Controllable Text Generation for All Ages: Evaluating a Plug-and-Play Approach to Age-Adapted Dialogue

Jansen, L.; Laichter, Š.L.; Sinclair, A.; van der Goot, M.J.; Fernández, R.; Pezzelle, S.

DOI
https://aclanthology.org/2022.gem-1.14

Publication date
2022

Document Version
Final published version

Published in
2nd Workshop on Natural Language Generation, Evaluation and Metrics

Controllable Text Generation *for All Ages*: Evaluating a Plug-and-Play Approach to Age-Adapted Dialogue

Lennert Jansen[†], **Štěpán Lars Laichter**[†], **Arabella Sinclair**[‡], **Margot J. van der Goot**[†], **Raquel Fernández**[†], **Sandro Pezzelle**[†]
[†]University of Amsterdam, [‡]University of Aberdeen
{lennertjansen95|lars.laichter}@gmail.com
arabella.sinclair@abdn.ac.uk
{m.j.vandergoot|raquel.fernandez|s.pezzelle}@uva.nl

Abstract

To be trusted and perceived as natural and coherent, conversational systems must adapt to the language of their users. While personalized dialogue is a promising direction, controlling generation for fine-grained language features remains a challenge in this approach. A recent line of research showed the effectiveness of leveraging pre-trained language models toward adapting to a text's topic or sentiment. In this study, we build on these approaches and focus on a higher-level dimension of language variation: speakers' age. We frame the task as a dialogue response generation, and test methods based on bag-of-words (BoW) and neural discriminators (Disc) to condition the output of GPT-2 and DialoGPT without altering the parameters of the language models. We show that Disc models achieve a higher degree of detectable control than BoW models based on automatic evaluation. In contrast, humans can partially detect age differences in BoW but not Disc responses. Since BoW responses are deemed better than Disc ones by humans, simple controllable methods thus appear to be a better tradeoff between adaptation and language quality. Our work confirms the challenges of adapting to higher-level dimensions of language variation. Moreover, it highlights the need to evaluate natural language generation thoroughly.

1 Introduction

Developing dialogue systems that can hold human-like conversations has been a long-standing goal in Artificial Intelligence (AI) research. This includes the ability to mimic speakers' speaking styles and language traits, which is shown to be of crucial importance for systems to be trusted and perceived as natural and coherent (Shum et al., 2018; van der Goot and Pilgrim, 2019).

Current approaches in conversational models typically aim to improve dialogues by leveraging *persona*-specific traits—a speaker's age, gender, geographic location, etc. This is achieved by training systems with either implicit (Kottur et al., 2017; Li et al., 2016) or explicit (Qian et al., 2018; Zhang et al., 2018; Zheng et al., 2019) representations of a speaker. These approaches are generally shown to produce multi-turn conversations that are deemed of better quality by humans, but they pay little attention to understanding what factors determine human judgements. Recently, See et al. (2019) showed that linguistic aspects such as specificity, relatedness, and repetition play an important role, and that explicitly controlling for them during generation increases human engagement in a conversation. However, fine-tuning these large models to control the generation turns out to be a challenging task, which is further limited by the scarcity of annotated conversational datasets.

A recent growing interest in controllable text generation has been fostered by approaches leveraging large pre-trained language models (PLMs; see Sec. 2.2). In particular, one direction is to operate at the decoding stage while leaving the underlying PLM unaltered (Dathathri et al., 2020; Li et al., 2022), which was shown to be successful in generating texts that adapt to a specific topic, length or sentiment. Though these approaches are not typically aimed at modelling conversations, they have been shown to be also suitable to generate controlled responses to dialogue utterances (Madotto et al., 2020).

Building on this line of research, in this study we explore adaptation by large PLMs to a specific, yet unexplored dimension: language variation due to speakers' **age**. The relationship between a person's age and their use of language is a thoroughly studied subject in linguistics and psychology, and various differences between younger and older speakers have been reported at the level of both content and style (see Pennebaker and Stone, 2003). While a few studies showed that speakers' age can be predicted both in discourse and dialogue (see Sec. 2.1),

Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM), pages 172 - 188

no work to date has explored whether, and to what extent, age-related detectable features can be leveraged by controllable text generation models.

In this study, we explore this issue for the first time. Though previous work showed that some degree of adaptation can be achieved to a text's sentiment or topic, we argue that age-related traits are different since they involve subtle, fine-grained features lying at a more abstract level compared to other language dimensions. Therefore, we hypothesize that this task is more challenging and possibly requires complex adaptation strategies.

Following the approach by Madotto et al. (2020), we experiment with dialogue data and frame the controlled generation problem as the task of generating a response to a dialogue utterance. We opt for this setup since it allows us to genuinely investigate language adaptation while leaving aside the extra challenges of modelling full, multi-turn dialogues. Though generally short, single dialogue utterances are shown to contain a fair amount of age-related language signal (Jansen et al., 2021).

We employ the Plug-and-Play Language Model (PPLM) method by Dathathri et al. (2020) and condition the generation of two large PLMs, GPT-2 (Radford et al., 2019) and DialoGPT (Zhang et al., 2020a), by means of various age-specific attribute models. With this approach, generation is steered leaving the underlying PLM unaltered. We test two attribute models based on bag-of-words (BoW) methods or more complex neural discriminators, and perform extensive evaluation of the generated outputs.

Through automatic evaluation, we show that (1) some degree of detectable age adaptation is achieved by all tested models, with (2) discriminator methods outperforming simpler BoW strategies. At the same time, (3) BoW models turn out to produce more fluent and less repetitive responses compared to the more complex models. These results are partially disconfirmed when moving to human evaluation. Indeed, (1) humans can detect age-related differences in the generated language only to a very limited extent, and (2) this is restricted to responses by BoW but not discriminator models. As for the quality of the generated language, (3) outputs by BoW are deemed more fluent and human-like compared to discriminator ones, though this does not systematically correspond to a perceived *better* output. Based on these results, BoW-based controllable strategies appear to be a better tradeoff between adaptation and language quality compared to more complex methods.

Overall, our results confirm the challenges of adapting to higher-level dimensions of language variation, such as those due to speakers' age. Moreover, we highlight the need of complementing automatic analyses with fine-grained human evaluation. Data and code to reproduce our experiments can be found here: `https://github.com/lennertjansen/pplm-age-adapt-dialogue`.

2 Related Work

2.1 Language and Age

A wealth of studies in linguistics and psychology showed that age plays a role in affecting both the content and style of the speaker's language (for further references and discussion, see Pennebaker and Stone, 2003). These findings motivated NLP research aimed at predicting the age of a speaker based on their language. By training a feature-based classifier on a corpus of age-annotated blog posts, Schler et al. (2006) found that speakers' age is best predicted by a combination of content and style features. A similar pattern of results was reported by Nguyen et al. (2011), who extended the investigation to phone conversations and online posts, and by Nguyen et al. (2013), who focused on tweets. Rao et al. (2010) further showed the advantage of including sociolinguistic features when dealing with tweets, with Rosenthal and McKeown (2011) showing that including features of a speaker's online behavior is beneficial when experimenting with blog posts. Recently, Jansen et al. (2021) went beyond feature-based approaches and showed that BERT (Devlin et al., 2019) outperforms other methods when fine-tuned on a dataset of dialogue utterances. Again, both stylistic and lexical cues were reported to be relevant for distinguishing between age groups.

