A Plug-and-Play Method for Controlled Text Generation

Conference Paper

Author(s):
Pascual, Damian; Egressy, Béni; Meister, Clara Isabel ⓘD; Cotterell, Ryan; Wattenhofer, Roger

A Plug-and-Play Method for Controlled Text Generation

Damián Pascual[1] **Béni Egressy**[1] **Clara Meister**[1], **Ryan Cotterell**[1,2] **Roger Wattenhofer**[1]
[1]ETH Zürich [2]University of Cambridge
damianp@ethz.ch begressy@ethz.ch clara.meister@inf.ethz.ch
ryan.cotterell@inf.ethz.ch wattenhofer@ethz.ch

Abstract

Large pre-trained language models have repeatedly shown their ability to produce fluent text. Yet even when starting from a prompt, generation can continue in many plausible directions. Current decoding methods with the goal of controlling generation, e.g., to ensure specific words are included, either require additional models or fine-tuning, or work poorly when the task at hand is semantically unconstrained, e.g., story generation. In this work, we present a *plug-and-play* decoding method for controlled language generation that is so simple and intuitive, it can be described in a single sentence: given a topic or keyword, we add a shift to the probability distribution over our vocabulary towards semantically similar words. We show how annealing this distribution can be used to impose *hard* constraints on language generation, something no other plug-and-play method is currently able to do with SOTA language generators. Despite the simplicity of this approach, we see it works incredibly well in practice: decoding from GPT-2 leads to diverse and fluent sentences while guaranteeing the appearance of given guide words. We perform two user studies, revealing that (1) our method outperforms competing methods in human evaluations; and (2) forcing the guide words to appear in the generated text has no impact on the fluency of the generated text.[1]

1 Introduction

Having systems capable of automatically generating human-like text has been an objective pursued since the early days of artificial intelligence (Meehan, 1977; Lebowitz, 1987). The recent development of large pre-trained language models based on the transformer architecture has brought us closer to this goal. Indeed, current state-of-the-art models can produce impressively realistic text (Radford et al., 2019; Raffel et al., 2020;

Brown et al., 2020), facilitating their use across a wide range of language generation applications.

Yet, these models are probabilistic in nature, and in certain language generation tasks, i.e., those that place few **semantic constraints** on the output, it can be difficult to control the general theme or ensure the presence of specific words in generated text. We use story generation (Fan et al., 2018) as a running example; even in the presence of a prompt, there are many continuations that may result in a good story. Indeed, for this task, a generator that does not reflect this diversity is likely undesirable, since the number of stories it could produce would be limited. But what if you would like your story to fulfill certain criteria? Under existing methods, one would have to train (or at least fine-tune) a new model for a limited set of use-cases or hope that the model places sufficient probability mass on text meeting this criteria such that it shows up during search[2] (Hokamp and Liu, 2017; Post and Vilar, 2018; Ziegler et al., 2019; Keskar et al., 2019).

In this work, we introduce a better strategy for **controlled decoding**—where generated text must contain certain words—for semantically unconstrained generation tasks. In short, we shift the output distribution of a language generation model towards the semantic space of a given guide word. While this proposal sounds abstract at first, its realization is simple and intuitive: at each generation step, we modify the log-probability of the words in our vocabulary according to their semantic similarity to our guide word, quantified as the cosine similarity between their vector representations, i.e., word embeddings. This method not only encourages the explicit appearance of a guide word, but also encourages the model to generate appropriate context for the guide word to appear, i.e., words in the same semantic neighborhood (Hashimoto

---

[1]Code: https://github.com/dapascual/K2T

[2]Whilst this approach is often effective for semantically constrained tasks, such as machine translation, it becomes prohibitively expensive in practice for open-ended tasks.

et al., 2016). By annealing the strength of the probability shift, we can guarantee that all guide words appear in our text. Moreover, this method allows for guidance towards both ordered and unordered sets of words. Our decoding strategy is plug-and-play, i.e., it can be combined with any autoregressive language model and decoding algorithm with no additional training, differentiating it from other self-claimed plug-and-play methods that both (1) require additional discriminators and (2) lack the ability to ensure the appearance of specific words (Dathathri et al., 2020; Krause et al., 2020). In particular, this means it is out-of-the-box compatible with large pre-trained transformer models (Brown et al., 2020; Raffel et al., 2020).

We run extensive experiments with GPT-2 as our language model and find that our method produces text containing all specified guide words without an impact on the fluency and diversity of generated text. That is, we can control generation without harming text quality. Furthermore, we find that our strategy outperforms competing methods in human evaluations, and that generated text is close in fluency and overall quality to human text.

2 Preliminaries

2.1 Semantic Spaces

We model a semantic space as a vector space $\mathbb{R}^d$ over concepts where distances indicate semantic similarities (Hashimoto et al., 2016). In practice, this amounts to assigning a vector representation to each concept that reflects its semantic properties; arguably, the most basic unit that we can use to represent concepts are words (Padó and Lapata, 2003). While the notion of a semantic space is abstract in nature, prior research suggests that word embedding algorithms may in fact provide rough representations of these spaces (Hashimoto et al., 2016).

Word embeddings, i.e., continuous word representations produced by algorithms like word2vec (Mikolov et al., 2013) or GloVe (Pennington et al., 2014), have been studied in depth. Notably, it has been shown that cosine similarity between word embeddings learned by algorithms like GloVe provides a metric of semantic similarity (Erk, 2009; Pennington et al., 2014). Motivated by these results, we adopt word embedding vectors $\gamma(w)$ as points representing concepts in our semantic space—where $\gamma : \mathcal{V} \rightarrow \mathbb{R}^d$ is the mapping learned by an embedding algorithm from a word in our vocabulary $w \in \mathcal{V}$ to the vector space $\mathbb{R}^d$; we adopt cosine similarity $\cos(\gamma(w), \gamma(w'))$ for words w, w' as our notion of semantic similarity.

2.2 Text Generation

We consider probabilistic models p that assign a probability to all sequences $\mathbf{y}$ in the space of strings $\mathcal{Y}(\mathcal{V}, n_{\max})$, where $\mathcal{V}$ is the model's vocabulary and $n_{\max}$ is the maximum sequence length considered;[3] all sequences are padded with distinguished beginning-of- and end-of-sequence symbols BOS and EOS. In this work, we focus on the case of autoregressive locally normalized probabilistic models, for which the probability of a sequence $\mathbf{y}$ can be decomposed using the chain rule of probability:

$$p(\mathbf{y}) = \prod_{t=1}^{|\mathbf{y}|} p(y_t \mid \mathbf{y}_{<t}) \qquad (1)$$

In today's language generation tasks, p is typically parameterized by a neural network, e.g., a transformer or a recurrent neural network; these models have led to impressive results, producing language that is both fluent and coherent (Radford et al., 2019; Raffel et al., 2020; Brown et al., 2020).

Note that p may additionally be conditioned on some input $\mathbf{x}$, e.g., an image or a sentence in a source language.[4] Informally, we can view this conditioning as a shift of the mass of p towards the semantic space a model learns to map to from input $\mathbf{x}$; this shift may be strong, as in the case of machine translation, where the outcome should satisfy strict semantic constraints, or weak, as in topic-oriented story generation, where text should simply follow a general theme.

The task of **text generation** is to decode sequences of natural language, i.e., text, from p. There are myriad strategies for decoding, with no single de facto method used for all language generation tasks. Yet with few exceptions, all can be described using a framework consisting of two components: a **score function** and a **decoding algorithm**. Following the structure of Eq. 1, we define decoding algorithms as the general class of algorithms—which may be stochastic or deterministic in nature—that decode text autoregressively according to a score function. Common examples include nucleus sampling (Holtzman et al., 2020) and beam search (Reddy, 1977; Meister

---

[3] Typically, we have $\mathcal{Y}(\mathcal{V}, n_{\max}) = \mathcal{V}^{n_{\max}}$, i.e., the set $\mathcal{Y}$ is exponentially large in $\mathcal{V}$.

