UNIVERSITY OF AMSTERDAM

MSc ARTIFICIAL INTELLIGENCE
MASTER THESIS

A Plug-and-Play Approach to Age-Adaptive Dialogue Generation

---

by

LENNERT JANSEN
10488952

December 8, 2021

48EC
January 11, 2021 - December 8, 2021

Supervisors:
Dr SANDRO PEZZELLE
Dr RAQUEL FERNÁNDEZ
Dr ARABELLA SINCLAIR

Assessor:
Dr WILKER AZIZ

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

Acknowledgments

Abstract

Personalized interaction between users and dialogue systems is crucial to obtain systems that are trusted by users and perceived as natural. Because a person's use of language is known to change with age, personalized dialogue systems should be able to identify a user's age profile, and then generate age-adaptive responses, based on the identified signal. Recently, the development of Plug-and-Play language models (PPLM) has lead to computationally relatively inexpensive methods to control the writing style of text outputs generated by large pre-trained Transformer-based language models. However, their application to enforce abstract writing styles (like age-related language) on dialogue responses is virtually unexplored. This thesis therefore aims to study the detection and subsequent generation of age-related linguistic features in dialogue settings. First, I investigate to what extent various purely text-based NLP models can detect age-related linguistic patterns in dialogue. It is found that a fine-tuned BERT-model is able to distinguish between transcribed dialogue utterances of different age groups with reasonable accuracy, while much simpler models based on n-grams are able to do so with comparable performance, which suggests that, in dialogue, "local" features can be indicative of the language of speakers from different age groups. The presumed locality of age-indicative features identified during age-detection motivates the use of both simple bag-of-words (BoW) attribute models, in addition to more sophisticated neural discriminator attribute models for controlled dialogue generation using PPLM. The results show that age-adaptive dialogue generation with PPLM is possible to the extent that my text-based classifiers can reliably detect age-related linguistic patterns. Furthermore, it is observed that discriminator-based PPLM-setups achieve higher levels of detectable control than BoW-based setups, but generate significantly more perplexing and repetitive responses. I also carry out extensive quantitative and qualitative analyses, revealing the effects of prompt-induced biases, and the presence of previously studied age-related linguistic patterns in dialogue responses generated by PPLM-setups. I conclude that both structural and local features should be taken into account when adapting the output of dialogue generation models to certain age groups. Overall, I believe this work is a promising step towards the development of personalized, age-adaptive, dialogue systems.

Contents

List of Figures

List of Tables

List of Footnotes

Chapter 1

Introduction

Endowing dialogue systems with the capability to emulate users' speaking styles in order to generate human-like conversation is a long-standing goal in artificial intelligence (AI) research. Despite impressive recent progress in the development of dialogue systems [McTear, 2020], the development of systems that can adapt their language to that of a particular individual or group of users continues to pose more of a challenge. Recent examples of this line of research include adaptation at style level [Ficler and Goldberg, 2017], persona-specific traits [Zhang et al., 2018a], or other traits such as sentiment and topic [Madotto et al., 2020].

Personalized interaction between users and dialogue systems is of crucial importance to obtain systems that can be trusted by users and perceived as natural [van der Goot and Pilgrim, 2019], but most of all to be accessible to varying user profiles, rather than targeted at one particular user group [Zheng et al., 2019, Zeng et al., 2020].

In this thesis, I focus on a particular user property a dialogue system can detect and adapt its language to: the user's age group. The research presented in this thesis is comprised of two connected experiments that serve as important steps towards the development of age-adaptive dialogue systems. The first experiment, Experiment 1, focuses on the automated detection of age-related linguistic patterns in dialogue. Subsequently, the goal of the second experiment, Experiment 2, is to develop dialogue generation models that can produce conversational responses that possess linguistic features learned by Experiment 1's classifiers to be characteristic of certain age groups.

Experiment 1 The preliminary research objective of this thesis is to investigate whether the linguistic behavior of conversational participants differs across age groups by using state-of-the-art natural language processing (NLP) models on purely textual data, without considering vocal cues. I aim to detect age from characteristics of language use and adapt to this signal, rather than work from ground-truth metadata about user demographics. This is in the interest of preserving privacy, and from the perspective that while age and language use may have a relationship, this will not be linear [Pennebaker and Stone, 2003] and there are individual differences. [L: revise this last sentence] Previous work on age detection in dialogue has largely focused on combinations of auditory speech features or handcrafted features [Schler et al., 2006, Wolters et al., 2009, Li et al., 2013]. In contrast to this line of work, I investigate whether different age groups can be detected from textual linguistic information rather than voice-related cues, and train my models end-to-end. I explore whether, and to what extent, various state-of-the-art NLP models are able to capture such differences in dialogue data as a preliminary step to age-group adaptation by dialogue systems.

Experiment 2 In Experiment 2, I aim to test whether it is possible to generate dialogue responses that possess age-indicative features, identified and studied in Experiment 1. The main focus of this thesis is therefore on *controlled dialogue generation*, which entails enforcing a specific linguistic style on automatically generated dialogue responses. Specifically, I focus on controlling generated dialogue responses to possess stylistic features that are characteristic of certain age groups. Recent work on dialogue generation models often relies on the Transformer architecture [Vaswani et al., 2017], and the use of large-scale pre-trained Transformer-based architectures for dialogue generation has also become commonplace [Zhang et al., 2020, Brown et al., 2020]. However, retraining such large-scale pre-trained language models to generate text with a specific style is extremely computationally expensive, and infeasible for most developers. Plug-and-Play language models (PPLM) [Dathathri et al., 2020] circumvent these problems by providing a framework to control the output of large Transformer-based language models by using substantially smaller style-specific attribute models to perturb the underlying language model's activation space. While previous work on Plug-and-Play approaches to controlled text generation demonstrated adaptation to coarse and tangible linguistic style, like topic and sentiment [Madotto et al., 2020], more fine-grained and abstract linguistic styles, like age-related language, remains unexplored. This work therefore focuses on the use of Plug-and-Play Language Models for age-adaptive dialogue generation.

Key Contributions My age detection experiment builds on the work of Schler et al. [2006], who focus on age detection in written discourse using handcrafted features. I extend their work by: **(1)** eliminating the need for handcrafted features by learning end-to-end representations using state-of-the-art NLP models; **(2)** applying this approach to dialogue data, using a dataset of transcribed spontaneous open-domain dialogues; **(3)** showing that text-based models can indeed detect age-related differences, even in the case of very sparse signals at the level of dialogue utterances; **(4)** carrying out an in-depth analysis of the models' predictions to gain insight about which elements of language use are most informative. Furthermore, the age detection analyses motivate the use of local features (i.e., BoW-based attribute models) for controlled generation as a viable alternative to neural discriminator-based attribute models. My work on age detection from dialogue can be considered a preliminary step to the modeling of age-related linguistic adaptation by AI conversational systems. In particular, these results informed my work on controlled dialogue generation using PPLM.

My work on controlled dialogue generation using PPLM builds on previous work on controlled language generation by Dathathri et al. [2020], who focus on controlled story writing for concrete styles (e.g., sentiment, or topic). I extend their work in several important ways: **(1)** I control language generation for more abstract writing styles, i.e., age group specific linguistic style; **(2)** I use PPLM for dialogue response generation; **(3)** I propose methods for empirical development of BoW attribute models (as opposed to the manually curated BoWs used by Dathathri et al. [2020]) and demonstrate their applicability for controlled dialogue generation; **(4)** I thoroughly study the relationships between dialogue response quality, response style, and response length; and finally **(5)** I carry out an extensive analysis on the effects of prompt-induced biases on the quality and style-attribute adherence of generated language, which has been overlooked by previous work on Plug-and-Play generation. Despite previous work also focusing on conversational models, my work demonstrates a Plug-and-Play approach to controlled dialogue generation, without the need to generate attribute-specific dialogue datasets, or separately optimize residual adapter modules [Madotto et al., 2020].

Overview of Thesis The code for the experiments is available on GitHub.[1] The rest of this thesis is structured as follows: Chapter 2 is a two-part literature review that first provides the relevant theoretical background of important components involved with age-adaptive dialogue generation using PPLM (Section 2.1), and then compares my work to the most relevant related

[1] `https://github.com/lennertjansen/msc-ai-thesis`

work (Section 2.2). Then, in Chapter 3, I report my work on detecting age-related linguistic patterns in dialogue (i.e., Experiment 1), starting with the clearly stated research objectives and hypotheses, followed by a description of the datasets used throughout this thesis in Section 3.1, the methodology and experimental in Section 3.2, the results of Experiment 1 in Section 3.3, subsequent quantitative and qualitative analyses of these results in Section 3.4, and a short discussion and conclusion to Experiment 1 in Section 3.5. Chapter 4 reports my research about age-adaptive dialogue generation using PPLM (i.e., Experiment 2). I begin by more explicitly describing the research objectives and hypotheses, and connection to Experiment 1 in Section 4.1. This is followed by Experiment 2's methodology and experimental setup in Section 4.2. The results of the age-adaptive dialogue are discussed in Section 4.3, followed by various quantitative and qualitative analyses of these results in Section 4.4. The main contributions and key takeaways of this thesis are summarized in Chapter 5, along with an extensive discussion about the implications of the results, the research limitations, ethical considerations, and future research directions. Finally, the most important conclusions to be derived from this thesis are repeated in Chapter 6.

Chapter 2

Literature Review

This chapter is a two-part literature review. The first section, Background or Section 2.1, provides an overview of this thesis' central problem of controllable dialogue generation, and the components involved, i.e., modeling age in language, language models, controlled text generation, dialogue, dialogue systems, dialogue generation, Transformers, and Plug-and-Play language models. In the second section, Related Work or Section 2.2, I discuss previous approaches, relevant to my work, that have been proposed to tackle each of these components, either separately or jointly. Approaches to different, but strongly related, problems, like text style transfer, are also described in Section 2.2.

2.1 Background

2.1.1 Language and Age

The relationship between a person's age and use of language is a thoroughly studied subject with a decades-long history and inherent challenges [Pennebaker and Stone, 2003, Nguyen et al., 2014, Zheng et al., 2019]. A number of factors like community membership (e.g., gender, socioeconomic status, or political affiliation), experimental condition (e.g., emotional versus non-emotional disclosure), mode of disclosure (writing versus talking), and other confounding variables complicate the study of age's relation to language [Nguyen et al., 2011]. The relatively recent advent of widely available computational resources and vast amounts of textual data made it possible to leverage machine learning methods to help detect patterns in language that eluded previous conventional sociolinguistic research. Early computational investigations into the connection between a person's age and use of language is typically a combination of qualitative and statistical methods. For instance, using a mix between their proprietary count-based text analysis framework, Linguistic Inquiry and Word Count (LIWC) and sociolinguistic theory, Pennebaker and Stone [2003] study the changes in written and spoken language use with increasing age. They discuss four important areas of a person's character that have been found to change with age: emotional experience and expression, identity and social relationships, time orientation, and cognitive abilities. These four axes and their hypothesized relationships with language use and age can be interpreted in the following ways:

1. *Emotional experience and expression:* This is the relationship between increasing age and linguistically observable manifestations of a person's experienced emotions. In practical terms, this is framed as detectable instances of positive and negative affect in language. This complex relationship between age and emotional expression is characterized by decreased levels of negative affect and slightly non-decreasing levels of positive affect. This is also confirmed by the findings of Schler et al. [2006].

2. *Sense of identity and social relationships:* These terms refer to developmental tends in one's relation to self and others, as expressed in their language, e.g., as references to self (*I, me, my*, and *we, us, our*) or others (*they, them, theirs*). Pennebaker and Stone [2003] report that the *quantity* of social connections decreases and the *quality* of remaining relationships increases with age.

3. *Time orientation:* This relationship describes how people express their perception of and orientation towards time. For instance, this can be indicated by the use of time-related verb tenses. The authors suggest that older individuals tend to be more past-oriented than their younger future-oriented counterparts.

4. *Cognitive abilities:* This refers to markers of cognitive capacity in language. Aging is expected to be associated with less use of cognitively complex words after a certain mid-adulthood peak. Specifically, the relationship between markers of cognitive complexity in natural language (cognitive mechanisms, causal insight, and exclusive words) and age is hypothesized to be curvilinear, with markers of verbal ability only declining very late in life.

Pennebaker and Stone [2003] consider the following variables: positive and negative emotions, first-person singular and first-person plural pronouns, social references, time-related words (past-tense, present-tense, and future-tense verbs), big words ($>$ 6 letters), cognitive mechanisms, causal insight, and exclusive words. Their main findings suggest that increasing with age, people use more positive and fewer negative affect words, use fewer self-references, use more future-tense and fewer past-tense verbs, and exhibit a general pattern of increasing cognitive complexity.

Detectable linguistic differences between age-groups can often be attributed to the use of language fads or references to age-specific popular culture. For instance, Schler et al. [2006] find that the use of slang and neologisms (such as *lol* and *ur*) are strong indicators of youth. Similarly, words like 'facebook', 'instagram', and 'netflix' appear in the most frequently used words by younger participants of conversational data collection efforts, like that of the British National Corpus' spoken component [Love et al., 2017].

Despite the demonstrated utility of using handcrafted age-related linguistic features for age group detection from text [Schler et al., 2006], modern approaches leverage statistical language models that represent a probability distribution over sequences of words, often parameterized by neural network architectures [Zheng et al., 2019]. These approaches based on language models show impressive efficacy for many NLP tasks, and can be trained end-to-end, often not requiring domain-knowledge. The following section provides further details about language models.