Overall, these studies revealed that the language by younger and older speakers can be detected, among other aspects, by the use of slang and neologisms, pronouns, affect words, capitalizations, alphabetical lengthening, acronyms and verb tenses. Surprisingly, little attention has been paid to model age-related differences in language generation. One exception is represented by research on personalized conversational models, where age is typically considered as one of the speaker-specific traits (Li et al., 2016; Zheng et al., 2019). In these

approaches, however, age adaptation is neither explicitly enforced nor directly measured. We tackle this problem by leveraging recent methods from controllable text generation.

2.2 Controllable Text Generation

Broadly speaking, controllable text generation (CTG) refers to the problem of generating texts that meet certain controllable constraints, which are usually task-specific (for an overview, see Prabhumoye et al., 2020; Zhang et al., 2022). In the context of storytelling, for example, endowing a story with a plot and an ending is a CTG problem, as is the control of topic, sentiment or style in a discourse or dialogue response. The latter line of research, aimed at enforcing attribute-based generation, is particularly relevant to our work. Focusing on discourse data such as reviews or news, various studies demonstrated the effectiveness of RNN language models (Ficler and Goldberg, 2017), VAEs (Hu et al., 2017; Wang et al., 2019; Xu et al., 2020), and GANs (Wang and Wan, 2018) in controlling for attributes such as sentiment, theme, style or, more rarely, age (Lample et al., 2019). As for dialogue, early approaches showed the effectiveness of SEQ2SEQ models in capturing speaking style and background information of specific speakers (Li et al., 2016). However, all these approaches heavily rely on large-scale datasets, which is a challenge for supervised and cross-domain text generation tasks (Zhang et al., 2022).

To alleviate this limitation, approaches that leverage large pre-trained language models (PLMs) such as GPT (Radford et al., 2019), GPT-3 (Brown et al., 2020), T5 (Raffel et al., 2020) or DialoGPT (Zhang et al., 2020a) were recently proposed. Some of them model CTG by fine-tuning the PLM parameters (Lin et al., 2021); others by changing the PLM architecture or training a large conditional model from scratch (Keskar et al., 2019; Zhang et al., 2020b; Wang et al., 2021; He, 2021; Zeng and Nie, 2021). While these methods have generally proven effective in controlling for the desired attribute in a dialogue, discourse, and even image captioning setting, they are often computationally expensive to train and involve fine-tuning or modifying the PLM for each desired attribute. To avoid these issues, a few approaches have proposed to operate at the decoding stage by steering the PLM outputs while leaving its parameters unaltered (Dathathri et al., 2020; Khalifa et al., 2020; Krause et al., 2021; Liu et al., 2021; Yang and Klein, 2021; Li et al., 2022).

One of the most successful and popular methods is the Plug-and-Play Language Model (PPLM; Dathathri et al., 2020). Using a previously trained attribute-based classifier (with 100,000 times fewer parameters than the PLM) to guide text generation by the PLM, this approach was shown to achieve a good degree of CTG for topic and sentiment in a discourse setting while being very inexpensive to train. Motivated by this, Madotto et al. (2020) extended the approach to model dialogue response generation and demonstrated its portability to the conversational domain, where a high degree of control for sentiment and topic was achieved while ensuring fluency. In this work, we build on Madotto et al. (2020) and make a step forward by controlling a more abstract, higher-level dimension compared to sentiment or topic: language variation due to speakers' age.

3 Problem Formulation

In general terms, the problem we tackle is the following: given a dialogue utterance (*prompt*), we generate a dialogue response (*output*).

3.1 Non-Adaptive Setting

In the non-adaptive setting, we tackle the task as a plain text generation problem. Given a prompt, a Transformer-based pre-trained language model $p(\mathbf{x})$ generates an output $\mathbf{x}$ by sampling from the distribution of words that are assigned the highest likelihood of following the prompt. This can be seen as sampling from a conditional distribution, $p(\mathbf{x}|\texttt{prompt})$.

3.2 Age-Adaptive Setting

In the age-adaptive setting, the task is an instance of controllable text generation (CTG). Given a prompt, we seek to generate an output that is controlled for age, i.e., that resembles a response by a *younger/older* speaker. This can be seen as a subproblem of vanilla text generation: the conditioning factor for the generated text is further constrained to also include some predefined attribute, a (in our case, age). CTG is then analogous to sampling from the conditional distribution, $p(\mathbf{x}|\texttt{prompt}, a)$.

PPLM To control generation, we use the PPLM method (Dathathri et al., 2020). PPLM builds on a text classifier or attribute model, $p(a|x)$, that represents the degree of adherence of text x to a certain attribute a, e.g., age. Since the attribute model,

$p(a|x)$, is used to control the generation by a pre-trained Transformer-based language model, $p(x)$, PPLM can be seen as modeling the conditional distribution of generated text x given a, i.e., $p(x|a)$.

In simple terms, the attribute model perturbs the activation space of the underlying language model by making it more likely to generate text that aligns with the predefined attribute. This is achieved by leaving the parameters of the underlying language model unaltered. More formally, PPLM perturbs the generated output one token at a time in the direction of the sum of two gradients: (1) by maximizing the loglikelihood of a under the conditional attribute model $p(a|x)$ (to enforce control); (2) by ensuring high loglikelihood of the generated text under the unaltered language model $p(x)$ (to enforce fluency). The gradient updates only affect the activation space, i.e., the original model parameters are preserved. Sampling is done by following gradients in the latent representation space by approximately implementing the Metropolis-adjusted Langevin sampler (Roberts and Tweedie, 1996) deployed in Nguyen et al. (2017).[1]

4 Method

In our experimental pipeline, we condition the generation of two large pre-trained language models by means of two age-specific attribute models. In particular, we generate responses to a number of dialogue utterances used to prompt text generation. We then evaluate the extent to which the generated responses contain age-related features that can be detected by automatic metrics.

4.1 Data

To train/initialize our attribute models, we use the data introduced by Jansen et al. (2021). This data comes from the spoken partition of the British National Corpus (BNC; Love et al., 2017) and includes dialogue utterances by users from either of two age groups: a *younger* group (age: 19-29) and an *older* group (age: 50 or more). In total, the data consists of 172,303 utterances, i.e., 138,662 *younger* utterances and 33,641 *older* ones. In addition to the full dataset, Jansen et al. (2021) also use a partition of it which is balanced per age group and includes 67,282 total utterances. This is the split of the data they employ to train their *younger/older* classifiers (Sec. 4.5). As described in Sec. 4.3, we use both the full and balanced version of the data.

[1] See Dathathri et al. (2020) for further details.

younger-specific words	**older-specific words**
um, sh*t, cool, f*cking, friends, literally, weekend, amazing, friend, ha, huh, hate, fun, blah, uni, massive, Friday, parents, mate, hell, annoying, wait, ridiculous, crazy, horrible	may, mother, perhaps, huge, business, although, certainly, email, along, often, possibly, wonderful, dear, supposed, otherwise, asked, gosh, bits, almost, particularly, decided, finished, across, near, flat

Table 1: Some of the younger- (age 19-29) and older-specific (age 50+) words used by the BoW method.