[4] We leave the dependence on $\mathbf{x}$ implicit when obvious from context.

et al., 2020). Formally, we define a score function, $\text{score}(\cdot \mid \mathbf{y}_{<t}) \colon \mathcal{V}^t \to \mathbb{R}$, as a map from strings generated under a model's vocabulary to a real number. As is clear from notation, we assume a dependency of this score on previously generated text $\mathbf{y}_{<t} = \langle y_1, \ldots, y_{t-1} \rangle$. For probabilistic text generators, the default score function is $\text{score}(\cdot \mid \mathbf{y}_{<t}) = \log p(\cdot \mid \mathbf{y}_{<t})$.[5] Other examples include mutual information (only applicable in the presence of an input $\mathbf{x}$): $\text{score}(\cdot \mid \mathbf{y}_{<t}) = \log(p(\cdot \mid \mathbf{y}_{<t}, \mathbf{x})/p(\mathbf{y}_{<t}))$ (Li et al., 2016). We note that for sampling-based decoding algorithms, the distribution over $\mathcal{V}$ output by $\text{score}(\cdot \mid \mathbf{y}_{<t})$ is projected onto the $\Delta^{|\mathcal{V}|-1}$ probability simplex—typically via the softmax transformation—such that it can be sampled from.[6]

2.3 Controlling Generation

Two different types of control can be applied over language generation models: **soft control** and **hard control**. Soft control aims at directing the mood or the general topic of the generated text, whilst hard control aims at ensuring that some explicit constraints are met, e.g., specific words are contained in the text. Note that soft control can also be reached via hard control, i.e., text that contains a set of words related to a certain topic should arguably revolve around that topic.

Some recent work has approached the problem of soft control on unconstrained language generation by training or fine-tuning language models (Ziegler et al., 2019; Yu et al., 2017; Keskar et al., 2019). However, given the existing trend towards using out-of-the-box pre-trained language models, it is desirable to develop control methods that are plug-and-play, i.e., that can be applied on an existing model without additional training. Yet currently, even methods that are termed plug-and-play require the training of an external discriminator (Dathathri et al., 2020; Krause et al., 2020).

While hard control of constrained generation, e.g., machine translation, can be attained with grid beam search methods (Hokamp and Liu, 2017; Post and Vilar, 2018; Hu et al., 2019), it is impractical to use the same approach for hard control of unconstrained generation. Methods such as grid beam search rely on the assumption that there

exists a core set of plausible candidates fulfilling the desired criteria. While typically true for machine translation—where a well-calibrated model places most of its probability mass on a (relatively) small subset of natural language sequences—this is not often the case for open-ended generation tasks. Recent work on stochastic search (Miao et al., 2019; Sha, 2020) has approached this problem by performing bidirectional search during generation and editing the text until the constraints are fulfilled. Although stochastic search is suitable for bidirectional RNN models, it is not yet clear if it can be applied to forward generation models, e.g., transformer-based models.

3 Keyword2Text

In this work, we propose *Keyword2Text* (K2T), a new and simple plug-and-play method for exerting hard control during text generation. By modifying the score function, we can incorporate a semantic shift at decoding time, without additional models or fine-tuning. This method is model agnostic—it works with any autoregressive language model, including pre-trained transformers, and can be combined with other decoding methods and objectives. Further, we show that K2T (1) does not require a pre-defined ordering of constraints, and (2) can be used for guaranteeing hard constraints.

3.1 A Controlled Generation Objective

We consider a probabilistic language generator p and a word w, which is either a specific word we would like to appear in the generated text or a topic we would like to to steer generation towards. We refer to $\gamma(w) \in \mathbb{R}^d$ as our topic vector, i.e., the point in our semantic space associated with the word w. We propose a simple modification to the score function $\text{score}(\cdot \mid \mathbf{y}_{<t})$ to guide generation towards w:

$$\begin{aligned} \text{score}'(y_t, w \mid \mathbf{y}_{<t}) &= \text{score}(y_t \mid \mathbf{y}_{<t}) \qquad (2) \\ &+ \lambda \cdot \max\Big(0, \cos(\gamma(y_t), \gamma(w))\Big) \end{aligned}$$

where λ is a hyperparameter indicating the strength of the shift. As we take $\lambda \to 0$, we recover our original score function; as $\lambda \to \infty$, the word w will be assigned increasingly more weight, until it becomes the dominating term in Eq. 2.

Note that we only consider positive similarities to avoid explicitly decrementing the scores of words that are favorable according to $\text{score}(\cdot \mid$

---

[5]Log-probability is often used for numerical stability.

[6]Algorithms that decode sets, like beam search, may additionally include a *re-ranking* process with its own separate score function. This process allows a decoding algorithm to choose among a set of candidates, potentially taking into account global qualities of a sequence, such as length (Murray and Chiang, 2018) or coverage (Wu et al., 2016).

$\mathbf{y}_{<t}$). Including negative cosine similarities could be used to control generation *away* from certain words or concepts, e.g., text detoxification (Dathathri et al., 2020; Krause et al., 2020).

3.2 An Algorithm for Guided Generation

Using the framework given in §2, the proposed score function can be used with any standard decoding algorithm, e.g., nucleus sampling or beam search. Yet incorporating this objective on its own will not guarantee that all the desired criteria for controlled decoding are met: it must be used in a principled way if we wish to work with multiple guide words or enforce the appearance of a word while still generating fluent text. We now present a general set of decoding algorithm modifications for controlled generation towards a list of guide words $W = [w_1, \ldots, w_N]$.

3.2.1 Ordered and Unordered Control

Given a list of guide words W, we propose two approaches for both the case where we need words w_n to appear in a fixed order as well as when any order suffices.

Fixed Order. We guide towards each of the words w_n in turn, i.e. we start at $n = 1$ and take $\text{score}'(y_t, w_1 \mid \mathbf{y}_{<t})$ as our score function until the word w_1 appears in the generated text. Then, we switch to $\text{score}'(y_t, w_2 \mid \mathbf{y}_{<t})$ until the word w_2 appears in the generated text. We repeat this process until all N words appear in the generated text.

Guide Closest. We now treat W as a set, i.e., the ordering of guide words is no longer important. At any given decoding step, we shift the score function by the highest cosine similarity across all words $w \in W_t$.[7] Explicitly, we score y_t as

$$\text{score}'(y_t, W_t \mid \mathbf{y}_{<t}) = \text{score}(y_t \mid \mathbf{y}_{<t}) + \quad (3)$$
$$\lambda \cdot \max\left(0, \max_{w \in W_t} \cos(\gamma(y_t), \gamma(w))\right)$$

where we overload $\text{score}'(\cdot \mid \mathbf{y}_{<t})$ to take a set $W_t \subseteq W$—the guide words that have *not* appeared before step t—as input. Notably, this implies that the guide words do *not* need to be ordered in our approach. Previously proposed decoding algorithms, e.g. Hokamp and Liu (2017) and Post and Vilar (2018), run in exponential time without an ordering

of the guide words. Note that in either of our approaches discussed above, once all $w \in W$ appear in $\mathbf{y}_{\leq t}$, we may revert back to the original score function $\text{score}(\cdot \mid \mathbf{y}_{<t})$.

3.3 Guaranteeing Word Appearance

We can guarantee the appearance of guide words when generating text by controlling the shift parameter λ of Eq. 2. As said, this parameter regulates the spectrum ranging from uncontrolled generation to forcing the next word to be a guide word. We propose to increase λ on an exponential schedule. In words, as the generated sequence increases in length, so does the strength of the semantic shift, until we deterministically choose the guide word. [8]

Formally, suppose the maximum length of the sequence is T and the previous guide word appeared at t_n (we define $t_0 = 0$), then the weight at time t, where $t_n < t < T - |W_t|$, is given by

$$\lambda_t = \begin{cases} \lambda_0 \exp\left\{\frac{c(t-t_n)}{T-|W_t|-t_n}\right\} & \textbf{if } t < T - |W_t| \\ \infty & \textbf{otherwise} \end{cases}$$
$$(4)$$

Thus, as we approach the maximum length, T, we exponentially increase the shift parameter. When we have only enough space for the remaining guide words, i.e. $t = T - |W_t|$, we explicitly force the remaining guide words to appear by setting $\lambda_t = \infty$. We use the hyperparameters λ_0 and c to control the initial value and growth of λ, respectively.[9]

4 Experimental Setup

To evaluate our controlled decoding strategy, we run three sets of experiments: (1) analysis of hyperparameters; (2) comparison to competing methods; (3) comparison to human text. In each of these experiments, the task consists of generating text that contains certain keywords. We use GPT-2 large (774 million parameters; Radford et al., 2019) as our language model and GloVe as the embedding algorithm that generates $\gamma(\cdot)$. For experiments (2) and (3), we additionally run user studies; in all experiments we calculate the following automatic metrics, which in combination, we take as an (imperfect) estimate of text quality (Welleck et al., 2020; Martins et al., 2020).