2.1.2 Language Models

Generally speaking, language modeling is central to many NLP tasks. A language model (LM) is a probability distribution over words in a sentence or document. Language models are trained to predict the probability of the next word in an sentence, given the preceding sequence of words. The language modeling task is formulated as an unsupervised distribution estimation problem of datapoints $\{\mathbf{x}_1, ..., \mathbf{x}_N\}$ (e.g., documents), each representing sequences (of e.g., symbols or tokens) of varying lengths $(s_{i,1}, ..., s_{i,n}), i \in \{1, ..., N\}$. Note that N denotes the corpus size, and n the sequence length of datapoint i. To avoid cluttered notation, the subscript i will sometimes be omitted when discussing an arbitrary datapoint. The probability distribution over an observation **x** (i.e., the joint probability of an ordered sequence) can then be factorized as the product of its constituent conditionals [Radford et al., 2019]:

$$p(\mathbf{x}) = \prod_{j=1}^{n} p(s_j | s_1, ..., s_{j-1}). \tag{2.1}$$

This formulation allows language models to detect and learn patterns in language. The learned representations of these patterns can then be used for a plethora of applications, such as classification, and text generation. Moreover, this results in a framework for tractable sampling from the unconditional language model $p(\mathbf{x})$. $p(\mathbf{x})$ can therefore be seen as a base generative model that can generate sample sentences [Dathathri et al., 2020].

In recent years, the attention-based models, Transformers [Vaswani et al., 2017], have replaced recurrent neural networks (RNNs) as the dominant architecture for LMs, with major improvements in distribution estimation, long-range dependency handling, sample diversity, and parallel processing. Another recent development in language modeling is that of pre-training LMs on massive corpora. So-called large-scale general purpose LMs have demonstrated significant improvements in downstream tasks, i.e., other NLP tasks for which the model was not specifically trained or fine-tuned. Most famously the OpenAI's series of Generative Pre-trained Transformer (GPT) models have improved numerous NLP benchmarks [Radford et al., 2018, 2019, Brown et al., 2020]. In particular, this series of autoregressive language models has demonstrated impressive advancements in the task of text generation, to produce human-like text.

2.1.3 Text Generation and Controlled Text Generation

In *text generation*, a language model $p(\mathbf{x})$ is asked to produce text $\mathbf{x}$ given a prompt by sampling from the distribution of words that are assigned the highest likelihood of following the prompt [Radford et al., 2019]. More concisely, text generation in itself is the task of generating a piece of text given an input text. This process can be seen as sampling from a conditional distribution, $p(\mathbf{x}|\texttt{prompt})$. *Controlled text generation* refers to the more restrictive problem of enforcing higher-level linguistic features on the generated text during sampling [Dathathri et al., 2020, Prabhumoye et al., 2020]. This can be seen as a sub-problem of vanilla text generation, because the conditioning factor for the output text is further constrained to also include some predefined textual attribute, a. That is, controlled text generation is analogous to sampling from the conditional distribution, $p(\mathbf{x}|\texttt{prompt}, a)$, where the attribute, a, represents a linguistic characteristic of the text, like sentiment, topic, or writing style.

Controlled text generation or CTG is a more challenging problem than vanilla text generation for a number of reasons. First, defining the desired attribute to be controlled for in a manner that it is intelligible for a machine is a challenge in itself [Zheng et al., 2019]. Second, like many NLP problems, there are not many parallel corpora [Dai et al., 2019]. In the context of controlled generation, parallel corpora are datasets of target and source texts that only differ with respect to some attribute. Furthermore, the measure of attribute adherence is a very vague and ambiguous concept [Dathathri et al., 2020, Dai et al., 2019]. Namely, a text can be written to have an extremely positive sentiment in multiple formulations, all of which adhere to the positive sentiment. Another important hurdle for controlled text generation, especially when CTG is combined to leverage the linguistic power of large-scale language models, is that the cost of fine-tuning or pre-training a model to control for a linguistic attribute can be very high [Dathathri et al., 2020, Madotto et al., 2020].

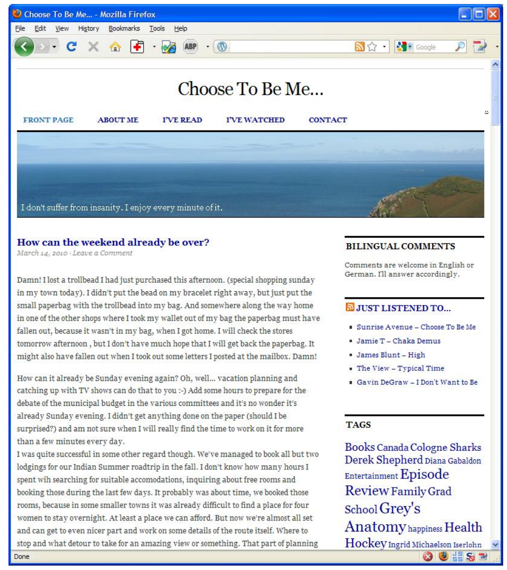
2.1.4 Dialogue and Dialogue Systems

Because the focus of this thesis is on controlled text generation in the context of dialogue, I provide a section explaining key concepts about dialogue in general and in NLP. This section covers the relevant definition of dialogue, what dialogue systems are and why they are developed and studied, and dialogue generation as a language modeling problem.

A dialogue is a written or spoken conversational exchange between two or more interlocutors (e.g., two people, or a person and a virtual assistant like Amazon Alexa) [Merriam-Webster, n.d.], and it is the most natural form of interaction between people [Burtsev et al., 2018]. Generally speaking, a dialogue's purpose is to exchange information or build relationships between interlocutors [Bohm and Nichol, 2013], and it typically consists of interlocutors exchanging utterances in turns. In the context of spoken language analysis, an utterance is considered the smallest unit of speech, that is followed by a change in speaker or the end of a conversation, thus representing a dialogue turn [Traum and Heeman, 1996].

In the age group detection experiment presented in Chapter 3 (i.e., Experiment 1), I also experiment with a so-called discourse dataset (i.e., a dataset of blog posts), in addition to dialogue data. It is therefore useful to establish a distinction between dialogue and discourse. Discourse has multiple definitions, but for the purpose of this thesis, I adopt the definition of discourse as a linguistic unit (e.g., a long talk or a piece of writing), typically longer than a sentence, whose purpose is to convey thoughts or ideas [Merriam-Webster, n.d.]. Dialogue therefore distinguishes itself from discourse in that it necessarily involves two or more participants exchanging information and contributing to the conversation, whereas discourse can be a one-way exchange of information, like a lecture or blog-post. See Figure 2.1 for an example of a dialogue between people, and discourse in the form of a blog post.

(a) *Original image source:* Figure 1 of Liu et al. [2020].



(b)

Figure 2.1: Examples of a dialogue ((a) / left), and discourse in the form of a blog post ((b) / right).

It remains an open challenge for NLP researchers to endow a machine with the capability to engage in meaningful dialogue [Burtsev et al., 2018], and the development of so-called dialogue systems is an active field of study. A dialogue system is a computer program that supports spoken, text-based, or multi-modal conversational interactions with humans [McTear, 2020]. Dialogue systems typically fall into one of two categories: task-oriented and non-task-oriented dialogue systems. The former engages in an interaction with the user to complete some task, whereas the latter engages in a general conversational interaction with the user [Kushneryk et al., 2019]. The focus of this thesis is on non-task-oriented dialogue. The development of dialogue systems distinguishes itself from most NLP domains due to the inherent challenges associated with modeling human conversation: informal, noisy, unstructured, and even erroneous real-world responses, possibly competing goals of interlocutors, and extremely diverse sets of acceptable responses to prompts.

Nevertheless, McTear [2020] suggests that there are three main motivations for researches to investigate and develop dialogue systems:

1. Dialogue systems can provide users with a convenient and intuitive way to interact with technological services. Furthermore, they can help providers of these services by taking over simple and mundane tasks. However, it must be noted that the replacement of

human workers by dialogue systems has ethical and societal consequences that should be taken into account by researchers [Ivanov, 2020].

2. The development of dialogue systems requires researchers to model human conversational dynamics, and doing so can broaden our understanding of human behavior.

3. Research about dialogue systems might one day result in human conversational behavior being simulated so accurately that users could be convinced they are interacting with another human (e.g., by passing the Turing test [Oppy and Dowe, 2003]). However, it is important to note that convincing human users that they are conversing with another human, while it is in fact a dialogue system, is by no means a requirement for an effective dialogue system. And despite this being a core goal of AI [Zheng et al., 2019], the ethics of deceiving humans into believing they are talking to another human are highly debatable.

Developing dialogue systems is a very complex task and many approaches have been proposed for this [McTear, 2020]. These approaches can differ greatly with respect to their methodologies, ranging from the use of recurrent neural networks [Li et al., 2016b], to reinforcement learning [Mo et al., 2018], variational auto-encoders [Ruan et al., 2019], Transformers [Madotto et al., 2020], and multi-modal models [Shuster et al., 2021]. Despite the many types of modules that can comprise a dialogue system, generally, the basic components of a dialogue system are automatic speech recognition (ASR), language understanding (LU), dialogue management (DM), natural language generation (NLG), and text-to-speech synthesis (TTS) [Chen et al., 2017]. Figure 2.2 shows a schematic overview of how information is routed through such a pipeline. Note that the first and last components (i.e., ASR, and TTS) are omitted if the dialogue system does not support auditory signals and responses. The focus of this thesis is on dialogue generation, a particular operationalization of dialogue in NLP, which falls under the "*Natural Language Generation (NLG)*" category of Figure 2.2.

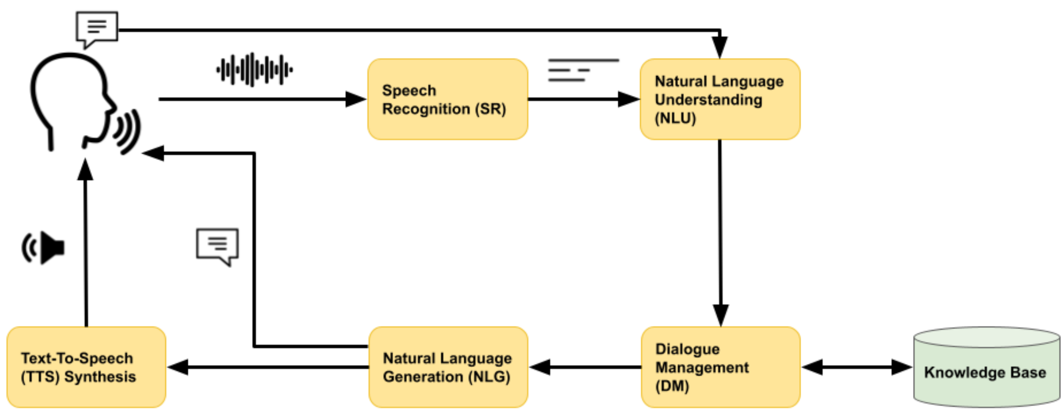


Figure 2.2: Schematic overview of a pipeline for a dialogue system.

2.1.5 Dialogue Generation and Controlled Dialogue Generation

Dialogue response generation, also referred to simply as dialogue generation, is the task of automatically generating a response given a prompt [Madotto et al., 2020]. Dialogue generation can be formulated as a language modeling task [Welleck et al., 2019], where the next utterance is conditioned on the dialogue history. In this framework, an utterance is a sequence of tokens representing a dialogue turn. The next utterance, $\mathbf{x}_{t+1}$, is predicted conditioned on a dialogue prefix or prompt, $\mathbf{x}_{\leq t}$ (e.g., a single previous dialogue turn, or the entire conversation history). When developing dialogue generation models, the language modeling training objective can be formulated as a product of source-target pair probabilities [Zhang et al., 2020]. This consists of concatenating all dialogue turns in a multi-turn dialogue session into a long text: $\mathbf{x}_1, ..., \mathbf{x}_N$. The source sequence (or dialogue history) is then denoted as $\mathcal{D} = \mathbf{x}_1, ..., \mathbf{x}_m$ and the target sequence (or ground truth dialogue continuation) as $T = \mathbf{x}_{m+1}, ..., \mathbf{x}_N$. The conditional probability of dialogue continuation given its history $P(T|\mathcal{D})$ can be then expressed as

$$p(T|\mathcal{D}) = \prod_{n=m+1}^{N} p(\mathbf{x}_n|\mathbf{x}_1, ..., \mathbf{x}_{n-1}). \tag{2.2}$$

A multi-turn dialogue session $T_1, ..., T_K$ can be written as $p(T_K, ..., T_2|T_1)$ which is essentially the product of all source-target pairs probabilities $p(T_i|T_1, ..., T_{i-1})$. This formulation also shows that optimising the single objective $p(T_K, ..., T_2|T_1)$ is equivalent to optimising all source-target pair probabilities. Sampling from such a dialogue generation model, i.e., generating a next utterance $\mathbf{x}_{t+1}$ based on a dialogue history at turn t, $\mathcal{D}_t$, entails sampling from the conditional distribution $p(\mathbf{x}_{t+1}|\mathcal{D}_t)$.

Analogous to the definition of controlled text generation given earlier in Section 2.1.3, *controlled dialogue generation* refers to the process of enforcing a particular linguistic style on an automatically generated dialogue response, and it is the core task of this thesis. Controlled dialogue generation can be seen as sampling from the conditional distribution $p(\mathbf{x}_{t+1}|\mathcal{D}_t, a)$, where $\mathbf{x}_{t+1}$ is a sequence of tokens representing the next dialogue turn or utterance, $\mathcal{D}_t$ is the dialogue history at dialogue turn t, and a denotes a specific linguistic style which utterance $\mathbf{x}_{t+1}$ is desired to possess. Similar to the distinction made between text generation and controlled text generation in Section 2.1.3, controlled dialogue generation can also be seen as a more restrictive version of dialogue generation, as the probability space being conditioned on is further constrained to also contain the style attribute a. This style attribute typically represents concrete linguistic styles, like sentiment or topic [Madotto et al., 2020], and rarely a more abstract style, like the speaking style of a particular age group. However, recent efforts to develop personalized dialogue systems require researchers to develop controlled dialogue generation models that control for more abstract linguistic styles.

2.1.6 Personalized Dialogue Generation Models

Personalized dialogue generation models are special cases of controlled dialogue generation models in that they attempt constrain generated dialogue responses to possess certain linguistic characteristics that align with a user's persona (e.g., their age, gender, or geographical region). Although this thesis solely focuses on a particular aspect of a user's profile, i.e., their age group, discussing general personalization in dialogue generation models is still useful to develop a broader understanding of controlled dialogue generation. Namely, the distinctions between types of personalization approaches for dialogue generation models, and the associated challenges often also apply to controlled dialogue generation models more broadly.