4.2 Pre-Trained Language Models

We experiment with two large pre-trained language models: GPT-2 (Radford et al., 2019) and DialoGPT (Zhang et al., 2020a). We generate responses using these two models both in a non-adaptive (Sec. 3.1) and age-adaptive setting (Sec. 3.2). In the age-adaptive setting, the models are conditioned by an attribute model. Similarly to Dathathri et al. (2020), we experiment with two attribute models.

4.3 Age-Controlled Language Models

Below, we describe the attribute models used in our study. For both models, we experiment with the same hyperparameters, reported in Appendix A.

BoW-based attribute model This method relies on lists of words that are representative of each age group's language. We automatically extract them from the full version of the dataset via a frequency-based approach. In particular, for each age group, we (i) order all unique words by frequency; (ii) keep the most frequent words—the ones that make up for at least 85% of the cumulative occurrences; (iii) remove words that are in both age groups; (iv) keep, for each group, only the words that account for at least 85% of the respective cumulative occurrences.[2] Our final lists include 56 younger- and 92 older-specific words. A few examples can be found in Table 1. The BoW-based attribute model gives the log of the sum of likelihoods of each word in the list. Given a bag-of-word $\{w_1, ..., w_k\}$ that represents a given age group a, and the output distribution of the language model p_{t+1}, the attribute model's log-likelihood is:

$$\log p(a|x) = \log \left(\sum_{i}^{k} p_{t+1}[w_i] \right) \quad (1)$$

[2] The 85-th percentile cutoff points are used to yield wordlists of similar lengths as those by Dathathri et al. (2020).

Ascending $\nabla \log p(a|x)$ increases the likelihood of generating words that are either in the BoW or not in the BoW, but semantically related.

Neural discriminator attribute model We randomly split the balanced version of the dataset into a training (90%) and test (10%) set and train a neural classifier to distinguish between dialogue utterances from the two age groups. The classifier receives the representation of the sentence from the last layer of a frozen pre-trained language model and performs the binary task via a single linear layer. The size of both the input and linear layer is equal to the size of the LM's output layer. The discriminator is trained using Adam (Kingma and Ba, 2015) with a learning rate of $1 \cdot 10^{-4}$ and default values for all other parameters from PyTorch's implementation of Adam, with a maximum sequence length of 512 tokens, for 20 epochs, and a batch size of 64. The discriminator parameters that are used in the age-adaptive setting come from the epoch with the highest test accuracy (67.4% accuracy for GPT-2, 67.6% for DialoGPT).

4.4 Prompts

In both the non-adaptive and age-adaptive settings, we prompt the models with handcrafted dialogue utterances. This allows us to devise dialogue utterances that are neither younger- nor older-sounding,[3] so as to genuinely explore age adaptation of the tested methods while minimizing bias effects. We experiment with the following 5 prompts: (i) *Good weather we're having*; (ii) *Can we talk?*; (iii) *Hi, how's it going?*; (iv) *Hey*; (v) *Hello, tell me about your latest holiday*.[4] For each prompt, we let models generate 6 outputs of a given token length. Since we experiment with 9 output lengths (6, 12, 24, 30, 36, 42, 48, 54, and 60 tokens), each model is evaluated over a total of 270 dialogue responses, i.e., 5 prompts $\times$ 9 lenghts $\times$ 6 outputs. [5]

4.5 Evaluation

We evaluate model outputs along two dimensions: age adaptation and quality of generated language.

[3]We verify this by feeding the prompts to the best-performing BERT-based younger/older classifier by Jansen et al. (2021). We consider them *neutral* if the classifier assigns a probability of 0.6 or lower to both age groups.

[4]For comparison, we have also experimented with prompts that are classified as either younger- or older-sounding. Results are in Appendix C.

[5]An exhaustive exploration of the effect of various prompts and output lengths is beyond the scope of this study. We leave it for future work.

Age adaptation To quantify age adaptation, we leverage the best-predictive younger/older classifier by Jansen et al. (2021). This model adds a dropout layer and a linear layer on top of BERT embeddings (Devlin et al., 2019), which are fine-tuned on the age classification task. In particular, we use the weights of the best-performing run of their model (achieving 73% accuracy) and report accuracy in predicting the expected age, i.e., the one which the model has been adapted to. Note that we do not use this classifier as an attribute model to condition generation for 2 main reasons: (1) BERT and *GPT models differ on several levels, which would make the implementation technically challenging; (3) using BERT as an attribute model would go against the overall goal of PPLM, which is to use tiny models to condition large models.

Language quality Following standard practice in NLG, we take perplexity of an external LM as a proxy for fluency of the generated language: the lower the perplexity, the higher the fluency of the generated output. Perplexity (ppl) is expressed as:

$$\text{ppl}(\mathbf{x}) = \exp\left\{-\frac{1}{t}\sum_{i}^{t} \ln p_\theta(x_i|x_{<i})\right\} \quad (2)$$

where $\mathbf{x}$ represents a sequence of tokens, t is sequence length, x_i is the i-th token, and θ denotes the LM's parameters. Following Dathathri et al. (2020), we obtain perplexity scores by GPT-1 (Radford et al., 2018).

Furthermore, we evaluate the degree of text diversity by considering the normalized number of distinct unigrams (Dist-1), bigrams (Dist-2), and trigrams (Dist-3) in the generated output. The higher the score, the less repetitive the language is.

5 Results

Age adaptation Tables 2a and 2b report the results by the younger-adapted and older-adapted models, respectively. As can be seen, discriminator-based models (Disc) achieve a higher degree of age control as detected by automatic means compared to bag-of-words (BoW). This is particularly the case for the older setting, where both GPT2-based and DialoGPT-based Disc models outperform their BoW counterparts by more than 30 accuracy points, with BoW models being far below chance level. As for the younger setting, BoW models perform comparably better by slightly underperforming (DialoGPT) or outperforming (GPT2) their Disc coun-

Model	ppl ↓ better	Dist-1 ↑ better	Dist-2 ↑ better	Dist-3 ↑ better	Acc. ↑ better
GPT-2 (G)	**27.50** (6.58)	0.87 (0.09)	**0.94** (0.04)	**0.90** (0.06)	-
G-BoW	27.91 (7.18)	0.87 (0.10)	0.93 (0.05)	**0.90** (0.06)	70.4%
G-Discrim	32.09 (18.98)	0.77 (0.20)	0.86 (0.13)	0.84 (0.15)	67.8%
DialoGPT (D)	37.52 (12.06)	0.86 (0.13)	0.90 (0.08)	0.85 (0.10)	-
D-BoW	38.53 (12.64)	0.87 (0.12)	0.90 (0.08)	0.86 (0.10)	83.0%
D-Discrim	42.01 (16.94)	**0.90** (0.12)	0.86 (0.14)	0.77 (0.22)	**85.9%**