---

[7]Note that we do not shift the tokens towards all guide words additively; this corresponds to shifting towards the mean of the guide word embeddings, which may correspond to a different area of the semantic space.

[8]Note that if all the keywords have not appeared and the *end-of-sequence* token is generated the algorithm discards it and samples again; i.e., *end-of-sequence* is not allowed until all the keywords have been generated.

[9]Empirically, we find that any reasonably large value of c works well, e.g., $c = 100$.

- *Perplexity*: exponentiation of the negative average per-token log-probability under a language model, lower is typically considered better. We use a separate model Distil-GPT-2 (Wolf et al., 2020) to calculate perplexity to avoid inflated estimates (Liu et al., 2016).

- *Repetition*: we use the sequence repetition score from Welleck et al. (2020), which computes the proportion of repeated 4-grams in the text; a lower score implies less repetitive text. As reference, Welleck et al. (2020) find the repetition of human text from a subset of WikiText-103 (Merity et al., 2017) to be 0.6%.

- *Success Rate*: Percentage of total number of keywords that appear in the text.

We use word stemming to check for the occurrence of the target words, i.e., if a generated word has the same stem as the keyword, it is counted as an occurrence; this way special cases such as plurals can be handled automatically, whilst keeping the text coherent by avoiding semantic redundancies, such as "protective protection".

4.1 Hyperparameter Analysis

To analyse hyperparameter choices, we devise a task that consists of generating a short piece of text that contains five keywords (cf. App. A). We use solely the *start of sequence* token as initial context for GPT-2 to start generating text, and we stop after 90 tokens have been generated. Since the keywords in each set bear no relation to each other, it is very challenging to include all of them in a coherent and fluent piece of text; in fact, generated samples tend to have high perplexity scores. Therefore, this task is highly demanding, making it a tough benchmark for controlled language generation.

4.2 Comparison to Alternative Methods

Using the ROC story dataset, we compare against two methods for hard control in language generation, CGMH (Miao et al., 2019) and Plan-and-write (Yao et al., 2019); as a trainable baseline we fine-tune GPT-2. See App. A for model details.

CGMH. CGMH is a stochastic search method that at each step samples a word in the generated sentence and an operation (replace, delete, insert). This method is plug-an-play but it works only with bidirectional language models. Miao et al. show that CGMH clearly outperforms grid

beam search (Hokamp and Liu, 2017) and the backward-forward model (Mou et al., 2015).

Plan-and-write. Plan-and-write is an end-to-end model for story generation, which takes a title as input and outputs a story. It proceeds in two steps: first, it generates a storyline represented by a sequence of keywords; then it generates a story based on the title and the storyline.

Fine-tuned GPT-2. To fine-tune GPT-2, for each training example we extract five keywords w_n from the story text using the publicly available YAKE library. Then, we provide these five keywords as a prompt together with the story text: "BOS w_1, w_2, w_3, w_4, w_5 = *<text>* EOS", where EOS is the *end-of-sequence* token. We fine-tune GPT2 on a Titan RTX GPU (24GB) for five epochs on causal language modeling, which takes around 26 hours. At evaluation time we give as initial context a prompt with the same format as during training and we use nucleus sampling with $p = 0.9$.

The objective of the ROC story dataset is to generate stories given a title.[10] To conform the dataset to our task, i.e., keyword to text, we build a test set of 20 titles randomly selected from the dataset and use the plan component of Plan-and-write (Yao et al., 2019) to generate five keywords from these titles. Using these keywords we generate one story with each of the competing methods, resulting in 20 examples per method. We perform a human evaluation comparing *GPT-2+K2T*, *CGMH* and *Plan-and-write* in which 30 evaluators are presented each of the three texts and asked to evaluate them relative to each other in terms of 4 criteria: fluency, logical consistency, creativity and best overall (see App. A and B for details).

4.3 Comparison to Human Text

To assess the quality of text generated using K2T with respect to human text, we generate news articles from keywords. We employ the 500N-KPCrowd dataset from Marujo et al. (2011), which consists of pieces of news written by professional journalists with keywords assigned by human annotators. For our evaluation, we randomly select ten keyword-article pairs from the test set. At generation time, the language model receives the first

---

[10]We do not to evaluate on tasks like CommonGen (Lin et al., 2020) since texts for those tasks are expected to be very simple; thus performance metrics would not reveal how K2T affects aspects such as fluency, which are vital in the tasks we consider, e.g., generating stories from models like GPT-2.

Method	SR (%)	PPL	Rep. (%)
No control	0.6 ± 0.5	34.6 ± 3.2	$\mathbf{2.4} \pm 0.7$
W. $\lambda = 5$	4.4 ± 0.9	34.5 ± 2.8	3.7 ± 0.8
W. $\lambda = 10$	52.0 ± 3.4	46.7 ± 3.3	7.2 ± 1.3
W. $\lambda = 20$	84.35 ± 1.2	225.9 ± 132.6	33.0 ± 1.5
C. $\lambda = 5$	12.2 ± 2.1	$\mathbf{29.8} \pm 1.3$	3.3 ± 1.4
C. $\lambda = 10$	72.6 ± 2.8	44.75 ± 3.7	8.7 ± 1.3
C. $\lambda = 20$	$\mathbf{95.1} \pm 2.3$	99.3 ± 20.1	13.4 ± 2.1

Table 1: Comparison of the *No control*, *Guide words only (W)* and *Guide Context (C)* strategies.

λ_0	PPL	Rep. (%)
5	$\mathbf{58.4} \pm 4.5$	$\mathbf{3.5} \pm 1.1$
10	70.5 ± 7.1	6.4 ± 2.2
15	109.5 ± 24.2	10.5 ± 3.2
20	235.8 ± 352.2	10.6 ± 1.7
25	135.8 ± 44.9	9.9 ± 2.4
30	310.3 ± 366.8	9.5 ± 2.1

Table 2: Evaluation of the shift strength λ.

30 words of the original article as initial context. In our baseline, the model receives no guidance, i.e., $\lambda = 0$. We compare this to text generated from the same model albeit controlled by K2T.

We design our human evaluation based on best practices for evaluation of generated text (van der Lee et al., 2019). Specifically, we prepare three sets of ten articles, each consisting of a combination of original articles (written by humans), articles generated by our method and articles generated by GPT-2 without control. We create three separate surveys so that evaluators will only be exposed to one version of each story (they do not know the origin). Participants are asked to evaluate how coherent, fluent and natural (human-like) each article is, as well as its overall quality, on a 7-point Likert scale (van der Lee et al., 2019), i.e., from 1 to 7 where higher is better (cf. App. C). Each survey is shown to 30 evaluators. In App. C.2 we show all the articles used in this study.

5 Results

5.1 Hyperparameter Analysis

We perform the keyword to phrase task specified in §4.1. We run each experiment ten times with different seeds and report the mean and standard deviation across the runs.

Controlled Generation. First, we assess the baseline effectiveness of our generation objective in Eq. 2 for generating text that contains a set of guide words W. We compare three approaches.

- *No control*: language generation without guidance, i.e., $\lambda = 0$;

- *Guide words only*:[11] shifting the scores only for

the tokens corresponding exactly to the guide words W and not for similar words;

- *Guide context*: shifting the score for the guide words W in addition to semantically similar words.

In this evaluation we do not anneal λ; thus, guide words are *not* guaranteed to appear. It follows that we can only evaluate the performance of our method under soft constraints. Unless otherwise stated, here and in the following, we apply unordered control with the *Guide Closest* strategy, we use $\text{score}(\cdot \mid \mathbf{y}_{<t}) = \log p(\cdot \mid \mathbf{y}_{<t})$ as our original score function and we use nucleus sampling with $p = 0.9$ as the decoding algorithm. We present the results of this comparison in Table 1.

When the score function is not shifted (*No control*), the average success rate is only 0.6%, which serves as a random baseline. We see that guiding towards the context is more effective than guiding only towards the exact guide words, i.e., for the three values of λ considered here, both success rate and perplexity are significantly better. Furthermore, the repetition score of *Guide Context* is clearly better for $\lambda = 20$, the only value where both approaches reach a high success rate. Note that encouraging the appearance of specific keywords also helps keep the generated text on-topic; this may explain the lower perplexity of the best performing configuration versus *No control*. These results validate *Guide Context* as the best strategy.