Assuming an encoder-decoder setup, personalized dialogue generation models can be classified as one of two types: implicit and explicit personalisation models Zheng et al. [2019]. For implicit personalization models, each speaker has its own vector representation, which implicitly captures the speaking style of the speaker in the decoding process [Kottur et al., 2017, Li et al., 2016a].

These models enjoy the benefit of having a more granular and realistic representation of speaking style, as opposed to a simple discrete set of traits (as is the case for explicit personalization models). On the other hand, it is unclear how speaker style is captured and should be interpreted, as all the information about a speaker's style is encoded in a real-valued vector. Furthermore, these methods suffer from a data sparsity issue, because each dialogue should be tagged with a speaker identifier and there should be sufficient dialogues from each trait-group to train a reliable trait-adaptive model.

When generating responses, explicit personalization models are conditioned either on a given personal profile [Qian et al., 2018], text-described persona [Zhang et al., 2018a], or simply an attribute label [Madotto et al., 2020]. That is, speaker traits are represented as key-value pairs or descriptions about age, gender, etc. This can be seen as conditioning the decoder's output on an attribute a (as is the case in this thesis). Speakers with same set of personality traits can share attribute representations, so it does not require a speaker-specific representation vector. Such structured character descriptions are more explicit, straight-forward, and interpretable. However, explicit personalization models require manually labeled or crowdsourced datasets for development, making it difficult to scale these models to large-scale dialogue datasets [Zheng et al., 2019, Madotto et al., 2020].

2.1.7 Transformers

The Transformer architecture plays a central role in most of the recent advances in NLP. The same holds for the methods used in this thesis to investigate controlled dialogue generation and automated detection of age-related linguistic patterns in dialogue utterances. For a more detailed review of the model architecture, the reader is referred to the original paper [Vaswani et al., 2017], the annotated and replicated version of the original paper [Rush, 2018], or this blog post[1].

The Transformer, like most neural sequence processing models, has an encoder-decoder structure. On a high level, the encoder receives an input sequence $\mathbf{x} = (x_1, ..., x_n)$ (e.g., a sentence), and maps this to a sequence of latent continuous variables $\mathbf{z} = (z_1, ..., z_n)$. The decoder then takes $\mathbf{z}$ as input, and maps this to an output sequence $\mathbf{y} = (y_1, ..., y_m)$. Note that the use of positional encodings of the input and output embeddings enables the Transformer to process and generate sequences in arbitrary order, allowing for a high degree of parallelization. The generation of $\mathbf{y}$

[1] `https://jalammar.github.io/illustrated-transformer/`

happens element-by-element in an auto-regressive fashion, where at step t, element y_{t-1} is also taken as input.

Both the encoder and decoder are comprised of N identical layers (denoted by the 'N $\times$' in the left part of Figure 2.3). Every sub-layer performs a succession of transformations using multi-head self-attention mechanisms and point-wise, fully connected layers, along with residual connections [He et al., 2016] around every sub-layer followed by layer normalization [Ba et al., 2016]. The decoder's first self-attention sub-layer is masked to ensure that the output predictions at sequence position i cannot depend on output positions greater than i. Finally, the decoder passes its output through a linear and softmax layer to produce a probability distribution over the problem space (e.g., the vocabulary) from which the most likely symbols for the generated output sequence **y** can be sampled.

A key aspect of the Transformer architecture is its use of attention [Bahdanau et al., 2015]. This allows the encoder-decoder architecture to selectively focus on parts of the input sequence to produce a more informative hidden representation. Vaswani et al. [2017] formulate an attention function as a mapping of queries and sets of key-value pairs to an attention output, where matrices represent the queries Q, keys K, and values V. The attention output is a weighted sum of the values, based on the relevance of the corresponding keys to a query. In particular, they employ scaled dot-product attention:

$$\texttt{Attention}(Q, K, V) = \texttt{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{2.3}$$

Furthermore, Vaswani et al. [2017] propose to use multi-head attention by using learned linear projections to project the queries, keys and values h times, and apply the aforementioned attention function to these projections in parallel. The concatenation of these attention outputs, passed through a linear layer, ultimately produces the final output of the Transformer's attention sub-layers. This allows the model to attend to the relevant information from all representation sub-spaces at various sequence positions. See Figure 2.3 for a schematic illustration of the Transformer's structure described above.

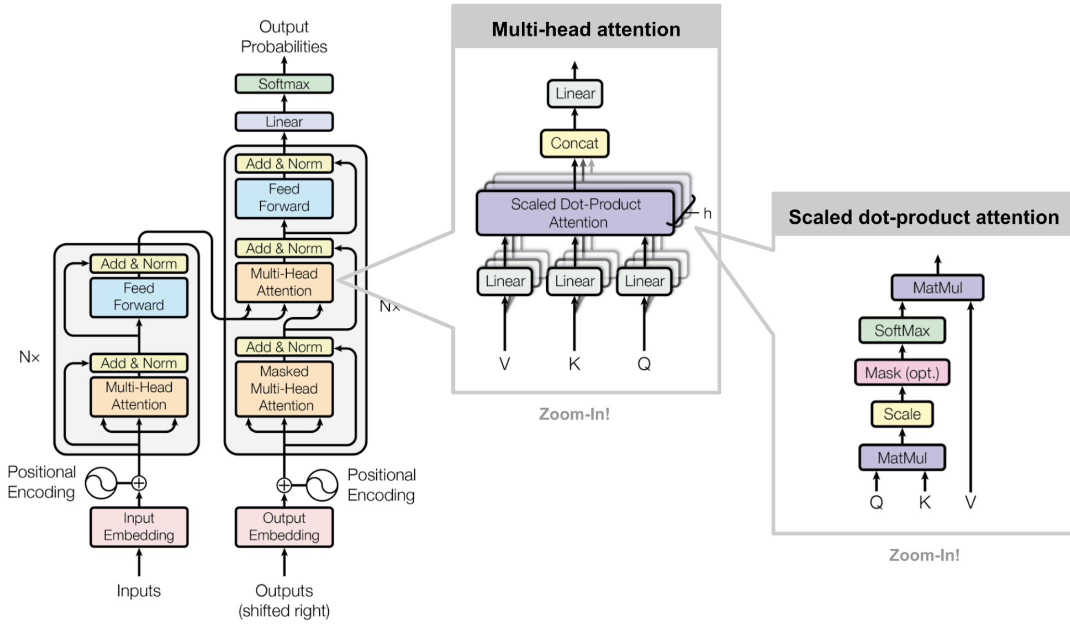


Figure 2.3: An overview of the full Transformer model architecture. *Collated image source:* Fig. 17 in this blog post `https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html`. *Original image source:* Figures 1 and 2 in Vaswani et al. [2017]

2.1.8 Causal Language Modeling with Transformers for Dialogue Generation

As previously described in Section 2.1.4, a dialogue is comprised of multiple alternating turns, i.e., utterances, between more than one speaker. Furthermore, as described in Section 2.1.5, the concatenation of the sequences representing utterances is defined as the dialogue history. Because the focus of this thesis is on dialogues between two speakers, the dialogue history at turn t is defined as $\mathcal{D}_t = \{S_1^{(1)}, S_1^{(2)}, ..., S_t^{(1)}\}$, where $S_t^{(j)}$ is speaker j's utterance at time t.

A Transformer-based language model (denoted `LM`) is used in this thesis to model the distribution of dialogues, using dialogue history at time t, $\mathcal{D}_t$, as a prompt to auto-regressively generate the dialogue continuation S_t. More specifically, let the concatenation of the dialogue history at t and its continuation, $\{\mathcal{D}_t, S_t\}$, be represented as a sequence of tokens $\mathbf{x} = \{x_0, ..., x_n\}$. Then, by recursively applying the product rule of probability (Bishop [2006]), the unconditional probability of the sequence $p(\mathbf{x})$ can be expressed as:

$$p(\mathbf{x}) = \prod_{i=1}^{n} p(x_i|x_0, ..., x_{i-1}). \tag{2.4}$$

Dathathri et al. [2020] and Madotto et al. [2020] define the Transformer's decoding process in a recursive fashion. Let H_t denote the conversation history's key-value pairs, i.e., $H_t = \left[(K_t^{(1)}, V_t^{(1)}), ..., (K_t^{(l)}, V_t^{(l)})\right]$, with $(K_t^{(i)}, V_t^{(i)})$ representing the key-value pairs from the LM's i-th layer generated at all time steps 0 through t. This results in the recurrent dedocing process being expressed as:

$$o_{t+1}, H_{t+1} = \texttt{LM}\left(x_t, H_t\right), \tag{2.5}$$

where o_{t+1} is the hidden state of the last layer. Finally, after applying a softmax transformation, the next token x_{t+1} is sampled from the resulting probability distribution, i.e., $x_{t+1} \sim p_{t+1} = \texttt{softmax}\left(W o_{t+1}\right)$, where W is a linear mapping from the model's last hidden state to a vector of vocabulary size. This recursive formulation allows for efficient text generation by leveraging cached memories, without repeated forward passes.

2.1.9 Plug-and-Play Language Modeling

The Plug-and-Play approach to controlled text generation, proposed by Dathathri et al. [2020], is a core method of the research presented in this thesis. It provides a framework to control the writing style of large pre-trained Transformer-based language models (like GPT-2 [Radford et al., 2019]), without incurring relatively high computational costs. What follows is a detailed explanation of the PPLM-method. For additional information, the reader is referred to the original paper [Dathathri et al., 2020].

Relationship to Bayes' Theorem The Plug-and-Play Language Model (PPLM) works by using a text classifier, referred to as an attribute model, to control the text generated by a language model. Let $p(X)$ denote the output distribution of a Transformer-based language model (e.g., GPT-2 or DialoGPT), where X represents the generated text. And $p(a|X)$ denotes the attribute model (e.g., a single-layer or BoW classifier) that represents the degree of adherence of text X to a certain attribute a (e.g., topic or sentiment (previous work), or age-group characteristics (this work)). Then PPLM can be seen as modeling the conditional distribution of generated text X given attribute a, i.e., $p(X|a)$. Note that Bayes' theorem ties these three definitions together as follows:

$$p(X|a) \overbrace{=}^{\text{Bayes' theorem}} \frac{p(X)p(a|X)}{p(a)} \propto p(X)p(a|X). \tag{2.6}$$

PPLM Gradient Update Rule To control the generated text, PPLM shifts the previously mentioned history H_t (see Section 2.1.8 and Equation 2.5), i.e., all Transformer key-value pairs generated up to time t, in the direction of the sum of two gradients:

1. Ascending $\nabla \log p(a|X)$: maximizing the log-likelihood of the desired attribute a under the conditional attribute model. This enforces attribute control.
2. Ascending $\nabla \log p(X)$: maximizing the log-likelihood of the generated language under the original (possibly conversational) language model. This promotes fluency of the generated text.

These two incentive-representing gradients are combined with various coefficients, yielding a set of tunable parameters to steer the generated text in the direction of the desired fluency, attribute control, and length.

Let's first focus on the first of the two gradients, i.e., the attribute control promoting $\nabla \log p(a|X)$. ΔH_t represents the update to history H_t that pushes the distribution of the generated text X in the direction that has a higher likelihood of adhering to desired attribute a. The gradient update rule can be expressed as:

$$\Delta H_t \leftarrow \Delta H_t + \alpha \frac{\nabla_{\Delta H_t} \log p(a|H_t + \Delta H_t)}{\|\nabla_{\Delta H_t} \log p(a|H_t + \Delta H_t)\|^{\gamma}} \tag{2.7}$$

where α is the step size, and γ denotes the normalization term's scaling coefficient. Both step size (α) and the scaling coefficient (γ) influence attribute control. Attribute control can be softened by either decreasing α or increasing γ and vice versa. Note that $\alpha = 0$ recovers the original uncontrolled underlying language model (e.g., GPT-2 or DialoGPT). In practice, ΔH_t is initialized at zero, and the update rule in Equation 2.7 is applied m times (usually 3 to 10), resulting in the updated key-value pair history $\tilde{H}_t = H_t + \Delta H_t$. Then the updated history $\tilde{H}_t$ is passed through the language model, yielding the updated logits (final Transformer-layer): $\tilde{o}_{t+1}, H_t = \texttt{LM}(x_t, \tilde{H}_t)$. And finally the shifted $\tilde{o}_{t+1}$ is linearly mapped through a softmax layer to produce a new, more attribute-adherent, distribution from which to sample, i.e., $x_{t+1} \sim \tilde{p}_{t+1} = \texttt{softmax}\,(W\tilde{o}_{t+1})$.

Maintaining Fluency of Generated Text The method described until now will generate attribute-adherent text, but will likely yield fooling examples [Nguyen et al., 2015] that are gibberish to humans, but get assigned high $p(a|x)$ by the attribute model [Dathathri et al., 2020].

That is why Dathathri et al. [2020] apply two methods to ensure fluency of the generated text. The first is to update ΔH_t such to minimize the Kullback-Leibler (KL) divergence [Kullback and Leibler, 1951] (denoted D_{KL}) between the shifted and original distributions. In practice, D_{KL} is scaled by a coefficient λ_{KL}, typically found to work well for most tasks when set to 0.01. Repetitive text generation (i.e., high $p(a|x)$ but low $p(x)$) can therefore sometimes be avoided by increasing λ_{KL}. The second method to ensure fluency is Post-norm Geometric Mean Fusion [Stahlberg et al., 2018] which, instead of directly influencing ΔH_t like minimizing D_{KL}, fuses the altered generative distribution $\tilde{p}_{t+1}$ with the unconditional language distribution $p(x)$. This is done during generation by sampling the next token as follows:

$$x_{t+1} \sim \frac{1}{\beta}\left(\tilde{p}_{t+1}^{\gamma_{gm}} p_{t+1}^{1-\gamma_{gm}}\right) \tag{2.8}$$

where β is a normalization constant, p_{t+1} and $\tilde{p}_{t+1}$ denote the original and modified distributions, respectively, and γ_{gn} is a scaling term that interpolates between the two distributions. Because the new sampling distribution in Equation 2.8 converges towards the unconditional language model as $\gamma_{gm} \to 0$, repetitive text generation can be avoided by decreasing the scaling term.

