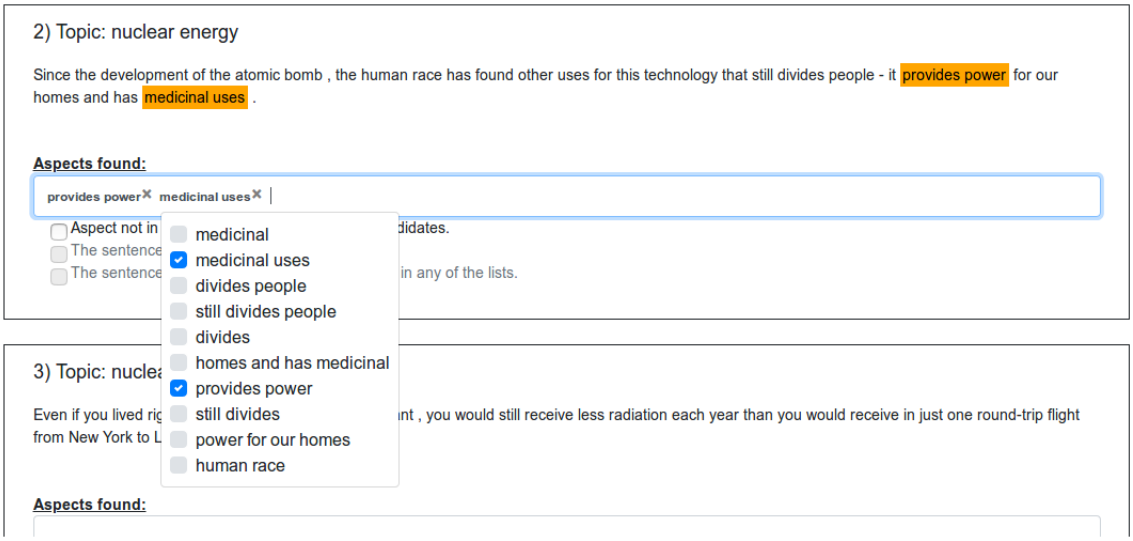

Figure 4: Example sentence of a HIT with the list of ranked aspect candidates.

Topic	Search query
Marijuana legalization (CC and REDDIT)	((marijuana legalization) OR (legalization of marijuana) OR (legalization of cannabis)) OR (((marijuana) OR (dope) OR (cannabis) OR (weed)) AND ((law) OR (legal) OR (legalization)))
School uniforms (CC and REDDIT)	(school uniform) OR (college uniform) OR (school outfit) OR ((school) AND (uniform)) OR ((school) AND (outfit)) OR ((school) AND (jacket)) OR ((school) AND (cloth)) OR ((school) AND (dress)) OR ((college) AND (dress))
Cloning (REDDIT)	(clone) OR (cloning) OR (clones) OR (cloned) OR (clones) OR (genetically identical) OR (asexual reproduction)
Death penalty (REDDIT)	(capital punishment) OR ((execution) AND (death)) OR ((punishment) AND (death)) OR ((punishment) AND (execution)) OR (electric chair) OR (death penalty)
Gun control (REDDIT)	(gun control) OR (second amendment) or (2nd amendment) OR ((gun) AND (ownership)) OR (gun ownership) OR (arms reduction) OR (arms limitation) OR (gun politics) OR ((gun) AND (lobby))
Nuclear energy (CC and REDDIT)	nuclear AND (energy OR fission OR power OR plant)

Table 8: Synonyms added to the topic query to gather initial documents from ElasticSearch. For combinations of topics and data sources (i.e. REDDIT and CC) that are not listed, we only used the topic as search query.

Topic	Synonyms
School uniforms	uniform, college, outfit, dress, suit, jacket, cloth
Nuclear energy	fission, fusion, atomic energy, nuclear power, atomic power, radioactive, radioactivity
Marijuana legalization	cannabis, legalization of marijuana, legal, illegal, law, weed, dope
Cloning	clone, cloned, duplicate, copy, reproduct, asexual
Death penalty	capital punishment, execution, electric chair, punishment, punish
Minimum wage	living wage, base pay, average wage, low income
Abortion	abort, termination, misbirth, birth control
Gun control	second amendment, ownership, arms reduction, arms limitation

Table 9: Topic synonyms to pre-filter sentences prior to argument and stance classification.

Topic	Sentence-Level Args		Source
	pro	con	
Abortion	57	47	procon.org
Gun control	49	50	procon.org
School uniforms	60	55	procon.org
Minimum wage	59	64	procon.org
Marijuana legalization	63	71	procon.org
Death penalty	22	27	idebate.org
Cloning	23	24	idebate.org
Nuclear energy	16	17	idebate.org
Total	**349**	**355**	/

Table 10: Sources and number of arguments for each topic from the reference dataset.