Initial Shift Strength λ_0**.** Next, we investigate the effect of varying the initial shift strength λ_0; here and in the remaining experiments we guarantee the appearance of all keywords. In Table 2 we report perplexity and repetition for different values of λ_0 (success rate is 100% in all cases). We see that increasing λ_0 results in a worsening of perplexity and repetition scores, with strong variability across different runs, i.e., high standard deviation.

---

[11]This is similar to the weighted decoding (WD) baseline used in (Dathathri et al., 2020).

Strategy	PPL	Rep. (%)
Guide Closest	58.4 ± 4.5	3.5 ± 1.1
Guide All	$\mathbf{39.7} \pm 2.7$	30.0 ± 2.3
Guide Random	66.9 ± 3.7	$\mathbf{1.5} \pm 0.5$
Fixed Order	61.7 ± 4.2	3.4 ± 1.2

Table 3: Comparison of different approaches for control towards unordered constraints with $\lambda_0 = 5$.

For $\lambda_0 = 5$ the average perplexity (58.4) and repetition score (3.5%) are the best among the considered values; we use this value of λ_0 in the remaining experiments.

Unordered Control. Given a set of words, it is not trivial to devise a strategy to guide text generation towards those words without a pre-specified order. We explore two additional strategies on top of *Guide Closest*, described in §3.2:

- *Guide All*: We shift the scores towards all guide words at once, by adding the sum of the cosine similarities.
- *Guide Random*: At each generation step we choose the next guide word uniformly at random from the remaining guide words and shift the scores towards this word.

In Table 3 we present the evaluation of these strategies. For reference, we also report results for guiding the guide words in order. We see that guiding towards *all* the words at the same time performs poorly in terms of repetition score (30.0%). On the other hand, *Guide Random* and *Guide Closest* produce similar results, with smaller perplexity for the *Guide Closest* strategy and thus, in the following we adopt the *Guide Closest* strategy.

Decoding Algorithm. Finally, we evaluate our method in conjunction with different decoding algorithms:

- Nucleus sampling (NS; Holtzman et al., 2020) with $p = 0.9$.
- Beam search (BS) with 4 beams and length normalized re-ranking, i.e. taking the top-K candidates according to $\text{score}(\mathbf{y}) = \log p(\mathbf{y})/|\mathbf{y}|$.
- BS with word count (BS+WC): we increment the above re-ranking function if the generated token y_t corresponds to one of the remaining

Strategy	PPL	Rep. (%)
NS	58.4 ± 4.5	$\mathbf{3.5} \pm 1.1$
BS	$\mathbf{9.8} \pm 0$	46.5 ± 0
BS+WC	11.9 ± 0	38.7 ± 0
BS+WC+NS	21.2 ± 1.7	13.4 ± 2.2

Table 4: Comparison of different decoding algorithms.

Words: *Guide, Jump, Eight, Row, Settle*

Text: *One of the most common questions I get from new **rowers** is how to warm up before a race. **Eight** months ago, I wrote a **guide** to warm up and how to **jump** in the water, but I never got around to updating it. So I decided to **settle** the question once and for all, and try to write the best possible course through it.*

Words: *Search, Major, String, Cost, Village*

Text: *A **major string** of high-profile **cost**-cutting measures announced by the Ontario government this week is expected to result in a significant drop in the **village**'s operating budget by the end of the year. The **search** for ways to lower the village's pension plan contribution from 7.5 per cent to 6.9 per cent is expected to save an additional \$80 million in the current fiscal year.*

Table 5: Examples of text generated by K2T.

guide words:

$$\text{score}'(\mathbf{y}_{<t}) = \text{score}(\mathbf{y}_{<t}) \qquad (5)$$
$$+ \left(|W| - |W_{t+1}| \right)$$

where $|W| - |W_t| \geq 0$ equals the number of guide words that have appeared by step t.

- BS+WC with NS (BS+WC+NS): we consider a hybrid of beam search and nucleus sampling where the words are sampled using nucleus (instead of deterministically picking the top words).

Table 4 shows that while NS produces diverse text, its perplexity is the highest by a margin. Conversely, BS with and without word count reward, generates text with low perplexity but high repetition. The combination of both reaches the best trade-off.

In Table 5 we show two examples generated with the best-performing configuration: $\lambda_0 = 5$, *Guide*

Method	SR (%)	PPL	Rep. (%)
Plan-and-Write	96.0	**33.9**	25.7
CGMH	97.0	127.8	1.6
GPT-2 fine-tuned	72.0	89.4	1.8
GPT-2 + K2T	**100.0**	48.8	**1.5**

Table 6: Comparison to competing methods.

Text	SR (%)	PPL	Rep. (%)
Original	100.0	15.2	1.3
GPT-2	0.0	8.8	11.5
Ours	100.0	12.5	1.0

Table 7: Comparison of the *original* articles and text generated by *GPT-2* without and with control (*ours*).

Closest, *BS+WC+NS* decoding, which we use for the experiments in §5.2 and §5.3.

5.2 Comparison to Alternative Methods

In Table 6, we report metrics for each of the three methods (K2T, Plan-and-write and CGMH) on the keyword-to-text task based on the ROC dataset. Neither Plan-and-write (P&W), CGMH nor fine-tuned GPT-2 reach a 100% success rate and, while P&W presents the lowest perplexity, its repetition score is by far the highest (25.7). Conversely, although the repetition score of CGMH and GPT-2 fine-tuned are comparable to our method, they generate text with very high perplexity, 127.8 and 89.4 respectively. In light of these results, K2T seems to provide the best trade-off between perplexity and repetition, on top of reaching 100% success rate.

In Figure 1 we see the results of the user study comparing CGMH, P&W and GPT-2+K2T. For each method, we report the proportion of times that it is chosen as the best of the three across four different axes: fluency, consistency, creativity and overall quality. K2T outperforms the other two in all aspects, in particular in creativity and overall quality. These results suggest that our method is superior to the baseline hard control methods on both objective and subjective metrics.

5.3 Comparison to Human Text

In Table 7 we report perplexity, repetition score and success rate on the news article task. We compare the original news articles with those generated by uncontrolled GPT-2 and with GPT-2 controlled by K2T. We see that without controlling GPT-2, the keywords do not appear in the generated texts, despite the 30 words of initial context. This suggests that methods such as grid beam search (Hokamp and Liu, 2017; Post and Vilar, 2018) would perform poorly. We also see that when GPT-2 is controlled by K2T, the text it generates is close to the original text in terms of perplexity and repetition score. To explore the perceived impact of the differences, we look at the results of the human evaluations.

Figure 2 shows the results of the human evaluations on the news article task. As expected, news articles written by professional journalists are assessed superior to those written by GPT-2 when controlled by K2T. Although the difference is statistically significant ($p < 0.05$; t-test), it is small: less than 0.6 points for each of the evaluated properties on a scale from 1 to 7. Furthermore, the text generated when controlling GPT-2 is comparable in terms of naturalness and fluency, and significantly better ($p < 0.05$; t-test) in terms of consistency and overall quality, than the text from uncontrolled GPT-2. These results demonstrate that K2T does not compromise the quality of the generated text. Further, the resulting text is close in perceived quality to human text.

6 Related Work

Various other approaches exist for controlled language generation; here we review those that are not discussed in §2.

Hard Control. Hard control of autoregressive language models in unconstrained settings has remained elusive thus far. Xu et al. (2020) propose a framework to exert hard control over language generation which, nevertheless, requires training three large transformer models.

Soft Control. Related to our approach of using semantic spaces to control generation, Chang et al. (2021) use the space of GloVe embeddings to define control topics. However, this method is not plug-and-play since it requires fine-tuning a GPT-2 encoder to generate text aligned with the control topics. A number of plug-and-play methods for soft control exist that use external discriminators to steer language generation. Holtzman et al. (2018) train discriminators on different linguistic properties to improve the quality of generated text. Dathathri et al. (2020) use the gradients of an external discriminator to direct the generation of a
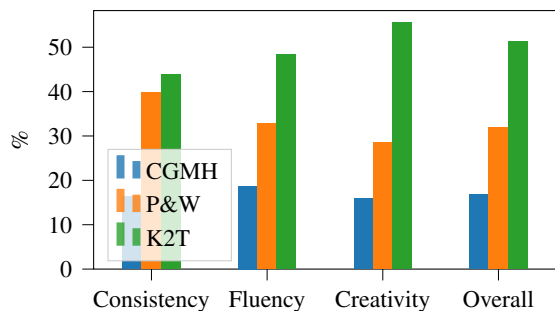
Figure 1: User study comparing the three methods; bars indicate how often each method was picked as the best.