(a) Younger-adapted models

Model	ppl ↓ better	Dist-1 ↑ better	Dist-2 ↑ better	Dist-3 ↑ better	Acc. ↑ better
GPT-2 (G)	**27.50** (6.58)	**0.87** (0.09)	**0.94** (0.04)	**0.90** (0.06)	-
G-BoW	27.58 (7.07)	0.86 (0.10)	0.93 (0.04)	**0.90** (0.06)	43.0%
G-Discrim	47.15 (47.56)	0.73 (0.24)	0.75 (0.28)	0.75 (0.27)	**74.3%**
DialoGPT (D)	37.52 (12.06)	0.86 (0.13)	0.90 (0.08)	0.85 (0.10)	-
D-BoW	37.85 (11.17)	**0.87 (0.12)**	0.90 (0.08)	0.86 (0.09)	21.5%
D-Discrim	41.17 (20.72)	**0.87 (0.12)**	0.89 (0.13)	0.83 (0.16)	56.7%

(b) Older-adapted models

Table 2: Results of age-controlled dialogue generation. Format: *average metric (standard error)*. **ppl** is perplexity wrt GPT-1. **Dist-**n (for $n = 1, 2, 3$) is the number of distinct n-grams normalized by text length. **Acc.** stands for accuracy of the younger/older classifier. Values in **bold** are the best in the column; the second-best are underlined.

terparts. Overall, these results show that Disc models are more effective than simple BoW ones to control for age-related language features, with this advantage being particularly evident in the older setting.

Striking differences in performance can be observed between GPT2- and DialoGPT-based models. While the latter clearly outperform the former in the younger setting (+13-18 acc. points), an opposite pattern is observed in the older setting, with GPT2-based models gaining 18-22 points over their DialoGPT-based counterparts. This divergent pattern is interesting, and could be due to a younger-language bias of DialoGPT (fine-tuned on Reddit threads, where the majority of users are in the age range 20-29), which would limit adaptation toward the older group.[6] On average, BoW models are more effective in GPT-2 than DialoGPT (56.7 vs 52.3), while Disc results are on par (71.1 vs 71.3).

Taken together, these results show that the PPLM approach is effective in controlling for age-related language features that can be detected by a trained classifier, and that adaptation is stronger when using a neural discriminator attribute model.

Language quality Moving to measures of language quality, we observe that the base GPT-2 is either the best or the second-best with respect to both perplexity and number of distinct n-grams. As for age-adapted models, BoW ones are generally shown to outperform Disc models. GPT2-based BoW models, in particular, appear to be the best overall age-adapted models: they compare to GPT-2 in terms of both fluency and diversity, which confirms that conditioning generation through wordlists does not negatively impact on the quality of the generated language. In contrast, Disc models perform comparably worse on these metrics, which suggests a much bigger impact.

Since automatic metrics are known to have their own limitations (see, e.g., the case of perplexity in capturing language fluency; Mir et al., 2019), in the next section we complement the results by the automatic metrics via extensive human analysis.

[6] This is supported by the mean probabilities assigned by the BERT-based classifier to the younger class on DialoGPT and GPT-2 outputs: 0.76 for DialoGPT, 0.62 for GPT-2.

6 Analysis

We run 3 crowdsourcing studies with human participants aimed at (1) exploring whether age differences detected by a classifier correspond to human intuitions on age-related language features; (2) assessing the quality of each model's generated language; (3) testing which age-adapted model produces the best outputs. Based on their overall better age control and language quality, we choose to focus on GPT2-based models: GPT2 (hence, *base*), BoW-younger, BoW-older, Disc-younger, and Disc-older. Data collection is performed on Appen (appen.com). Participants are paid 0.08$ per judgement (which corresponds to around 10$/hour considering a conservative rate of 2 judgements/minute). In total, the full data collection costed around 2.5K$. The instructions given to the participants in the three studies we describe below are available in Appendix D. Participants were restricted to be from English-speaking countries and we used test-questions for quality control: only participants who correctly answered at least 70% of test-questions were considered trustworthy.

6.1 Are Age-Related Differences Detectable?

We aim at testing whether the age-adapted outputs by a model (e.g., BoW-younger) are perceived as sounding more like their target age group (*younger*) than those by both its counterpart (BoW-older) and the base model. Therefore, we set up three comparisons of outputs by BoW and Disc models, respectively: younger vs older, younger vs base, and older vs base. In particular, we experiment with 300 outputs per model,which sums up to 900 unique outputs within BoW and 900 within Disc models. Outputs from various models are paired based on the same prompt and if they have similar length.[7] We ask 5 participants to judge which of the two outputs in a given comparison pair sounds younger/older than the other. In total, we collect 9K judgements by 467 different participants. Table 3 shows some examples.

Results We consider the assessment for a pair as correct if at least 3 out of 5 participants converge on the target age group; otherwise, we deem it wrong. In Figure 1, we report the results of this analysis. As can be seen, human 'accuracy' lags well behind the accuracy by the classifier in 3 models out of 4. This is not the case only for BoW-older, where

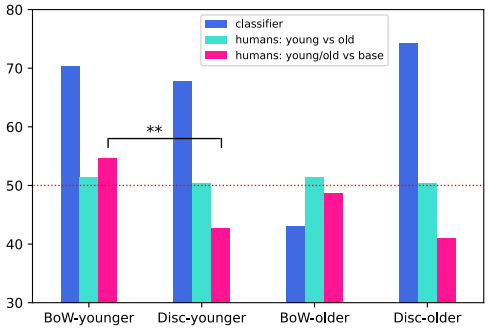


Figure 1: Accuracy by humans and the classifier in detecting age-adapted outputs. The dotted red line indicates chance level; ** stands for statistical significance at $p < 0.01$. Best viewed in color.

the classifier's accuracy is below chance level. One striking observation is that, overall, the degree of age control detected by humans is very limited, and indeed never significantly outperforming chance level (50%) for $p < 0.05$.[8] This suggests that what makes an output sound as younger or older for a text classifier is not something that is clearly detectable by humans. This could be due to the different type of language features that human speakers and the models leverage when making this assessment. For example, a classifier could exploit regularities on topics or domains that are present in the training data, while human participants solely rely on information from their language competence.[9]

At the same time, some age-related features appear to be present in the outputs by BoW-younger, where human accuracy in detecting younger from base outputs reaches 55% – though this comparison is not significantly different from chance according to conventional statistical criteria.[10] Moreover, we find that the difference between BoW-younger (M = 0.55, SD = 0.5) and Disc-younger (M = 0.43, SD = 0.5) is statistically significant via an unpaired t-test, t(587) = 2.9, p = .003. In contrast, no adaptation at all is detected by humans in the outputs by Disc models.

These results indicate that the adaptation to age brought by a PPLM approach can be detected by human speakers only to a limited extent. At the same time, BoW models are generally better than Disc ones, and this difference is statistically significant

[7] See Appendix E for more details on data preprocessing.

[8] We test this by means of a one-sample t-test.

[9] What are the cues that guide this assessment is in itself an interesting question, which deserves further investigation.