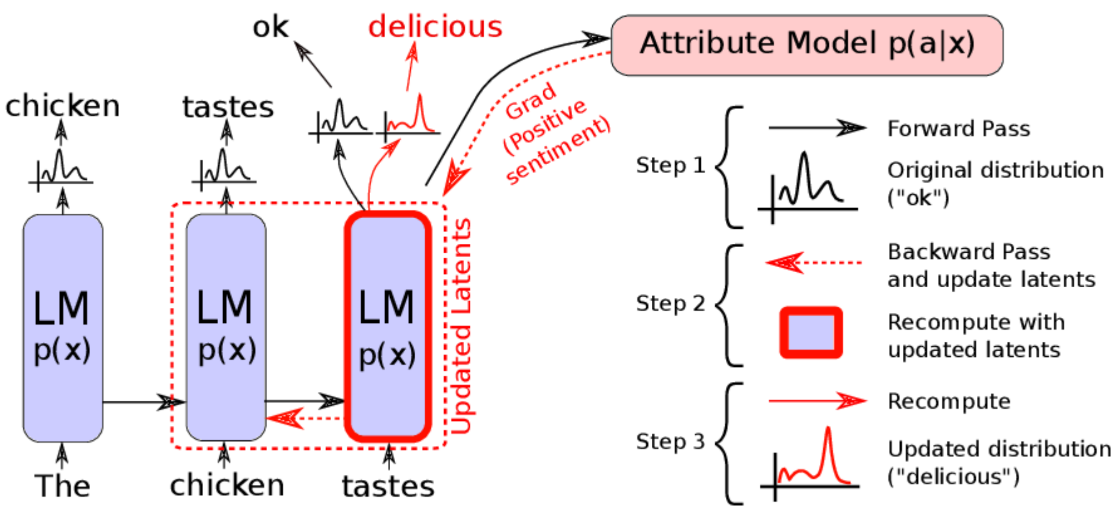


Figure 2.4: A schematic overview of the plug-and-play interaction between attribute model $p(a|\mathbf{x})$ and language model $p(\mathbf{x})$. *Original image source:* Figure 1 of Dathathri et al. [2020]

Attribute Models As mentioned at the beginning of this section (i.e., Section 2.1.9) and shown in Equation 2.7, the PPLM-method uses gradient updates that maximize the attribute model's log-likelihood $\log p(a|X)$ to enforce an attribute-specific style on a language model's generated output. The simplest attribute model is a bag-of-words (BoW), i.e., a list of words that represents the stylistic attribute desired to be enforced on the generated text. The log-likelihood of a BoW attribute model is the log of the sum of likelihoods of every constituent word of the BoW. Given a bag-of-words $\{w_1, ..., w_k\}$ that represents an attribute a, and the output distribution of the language model p_{t+1}, the attribute model's log-likelihood is

$$\log p(a|x) = \log \left(\sum_{i}^{k} p_{t+1}[w_i] \right) \tag{2.9}$$

Ascending $\nabla \log p(a|x)$ therefore increases the likelihood of words in the BoW, $\{w_1, ..., w_k\}$, being generated. Furthermore, this implies that BoW-based control is exerted most noticeably at the lexical level.

BoW attribute models are known to be capable of capturing accurate representations of attributes like sentiment and topic [Li et al., 2018]. However, it can be desirable to use a more complex attribute model, i.e., a neural discriminator, when the attribute is more abstract and not easily easily represented as a BoW (e.g., age-specific language) [Dathathri et al., 2020]. In the PPLM-method, a discriminator attribute model $f := p(a|x)$ is developed by training a singly layered linear discriminator on (presumably attribute-adherent) input sentences x and their corresponding labels y_x. More specifically, for an input sentence x of length (in tokens) t, the set of hidden states $o^x_{:t}$ is collected from the Transformer-based `LM` (see Equation 2.5). The set of hidden states is then averaged over all time steps, yielding $\bar{o}^x_t$. The discriminator f is then trained using the cross-entropy between the discriminator's output $f(\bar{o}^x_t)$ and the label distribution y. Because the discriminator attribute model is trained on $\bar{o}^x_t$, it is possible to capture and propagate a more comprehensive representation of a, shifting `LM`'s output distribution towards one that favors sentences that align with a.

Computational Costs of PPLM Although the discriminator attribute model requires a separate training schedule, the computational cost of training and using such an attribute model in a PPLM-setup is negligible compared to fine-tuning a typical choice of `LM` (e.g., GPT-2-medium [Radford et al., 2019]) for controlled generation. Namely, the number of trainable parameters comprising the discriminator attribute model is equal to $(L_o \cdot N_a) + N_a$, where L_o is the length of the hidden state vector, and N_a is the number of classes defined in attribute a (e.g., $N_a = 2$ if a represents age, defined by two age groups, younger and older). Note that this is a substantially smaller number of trainable parameters than `LM` typically has. For instance, training a PPLM-setup consisting of GPT-2-medium as its underlying language model, a discriminator attribute model, and an attribute a with two classes amounts to training $(1024 \cdot 2) + 2 = 2050$ parameters, whereas fine-tuning the same language model to generate attribute-adherent text for *both* attribute classes entails training $2 \cdot 345$M parameters (i.e., one fully fine-tuned model per attribute class).

It is worth mentioning that the Plug-and-Play method applied by Dathathri et al. [2020] and Madotto et al. [2020] is different from fine-tuning. Note that in Equation 2.7 the gradient updates are restricted to the history H_t, and do not affect the model's parameters. Because the key-value pairs $(K_t^{(i)}, V_t^{(i)})$ that comprise H_t are activations and not model-weights, the updates are only made to the activation-space. This means that PPLM leaves the underlying (conversational) language model untouched.

Contrary to fine-tuning often massive LMs, PPLM does not incur a significant training cost (depending of course on the complexity of the discriminator or attribute model). However, Madotto et al. [2020] show that PPLM needs a fixed number of m update-steps to for every generated token. This makes the original PPLM setup unsuitable for online interactive applications, like conversational systems. Addressing this problem, they introduce plug-and-play conversational model (PPCM), which extends PPLM by using the original model setup to generate dialogue datasets with the desired attribute a, and then use optimized residual adapters [Bapna and Firat, 2019] to control `LM`'s output distribution. More specifically, Madotto et al. [2020] use PPLM to generate n attribute-adherent dialogue datasets $\mathscr{D}^a = \{\mathcal{D}^1, ..., \mathcal{D}^n\}$, for attribute a. These generated dialogue datasets are then used to train the residual adapter modules that control the language model's output distribution.

2.2 Related Work

2.2.1 Automated Age Detection from Language

Automated (speaker or author) age detection from (spoken or written) language is a challenging task, for which many approaches have been suggested [Nguyen et al., 2011]. The work of Schler et al. [2006] focuses on age detection in written discourse using a corpus of blog posts. The authors learn a Multi-Class Real Winnow classifier leveraging a set of pre-determined style- and content-based features, including part-of-speech categories, function words, and the 1000 unigrams with the highest information gain in the training set. They find that content features (lexical unigrams) yield higher accuracy ($\sim 74\%$) than style features ($\sim 72\%$), while their best results are obtained with their combination ($\sim$76%). Previous work on age detection in dialogue has focused on speech features, which are known to systematically vary across age groups. For example, Wolters et al. [2009] learn logistic regression age classifiers from a small dialogue dataset using different acoustic cues supplemented with a small set of hand-crafted lexical features, while Li et al. [2013] develop SVM classifiers using acoustic and prosodic features extracted from scripted utterances spoken by participants interacting with an artificial system.

More recent studies, like that of Nguyen et al. [2011], Nguyen et al. [2014], Zheng et al. [2019], and Abdallah et al. [2020], frame age prediction from text as traditional machine learning problems, like linear regression, support vector machines, or neural architectures. These modeling approaches tend to reveal that strong indicators of age lie at the syntactic or structural level of language use, as opposed to the more content-based lexical level. Furthermore, this could explain why automatic detection from text of more content-based traits, like topic or sentiment, tend to be easier problems to solve than age prediction from text. To emphasize one such complicating factor, Nguyen et al. [2014] argue that differences in language use are often relation to the speaker's social identity, which could differ from their biological identity.

2.2.2 Controlled Text Generation

Previous approaches to controlled text generation require fine-tuning large Transformer-based language models or training conditional generative LMs from scratch. Most notably CTRL [Keskar et al., 2019], which achieves controlled generation by training a generative Transformer for a number of control codes. CTG models that require fine-tuning for control, like CTRL, can produce high quality fluent text because they are specifically trained to maximize the likelihood of generated sequences, given an attribute (denoted $p(\mathbf{x}|a)$), but require training massive language models with computational costs.

Other recent examples of controlled text generation models that are not Transformer-based also exist. Li et al. [2020] introduce OPTIMUS, a large pre-trained Variational Autoencoder (VAE) [Kingma and Welling, 2014] that can be fine-tuned for specific natural language tasks, like guided sentence generation. They demonstrate OPTIMUS' ability to perform controlled text generation from latent style-embeddings, with fluency at par with GPT-2. They also show how OPTIMUS generalizes better for low-resource languages than BERT [Devlin et al., 2019]. Nevertheless, much like the previously mentioned CTG models, OPTIMUS still incurs a significant computational cost for fine-tuning per NLP task.

The Plug-and-Play language model (PPLM) [Dathathri et al., 2020] is a recent solution to the problem of high re-training costs of controlled text generation. This approach, inspired by a similar technique for style-control of generated images [Nguyen et al., 2017], leverages the fluency of large-scale language models when controlling them for a specific linguistic attribute, while avoiding incurring significant costs of fine-tuning these massive language models. The main benefit of this setup is its low-cost extensibility. Namely, such large-scale language models are often open-source and available online, and can now be tailored to users' specific needs using a significantly easier to train attribute model. The original architecture proposed by Dathathri et al. uses GPT-2 as a base language model which provides grammatical fluency, combined with a significantly easier to train attribute model (i.e., a simple BoW or single-layer classifier). Using gradient updates to the activation space of the much smaller attribute model, they manage to generate language that combines (some of) the fluency of GPT-2 with the stylistic control of the attribute model, without the cost of retraining a specialised architecture. They demonstrate that PPLM achieves desirable fluency (i.e., perplexity measured with GPT(-1) [Radford et al., 2018]), as well as measurable attribute control. Their architecture's applicability is also demonstrated on tasks such as controlled story writing and language detoxification. They also show a clear trade-off between attribute control and grammatical correctness and diversity.

2.2.3 Text Style Transfer

Text style transfer is the task of changing a text's stylistic properties, while retaining its style-independent properties, like content and fluency [Dai et al., 2019]. Text style transfer is a closely related problem to controlled text generation. Its similarity lies in trying to modify the output distribution of a text generation model, such that stylistic characteristics of the produced text are controllable, keeping content and fluency preserved. It involves rewriting an input text with a specific style. More formally, given a text $\mathbf{x}$, its corresponding style-representing vector $\mathbf{s}^{(i)}$, the number of different styles K over which there exists a distribution, and a desired style $\hat{\mathbf{s}} \in \{\mathbf{s}^{(i)}\}_{i=1}^{K}$, the goal of text style transfer is to produce output text $\hat{\mathbf{x}}$ with style $\hat{\mathbf{s}}$, and the style-independent properties of $\mathbf{x}$.

Previous approaches to text style transfer involve passing input text through an RNN-based encoder, yielding a style-dependent latent representation $\mathbf{z}$ [Zhang et al., 2018b]. Typically, these approaches then attempt to "disentangle" $\mathbf{z}$ into a style-independent content representation and a latent representation of the stylistic properties of the input text. The subsequent decoder then receives the content representation and a new latent style variable as input, to ultimately produce a style-altered output text with unchanged content. This style-disentanglement approach has a number of drawbacks: **(1)** It is difficult to evaluate the quality of disentanglement of the latent space. **(2)** It is hard to capture rich semantic information in the latent representation due to limited capacity of vector representations (especially for long texts). **(3)** To disentangle style and content in the latent representations, all previous approaches have to assume all input texts can be encoded by a fixed-size latent vector. **(4)** Since most previous approaches use RNN-based encoder-decoder frameworks, they have problems capturing long-range dependencies in the input sentences. Furthermore, disentanglement might be unnecessary, as Lample et al. [2019] have shown a proper decoder can perform controlled text generation from an entangled latent representation by "overwriting" the original style.

To address these drawbacks, Dai et al. [2019] propose Style Transformer, a Transformer-based alternative encoder-decoder framework for text style transfer. The authors' approach does not require any manipulation (i.e., disentanglement) of the latent space, eliminates the need for a fixed-size vector representation of the input, and handles long-range dependencies better due to Transformers' attention mechanism. Aside from this being the first application of Transformers for text style transfer, Dai et al. [2019] contribute a novel training algorithm for such models, that boasts significant improvements of results on two text style transfer datasets.

2.2.4 Dialogue Generation

As explained earlier in Section 2.1.4, dialogue generation is task of automatically generating a response given a user's prompt. Zhang et al. [2020] introduce DialoGPT, a tunable large-scale language model for generation of conversational responses, trained on Reddit discussion chain data. DialoGPT therefore extends GPT-2 [Radford et al., 2019] to address a more restrictive sub-category of text generation, i.e., conversational response generation. DialoGPT inherits from GPT-2 a 12-to-48 layer transformer with layer normalization, a custom initialization scheme that accounts for model depth, and byte pair encodings [Sennrich et al., 2016] as a tokenizer. The generation task remains framed as language modeling, where a multi-turn dialogue session is modeled as a long text.