Figure 2: User study comparing our control method to *original* and *GPT-2* articles.

pre-trained language model towards the target topic. Similarly, Krause et al. (2020) use a contrastive strategy to soft-control language generation. Yang and Klein (2021) directly modify the output probabilities of a (pre-trained) language model using the output of a discriminator that determines whether future text will contain the desired attribute, e.g., formality. Unfortunately, the use of external discriminators limits the applicability of these methods since they require training data for each of the target topics or attributes. Our method elegantly dispenses with the need for a discriminator by using the geometric properties of semantic spaces.

7 Conclusion

In this work, we present K2T, a simple and intuitive plug-and-play decoding method that can be used to impose controls on any autoregressive model for language generation, including large pre-trained transformers, like GPT-2. Our method guarantees the appearance of guide words and requires neither re-training nor the use of external discriminators. Our two user studies reveal that K2T is superior to competing hard control methods and that there is no statistical difference in perceived fluency, consistency and overall quality between news articles generated by our method and by GPT-2 without control. In future work, we plan to investigate in more detail the application of this decoding method to soft control tasks and to text detoxification.

Ethics Statement

We recognize that controllable language generation can potentially be used to produce misinformation or offensive text. However, we believe that further research on controllable generation is necessary to ensure that we have at our disposal the tools needed to prevent automatic language generation
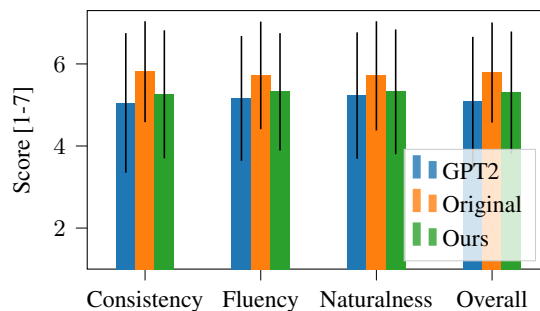
techniques from being used for malicious purposes. Methods for controlling large pre-trained language models are also a promising tool to mitigate the generation of biased text, e.g., steering generation towards the semantic space of both "woman" and "doctor" may mitigate the bias typically seen surrounding stereotypically male-oriented professions. We believe this is an important future direction for our work. Further, we consider the environmental impacts of our method. As our method is completely plug-and-play, i.e., it is able to make use of pre-trained language models and word embeddings, we hope that it can be utilized in place of techniques that require extensive model training. This would in turn reduce the energy consumption required to setup a controllable language generation system, which for large language generation models, can be quite significant.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Haw-Shiuan Chang, Jiaming Yuan, Mohit Iyyer, and Andrew McCallum. 2021. Changing the mind of transformers for topically-controllable language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2601–2611, Online. Association for Computational Linguistics.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *8th International Conference on Learning*, Addis Ababa, Ethiopia.

Katrin Erk. 2009. Representing words as regions in vector space. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 57–65, Boulder, Colorado. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Tatsunori B. Hashimoto, David Alvarez-Melis, and Tommi S. Jaakkola. 2016. Word embeddings as metric recovery in semantic spaces. *Transactions of the Association for Computational Linguistics*, 4:273–286.

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

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations*, Addis Ababa, Ethiopia.

Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine Bosselut, David Golub, and Yejin Choi. 2018. Learning to write with cooperative discriminators. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1638–1649, Melbourne, Australia. Association for Computational Linguistics.

J. Edward Hu, Huda Khayrallah, Ryan Culkin, Patrick Xia, Tongfei Chen, Matt Post, and Benjamin Van Durme. 2019. Improved lexically constrained decoding for translation and monolingual rewriting. 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 839–850, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2020. GeDi: Generative discriminator guided sequence generation. *CoRR*, abs/2009.06367.

Michael Lebowitz. 1987. Planning stories. In *Proceedings of the 9th Annual Conference of the Cognitive Science Society*, pages 234–242.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840, Online. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Pedro Henrique Martins, Zita Marinho, and André F. T. Martins. 2020. Sparse text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4252–4273, Online. Association for Computational Linguistics.

Luís Marujo, Márcio Viveiros, and João Paulo Neto. 2011. Keyphrase cloud generation of broadcast news. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association*, pages 2393–2396. ISCA.

James R. Meehan. 1977. TALE-SPIN, an interactive program that writes stories. In *Proceedings of the 5th International Joint Conference on Artificial Intelligence*, pages 91–98.

Clara Meister, Tim Vieira, and Ryan Cotterell. 2020. If beam search is the answer, what was the question? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Online. Association for Computational Linguistics.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *5th International Conference on Learning Representations*, Toulon, France.

Ning Miao, Hao Zhou, Lili Mou, Rui Yan, and Lei Li. 2019. CGMH: constrained sentence generation by metropolis-hastings sampling. In *The Thirty-Third*

AAAI Conference on Artificial Intelligence, pages 6834–6842. AAAI Press.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations (Workshop Track)*, Scottsdale, Arizona, USA.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Lili Mou, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. 2015. Backward and forward language modeling for constrained sentence generation. *CoRR*, abs/1512.06612.

Kenton Murray and David Chiang. 2018. Correcting length bias in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 212–223, Brussels, Belgium. Association for Computational Linguistics.

Sebastian Padó and Mirella Lapata. 2003. Constructing semantic space models from parsed corpora. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 128–135, Sapporo, Japan. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

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

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

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

Raj Reddy. 1977. Speech understanding systems: A summary of results of the five-year research effort at carnegie mellon university.

Lei Sha. 2020. Gradient-guided unsupervised lexically constrained text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8692–8703, Online. Association for Computational Linguistics.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. In *8th International Conference on Learning Representations*, Addis Ababa, Ethiopia.

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.

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

Peng Xu, Mostofa Patwary, Mohammad Shoeybi, Raul Puri, Pascale Fung, Anima Anandkumar, and Bryan Catanzaro. 2020. MEGATRON-CNTRL: Controllable story generation with external knowledge using large-scale language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2831–2845, Online. Association for Computational Linguistics.

Kevin Yang and Dan Klein. 2021. FUDGE: Controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3511–3535, Online. Association for Computational Linguistics.

Lili Yao, Nanyun Peng, Ralph M. Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In

The Thirty-Third AAAI Conference on Artificial Intelligence, pages 7378–7385. AAAI Press.

Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2017. Seqgan: Sequence generative adversarial nets with policy gradient. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 2852–2858. AAAI Press.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *CoRR*, abs/1909.08593.

A   Evaluation Details

In this appendix we provide additional details on different aspects of our evaluation to ease reproducibility.

A.1   Keywords for Hyperparameter Analysis

To generate the keyword sets we use in our hyperparameter analysis we use a list of 1000 common English words[12]. From this list, we discard the first 500 words, which we found to be too common and from the remaining 500 words, we additionally discard stop words. Then, from the resulting list we sample 50 sets of 5 words each, which constitutes our keyword sets.

A.2   Details of Alternative Methods

For CGMH (Miao et al., 2019), we use the same model as in the original work: a two-layer LSTM (63M parameters); and train it using a TITAN Xp 12GB GPU following the instructions from the original repository[13]. For generating the samples, we run the model for 1000 updates with a minimum sequence length of 30 words. Similarly, Plan-and-write (Yao et al., 2019) uses an LSTM based sequence-to-sequence model with 62M parameters. We train the model for 357 epochs on a TITAN Xp 12GB GPU with the ROC story dataset (Mostafazadeh et al., 2016).

A.3   Human Evaluation

We perform both user studies using Amazon MTurks. In each of them, we include simple comprehension questions as a control to ensure that the evaluators read the texts carefully. We discard and replace those evaluators who do not spend the minimum expected time on the study or who do not answer enough control questions correctly.

A.4   Comparison to Human Text

To generate the news articles using the 500N-KPCrowd dataset from Marujo et al. (2011), we randomly select ten keyword-article pairs from the test set with a length ranging from 81 to 217 words (so that the articles are neither too short nor too long for human assessment). Then, from the given keywords, we first remove proper nouns, e.g., names and cities. As the number of keywords provided in some articles is quite large, we further pare this list down by randomly removing selected keywords until the proportion is one keyword per ten words of text, such that all article-keyword pairs have present the same ratio of keywords per length.

---

[12]https://github.com/powerlanguage/word-lists
[13]https://github.com/NingMiao/CGMH

Figure 3: Screenshot of the ROC story user study.

B ROC Story Survey

Here we provide a screenshot to show how our survey for method comparison looks.