[10] One-sample t-test between BoW-younger (M = 0.55, SD = 0.5) and chance (M = 0.5), t(288) = 1.6, one-tailed p = .056

model	age group	output
BoW	younger	*We have the best weather in the world. I*
BoW	older	*The weather is good and I think you're all going to love it. I'm happy to announce that I have a new home*
Disc	younger	*This is great. It gives us more fun than ever before, and we can enjoy a great coffee. Happy birthday guys. ...*
Disc	older	*The sun was setting when we were getting up with a huge rain and we got stuck in on one one of the three*

Table 3: One example generated for the prompt *'Good weather we're having.'* by each age-adapted model for which at least 4 / 5 participants agreed the response sounds like language from the target age group, both against the other age model and the base model. Some outputs are truncated due to the fixed-length criterion used.

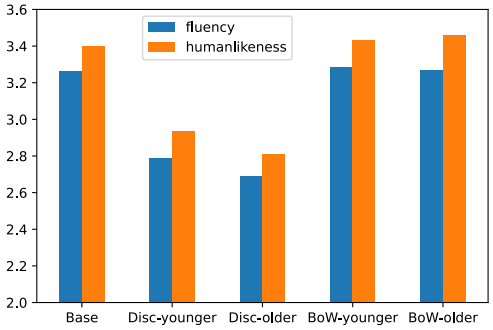


Figure 2: Human judgements on a scale from 1 to 5 on fluency and human-likeness of outputs by the base and the age-adapted models. Best viewed in color.

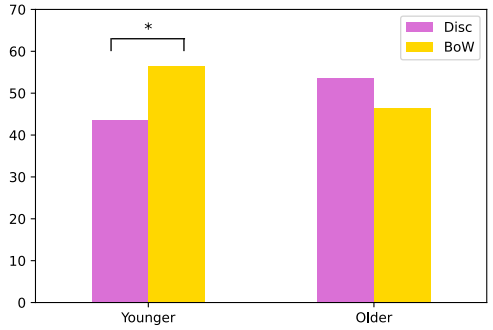


Figure 3: Percentage of human participants who judged an output by a model to be better than an output by another model. Best viewed in color.

for the younger age group. Here, people appear to be better than chance in detecting age-related differences, though this is only a trend without statistical significance (possibly due to sample size).

6.2 Is the Generated Language Good?

We test whether, and to what extent, the language generated by the 4 age-adapted models and the base model, that we use as a control, is deemed good by humans. We consider all the outputs generated by the 5 models, i.e., 1.5K outputs in total. We then ask 5 participants to judge, on a 5-point scale, the degree of *fluency* and *human-likeness* of the output. We define fluent language as having few repetitions and a good flow; human-like as being likely to be produced by a human speaker. We collect a total of 15K judgements, i.e., 7.5K per property, by 278 unique participants.

Results We compute the average score obtained by an output for a property, and then average over all the samples. Results are reported in Figure 2. As can be seen, while BoW models are on par with the base model with respect to both properties, Disc models are assigned much lower values. That is, the outputs by Disc models are deemed much less fluent and less human-like than those by BoW models and the base model. One interesting observation is that judgements of human-likeness are systematically higher than fluency, and this holds for all models. This reveals that the two properties capture different and possibly complementary aspects (a text can be human-like though not perfectly fluent), which highlights the need of using multiple metrics to assess the quality of NLG systems.

Taken together, these results suggest that the perturbations operated by Disc models on the decoding of the LM are heavier than those by BoW models, and that Disc models achieve a level of control that is more detectable by automatic classifiers at the cost of being less fluent and less human-like. It is worth noting that this pattern closely mirrors the perplexity and Dist-n scores reported in Tables 2, where Disc models are shown to be systematically behind GPT2 and BoW-based models.

6.3 Which Models Produce Better Outputs?

We focus on the two younger-adapted and older-adapted models and test which of the two produces an overall better output according to humans. We pair each output by a model with an output by the other model for the same target age group: i.e., BoW-younger vs Disc-younger and BoW-older vs

Disc-older. We end up with 300 pairs per age group, i.e., 600 pairs in total. We then ask 5 participants to judge which of the two outputs in the pair is *overall better*. In total, we collect 3K judgements by 230 distinct participants.

Results For each pair, we take the output with the majority of votes (3 or more). We then compute the proportion of cases in the data where the output by BoW/Disc was chosen. Results are reported in Figure 3. For the younger group, BoW-based outputs (M = 0.56, SD = 0.5) are deemed better than those by Disc (M = 0.44, SD = 0.5), and this difference is statistically significant as per a paired t-test, t(293) = 2.2, p = .026. Surprisingly given the results of the previous analysis (where BoW neatly outperforms Disc in terms of fluency and human-likeness), an opposite pattern is observed for the older group, though the difference between BoW and Disc is not significant (p = .222). We hypothesize that this dissociation could be due to the different types of evaluation (rating vs. binary choice), which deserves further investigation.

7 Conclusion

We focused on age-related language variation and tested whether current approaches to controllable text generation can capture it in a dialogue response. We showed that models achieve substantial adaptation based on automatic metrics, while age-related differences can be detected only to a limited extent by humans. At the same time, simple controllable methods based on BoW appear to be a good trade-off between control and quality. From a broader perspective, our case-study on age adaptation reveals that controlling for subtle, fine-grained language features remains an open challenge. Moreover, we show that human evaluation is crucial to assess the degree of achieved control since it provides different insights compared to automatic metrics (Li et al., 2018; Sudhakar et al., 2019).

Limitations

On the need for age adaptation This work starts from the assumption that users of language technologies, such as dialogue systems or chatbots, would appreciate an age-adaptive system, i.e., would perceive age-adaptation as positive. This is motivated by evidence from psychology and sociolinguistics showing that age-driven linguistic variation is typically in play. Nevertheless, this assumption remains to be validated by means of user studies.

On the impact of prompts While we experiment with both age-neutral and age-adapted prompts, texts are generated based on a limited number of prompts. Further attention should be paid to investigating the impact of prompts (and prompt features) on the resulting outputs.

On the experimental setting While we formulate the problem as dialogue response generation, dialogue features are not exploited. A simple step in this direction could be to experiment with other ways of prompting the model, e.g., by providing a signal of which dialogue participant is speaking (A, B) and whether there is a turn transition between prompt and generated response.

On the use of other CTG methods We experiment with only one CTG method, namely PPLM. In future work, we plan to address this limitation by extending our investigation to other approaches.

Ethics Statement

Broader impact As for most technologies, ours can have a positive impact on society, e.g., by promoting the development of more inclusive systems that speak the language of their users, independently of their age; or, by informing work on debiasing language models, that appear to be biased toward the language of younger groups of users. On the other hand, we acknowledge and are aware of possible harmful or undesirable uses of this technology, e.g., toward amplifying biases or explicitly/implicitly discriminating people based on their age (Rosales and Fernández-Ardèvol, 2020; Stypińska, 2021; Noble, 2018). We advocate for a responsible, rigorous use of the methodology and materials described in this study.

Privacy and discrimination Age is personal data or privately identifying information according to the EU or US defintion, respectively. As such, it is a protected class in US and various other anti-discrimination regulations. In the present sttudy, we experiment with anonymous textual data aggregated at the level of two macro age classes: younger vs older speakers. For a given utterance, we only consider the age range of the speaker who uttered it, 19-29 vs 50+. No info regarding speaker identity (ID, previous dialogues, etc.) or their demographics

(gender, location, social status, etc.) is ever considered. We argue this is a valuable way to limit as much as possible any privacy and discrimination risks. We thank the anonymous ethics reviewer for providing valuable input on this and for pointing to the studies cited in the paragraph above.