To address the well-known problem of open-domain text generation models producing bland and uninformative samples, Zhang et al. [2020] implement a maximum mutual information (MMI) scoring function. MMI uses a pre-trained backward model to predict $p(\texttt{source}|\texttt{target})$: i.e., the source sentences (dialogue history) given the target (responses, dialogue continuation). First, top-K sampling is used to generate a set of hypotheses. Then the probability $p(\texttt{source}|\texttt{hypothesis})$ is used to re-rank all hypotheses. As frequent and repetitive hypotheses can be associated with many possible queries/sources (i.e., a hypothesis that frequently occurs is one that is apparently applicable to many queries), and maximizing backward model likelihood penalizes repetitive hypotheses, MMI yields a lower probability for highly frequent hypotheses, thereby reducing blandness and promoting diversity.

DialoGPT is evaluated on the Dialog System Technology Challenge (DSTC) 7 track, an end-to-end conversational modeling task in which the goal is to generate conversation responses that go beyond chitchat by injecting information that is grounded in external knowledge. The model achieves state-of-the-art results on both the human and automatic evaluation results, by achieving near human-like responses that are diverse, relevant to the prompt, much like GPT-2 for open-domain text generation. They train 3 models of small (117M), medium (345M), and large (762M) parameter sizes. The medium-sized 345M model achieves the best automatic evaluation results across most metrics, and is used as one of the baselines in later experiments in this thesis. Their Hugging Face PyTorch implementation can be tested here: `https://huggingface.co/microsoft/DialoGPT-medium`.

2.2.5 Controlled Dialogue Generation

As mentioned previously in Section 2.1.5, controlled dialogue generation is the task of steering automatically generated conversational responses to possess desired attributes, like sentiment, topic, or more abstract writing style characteristics. Zeng et al. [2020] explore the applications of fine-tuning large language models, like GPT, on (Mandarin and English) medical consultation data. The resulting dialogue systems succeed at generating clinically correct and human-like responses to patients' medical questions. Medical dialogue systems like these can help make healthcare services more accessible and aid medical doctors to improve patient care.

Zheng et al. [2019] investigate the problem of incorporating explicit personal characteristics in dialogue generation to deliver personalized conversation. They introduce a dataset `PersonalDialog`, which is a large-scale multi-turn dialogue dataset with personality trait labeling (i.e., `Age`, `Gender`, `Location`, `Interest Tags`, etc.) for a large number of speakers. Zheng et al. [2019] also propose persona-aware models that include a trait fusion module in the encoder-decoder framework to capture and address personality traits in dialogue generation.

Persona-aware attention mechanisms and bias are used to incorporate personality information in the decoding process. All their tested classification and dialogue generation models are either variations of RNNs (such as LSTMs or gated recurrent units (GRUs)), convolutional neural networks (CNNs), or hybrids of these systems (LSTM-outputs fed into a CNN, known as recurrent convolutional neural networks (RCNNs)). The authors study the influence of age, gender, and location on dialogue classification and generation, and use both automatic (perplexity, trait accuracy, and generated response diversity measures) and human evaluation. They find dialogues to be distinguishable by gender (about 90.61% test accuracy), then age (78.32% test accuracy), and finally location (62.04% test accuracy). Both automatic and human evaluation of the generated responses show that the best performing models benefit greatly from the persona-aware attention mechanism, possibly making a case to consider more attention-based architectures instead of RNNs.

Although the previously mentioned architectures are able to produce human-like conversational responses, sometimes even leveraging the fluency of large pre-trained LMs, they all suffer from the same computational drawback. They all require massive amounts of computational power to adapt their language styles, because in their cases, guided generation implies fine-tuning (or even retraining) large attribute-specific dialogue datasets. For general controlled text generation, this obstacle is overcome by Dathathri et al. [2020]'s previously mentioned PPLM setup. The conversational analog of this idea, plug-and-play conversational model (PPCM), is proposed by Madotto et al. [2020]. Similar to PPLM, PPCM achieves guided dialogue generation via activation-space perturbations using easy to train attribute models. Due to the computational complexity of PPLM's decoding process, PPLM is unusable as practical conversational system. PPCM solves this problem by using residual adapters [Bapna and Firat, 2019] to tweak the decoding procedure such that it does not require more computational resources. Madotto et al. [2020] show, using both human and automatic evaluation, that PPCM can balance grammatical fluency and high degrees of attribute-adherence in its generated responses. PPCM uses DialoGPT as its base language model, and is tested for topical or sentimental attributes (i.e., positive, negative, sports, business, or science & tech). Previous work on controlled text generation focuses on content (e.g., topical attributes, or sentiment), rather than more abstract linguistic features, which I hypothesize are more challenging to model and control. The previously mentioned work by Zheng et al. [2019] is a notable exception, as their approach deals with controlling dialogue systems for linguistic features, like age, gender, and geographical region.

However, Zheng et al. [2019] still suffers from significant computational costs, because control is achieved by fine-tuning a large system for every specific set of attributes.

The work presented in this thesis extends the applicability of the Transformer-based Plug-and-Play controlled generation model to more abstract writing styles, namely the linguistic characteristics associated with certain age groups. Furthermore, I apply this adaptation to the task of dialogue generation. As a preliminary research objective, I aim to use text-based NLP models to detect age-related linguistic features from dialogue and discourse data. This preliminary experiment is presented in the next chapter.

Chapter 3

Experiment 1: Detecting Age-Related Linguistic Patterns in Dialogue

In this chapter I report experiments aimed at age detection from text, and the components involved. While previous work showed differences at various linguistic levels between age groups when experimenting with written discourse data (e.g., blog posts), previous work on dialogue has largely been focused on acoustic information related to voice and prosody [Wolters et al., 2009, Li et al., 2013]. Detecting age-related linguistic properties of human dialogues is of crucial importance for developing AI based conversational systems which are able to adapt to the age-specific speaking style their human interlocutors. In particular, being able to detect and investigate these linguistic differences is important for controlled dialogue generation (the main topic of this thesis), as it can provide insights about which linguistic features are most salient for distinguishing between, and adapting to, different age groups. More generally, it is interesting in itself to understand which age-related linguistic features are most characterizing of age groups.

I therefore aim to investigate whether, and to what extent, current text-based NLP models can detect such linguistic differences, and what the features driving their predictions are. The age group detection experiment is performed on dialogue (i.e., transcribed open-domain spontaneous dialogue) and discourse (i.e., blog posts) text data, and on the former, an in-depth analysis is carried out about the saliency of identified age-related linguistic features.

Based on previous work on automated detection of age-related linguistic differences between age groups from written discourse (i.e., blog posts [Schler et al., 2006]) and open-domain texts [Abdallah et al., 2020], I expect that the classification models are able to reliably detect age-related differences in both transcribed dialogue and discourse, and the most informative differences to lie at the syntactic-level. However, I also hypothesize age group detection to be more challenging from the dialogue data than from the discourse data, based on earlier work on age group detection from short-form text (i.e., tweets [Nguyen et al., 2014]) and the shorter and noisier data-entries that constitute the dialogue data, when compared to the discourse data. Furthermore, I believe that the results about detection of age-related characteristics from dialogue (i.e., Experiment 1) can inform my work on age-adaptive dialogue generation, reported in the next chapter (Experiment 2).

The rest of this chapter is structured as follows: The following section describes the two datasets used for these experiments. There, I provide descriptive statistics, examples, and comparisons between the datasets. Section 3.2 covers the problem description in more detail, along with the models used, and the experimental setup. The classification results are presented in Section 3.3. Then for the dialogue classification models, Section 3.4 contains both quantitative and qualitative analyses of the results.

3.1 Data

One of the used datasets is one of dialogue data where information about the age of the speakers involved in the conversation is available (see the dialogue snippets in Figure 3.1), i.e., the spoken partition of the British National Corpus [Love et al., 2017]. It is henceforth referred to as the *dialogue* dataset. For comparison with previous work, and to explore commonalities and differences between various types of language data, I also experiment with a dataset of discourse data, i.e., the Blog Authorship Corpus used by Schler et al. [2006], that is henceforth referred to as the *discourse* dataset. Below, I briefly describe the two datasets along with the pre-processing steps taken to make the data suitable for the experiments.

dataset	# age groups	# samples	# tokens	mean length (±std)	min - max length	# topics
dialogue	2	67,282	787,352	11.7 (±19.0)	1 - 1246	790
discourse	3	678,165	137M	201.7 (±415.9)	1 - 131,169	40

Table 3.1: Descriptive statistics of the datasets used in my experiments. Length is the number of tokens in a sample.

age 19-29
A: oh that's cool
B: different sights and stuff
A: oh

age 50+
A: well quite and I'd have to come back as well
B: that's of course
A: and make up for you know

Figure 3.1: Example dialogue snippets from speakers of different age groups (19-29 vs. 50+) in the British National Corpus. I conjecture that stylistic and lexical differences between age groups can be detected. In my approach, experiments are conducted at the level of the single utterance.

3.1.1 Dialogue Dataset

This partition of the British National Corpus includes spoken informal open-domain conversations between people that were collected between 2012 and 2016 via crowd-sourcing, and then recorded and transcribed by the creators. Dialogues can be between two or more interlocutors, and are annotated along several dimensions including age and gender together with geographic and social indicators. Speaker ages in the original dataset are categorized in the following ten brackets: 0-10, 11-18, 19-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, and 90-99.

The focus is on conversations in the British National Corpus that took place between two interlocutors, and only dialogues between people of the same age group are considered. Furthermore, only dialogues are considered between speakers belonging to two age groups: 19-29 and 50+, in which the conversations are grouped from five original brackets: 50-59, 60-69, 70-79, 80-89, and 90-99. The intermediate age brackets are omitted to allow for clearer differentiation.

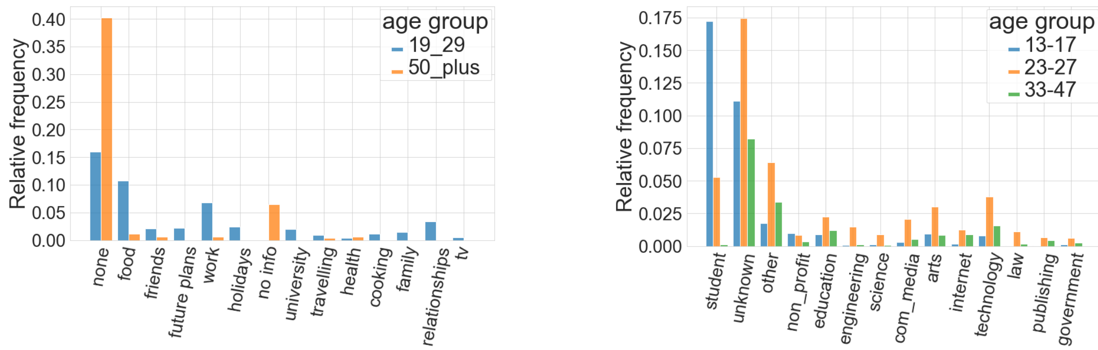
The dialogues are split into their constituent utterances (e.g., from each dialogue snippet in Figure 3.1 three utterances are extracted), and further pre-process them by removing non-alphabetical characters. Only samples which were not empty after pre-processing were kept. The resulting dialogue dataset, that is used for the experiments, includes around 67K utterances with an average length of 11.7 tokens. Descriptive statistics of it are reported in Table 3.1.

Each conversation in the British National Corpus is annotated with a list of *topics* provided by the speakers during data collection. However, it is desirable to obtain a better picture of how single representative dialogue topics are distributed among the age groups. Thus, to extract a single representative topic from this list, I first compute the frequency of all topic labels in the whole dataset. Then, for each utterance, I take the label in the conversation with the highest frequency in the ranking. In total, the final dataset includes 790 unique topic labels. The distribution of the most frequent ones is reported in Figure 3.2a. As can be seen, frequent topics (besides the frequent *none* label) are *food*, *work*, and *holidays*, which reveals the colloquial and everyday nature of the dialogues in this dataset.

3.1.2 Discourse Dataset

The Blog Authorship Corpus [Schler et al., 2006] is a collection of blog posts written on `https://www.blogger.com`, gathered in or before August 2004. Each blog entry is written by a single user whose age, gender, and astrological sign are reported. The corpus contains almost 700,000 posts by 19,000 unique bloggers (i.e., $\sim$35 posts per blogger on average). For my experiments, similar to Schler et al. [2006], three age groups are considered: 13-17, 23-27, and 33+. The data are pre-processed in the same way as described above, namely by removing non-alphabetical characters. The resulting dataset, that is use for the experiments, includes slightly more than 678K samples with an average length of 201.7 tokens. Descriptive statistics of it are reported in Table 3.1.

Each sample in the Blog Authorship Corpus is annotated with one topic. In my final discourse dataset, the unique topics present are 40. Figure 3.2b reports the distribution of the most frequent ones. As can be noted, frequent topics are *student*, *arts*, and *technology*, which reveals that this and the dialogue dataset are rather different with respect to the topic themes, input text length, and how well the topics are organized. Namely, the discourse dataset has a clear over-representation of topics related to student life, compared to the more colloquial and everyday subject matter of the dialogue dataset. Furthermore, the descriptive statistics in Table 3.1 show that the average data-entry length in tokens of the discourse dataset is almost 20 times larger than that of the dialogue dataset. Additionally, the differences number of unique topics (790 in dialogue dataset versus 40 in discourse dataset) between the two datasets illustrates the increased noisiness of the dialogue dataset compared to the discourse dataset. The shorter and noisier form of the dialogue dataset could make it more challenging to perform automated age group detection on its texts, as they are likely to carry less and more distorted discriminative signal.



(a) (b)

Figure 3.2: Distribution of most frequent topics shown by age group, in the **dialogue dataset** ((a) / left) and **discourse dataset** ((b) / right). Best viewed in color.

3.2 Methodology and Experimental Setup

The current section describes the methodology and experimental details of the automated age-detection experiments. I frame the problem as a N-class classification problem: given a fragment of text X, I seek to predict the age class of its speaker/writer. For the dialogue dataset, $N = 2$, while $N = 3$ for the discourse dataset. I experiment with various models, that are briefly described here below. Details on the training and evaluation of models are given at the end of the sub-section.