C News Articles

C.1 Survey

As above, here we provide a screenshot that shows how our survey for comparison to human text looks.

News Article Text Quality Survey

*Required

Read the text below and answer the questions

On Wednesday, Egypt's Supreme Military Council issued a declaration which will temporarily replace the country's former constitution, dating back to 1971. The declaration is based on amendments to the old constitution which were backed by the majority of people at a referendum on March 19. The amendments mostly concern procedures for presidential and parliamentary elections, and the terms of presidency. The old constitution was abrogated by the country's military on February 13, soon after President Mubarak's resignation. Parliamentary elections in Egypt are planned for this September.

Definitions (reminder)
Fluency: the text can be understood easily, well, and quickly
Logical Consistency: there are no contradictions in the text
Naturalness: the text seems to have been written by a human

There was a referendum about the constitution in February. *

○ True

○ False

This text is (logically) consistent. *

	1	2	3	4	5	6	7	
strongly disagree	○	○	○	○	○	○	○	strongly agree

This text is fluent. *

	1	2	3	4	5	6	7	
strongly disagree	○	○	○	○	○	○	○	strongly agree

Figure 4: Screenshot of the news article user study.

C.2 Texts

Below we show the original news articles and the news articles generated by our method that were used in the human evaluation study. Detailed results of the study are included in Appendix C.3.

Text 1

Context: Johnny Depp will be asked to make a fifth Pirates Of The Caribbean film if the fourth instalment is a success. Producer Jerry Bruckheimer said he already has a screenplay

Keywords: fifth, audience, embraces, character, original, trilogy, continue, story, digital, cameras, pirates, fresh, new, carry

---

OURS + GPT2-774M: *Johnny Depp will be asked to make a fifth Pirates Of The Caribbean film if the fourth instalment is a success. Producer Jerry Bruckheimer said he already has a screenplay with* **Pirates new trilogy character** *Johnny Depp and* **digital** *animation studio Story Studio. The* **fifth** *film will* **continue** *the* **original story** *from the first three films and will be set in the Caribbean. "We are very excited to* **carry** *on the story of the* **audience**'s *favourite Pirates of the Caribbean characters," said Bruckheimer. "We are* **fresh** *off the success of Pirates of the Caribbean: On Stranger Tides and we are looking forward to* **embracing** *the new generation of fans with a new film that will be even more exciting than the first three films." The Pirates of the Caribbean: On Stranger Tides* **cameras** *have been filming in Barranquilla since April 11. Filming will continue for two months to reach completion by May 23.*

HUMAN: *Johnny Depp will be asked to make a fifth Pirates Of The Caribbean film if the fourth instalment is a success. Producer Jerry Bruckheimer said he already has a screenplay in the works for a* **fifth** *Pirates tale, which would follow the forthcoming Pirates Of The Caribbean: On Stranger Tides. "As long as the* **audience embraces** *this one, we'll certainly try to make another one. It's really up to Johnny. He loves the* **character**," *he said. The* **original** *three Pirates blockbusters ended up as a* **trilogy continuing** *the same key characters and* **story** *line. But Jerry said On Stranger Tides - the first Pirates flick shot with* **digital 3D cameras** *- and future* **Pirates** *films will be stand-alone stories continuing the adventures of Johnny's woozy buccaneer Captain Jack Sparrow. "The audience told us (at test screenings of On Stranger Tides) what they loved about it is that it was* **fresh**, *it was new, it was a whole* **new** *story, "Jerry said. "So that will* **carry** *over into the next one, too, to give it something fresh and different."*

Text 2

Context: On Wednesday, Egypt's Supreme Military Council issued a declaration which will temporarily replace the country's former constitution, dating back to 1971. The declaration is based on amendments to the old

Keywords: constitution, majority, elections, military, resignation

OURS + GPT2-774M: *On Wednesday, Egypt's Supreme Military Council issued a declaration which will temporarily replace the country's former constitution, dating back to 1971. The declaration is based on amendments to the old **constitution** which were approved by the **majority** of **military** members in **elections** earlier this year. The new constitution will replace the previous one, which was approved by a referendum in June. The **resignation** of President Mohammed Morsi in July 2011 was a prerequisite for the current constitutional amendment. Civil society activists point to the judiciary, a particularly divisive institution. They accuse the military council and authorities of targeting the judiciary to weaken the role of the Egyptian people.*

HUMAN: *On Wednesday, Egypt's Supreme Military Council issued a declaration which will temporarily replace the country's former constitution, dating back to 1971. The declaration is based on amendments to the old **constitution** which were backed by the **majority** of people at a referendum on March 19. The amendments mostly concern procedures for presidential and parliamentary **elections**, and the terms of presidency. The old constitution was abrogated by the country's **military** on February 13, soon after President Mubarak's **resignation**. Parliamentary elections in Egypt are planned for this September.*

Text 3

Context: A group of moderate House Democrats have initiated conversations with Republican lawmakers in a bid to try to reach a deal to tackle burgeoning federal spending. The group is being

Keywords: leaders, coalition, conservative, lawmakers, goals

OURS + GPT2-774M: *A group of moderate House Democrats have initiated conversations with Republican lawmakers in a bid to try to reach a deal to tackle burgeoning federal spending. The group is being led by **conservative lawmakers** and **leaders** of the **coalition**, including Reps. Chris Van Hollen (Md.), Jim McGovern (Mass.), and Jim McDermott (Wash.). The **goals** are to present a united front and try to provide some bipartisan credibility to the nascent spending talks. That effort will likely fail, however, as the Democratic leaders have been unable to reach agreement on a spending plan that is both ambitious and bipartisan.*

HUMAN: *A group of moderate House Democrats have initiated conversations with Republican lawmakers in a bid to try to reach a deal to tackle burgeoning federal spending. The group is being led by Reps. Jim Cooper (D., Tenn.) and Kurt Schrader (D., Ore.), both of whom are **leaders** of the Blue Dogs, a **coalition** of fiscal **conservative** Democrats. Tuesday, the **lawmakers** released a series of fiscal discipline **goals**, many of which are similar to the targets set in place by the deficit commission established...*

Text 4

Context: Oscar nominated actress Hailee Steinfeld is set to play Sleeping Beauty in a re-visioning of the classic fairytale. The 14-year-old star has signed on to play the unfortunate princess in

Keywords: movie, version, story, star, sleep, dream, role, novel, upcoming, youngest, actress, nominated, television, adaptation, classic

OURS + GPT2-774M: *Oscar nominated actress Hailee Steinfeld is set to play Sleeping Beauty in a re-visioning of the classic fairytale. The 14-year-old star has signed on to play the unfortunate princess in the **adaptation** of **classic novel Sleeping** Beauty **starring actress** Emma Watson. The **movie** will be directed by Story **television** and produced by **DreamWorks** Animation. The **role** of Sleeping Beauty was originally played by Australian actress Sophie Okonedo in the original **version** of the **story**. The **upcoming** remake will be directed by Oscar **nominated** director Rob Marshall. Steinfeld, who is the **youngest** of four siblings and a native of Melbourne, Australia, has already landed a pair of teen movies with her younger sister Zoe in Zoe and Me and Ice Age. "It's exciting to be working with DreamWorks again, and on a film like Sleeping Beauty," Steinfeld said.*

HUMAN: *Oscar nominated actress Hailee Steinfeld is set to play Sleeping Beauty in a re-visioning of the classic fairytale. The 14-year-old star has signed on to play the unfortunate princess in the new **movie**. It will be a spin of a **version** of the **story** which is being written by screen writer Lindsay Devlin. According to Deadline.com the new version will give the **star** more to do than just **sleep** as it will follow her as she enters a **dream** world and has to find her way out. Hailee is also currently being considered to play the lead **role** in the **novel**-based film Forgotten, from the **upcoming** book of the same name by Cat Patrick. After being one of the **youngest** ever **actresses** to be **nominated** for an Academy Award at this years Oscars she is set to have a busy year with several other offers on the table. The young California native had only starred in several **television** bit parts before landing the lead role in the Cohen Brothers **adaptation** of the **classic** John Wayne western True Grit alongside Jeff Bridges.*

Text 5

Context: Pinch-hitter David Murphy delivered a tiebreaking, two-run double in the eighth inning and the Texas Rangers rallied for a 9-5 victory over the Boston Red Sox on Friday after raising

Keywords: championship, kicked, land, newcomers, home, score, played, season, victory, hits, runs, debut, strikeouts