Human evaluation We ensure human participants taking part in our human evaluation are paid properly according to the standards of our institution's country/countries. To avoid any harm, we carefully remove any offensive or inappropriate language from the samples. Participants were given the opportunity to report any problem when participating in the evaluation. No issues were reported.

Pretrained language models There are serious risks associated with the development and use of large PLMs (Bender et al., 2021), which we leverage in this research. Such risks include the environmental impact of the computational resources required for training and the encoding and possible amplification of biases present in the massive amounts of un-curated data the models learn from. The PPLM approach we explore in the present work provides an alternative to re-training or finetuning the model and in this sense it does not incur further environmental cost. Nevertheless, the approach does rely on a large PLM, with all the lack of transparency regarding the pre-training data that this involves. Our attribute models are trained on the BNC, a carefully curated and documented dataset. Yet, we acknowledge that the lack of control over the PLM pre-training data is likely to occasionally lead to undesirable outputs.

Acknowledgements

We would like to thank the four anonymous GEM reviewers for their valuable feedback and the participants of our crowdsourcing experiments. The work received funding from the University of Amsterdam's Research Priority Area Human(e) AI and from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 819455).

References

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

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. Curran Associates, Inc.

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 *International Conference on Learning Representations*.

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.

Jessica Ficler and Yoav Goldberg. 2017. Controlling linguistic style aspects in neural language generation. In *Proceedings of the Workshop on Stylistic Variation*, pages 94–104, Copenhagen, Denmark. Association for Computational Linguistics.

Xingwei He. 2021. Parallel refinements for lexically constrained text generation with bart. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8653–8666.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *International conference on machine learning*, pages 1587–1596. PMLR.

Lennert Jansen, Arabella Sinclair, Margot J van der Goot, Raquel Fernández, and Sandro Pezzelle. 2021. Detecting age-related linguistic patterns in dialogue: Toward adaptive conversational systems. In *CLiC-it*.

Nitish Shirish Keskar, Bryan McCann, Lav Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL - A Conditional Transformer Language Model for Controllable Generation. *arXiv preprint arXiv:1909.05858*.

Muhammad Khalifa, Hady Elsahar, and Marc Dymetman. 2020. A distributional approach to controlled text generation. In *International Conference on Learning Representations*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

Satwik Kottur, Xiaoyu Wang, and Vitor Carvalho. 2017. Exploring personalized neural conversational models. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3728–3734.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2021. Gedi: Generative discriminator guided sequence generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4929–4952.

Guillaume Lample, Sandeep Subramanian, Eric Michael Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-attribute text rewriting. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, Berlin, Germany. Association for Computational Linguistics.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. 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 1865–1874, New Orleans, Louisiana. Association for Computational Linguistics.

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B Hashimoto. 2022. Diffusion-LM improves controllable text generation. *arXiv preprint arXiv:2205.14217*.

Zhaojiang Lin, Andrea Madotto, Yejin Bang, and Pascale Fung. 2021. The adapter-bot: All-in-one controllable conversational model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 16081–16083.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A Smith, and Yejin Choi. 2021. Dexperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6691–6706.

R Love, C Dembry, A Hardie, V Brezina, and T McEnery. 2017. The spoken bnc2014: designing and building a spoken corpus of everyday conversations. *In International Journal of Corpus Linguistics*, 22(3):319–344.

Andrea Madotto, Etsuko Ishii, Zhaojiang Lin, Sumanth Dathathri, and Pascale Fung. 2020. Plug-and-play conversational models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2422–2433, Online. Association for Computational Linguistics.

Remi Mir, Bjarke Felbo, Nick Obradovich, and Iyad Rahwan. 2019. Evaluating style transfer for text. 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 495–504.

Anh Nguyen, Jeff Clune, Yoshua Bengio, Alexey Dosovitskiy, and Jason Yosinski. 2017. Plug & play generative networks: Conditional iterative generation of images in latent space. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4467–4477.

Dong Nguyen, Rilana Gravel, Dolf Trieschnigg, and Theo Meder. 2013. "how old do you think i am?" a study of language and age in twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7.

Dong Nguyen, Noah A. Smith, and Carolyn P. Rosé. 2011. Author age prediction from text using linear regression. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 115–123, Portland, OR, USA. Association for Computational Linguistics.

Safiya Umoja Noble. 2018. Algorithms of oppression. In *Algorithms of Oppression*. New York University Press.

James W Pennebaker and Lori D Stone. 2003. Words of wisdom: Language use over the life span. *Journal of Personality and Social Psychology*, 85(2):291–301.

Shrimai Prabhumoye, Alan W Black, and Ruslan Salakhutdinov. 2020. Exploring controllable text generation techniques. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1–14, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Qiao Qian, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Assigning personality/profile to a chatting machine for coherent conversation generation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial*

Intelligence, IJCAI-18, pages 4279–4285. International Joint Conferences on Artificial Intelligence Organization.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

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*, 21:1–67.

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 37–44.

Gareth O Roberts and Richard L Tweedie. 1996. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, pages 341–363.

Andrea Rosales and Mireia Fernández-Ardèvol. 2020. Ageism in the era of digital platforms. *Convergence*, 26(5-6):1074–1087.

Sara Rosenthal and Kathleen McKeown. 2011. Age prediction in blogs: A study of style, content, and online behavior in pre-and post-social media generations. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 763–772.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 6, pages 199–205.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. 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 1702–1723.

Heung-Yeung Shum, Xiao-dong He, and Di Li. 2018. From Eliza to XiaoIce: Challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, 19(1):10–26.

Justyna Stypińska. 2021. Ageism in AI: new forms of age discrimination in the era of algorithms and artificial intelligence. In *CAIP 2021: Proceedings of the 1st International Conference on AI for People: Towards Sustainable AI, CAIP 2021, 20-24 November 2021, Bologna, Italy*, page 39. European Alliance for Innovation.

Akhilesh Sudhakar, Bhargav Upadhyay, and Arjun Maheswaran. 2019. "transforming" delete, retrieve, generate approach for controlled text style transfer. 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 3269–3279.

Margot J van der Goot and Tyler Pilgrim. 2019. Exploring age differences in motivations for and acceptance of chatbot communication in a customer service context. In *International Workshop on Chatbot Research and Design*, pages 173–186. Springer.

Ke Wang and Xiaojun Wan. 2018. SentiGAN: Generating sentimental texts via mixture adversarial networks. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4446–4452.

Wenlin Wang, Zhe Gan, Hongteng Xu, Ruiyi Zhang, Guoyin Wang, Dinghan Shen, Changyou Chen, and Lawrence Carin. 2019. Topic-guided variational auto-encoder for text generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 166–177.

Yufei Wang, Ian Wood, Stephen Wan, Mark Dras, and Mark Johnson. 2021. Mention flags (MF): Constraining transformer-based text generators. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 103–113, Online. Association for Computational Linguistics.