abortion CON violence : Abortion is not only morally wrong but also physically dangerous for both mother and baby ; it increases breast cancer risk by 50 % and causes mental illness including schizophrenia and psychosis ; it leads to infertility and premature birth in subsequent pregnancies ; it contributes significantly to maternal mortality due to hemorrhage during childbirth ; complications arising from unsafe abortion account for 13 percent of all maternal deaths worldwide ; and post-abortion syndrome affects up to one quarter of pregnant women suffering from this condition . (0.99) **abortion CON lives .** ..I 'm sorry , but your logic does n't pass the laugh test . (0.49)
cloning CON unrespectable . Cloning humans for reproductive purposes is unethical and unacceptable , but creating cloned embryos solely for research - which involves destroying them anyway - is downright criminal . (0.97) **cloning CON disfavored .** , cliques) to them . (0.36)
nuclear energy PRO safe . In addition , we must continue developing safer technologies like small modular reactors which will help us meet our nation 's need for reliable , emission-free sources of low-emission energy while also creating jobs and providing solutions to some of the world s most pressing problems : climate change , food security and sustainable development . (0.96) **nuclear energy CON leak .** " We are concerned about the possibility of further releases of radioactivity due to possible melting or cracking of fuel rods at the No . (0.47)
death penalty CON inhuman . Amnesty International opposes the death penalty in all cases as the ultimate form of cruel , inhuman or degrading punishment and a violation of fundamental rights - the right to life and the prohibition of torture . (1.00) **death penalty CON god .** And yet , while the Church exhorts civil authorities . (0.41)
school uniforms CON resources . The lack of a uniform system for the collection and use of data , as well as insufficient funding to collect it , are major barriers that limit access to information on student achievement in schools with high concentrations of students from low-income families . (0.92) **school uniforms PRO fun :** The kids are having a lot of fun . (0.38)
gun control PRO homicides . In addition to being an effective crime deterrent and reducing suicides , research has shown that defensive firearm use prevents more than 80 percent of all gun-related violent crimes including homicides , robberies , assaults , rapes , and aggravated assaults [1] . (0.96) **gun control PRO prevent ;** You liberals need to wake up and smell the coffee . (0.43)
marijuana legalization PRO safer : Legalizing cannabis will help reduce crime rates (especially violent crimes) and make society safer overall . (0.96) **marijuana legalization PRO benefits .** Decrease amount of police officers needed 6 . (0.37)
minimum wage PRO poor : Raising the minimum wage will not only benefit those working full time but also reduce government expenditures on social services such as food stamps and Medicaid which disproportionately affect the poor . (0.97) **minimum wage CON cost :** If you raise the price of a Big Mac to $ 10 and then pay an extra dime or two per burger so that it 's still only $ 9 ... well , maybe your business is n't worth saving at all because there just wo n't be anyone willing to eat out after they 've had their fill (unless perhaps they 're on welfare) and now have no job and ca n't afford health care either - but I doubt very seriously that most people making those kinds of wages/salaries would suddenly decide not to work if forced to shoulder the added cost . (0.44)

Table 11: Generated arguments with the Arg-CTRL$_{\text{CC}}$. Text in bold shows the given control code, text afterwards represents the generated argument. Numbers in brackets after the text show the quality score as predicted by the argument quality model.

abortion PRO reduce abortions : It 's been shown time after time that comprehensive sexual health services (including free or cheap long term hormonal birth controls) greatly reduces abortions by preventing unplanned pregnancies from happening in the first place . (0.99) **abortion PRO crime .** _r=0 & amp ; pagewanted=print & amp ; oref=slogin) . (0.40)
cloning PRO reproduction . The only way to increase the number of clones is through sexual reproduction , which increases genetic diversity and therefore reduces extinction rates . (0.85) **cloning PRO awesome .** But yeah , the clone skins look fucking awesome . (0.36)
nuclear energy PRO safe . Nuclear is the only viable option for a large scale , reliable and safe form of energy production that can replace fossil fuels as our main energy source . (0.97) **nuclear energy CON leak .** Biofuel does n't need batteries 6 . (0.41)
death penalty PRO save . The only way we can possibly make sure no innocents are executed is by abolishing the death penalty altogether - there 's just too much chance that at least one innocent person will die before their execution date was up and they were able to prove their innocence with DNA evidence and/or other exonerating circumstances . (0.95) **death penalty PRO innocent person .** Innocent people do n't deserve to live 2 . (0.43)
school uniforms PRO fit . Dress codes exist to prevent distractions from other students while trying to teach kids appropriate attire which helps them learn proper social skills and fitting into society . (0.83) **school uniforms PRO nice :** It looks really nice on my college application . (0.37)
gun control PRO prevent . Guns also help prevent tyranny by removing checks against government overreach into areas where the populace has little power . (0.95) **gun control CON problem ;** the guns are n't the real problems . (0.32)
marijuana legalization CON bad : Alcohol is also very addictive and has been shown time after time to have negative effects on health yet it remains completely legal while cannabis gets demonized by law enforcement and politicians alike despite being less harmful than many prescription medications in every way imaginable . (0.93) **marijuana legalization PRO buy .** Get busted by police 5 . (0.36)
minimum wage PRO poverty : Raising the minimum wage helps alleviate poverty as well as increase demand for goods and services from consumers . (0.93) **minimum wage CON pay :** They ca n't pay below minimum wage either . (0.41)

Table 12: Generated arguments with the Arg-CTRL$_{\text{REDDIT}}$. Text in bold shows the given control code, text afterwards represents the generated argument. Numbers in brackets after the text show the quality score as predicted by the argument quality model.