***n*-gram** The simplest models are based on n-grams, which have the advantage of being highly interpretable. Each data entry (i.e., a dialogue utterance or blog post) is split into chunks of all possible contiguous sequences of n tokens. The resulting vectorized features are used by a logistic regression model to estimate the odds of a text sample belonging to a certain age group. Specifically, experiments are performed with unigram, bigram and trigram models. Note that a bigram model uses unigrams and bigrams, and a trigram model uses unigrams, bigrams, and trigrams.

LSTM and BiLSTM A standard Long Short-Term Memory network [LSTM; Hochreiter and Schmidhuber, 1997] is used with two layers, embedding size 512, and hidden layer size 1024. Batch-wise padding is applied to variable length sequences. The original model's bidirectional extension, the bidirectional LSTM [BiLSTM; Schuster and Paliwal, 1997], is also used. BiLSTM more thoroughly leverages forward and backward directed information by combining the hidden states from both directions. Padding is similarly applied to this model, and the following optimal architecture is found via guided grid search: embedding size 64, 2 layers, and hidden layer size 512. Both RNN models are found to perform optimally for a learning rate of 10^{-3}.

BERT I also experiment with a Transformer-based model, i.e., Bidirectional Encoder Representations from Transformers [BERT; Devlin et al., 2019] for text classification. BERT is pre-trained to learn deeply bidirectional language representations from massive amounts of unlabeled textual data. The base, uncased version of BERT, is used in two settings: (1) by using its pre-trained frozen embeddings (BERT$_{frozen}$) and (2) by fine-tuning the embeddings on the age classification task (BERT$_{FT}$). The BERT embeddings are followed by a dropout layer with dropout probability 0.1, and a linear layer with input size 768.

Experimental Details Both datasets are randomly split into a training (75%), validation (15%), and test (10%) set. Each model with a given configuration of hyperparameters is run 5 times with different random initializations. All models are trained on an NVIDIA TitanRTX GPU.

The n-gram models are trained in a One-vs-Rest (OvR) fashion, and optimized using the Limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) algorithm [Liu and Nocedal, 1989], with a maximum of 10^6 iterations. The n-gram models are trained until convergence or for the maximum number of iterations.

LSTMs and BERT-based models are optimized using Adam [Kingma and Ba, 2015], and trained for 10 epochs, with an early stopping patience of 3 epochs. The RNN-based models' embeddings are jointly trained, and optimal hyperparameters (i.e., learning rate, embedding size, hidden layer size, and number of layers) are determined using the validation set and a guided grid-search. BERT$_{FT}$ is fine-tuned on the validation set for 10 epochs, or until the early stopping criterion is met. BERT models have a maximum input length of 512 tokens. Sequences exceeding this length are truncated.

3.3 Detecting Age-Related Linguistic Patterns in Dialogue and Discourse

First, I report the results on *discourse* to check whether previous findings [Schler et al., 2006] are replicated. Then, I focus on *dialogue* to answer my research questions. The accuracy and F_1 of the classifiers are reported for each age group.

3.3.1 Classification Performance on Discourse

Table 3.2 reports the results. As can be seen, all models are well above the majority class baseline in terms of both accuracy (0.472) and F_1s (0.642). This overall confirms previous evidence [Schler et al., 2006] that language features of (written) *discourse* can predict, to some extent, the age group to which the person belongs. At the same time, BERT fine-tuned on the age classification task stands out as the best-performing model by achieving highest accuracy (0.742) and highest F_1 in all age groups. LSTM ranks second (0.663) in terms of accuracy, while the second best F_1 scores for all age groups are distributed among the bigram and trigram models. Overall, these results indicate that powerful neural models that are capable of representing the linguistic context have a great advantage on this dataset over simpler n-gram models, which are more than 10 accuracy points behind.

Finally, it should be noted that my best results are slightly lower than those obtained by Schler et al. [2006]. This could be due to two main reasons: First, they experiment with a differently pre-processed and smaller dataset than mine.[1] Second, while in my approach all models are trained end-to-end on the task, Schler et al. [2006] use handcrafted features that are specific to the dataset, which could constitute an advantage.

3.3.2 Classification Performance on Dialogue

Table 3.3 reports the results obtained on the dialogue dataset. As can be seen, BERT fine-tuned on the task is again the best-performing model in terms of accuracy (0.729), which confirms the effectiveness of this model in detecting age-related linguistic differences. At the same time, it can be noted that the model based on trigrams is basically on par with it in terms of accuracy (0.722) and well above both LSTM and BiLSTM (0.693 and 0.691, respectively). A similar pattern is shown for F_1 scores, where BERT fine-tuned and the trigram model achieve comparable performance, with LSTMs being overall behind.

Overall, my results indicate that predicting the age group to which a speaker belongs, using text-based models, is possible also for *dialogue* data, though the task appears to be somehow more challenging compared to when performed on discourse. Note that the improvement with respect to the majority/random baseline is lower in dialogue (22.9 and 27 accuracy point increase w.r.t baselines for best model for dialogue and discourse datasets, respectively).

[1] They are left with roughly 511K datapoints after pre-processing, while I experiment with around 677K.

At the same time, the different ranking of models observed between discourse and dialogue suggests possibly different strategies used by models to solve the task. In particular, the very good performance of the trigram model in *dialogue* suggests that leveraging 'local' linguistic features captured by n-grams is extremely effective in this setup. This could indicate that differences among various age groups are at the level of local lexical constructions. This deserves further analysis, which is carried out in the next section.

Model	**Accuracy** ↑ better	$\boldsymbol{F}_1^{(13-17)}$ ↑ better	$\boldsymbol{F}_1^{(23-27)}$ ↑ better	$\boldsymbol{F}_1^{(33+)}$ ↑ better
Majority class	0.472	*	0.642	*
Schler et al. [2006]**	0.762	0.860	0.748	0.504
unigram	0.601 (0.001)	0.764 (0.001)	0.704 (0.001)	0.498 (0.003)
bigram	0.625 (0.001)	0.790 (0.001)	0.712 (0.001)	0.518 (0.001)
trigram	0.623 (0.001)	0.790 (0.001)	0.712 (0.002)	0.498 (0.002)
LSTM	0.663 (0.005)	0.748 (0.003)	0.664 (0.010)	0.502 (0.004)
BiLSTM	0.618 (0.008)	0.732 (0.003)	0.579 (0.016)	0.509 (0.004)
BERT$_{frozen}$	0.623 (0.002)	0.658 (0.006)	0.678 (0.007)	0.256 (0.041)
BERT$_{FT}$	**0.742** (0.010)	**0.813** (0.007)	**0.749** (0.013)	**0.592** (0.009)

Table 3.2: Discourse dataset. Test set results averaged over 5 random initializations. Format: *average metric (standard error)*. Values in **bold** are the highest in the column; in blue, the second highest. *: F_1 is actually $0/0$. **: these results were obtained with a different final dataset.

Model	**Accuracy** ↑ better	$\boldsymbol{F}_1^{(19-29)}$ ↑ better	$\boldsymbol{F}_1^{(50+)}$ ↑ better
Random	0.500	0.500	0.500
unigram	0.701 (0.007)	0.708 (0.009)	0.693 (0.004)
bigram	0.719 (0.002)	0.724 (0.003)	0.714 (0.003)
trigram	0.722 (0.001)	0.727 (0.003)	0.717 (0.001)
LSTM	0.693 (0.003)	0.696 (0.005)	0.691 (0.007)
BiLSTM	0.691 (0.009)	0.702 (0.017)	0.679 (0.007)
BERT$_{frozen}$	0.675 (0.003)	0.677 (0.008)	0.673 (0.010)
BERT$_{FT}$	**0.729** (0.002)	**0.730** (0.011)	**0.727** (0.010)

Table 3.3: Dialogue dataset. Test set results averaged over 5 random initializations. Format: *average metric (standard error)*. Values in **bold** are the highest in the column; in blue, the second highest.

3.4 Age Detection Analyses

Predominantly with the goal of obtaining insight about age-related distinguishing features in dialogue that can inform the subsequent experiment on age-adaptive dialogue generation

actual age	both correct	both wrong	BERT$_{FT}$ correct ∣ trigram wrong	trigram correct ∣ BERT$_{FT}$ wrong
19-29	oh that's cool	A retrospective exhibition	what even on the green slope?	really?
19-29	a text and then I'll do it	chuck them in those pots	yeah you told me to do you told me to do	and she like won't eat any carbs and she's like
19-29	yeah	mm	somebody made the f***ing table	do you not like total greens?
50+	I said no I don't have them	yeah	really?	my under stairs in the kitchen
50+	that's of course	no no that's alright	it's still we we frequently walk that way	in the first place
50+	oh right	what a tragic life	since this this was new this house?	thank you very much

Table 3.4: Examples where both models are correct/wrong or only BERT$_{FT}$/trigram is correct. Illustrative example: the sentence, *A retrospective exhibition* was uttered by a speaker of age 19-29, but was incorrectly classified by both models (BERT$_{FT}$ and trigram) as coming from a person age 50+.

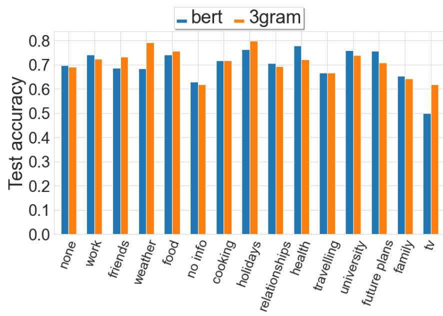


Figure 3.3: BERT$_{FT}$ and trigram test accuracies per topic for most frequent topics (including none/no info).

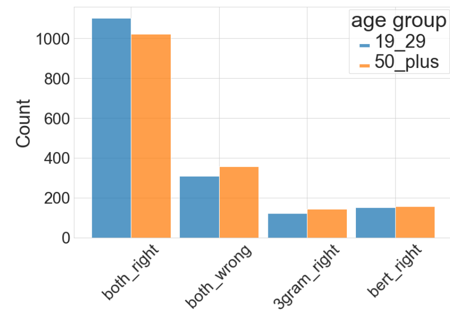


Figure 3.4: Distribution of predicted cases by trigram and BERT$_{FT}$ models for dialogue, split by age groups.

(Experiment 2 in Chapter 4), the analysis is focused on the classification results obtained on the dialogue dataset. In particular, I compare the two best-performing models, namely BERT$_{FT}$ and the one using trigrams, and aim to shed light on what cues they use to solve the task. First, I analyze how these models perform with respect to utterances of various topics. Secondly, I compare the prediction patterns of the two models, which allows for highlighting of easy and hard examples. Finally, I focus on the trigram model and report the n-grams that turn out to be most informative to distinguish between age groups.

3.4.1 Performance Against Topic

As described in Section 3.1.1, each utterance in the dialogue dataset is annotated with one label which is representative of its topic.[2] This information is not explicitly available to the models. To explore how the two models deal with utterances in different topical contexts, I compare the accuracy they achieve on the 15 most frequent topics. The results are shown in Figure 3.3. Two main observations can be made: Firstly, some topics turn out to be generally easier/harder than others, i.e., both models achieve higher/lower performance. To illustrate, both models achieve an accuracy well above 70% on topics like *food*, *holidays* or *university*, while topics such as *tv*,

[2]In particular, it represents one of the utterance's topic, i.e., the one most frequently used in the whole data.

family or *no info* appear to be generally more challenging for both models. While this could be due to (a combination of) various factors, one intuitive possibility is that certain topics allow for more discriminative language features, which could be at the level of the lexicon or the style used to talk about them.

Secondly, some topics appear to be easier for one model rather than the other, and *vice versa*. To illustrate, the trigram model outperforms BERT on the topics *weather*, *holidays* and *tv*, while an opposite pattern is observed for *work*, *health*, and *future plans*. I conjecture that these patterns could be indicative of different strategies and cues exploited by various models to make a prediction.

3.4.2 Comparing Model Predictions

The data for analysis is split by whether or not both models make the same correct or incorrect prediction, or whether they differ. Table 3.6 shows the breakdown of these results. As can be seen, a quite large fraction of samples are correctly classified by both models (63.17%), while in 19.78% cases neither of the models make a correct prediction. The remaining cases are comparably split between cases where only one of the two is correct, with BERT slightly outperforming the trigram model by 1.23 percentage points. As shown in Figure 3.4, the 19-29 age group appears to be be slightly easier compared to the 50+ group, where models are observed to make more errors: the trigram misclassifies 50+ utterances 1.12 times as often as 19-29 utterances, and 1.17 times as often by BERT$_{FT}$.

To qualitatively inspect what the utterances falling into these classes look like, in Table 3.4 a few cherry-picked cases are shown for each age group. It is observed that, not surprisingly, both models have trouble with backchanneling utterances consisting of a single word, such as *yeah*, *mm*, or *really?*, which are used by both age groups. For example, both models seem to consider *yeah* as a 'younger' cue, which leads to wrong predictions when *yeah* is used by a speaker in the 50+ group. As for the utterance *really?*, BERT$_{FT}$ assigns it to the 50+ group, while the trigram model makes the opposite prediction. This indicates that certain utterances simply do not contain sufficient distinguishing information, and model predictions that are based on them should therefore not be considered reliable.

This seems to be particularly the case for short utterances. Indeed, through comparing the average length of the utterances incorrectly classified by both models (rightmost column of Table 3.6), it is seen that they are much shorter than those belonging to the other cases. This is interesting, and indicates a key challenge in the analysis of dialogue data: on average, shorter utterances contain less signal. On the other hand, short utterances can provide rich conversational signal in dialogue; for example, backchanneling, exclamations, or other acknowledging acts. As a consequence, using length alone as a filter is not an appropriate approach, as it can remove aspects of language use key to differentiating speaker groups.