Peng Xu, Jackie Chi Kit Cheung, and Yanshuai Cao. 2020. On variational learning of controllable representations for text without supervision. In *International Conference on Machine Learning*, pages 10534–10543. PMLR.

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.

Yan Zeng and Jian-Yun Nie. 2021. A simple and efficient multi-task learning approach for conditioned dialogue generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4927–4939.

Han Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2022. A survey of controllable text generation using transformer-based pre-trained language models. *ArXiv*, abs/2201.05337.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020a. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Yizhe Zhang, Guoyin Wang, Chunyuan Li, Zhe Gan, Chris Brockett, and William B Dolan. 2020b. POINTER: Constrained progressive text generation via insertion-based generative pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8649–8670.

Yinhe Zheng, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. 2019. Personalized dialogue generation with diversified traits. *arXiv preprint arXiv:1901.09672*.

Appendices

A PPLM Hyperparameters

Table 4 reports the hyperparameters used for both the BoW and Disc models. Please refer to Dathathri et al. (2020) for further details on the hyperparameters. For comparison, the set of hyperparameters used in the paper by Dathathri et al. (2020) is given in their Table S18.

B Examples of Generated Outputs

In Table 5, we report a few more outputs generated by GPT2-based models.

C Results with Younger / Older Prompts

For comparison, we also experiment with prompts that are automatically classified as either younger- or older-sounding. Younger-sounding prompts: *What are your plans this week?; What do you wanna eat?; Do you have any hobbies?; Can I add you on Facebook?; When did you go? Awesome! I actually haven't been there.* Older-sounding prompts: *Tell me about your family.; Good afternoon.; I had a splendid weekend.; Hello, how are you?; Hello, tell me about yourself.* the results of the models using these prompts are reported in Tables 6 and 7.

D Instructions to Participants

The instructions given to the participants in the human evaluation studies can be found in Figures 4, 5 and 6.

E Preprocessing for Human Evaluation

The generated samples were checked for the presence of inappropriate language using a list of over 1300 English words that can potentially be offensive, available at `https://www.cs.cmu.edu/~biglou/resources/bad-words.txt`. The samples that contained terms in this list were manually scanned and those that were deemed actual instances of offensive language use were discarded. Altogether, 39 offensive samples were removed (BoW-younder: 3; BoW-older: 1; Disc-younger: 35). To counter the fact that some adapted models had fewer samples, some of the samples were reused with different pairings to arrive at 300 samples per model.

Outputs from different models being compared in the human evaluation always involve the same

method type	attribute	hyperparameters
BoW / Disc	younger / older	m = 3, $\lambda_{\text{kl}} = 0.01$, $\alpha = 0.02$, $\gamma = 1.5$, $\gamma_{\text{gm}} = 0.9$, r = 10, $\tau = 1.0$

Table 4: Hyperparameters used in our experiments. Please refer to Dathathri et al. (2020) for further details. An important hyperparameter is m, i.e., the number of perturbation steps/iterations.

model	age	prompt	output
BoW	Y	Can we talk?	*Yes, I can. And I hope you will join me in this discussion.*
BoW	O	Can we talk?	*What about the world? What is reality? Can we make it? Are we really all*
BoW	Y	Hey.	*My name is Alex. I am a programmer by trade. I've spent the past 3 years working,...*
BoW	O	Hey.	*I'm just getting my new laptop and I noticed that my wallpaper was getting a weird,...*
Disc	Y	Hey.	*I I I love my parents*
Disc	O	Hi, how is it going?	*The summer of 2015 has begun and it seems that I'm finally going home to Spain,...*
Disc	Y	Hello, tell me about your latest holiday.	*Do you wish you were you will be happy with it?You can use this technique,...*
Disc	O	Hey.	*ITheItIfThatYouWeAndItThatItIfItYouItItThereItItInThisItIfYouItYouThereIfThatYouItItIt,...*

Table 5: A few more examples of outputs generated by GPT2-based models. **Y** stands for younger, **O** for older.

Model	ppl. ↓ better	Dist-1 ↑ better	Dist-2 ↑ better	Dist-3 ↑ better	$\bar{P}_Y$ ↑ better	Acc. ↑ better
GPT2 (G)	**28.05** (±6.12)	0.85 (±0.13)	0.91 (±0.08)	0.88 (±0.08)	0.80 (±0.33)	-
G-BoW	28.81 (±7.09)	0.86 (±0.12)	**0.92** (±0.08)	**0.89** (±0.08)	0.82 (±0.32)	83.3%
G-Disc	39.32 (±37.49)	0.84 (±0.21)	0.61 (±0.40)	0.57 (±0.40)	0.70 (±0.40)	70.7%
DialoGPT (D)	36.69 (±9.11)	0.87 (±0.10)	0.91 (±0.06)	0.87 (±0.08)	**0.90** (±0.24)	-
D-BoW	37.35 (±8.60)	0.88 (±0.10)	0.91 (±0.06)	0.87 (±0.08)	0.90 (±0.26)	**90.0%**
D-Disc	39.22 (±14.96)	**0.89** (±0.12)	0.86 (±0.19)	0.79 (±0.23)	0.89 (±0.25)	91.1%

Table 6: Results of age-controlled dialogue generation: **younger**-targeted models, conditioned on **younger prompts**. Format: *average metric (standard error)*. **ppl.** is perplexity w.r.t. GPT-1. **Dist**-n (for $n = 1, 2, 3$) is the number of distinct n-grams normalized by text length, as a measure of diversity. $\bar{P}_Y$ is the sample's average probability to contain features learned to be younger by BERT-based classifier. **Acc.** is BERT-based classifier's accuracy when classifying the row's samples. Values in **bold** are the best in the column.

Model	ppl. ↓ better	Dist-1 ↑ better	Dist-2 ↑ better	Dist-3 ↑ better	$\bar{P}_O$ ↑ better	Acc. ↑ better
GPT2 (G)	29.34 (±10.30)	**0.86** (±0.09)	**0.94** (±0.04)	**0.90** (±0.06)	0.40 (±0.43)	-
G-BoW	**28.81** (±10.10)	**0.86** (±0.10)	0.93 (±0.05)	**0.90** (±0.06)	0.41 (±0.43)	41.1%
G-Disc	95.21 (±174.42)	0.65 (±0.27)	0.78 (±0.18)	0.78 (±0.18)	**0.90** (±0.25)	**90.3%**
DialoGPT (D)	38.18 (±12.03)	**0.86** (±0.12)	0.90 (±0.08)	0.86 (±0.09)	0.28 (±0.38)	-
D-BoW	37.80 (±11.74)	**0.86** (±0.12)	0.90 (±0.07)	0.87 (±0.08)	0.28 (±0.39)	29.3%
D-Disc	40.08 (±16.77)	0.85 (±0.14)	0.88 (±0.10)	0.83 (±0.14)	0.61 (±0.42)	61.1%

Table 7: Results of age-controlled dialogue generation: **older**-targeted models, conditioned on **older prompts**. Format: *average metric (standard error)*. **ppl.** is perplexity w.r.t. GPT-1. **Dist**-n (for $n = 1, 2, 3$) is the number of distinct n-grams normalized by text length, as a measure of diversity. $\bar{P}_Y$ is the sample's average probability to contain features learned to be younger by BERT-based classifier. **Acc.** is BERT-based classifier's accuracy when classifying the row's samples. Values in **bold** are the best in the column.

prompt. In addition, we make sure that the length of the generated outputs being compared is similar. We do this by always picking two relevant samples from the same length class.