OURS + GPT2-774M: *Pinch-hitter David Murphy delivered a tiebreaking, two-run double in the eighth inning and the Texas Rangers rallied for a 9-5 victory over the Boston Red Sox on Friday after raising their **season home** run total to 27. The **victory** was the Rangers' first since **hitting** 27 home **runs** in a season-opening win over the New York Yankees on April 1. The Rangers scored four runs in the eighth inning to take a 6-5 lead. Murphy **played** the role of a pinch hitter in the eighth inning and **kicked** a single to right field to **score** Mark Trumbo and make it 7-5. The Rangers scored four more runs in the ninth inning to **land** the win. Matt Bush hit a two-run homer to right field and Prince Fielder hit a two-run double to left field to make it 9-5. The **strikeouts** were the most the Rangers have allowed in a game since **debuting** with 26 in a win over the New York Yankees on April 1. The **championship** of the American League West will be decided on Sunday when the Rangers host the **newcomers** the Los Angeles Angels.*

HUMAN: *Pinch-hitter David Murphy delivered a tiebreaking, two-run double in the eighth inning and the Texas Rangers rallied for a 9-5 victory over the Boston Red Sox on Friday after raising their American League **championship** flag. Murphy's slicing liner off Daniel Bard (0-1) **kicked** up chalk when it **landed** on the left-field line. That sent Rangers **newcomers** Mike Napoli and Yorvit Torrealba **home** to break a 5-all tie. Murphy **scored** on a double by Elvis Andrus before another double by AL MVP Josh Hamilton. Napoli, Ian Kinsler and Nelson Cruz all homered off Jon Lester for Texas, which **played** its **season** opener exactly five months after a Game 5 loss to San Francisco at home ended its first World Series. Darren Oliver (1-0) got the **victory** after allowing a tying homer to David Ortiz in the eighth. Adrian Gonzalez had two **hits** and drove in three **runs** in his Boston **debut.** Carl Crawford went 0 for 4 with three **strikeouts** while leaving a runner in scoring position each at-bat.*

Text 6

Context: An Alaska lawmaker introduced an amendment Friday that would give oil companies a tax break provided they pledged to increase production. Alaska Gov. Sean Parnell, a Republican, unveiled plans for

Keywords: barrels, production, pipeline, companies, investments, oil, tax, message, legislators, strategy, future, amendment, restrictions, companies, tax, increase, percentage, benefit

OURS + GPT2-774M: *An Alaska lawmaker introduced an amendment Friday that would give oil companies a tax break provided they pledged to increase production. Alaska Gov. Sean Parnell, a Republican, unveiled plans for a **tax** credit for **oil** companies that **increase production**. **Pipeline companies** would **benefit** from **tax** breaks as well. The **amendment** would **restrict** oil **companies** from receiving a tax credit if they increase production by more than 10 percent a year for five years and **invest** more than $1 billion in **future** projects. The **legislation** is expected to be introduced in the House and Senate this week. Parnell said the **percentage** increase would be based on the amount of **barrels** of oil produced by the company. "This is a great opportunity for us to help the oil industry grow and create jobs," Parnell said. The **strategy** is expected to be controversial in Alaska, where the oil industry has been a major employer and a major contributor to the state's economy. The state has the nation's third-largest oil reserves and is home to the world's largest crude oil **messageboard**.*

HUMAN: *An Alaska lawmaker introduced an amendment Friday that would give oil companies a tax break provided they pledged to increase production. Alaska Gov. Sean Parnell, a Republican, unveiled plans for 1 million **barrels** of oil **production** per day through the Trans-Alaska **pipeline** system within the next decade. Parnell said he was proposing a tax break for oil **companies** to encourage **investments** and to erase declines in state **oil** production. "The time to reduce oil **taxes** is now and I am asking all Alaskans to send a clear **message** to **legislators** in Juneau that a "do-nothing" **strategy** is unacceptable because Alaskaś **future** is at stake," he said in a statement. But state Rep. Bob Miller, D-Fairbanks, introduced an **amendment** that would put certain **restrictions** on oil **companies**. Under Millerś plan, companies would get a **tax** break until January 2017. After that, they would need to have **increased** production by 10 percent of current levels to continue getting a break and increase production by another 2 **percent** each year beyond 2017. "We want to be sure that they are earning the breaks that we are giving," Miller said in a statement. "This amendment says hereś the **benefit**. If you do not make certain metrics for the benefit of Alaska, we withdraw those benefits."*

Text 7

Context: At a government laboratory in Alabama, workers in blue coats unload envelopes packed with small filters that trapped air particles in Hawaii, Alaska and elsewhere. The discs are placed in

Keywords: devices, traces, radioactive, materials, ocean, dangerous, workers, disaster, radiation, environmental, administration, federal, consumers, crisis, water

OURS + GPT2-774M: *At a government laboratory in Alabama, workers in blue coats unload envelopes packed with small filters that trapped air particles in Hawaii, Alaska and elsewhere. The discs are placed in a **water**-filled **device** called a " federal **environmental** control and **radiation** protection system," or E **consumers**, which **traces radioactive material** to the **ocean** floor. The **workers** put the discs in a machine that separates the **dangerous** particles from the rest of the water. The process is called " **crisis** management " and is used to deal with disasters such as the Fukushima nuclear **disaster** in Japan in 2011. The E consumers are also used to deal with the fallout from the Chernobyl nuclear disaster in 1986 in the Ukraine. The **federal** government has spent $5 billion on the system since the **administration** of President George W. Bush signed the U.S. Environmental Protection Agency Reactor Resilience Initiative, a 35-year program that aims to turn the Fukushima Daiichi nuclear plant's reactors into low-carbon and efficient power plants. It has provided $1 billion each year since it was established in the aftermath of the tragedy. If each E consumer is turned on once a year for 40 years, a certain percentage of radioactivity on the Pacific Ocean floor will be removed. This includes particulates in the ocean bottom, as well as radioactive materials such as uranium. It also includes contaminated groundwater.*

HUMAN: *At a government laboratory in Alabama, workers in blue coats unload envelopes packed with small filters that trapped air particles in Hawaii, Alaska and elsewhere. The discs are placed in lead-lined, barrellike **devices** for testing to make sure no **traces** of **radioactive materials** have wafted across the Pacific **Ocean** from Japan. So far, the sea breeze in places such as Honolulu is no more **dangerous** than the pollen-laden air of the Deep South, according to officials. Still, the 60 or so **workers** in the 72,000-square-foot building will be the first to know if the Japanese **disaster** spreads harmful amounts of **radiation** to the U.S. On Wednesday, the **Environmental** Protection Agency and the Food and Drug **Administration** said very low levels of radiation had turned up in a sample of milk from Washington state, but **federal** officials assured **consumers** not to worry. The FDA said such findings were to be expected in the coming days because of the nuclear **crisis** in Japan and that the levels were expected to drop relatively quickly. The EPA said it was increasing the level of nationwide monitoring of milk, precipitation and drinking **water**.*

Text 8

Context: A 25-year-old German man has been arrested for allegedly burying a cache of bombs near a German soccer stadium in a blackmail plot, authorities say. The unnamed German national was

Keywords: german, arrested, explosives, garage, stadium, federal, criminal, investigation, bombs, home, embassy, information, blackmail, criminal, terrorist, organizations, safe

OURS + GPT2-774M: *A 25-year-old German man has been arrested for allegedly burying a cache of bombs near a German soccer stadium in a blackmail plot, authorities say. The unnamed German national was **arrested** in the **German** city of Leipzig on Monday, **federal** prosecutors said. **Investigators** believe the **explosives** were hidden in a **home garage** and **bombs** were hidden in a safe in the garage, according to the **Criminal Criminal** Information Service. The suspect is suspected of blackmailing a German soccer **stadium** security guard and threatening to detonate a **terrorist** attack if he didn't comply with his demands, the prosecutors said. The suspect is also suspected of trying to obtain **information** about organizations and individuals involved in the soccer stadium security. The suspect is suspected of **blackmailing** a German soccer stadium security guard and threatening to detonate a terrorist attack if he didn't comply with his demands, the prosecutors said. The suspect is also suspected of trying to obtain information about organizations and individuals involved in the soccer stadium security. **Embassy** Security Blog discusses the "Security Update".*