3.4.3 Most Informative N-grams

	19-29		50+
coef.	**n-gram**	**coef.**	**n-gram**
-3.20	um	2.37	yes
-2.84	cool	2.12	you know
-2.58	s**t	2.09	wonderful
-2.12	hmm	1.90	how weird
-2.09	like	1.84	chinese
-2.02	was like	1.73	right
-1.96	love	1.71	building
-1.96	as well	1.66	right right
-1.88	as in	1.55	so erm
-1.84	cute	1.43	mm mm
-1.82	uni	1.41	cheers
-1.79	massive	1.39	shed
-1.79	wanna	1.37	pain
-1.79	f**k	1.36	we know
-1.72	tut	1.08	yeah exactly

Table 3.5: For each age group, top 15 most informative n-grams used by the trigram model. **coef.** is the coefficient (and sign) of the corresponding n-gram for the logistic regression model: the higher its absolute value, the higher the utterance's odds to belong to one age group. * indicates masking of foul language.

Analyzing the most informative n-grams used by the trigram model allows for qualitative comparison between the linguistic differences inherent to each age group. In Table 3.5, the top 15 most informative n-grams per group are reported. It appears that, primarily and intuitively, that colloquial language seems somewhat generational, with unigrams particularly indicative of younger speakers consisting of words such as *cool* and *massive*, and for older speakers, words like *wonderful*. These unigrams are both informative to the model and indicative of differences in both formality and 'slang' use across age groups.

These most informative n-grams also indicate differences in back-channeling use between age groups; younger speaker's language is more characterized by the use of *hmm*, *um*, *yeah course*, while the top n-grams in the older category will more likely use *yes*, *right*, *right right*. A feature of younger language also apparent from these examples is in their use of more informal language: *yeah course* rather than *yes*. This informal language use also extends to the use of foul language, which make up a percent of the most informative unigrams shown in Table 3.5.

Interestingly, while topic words make up many of the most informative n-grams for older speakers in Table 3.5, younger speakers are more defined by their use of slang words such as *wanna*, foul language, or adjectives such as *cute*, *cool*, and *massive*. A key finding from Schler et al. [2006] is in the sentiment of language playing an important role, something which some of the most informative n-grams suggest may also be true for the dialogue dataset. As Table 3.5 demonstrates, younger speakers use more dramatic language such as negative foul words, and positive *love, cute, cool*; all words with a strong connotative meaning. This prompts me to hypothesize that further inspection is needed to determine whether the same sentiment pattern will be true of dialogue as it has been reported to be in discourse.

	% cases	**avg. length (±std)***
both correct	63.17%	13.51 (±18.98)
both wrong	19.78%	5.82 (±8.33)
only Trigram correct	7.91 %	10.44 (±11.66)
only BERT correct	9.14 %	11.53 (±12.12)

Table 3.6: Percentage (% cases) of (non-)overlapping (in)correctly predicted cases between trigram and BERT$_{FT}$. *Utterance length measured in tokens.

3.5 Discussion and Conclusion

In this chapter, I studied the extent to which purely text-based NLP models can detect age-related linguistic features in dialogue and discourse data. The experiment on the discourse dataset confirmed and extended the work of Schler et al. [2006] by using end-to-end trained NLP models, instead of handcrafted features, and by applying this approach to transcribed spontaneous open-domain dialogues. Focusing on the dialogue dataset, I subsequently carried out and in-depth analysis of the best performing models' predictions to gain insights about the elements of language deemed most informative when classifying dialogue utterances into age groups. In line with what was observed for discourse, it is shown that state-of-the-art NLP models (in particular a fine-tuned version of BERT) are capable of distinguishing between dialogue utterances from different age groups with reasonable accuracy, in particular when the utterance is long enough to contain discriminative signal. However, different from what was observed for discourse data, it was found that much simpler models based on n-grams achieve comparable performance to fine-tuned BERT. This finding suggests that, in dialogue, 'local' features can be indicative of the language of speakers from different age groups. Qualitative inspection seems to confirm this postulation, as I found both lexical and stylistic features to be informative to both the BERT-based and trigram-based models.

Now that the feasibility of detecting age-related linguistic features in dialogue has been established, I aim to endow generated dialogue responses with similarly detectable linguistic features that align with target age groups. The results obtained on the dialogue dataset can inform the subsequent experiment about age-adaptive dialogue generation. Primarily, they provide insights about the age-related linguistic features that are most salient when distinguishing between, and subsequently adapting to, different age groups. Furthermore, the findings motivate the use of local, lexical level linguistic features, when controlling dialogue generation towards a target age group, as a comparison to more syntax-level adaptation.

Chapter 4

Experiment 2: Age-Adaptive Dialogue Generation Using PPLM

4.1 Introduction

4.1.1 Recap of Previous Chapter and Connection to Experiment 1

In Experiment 1, I studied the extent to which text-based NLP models are able to detect speaker and writer age-related linguistic features in dialogue and discourse data. Then, I investigated which features drive the predictions made by those models. It is shown that a fine-tuned version of BERT, BERT$_{FT}$, is capable of classifying language by speakers from different age groups based on linguistic features. it is also found that much simpler models based on n-grams achieve classification performance comparable to that of BERT$_{FT}$, suggesting that, in dialogue, local features (i.e., at the lexical level) can also be indicative of a speaker's age. This was shown to be the case, as both lexical and stylistic cues seem to be informative to these models when predicting speaker age from dialogue utterances.

Now, in Experiment 2, I aim to test whether it is possible to generate dialogue responses that possess these age-indicative features, identified in Experiment 1. More specifically, I seek to use Plug-and-Play language models (PPLM) [Dathathri et al., 2020] to develop controlled dialogue generation models that can produce conversation responses that contain linguistic elements of the speaking style of a certain age group to the extent that it would be classified as such by Experiment 1's best classifier. The use of the PPLM method is motivated by its capability to control the style of the text generated by a large pre-trained Transformer-based language model (e.g., GPT-2 or DialoGPT), without the computationally expensive requirement to re-train or fine-tune it. Namely, the Plug-and-Play approach uses substantially smaller and less costly to train attribute models to make perturbations to the large language model's activation space, thereby shifting its output distribution towards a desired style. It is important to realize that the PPLM method is fundamentally different from fine-tuning in that it leaves the parameters of the underlying language model unchanged.

The analyses of Experiment 1 largely concerned a comparison of performance between simpler n-gram-based discriminators and more sophisticated neural ones (i.e., BERT). This comparison is continued throughout Experiment 2, because now I aim to compare the applicability of PPLM for age-adaptation when adaptation is enforced by an aforementioned attribute model, which can be a simple bag-of-words (BoW) model, or a more complex neural discriminator. Throughout this thesis, using PPLM to control the writing style of generated output text using the former is referred to *BoW-based control*, and using the latter is referred to as *discriminator-based control*. The insights gained from Experiment 1 (e.g., about the applicability of n-gram-based representations of age-specific speaking style) are also used to inform decisions about the development of attribute models. Furthermore, in Experiment 1, the classifiers are trained on the spoken component of the BNC, which is a dialogue dataset. Now, in Experiment 2, I aim to generate something similar to a dialogue turn, i.e., a response to a dialogue prompt, in the style of a certain age group. Table 4.1 shows examples of how age-adapted dialogue responses to a prompt might look.

PROMPT	Can we talk?
Age 19-29 style response	Like about what?
Age 50+ style response	What would you like to share about?

Table 4.1: Examples of controlled dialogue responses to a prompt (in this case, a dialogue turn). The responses have been generated by DialoGPT after being controlled to generate younger sounding (red row) and older sounding (blue row) language.

It is also important to realize that the previously seen classifiers from Experiment 1 are not compatible with PPLM, because the method requires a bag-of-words or linear discriminator. Furthermore, one of the goals of Experiment 1 was to develop the best performing age-classifiers, which are ultimately used for (1) evaluation of the attribute adherence of the generated responses in Experiment 2, and (2) informing decisions about the choice of BoW-based attribute models by using a BoW comprised of the unigram classifier's most informative unigrams (see Section 4.2.1

for further details). It also improves the generalizability of the results to use a separate classifier for evaluation than the discriminators used as attribute models.

Lists of unigrams are used as BoWs (as opposed to lists of trigrams from the best performing n-gram-based classifier in Table 3.3), because the PPLM setup is not compatible with lists of n-grams for $n > 1$, as it relies on perturbations at the unigram-level. Making a PPLM-system compatible with, e.g., trigrams, would amount to retraining the entire underlying language model (like GPT-2), thereby defeating the purpose of PPLM, i.e., leveraging large LMs for controllability, without incurring significant re-training costs. However by allowing the best-performing classifiers to inform decisions about generation, the best unigram-based classifier's list of most informative features is used as an attribute model.

4.1.2 Research Objectives, Hypotheses, and Contributions

Experiment 2 concerns a Plug-and-Play approach to age-adaptive dialogue generation. The aim is to use PPLM for age-adaptive dialogue generation. Based on the previously demonstrated detectability of age-related linguistic patterns in dialogue by text-based NLP models, I hypothesize age-adaptive dialogue generation to be possible to the extent that similar patterns can be discerned from generated dialogue responses. Furthermore, based on previous work on modeling age-related characteristics in language [Pennebaker and Stone, 2003, Schler et al., 2006, Nguyen et al., 2014, Zheng et al., 2019], I presume age-related linguistic style to be a more abstract attribute than the sentiment and topic attributes previously studied by Dathathri et al. [2020] and Madotto et al. [2020]. This implies that it is probably more challenging to accurately express this attribute as a BoW, than it is to represent it in the latent space of a neural discriminator. Dathathri et al. [2020] suggest that for such attributes that are difficult to accurately express as wordlists, more sophisticated attribute models (i.e., discriminators) are desirable (see Section 2.1.9). It is therefore expected of discriminator-based control to be more detectable than BoW-based control, because a discriminator attribute model is probably able to capture a more comprehensive representation of age-related speaking style, and thus more accurately enforce the linguistic style during generation. However, I also expect this heightened degree of control exerted by discriminator attribute models, combined with the more lexical level control associated with the BoW attribute models (see Equation 2.9), to result in BoW-based control taking a smaller toll on the fluency of generated responses.

The Plug-and-Play approach has previously been used by Dathathri et al. [2020] and Madotto et al. [2020] for language generation, controlled for very concrete writing styles and topics (e.g., negative/positive sentiment, politics, religion, business, tech). I apply the PPLM method to a novel problem, that is, age-related language generation conditioned on dialogue data. More concretely, I extend the work of Dathathri et al. [2020] in several important ways: (1) I control the generated language for more abstract writing styles, i.e., age-related linguistic style; (2) I use PPLMs to generate dialogue responses; (3) I propose two empirical methods for age-specific BoW development, as opposed to the manually constructed BoWs used in previous work; (4) I carry out in-depth analyses about the relationships between evaluation measures of dialogue response quality and style; (5) finally, I address and study the problem of writing style biases induced by the prompt's style.

It must be noted that the activation perturbation steps necessary for PPLMs to exert influence on the style of generated text makes the PPLM-approach too slow to be suitable for online interactive conversational applications. This problem is addressed by Madotto et al. [2020], who propose Plug-and-Play Conversational Model (PPCM), an extension of PPLM that uses separately trained residual adapter modules to speed up the PPLM-method, making it suitable for real-time dialogue response generation. Despite their work also being about controlled dialogue generation, this work differs in important ways. They control their output for the same styles and topics as Dathathri et al. [2020], and do not experiment with more abstract writing styles, like age-related linguistic characteristics. Overall, this thesis is about establishing whether age-related linguistic features are automatically detectable, and if they can be enforced as a writing style of dialogue responses using PPLMs. The work of Madotto et al. [2020] addresses the problem of interactivity (i.e., speed of response generation), which requires solving a separate, very valuable, engineering problem, which is beyond the scope of this thesis. Nevertheless, the previously mentioned contributions still hold.

The rest of this chapter is structured as follows. Section 4.2 describes the methodology and experimental setup of the controlled dialogue generation experiments. It also covers details about attribute model development, the choice of prompts, and an explanation of the chosen evaluation metrics. The results of the language generation experiments are presented and interpreted in Section 4.3, followed by the outcomes of various quantitative and qualitative analyses in Section 4.4. A separate section about data is omitted in this chapter, because the previously mentioned dialogue dataset is used for these experiments. The reader is directed to Section 3.1.1 for a detailed description of that corpus, and the pre-processing steps that have been taken.

4.2 Methodology and Experimental Setup

This section describes and motivates the methodology, experimental details of my controlled dialogue generation experiments, and the most important components involved: attribute model development, dialogue response generation, choice of prompt, and evaluation of generated responses. All computations for attribute model development, dialogue response generation, and evaluation are performed on an NVIDIA TitanRTX GPU.

Controlled dialogue generation experiments are performed using PPLM-setups that differ with respect to **(1)** pre-trained Transformer-based language model (GPT-2 or DialoGPT), **(2)** type of attribute model (BoW or discriminator), **(3)** style attribute (younger, older, or uncontrolled), and **(4)** the style of the prompt on which generation is conditioned (younger, neutral, or older). Note that when the style attribute is "uncontrolled", it implies the unaltered original underlying language model (i.e., GPT-2 or DialoGPT) is recovered.

4.2.1 Attribute Model Development

As mentioned in the introduction to this chapter, in the PPLM-method, controlled generation is achieved by using an attribute model to make activation space perturbations to an underlying language model. This alters the output distribution of the language model to make it more likely to generate text that aligns with a predefined writing style. Such an attribute model is significantly smaller and less costly to train than the underlying large-scale language model (e.g., GPT-2 or DialoGPT). An attribute model can either be a simple bag-of-words, or a more complex neural discriminator. When developing an attribute model, either a discriminator is trained on the dialogue data, or a bag-of-words is statistically constructed from the same dialogue dataset (see Section 3.1.1).

Bag-of-Words A simple bag-of-words can also serve as an attribute model. I extend the approach of Dathathri et al. [2020] that relies on curated wordlists, by applying two empirical approaches to extract wordlists from the dialogue corpus. An empirical approach has the benefit of being more reproducible, and not requiring a domain expert to manually curate a list. In the first approach, the BoW consists of the 100 most informative unigrams of the unigram classifier used during the text classification experiments (see Table 3.3 for the results). The most informative unigrams per age group are deemed the most distinguishing features by the unigram classifier. They could therefore be used to make sensible perturbations to a language model's output, yielding more effective control.