Judging Language Produced By An Artificial Intelligence (A)

Overview

This job will require you to evaluate the outputs of different AI models that generate language.

We are interested in understanding how good the language generated by these systems is and in testing whether the generated language sounds like it is written by speakers belonging to **younger** or **older** age groups.

You will be asked to evaluate short pieces of text automatically generated by two AI models and compare them in this respect.

Your evaluation will help us to better understand the performance of different language generation models. In the future, this can help make technologies, such as virtual assistants and chatbots, more accessible to all people regardless of their age.

Instructions

You will see two short pieces of text generated by two AI models and you will be asked to compare them and decide which text is better or which text could have been produced by a **younger** or an **older** speaker.

Note: some of the texts may look unfinished or cut-off and this should **not** be taken into account in your assessment.

You will also be asked about your age, because this can have an effect on your evaluation. You only need to enter your age once per job.

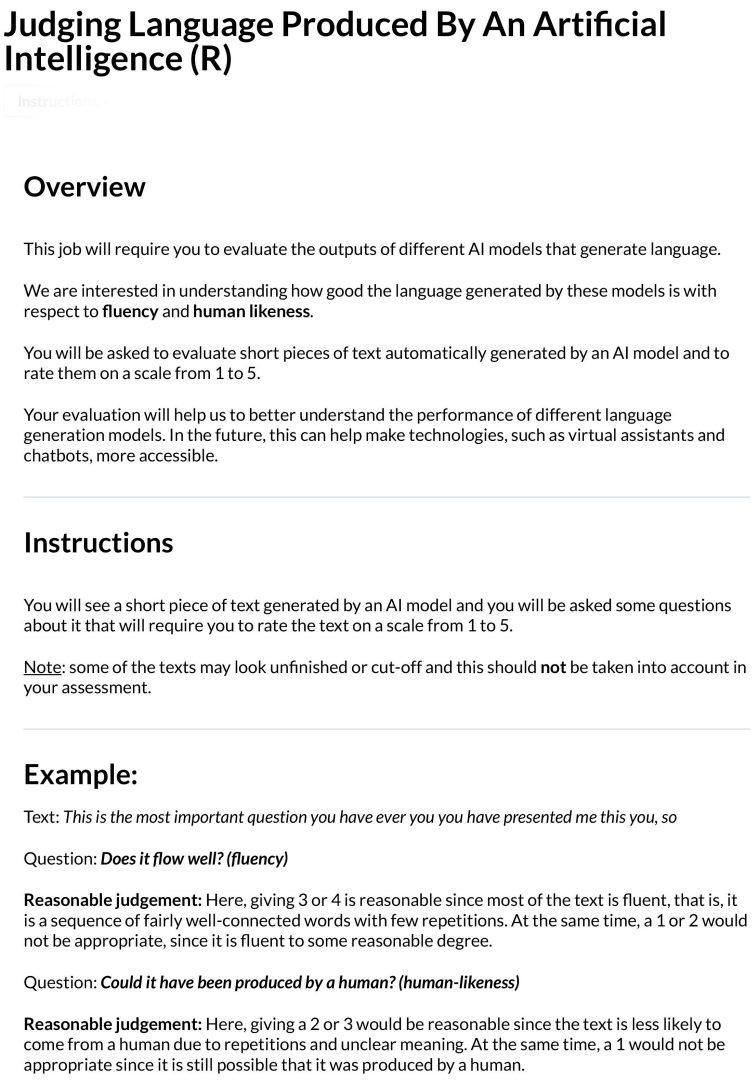
Example 1:

Question: ***Which one of the two outputs sounds like a text that could be written by a younger speaker?***

1) Can you tell what is being done to fix the water repair at the Water Treatment Plants. This is a couple months before a year and the first water damage is not visible to be noticed. The damage to some parts of the water treatment plants. In the winter this is a water plant at
2) A lot of fun, good vibes. I love the music and dancing. It's a great night out. Great atmosphere. Lots of energy.Good vibes. Great vibes! The crowd was great, there was a lot of energy and a lot

Reasonable response: It would be reasonable to select the second text as produced by a younger speaker. The text includes some stylistic traits and expressions, such as "good vibes" or "lot of fun", that are likely to be used by younger speakers.

Figure 4: Participant guidelines for the crowdsourcing study reported in Section 6.1.



Judging Language Produced By An Artificial Intelligence (R)

Instructions

Overview

This job will require you to evaluate the outputs of different AI models that generate language.

We are interested in understanding how good the language generated by these models is with respect to **fluency** and **human likeness**.

You will be asked to evaluate short pieces of text automatically generated by an AI model and to rate them on a scale from 1 to 5.

Your evaluation will help us to better understand the performance of different language generation models. In the future, this can help make technologies, such as virtual assistants and chatbots, more accessible.

Instructions

You will see a short piece of text generated by an AI model and you will be asked some questions about it that will require you to rate the text on a scale from 1 to 5.

Note: some of the texts may look unfinished or cut-off and this should **not** be taken into account in your assessment.

Example:

Text: *This is the most important question you have ever you you have presented me this you, so*

Question: ***Does it flow well? (fluency)***

Reasonable judgement: Here, giving 3 or 4 is reasonable since most of the text is fluent, that is, it is a sequence of fairly well-connected words with few repetitions. At the same time, a 1 or 2 would not be appropriate, since it is fluent to some reasonable degree.

Question: ***Could it have been produced by a human? (human-likeness)***

Reasonable judgement: Here, giving a 2 or 3 would be reasonable since the text is less likely to come from a human due to repetitions and unclear meaning. At the same time, a 1 would not be appropriate since it is still possible that it was produced by a human.

Figure 5: Participant guidelines for the crowdsourcing study reported in Section 6.2.

Judging Language Produced By An Artificial Intelligence (C)

Overview

This job will require you to evaluate the outputs of different AI models that generate language.

We are interested in understanding how good the language generated by these models is.

You will be asked to evaluate the overall goodness of short pieces of text automatically generated by two AI models.

Your evaluation will help us to better understand the performance of different language generation models. In the future, this can help make technologies, such as virtual assistants and chatbots, more accessible.

Instructions

You will see two short piece of text generated by two AI models and you will be asked to compare them.

Note: some of the texts may look unfinished or cut-off and this should **not** be taken into account in your assessment.

Example:

Question: *Which one of the two outputs is better overall?*

1) *Can we talk?CIWe,.w,,,k?,?k.k?WwW??.!?!w(?!!??!?)!WwW*
2) *This holiday, we celebrate the season by giving gifts*

Reasonable response: Although the second option constitutes an incomplete sentence, it would be reasonable to pick the second option as the better output, given that the first option contains a random string of characters.

Figure 6: Participant guidelines for the crowdsourcing study reported in Section 6.3.