HUMAN: *A 25-year-old German man has been arrested for allegedly burying a cache of bombs near a German soccer stadium in a blackmail plot, authorities say. The unnamed **German** national was **arrested** in Cologne on Tuesday after allegedly placing the **explosives** in a parking **garage** near the **Westfalenstadion** in Dortmund, home of the Borussia Dortmund team, the **Federal** Office of **Criminal Investigation** told The Local news agency. The **bombs** were safely defused, and three more were found at the man's **home** in Krefeld, officials said. Investigators said they began tracking the man after he e-mailed the German **Embassy** in Pakistan, offering **information** about two planned attacks in Germany by a group. The warning appeared to be a **blackmail** bid and was worded like an unsolved attempted blackmail case last year. "The suspect apparently acted alone with a general **criminal** motive," a federal spokesman said. "There are absolutely no ties to **terrorist** or Islamist **organizations.**" Authorities say he admitted placing the bombs. Dortmund police spokesman Michael Stein told the BBC: "We expect no security threat at all for the upcoming Bundesliga match on Saturday. Visitors are invited to come to Dortmund. They will be **safe** here."*

Text 9

Context: I'd enjoy reading the citation of constitutional authority for this: "The bill then says if the Senate does not act, then H.R. 1 [the House-passed bill that cuts $61 billion]

Keywords: senate, shutdown, pay, unconstitutional, prevention, legislation, president, sign, law, senate, compromise, problem, effort, attention, deal, option

OURS + GPT2-774M: *I'd enjoy reading the citation of constitutional authority for this: "The bill then says if the Senate does not act, then H.R. 1 [the House-passed bill that cuts $61 billion] will become **law** without **Senate** action. This **legislation** is **unconstitutional** because **Senate President** Pro Tempore (and the **signers** of the Constitution) have the **option** of not passing it. This **problem** does not exist in the House bill. This is the same **effort** that has been going on for years to get the Senate to act on the **shutdown deal**, and it has not worked. The **compromise** that was reached last week was not enough to get the **attention** of the Senate to act on the bill. So now they are **paying** the price for not acting **prevention** of a government shutdown happening, which all sides recognize will happen, eventually.*

HUMAN: *I'd enjoy reading the citation of constitutional authority for this: "The bill then says if the Senate does not act, then H.R. 1 [the House-passed bill that cuts $61 billion] will be the law of the land. In addition to that, it says that if all else fails, and the **Senate** brings about a **shutdown**, then members should not get their **pay**." That's House Majority Leader Eric Cantor describing his "Government Shutdown **Prevention** Act." The problem is, this would be blatantly **unconstitutional**: The Senate needs to pass the same piece of **legislation** the House does, and the president needs to either **sign** it or have his veto overturned. That's how deem-and-pass worked with the health-care **law**, for instance: Both the Senate and the House passed the same pieces of legislation, and then the **president** signed them. But it seems Cantor merely misspoke. I've clarified with both his office and Boehner's office that they believe the **Senate** and the president would still play their traditional roles. That means deem-and-pass isn't, as Cantor suggests, an alternative to actually striking a **compromise**. It's just an **effort** to message the shutdown that'll happen if a law isn't passed. The **problem** for Cantor is that by misdescribing how the gambit would work, he's drawing **attention** to the fact that it can't. At the end of the day, we need an actual **deal** here. There's no other **option**.*

Text 10

Context: We tend to think of Seattle in stereotypical ways: earthy, mellow, panoramic, rainy. But it's not all Patagonia and lattes out there, as the news yesterday about the Justice Department's

Keywords: investigation, police, force, discriminating, yesterday, robbery, suspect, detective, african, native, knife, chief, investigation, justice

OURS + GPT2-774M: *We tend to think of Seattle in stereotypical ways: earthy, mellow, panoramic, rainy. But it's not all Patagonia and lattes out there, as the news yesterday about the Justice Department's **investigation** of **police force** in Seattle shows. The **Justice** Department investigation, which was announced **yesterday**, is focused on the **Chief** of Police, who is African American, and the Seattle Police Department, which is **Native** American. The Seattle Police Department has been under federal **investigation** for a long time. The Justice Department has **suspected** that the Seattle Police Department has **discriminated** against **African** Americans and other minorities. The **detective** in charge of the investigation, who is white, is a member of the department's Professional Standards Bureau. The investigation has been going on for a long time. The **robbery** of a black man in Seattle in the early 1990s was one of the first cases that the Professional Standards Bureau investigated. The detective in charge of the investigation, who is white, was a **knife**-wielding undercover officer. There have been other violent incidents involving white Seattle police officers and Asian men.*

HUMAN: *We tend to think of Seattle in stereotypical ways: earthy, mellow, panoramic, rainy. But it's not all Patagonia and lattes out there, as the news yesterday about the Justice Department's **investigation** of Seattle **police** would indicate. The DOJ is looking into a possible pattern of the SPD using excessive **force** and **discriminating** against minorities, Justice announced here **yesterday**. At issue in the federal investigation are several high-profile incidents involving police violence. Last April, a **detective** was videotaped kicking a Latino **robbery suspect** and stating that he would beat the "Mexican piss" out of the suspect; In June, an officer was videotaped punching a 17-year-old **African** American girl who protested the arrest of a friend for jaywalking; and, In August, police shot a **Native** American woodcarver after he faied to drop his carving **knife**. Seattle Police **Chief** John Diaz told the Seattle Times that he welcomes the DOJ **investigation** and considers it like a "free audit." "We have nothing to hide," he said. "We've been open and transparent with the Department of **Justice**, which makes for a good working relationship."*
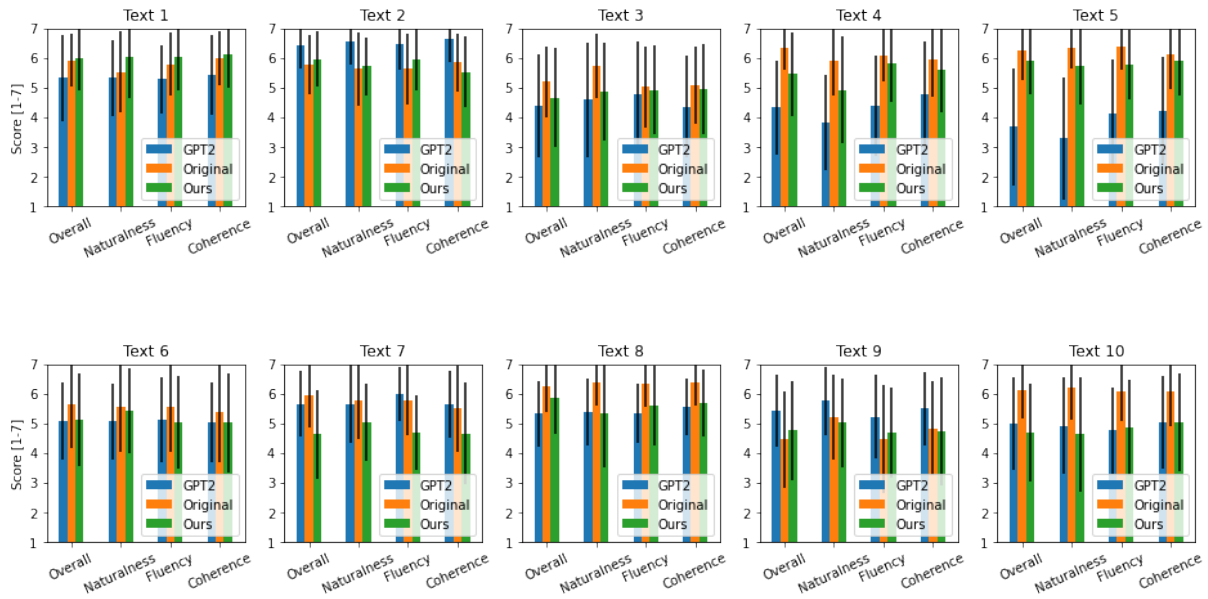
Figure 5: Results of the user study with mean scores and standard deviation error bars per text.

C.3 Detailed Results by News Article

In Figure 5 we show the results of the user study split by news article. As with the aggregated results, we see that per article the average scores for our method are very similar to the scores for uncontrolled GPT-2. However, there are some noteworthy differences. In particular, in the cases of texts 4 and 5, our method scores much higher than uncontrolled GPT-2 on all four scales. On the other hand uncontrolled GPT-2 performs exceptionally well on text 2, but without meeting the keyword requirements of course.

Although on average our method lags behind human-written text in all four categories, it performs almost on par for many of the articles, and even produces text of higher perceived quality than the original for the first article. On the other hand our method struggles to match the scores in all categories for text 10 and texts 7 and 8 to a lesser extent.

There is not much variation across the four categories even in the detailed results.