The second method of wordlist extraction is fully frequency-based. The goal of this extraction method is to yield two distinct sets of words that are representative of each age group's language. The frequency-based wordlists per age group are constructed from the *imbalanced* dialogue dataset. Recall that the 19-29 age group is over-represented in the original imbalanced dialogue dataset, and utterances from that class are discarded to balance the dataset when using it to train classifiers. The use of the original dataset for wordlist extraction is motivated by the fact that the original dataset contains more utterances, and a frequency-based wordlist from that corpus would be a better representation of the younger age group's speaking style. Wordlist extraction is carried out as follows: Per age group, all unique words are ordered by frequency of occurrence in all utterances. For both ordered lists of word counts, the most frequent words are kept that make up for at least 85% of the cumulative occurrences. Then, the words are removed that appear in both lists (i.e., the overlapping set of words is discarded). Of the resulting two non-overlapping ordered lists of words and their numbers of occurrences, only the words are kept that account for at least 85% of the respective wordlist's cumulative occurrences. The resulting lists consist of 56 (younger age group related), and 92 (older age group related) words. The 85-th percentile cutoff points are chosen to yield wordlists of similar lengths as those used by Dathathri et al. [2020]. Both pairs of wordlists are included in Section A.1.

Discriminators When training the discriminators, the dialogue dataset is randomly split into a training (90%) and test (10%) set. Note that there is no validation set, because the discriminator attribute models have no hyperparameters to be tuned. Namely, they are single-layer linear classifiers, whose input layer size is equal to the output layer size of the underlying language model (GPT-2 or DialoGPT). Furthermore, the dimensionality of the linear layer can not be varied once an underlying language model has been chosen, as it must be of equal size to the LM's output layer. Hence, the linear layer's dimensionality is not a changeable hyperparameter. Moreover, the effects of random initializations are found to be negligible, most likely due to the comparatively small size of the attribute model discriminator. The frozen embeddings of either GPT-2-medium [Radford et al., 2019] or DialoGPT-medium [Zhang et al., 2020] are fed into trainable linear classifiers, seeking to distinguish between transcribed dialogue utterances from younger (ages 19 to 29) and older (ages 50 and over) speakers. The discriminators are 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[1], with a maximum sequence length of 512 tokens, for 20 epochs, and a batch size of 64. The discriminator parameters that are used are those from the epoch with the highest test accuracy (67.4% for GPT-2-based discriminator, and 67.6% for DialoGPT-based discriminator).

4.2.2 Choice of Prompts

The prompt is the text on which a language generation model conditions its output. In the specific case of controlled language generation, given a prompt, `prompt`, a predefined style attribute a, and some controlled language generation model parameterized by θ, generating a style-controlled piece of text **x** entails modeling $p_\theta(\mathbf{x}|a, \texttt{prompt})$. The output distribution of the controlled generation model p_θ is therefore likely to be affected by `prompt`, which could manifest itself as the prompt's style influencing the output text's style. In line with this presumption, Fan et al. [2018] and Lester et al. [2021] show that the content and style of prompts can influence the output of neural text generation models.

Because this research is about the style of generated dialogue responses, it is desirable to take into account the effect that the style of the prompt might have. Therefore, a distinction is made between three categories of prompt style: younger, neutral, and older. Three random initializations of BERT$_{FT}$ classifiers from Experiment 1 (different from the one used for evaluation) are used to assign target probabilities to handcrafted prompts. A younger prompt is defined as having an average assigned "younger" probability (i.e., BERT$_{FT}$'s assigned probability of containing features learned to be associated with the younger age group) of at least 80%, with a standard deviation no greater than 0.15, to ensure that the prompt is deemed sufficiently characteristic of an age group's speaking style with certainty. The same holds for an older prompt, but for its assigned "older" probability. A neutral prompt must have an average assigned younger (or older, as they are complementary values) probability between 40% and 60%, with a standard deviation no greater than 0.15. Five prompts per style class are used, and they are shown, along with their average target probabilities and standard deviations, in Table 4.2.

The focus of the main quantitative results (covered in Section 4.3) is to see how much the writing style can be adapted towards that of an age group, so it uses neutral prompts, to avoid

[1] `https://pytorch.org/docs/stable/generated/torch.optim.Adam.html#torch.optim.Adam`

Prompts	Average P_Y	Average P_O	Standard deviation
Young prompts			
What are your plans this week?	0.938	0.062	0.056
What do you wanna eat?	0.958	0.042	0.026
Do you have any hobbies?	0.963	0.037	0.021
Can I add you on Facebook?	0.995	0.005	0.005
Awesome! I actually haven't been there. When did you go?	0.995	0.005	0.004
Neutral prompts			
Good weather we're having.	0.489	0.511	0.133
Can we talk?	0.412	0.588	0.116
Hi, how's it going?	0.437	0.563	0.079
Hello, tell me about your latest holiday.	0.496	0.504	0.059
Hey.	0.574	0.426	0.046
Old prompts			
Tell me about your family.	0.061	0.939	0.050
Good afternoon.	0.040	0.960	0.044
I had a splendid weekend.	0.030	0.970	0.034
Hello, how are you?	0.041	0.959	0.031
Hello, tell me about yourself.	0.025	0.975	0.026

Table 4.2: Younger (red rows), neutral (yellow rows), and older (blue rows) prompts used for controlled dialogue generation in Experiment 2. Average P_Y and P_O are the averaged young and old probabilities assigned to the prompts by three initializations of BERT$_{FT}$.

any unwanted biases induced by the prompts. The effects of using strongly styled prompts are analyzed in Section 4.4.3. It is worth mentioning that it is also possible to use PPLM in an unprompted setting, i.e., unconditioned language generation. In practice, this is done by conditioning the generated output on the <|endoftext|> token.

4.2.3 Experimental Setup of Dialogue Generation Models

Every PPLM-setup generates 30 sequences per output length 6, 12, 24, 30, 36, 42, 48, 54, and 60 tokens. The lengths are chosen to provide sufficiently distanced sub-samples for later analyses about the effects of response length (in tokens) on the quality and control of the PPLM-setups. Sub-sample sizes of 30 are chosen to satisfy the Central Limit Theorem (CLT), making it possible to assume the distribution of sub-sample averages $\bar{X}_n$ approximates the normal distribution with mean μ and variance $\frac{\sigma^2}{n}$ [CLT, 2008], where n is the sub-sample size. Note that perturbations of the base language model's output can affect the controlled sequence length, so the final sequence length may differ by a few tokens from the given one. All other PPLM-parameters are kept at their default values, as recommended by Dathathri et al. [2020].

Each reported PPLM-configuration represents the best initialization, if the term applies. The pre-trained language models are kept equal across configurations, as using different initializations of these large models is infeasible, and would defeat the purpose of PPLM. A BoW-based setup uses a single list of words as an attribute model, thereby not having random parameters to initialize. And finally, discriminator-based setups use comparatively small linear classifiers (a few hundred parameters), the initialization effects of which have been found to be negligible on performance.

4.2.4 Evaluation

The generated sequences are evaluated along two main axes: fluency and control. Fluency refers to the degree to which a text passage appears natural, grammatical, and non-repetitive. Control is the extent to which the produced language resembles that of the attribute being controlled for. Evaluation is done using a set of automated metrics. Fluency is measured automatically by perplexity (denoted ppl) expressed as

$$\mathrm{ppl}(\mathbf{x}) = \exp\left\{-\frac{1}{t}\sum_{i}^{t}\ln p_\theta(x_i|x_{<i})\right\} \tag{4.1}$$

where **x** represents a sequence of tokens, t is sequence length, x_i is the i-th token, and θ denotes GPT's parameters. Perplexity is a measure of a language model's uncertainty when posed with the task of predicting a succession of words. Assuming a language model to be a reliable representation of relationships within a natural language, low perplexity can serve as a rough proxy for fluency of a text. However, a major caveat of perplexity is that it only measures uncertainty w.r.t. one language model, making it less generalizable. To slightly reduce this effect, perplexity is evaluated by a different language model (GPT-1 [Radford et al., 2018]) than the one used for generation (GPT-2 or DialoGPT). Furthermore, the normalized number of distinct unigrams (Dist-1), bigrams (Dist-2), and trigrams (Dist-3), are used as measures of text diversity.

Experiment 1's best BERT-classifier's accuracy on a set of sequences generated by a single generation model is used as an automated measure of attribute control. It can be seen as a proxy for control, because it indicates how resemblant of an age group's speaking style a generated text is deemed to be.

Two types of baselines are used when evaluating text generation performance: an uncontrolled pre-trained model baseline (i.e., a GPT-2-baseline or DialoGPT-baseline), and a corpus-specific baseline. Therefore all controlled generation models that use GPT-2 as their language model, should be compared to the uncontrolled GPT-2 baseline. The same holds for DialoGPT. The second type of baseline combines the underlying language model with a bag-of-words consisting of the 100 most common words in the balanced dialogue corpus, irrespective of age. This setting is included to give an indication of how biased the balanced BNC's frequently occurring words might be towards a specific age group.

The deliberate choice to use BERT-based models for classification (and evaluation of generated output), and GPT-based models for generation is motivated by the following reasons. BERT's encoder-based bidirectional architecture makes it more suitable for sequence classification than for generation [Devlin et al., 2019]. By similar reasoning, GPT's decoder-based structure makes it a more suitable choice for generation tasks. Furthermore, Experiment 1's best classifier is a fully fine-tuned BERT model (ca 110M parameters). However, it is computationally highly expensive to fine-tune GPT-2-medium (ca 345M parameters) which makes it infeasible for this work, given the available computational resources. Finally, using a separate model class (i.e., BERT) for evaluation of GPT-based generation models makes the results more generalizable.

4.3 Controlled Dialogue Generation Results

The quantitative results of generating younger (ages 19 through 29) and older (ages 50 and over) sounding responses to *neutral* prompts are reported in Tables 4.3 and 4.4, respectively. The metrics per row are averaged over $N = 30 \cdot 9 = 270$ samples (see Section 4.2.3). In these tables, the underlying language model being used in a setup (i.e., a row in a table) is indicated by the prefixes G- for GPT-2 and D- for DialoGPT. Additionally, both tables report the results of the uncontrolled GPT-2 and DialoGPT baselines (labeled G-baseline and D-baseline, respectively), and those of the 100 most common age-independent bag-of-words setups for both GPT-2 and DialoGPT (labeled G-100MCW and D-100MCW, respectively). The accuracies for these two setups (i.e., baseline and 100 most common words) are omitted because they do not aim to generate responses that resemble any target age group. Moreover, two bag-of-words (BoW) setups are reported per underlying language model (GPT-2 or DialoGPT): the frequency-based BoW setup (indicated by the suffix, -BoW$_{FB}$), and the 100 most informative unigram-based setup (indicated by the suffix, -BoW$_{100MIU}$). Detailed descriptions of how and why these aforementioned wordlists are constructed are provided in Section 4.2.1. Finally, the discriminator-based setups are indicated by the suffix, -Discrim. The aforementioned reporting conventions also hold for the tables containing the results of response generation to younger and older sounding prompts (i.e., Tables 4.5, 4.6, 4.7, and 4.8). The results of those experiments are discussed in Section 4.4.3.

Uncontrolled Baseline Models As one would expect, Tables 4.3 and 4.4 show that the GPT-2 baseline consistently scores among the best on perplexity (best perplexity compared to younger generation models, and second best compared to older generation models) and diversity of generation (the Dist-$n|_{n=1,2,3}$ scores are almost consistently in the upper registers). Similarly, the uncontrolled DialoGPT baseline also scores best in terms of perplexity, when compared to other DialoGPT-based setups. This means that the responses generated by GPT-2 and DialoGPT are found to be among the least perplexing to GPT-1. This is unsurprising, as GPT-2, and thereby also DialoGPT, are pre-trained in similar fashion to GPT-1 [Radford et al., 2018, 2019, Zhang et al., 2020]. Additionally, the target probabilities ($\bar{P}_Y = 0.62$ in Table 4.3, and $\bar{P}_O = 0.38$ in Table 4.4) indicate that the GPT-2 baseline is biased towards generating younger language. That is, given a neutral prompt, GPT-2 is inclined to produce responses that are likely to contain features learned to be younger by BERT$_{FT}$. This could be attributable to GPT-2 being

Model	**ppl.** ↓ better	**Dist-1** ↑ better	**Dist-2** ↑ better	**Dist-3** ↑ better	$\bar{P}_Y$ ↑ better	**Acc.** ↑ better
G-baseline	**27.50** (6.58)	0.87 (0.09)	**0.94** (0.04)	0.90 (0.06)	0.62 (0.42)	-
G-100MCW	27.56 (6.60)	0.86 (0.10)	0.93 (0.04)	**0.90** (0.05)	0.63 (0.42)	-
G-BoW$_{FB}$	27.91 (7.18)	0.87 (0.10)	0.93 (0.05)	0.90 (0.06)	0.69 (0.41)	70.4%
G-BoW$_{100MIU}$	28.37 (7.31)	0.87 (0.09)	0.93 (0.04)	0.90 (0.06)	0.67 (0.41)	67.4%
G-Discrim	32.09 (18.98)	0.77 (0.20)	0.86 (0.13)	0.84 (0.15)	0.66 (0.43)	67.8%
D-baseline	37.52 (12.06)	0.86 (0.13)	0.90 (0.08)	0.85 (0.10)	0.76 (0.37)	-
D-100MCW	37.80 (10.89)	0.85 (0.14)	0.89 (0.10)	0.85 (0.10)	0.82 (0.33)	-
D-BoW$_{FB}$	38.53 (12.64)	0.87 (0.12)	0.90 (0.08)	0.86 (0.10)	0.82 (0.33)	83.0%
D-BoW$_{100MIU}$	38.67 (11.70)	0.88 (0.11)	0.91 (0.07)	0.86 (0.10)	**0.87** (0.28)	**88.5%**
D-Discrim	42.01 (16.94)	**0.90** (0.12)	0.86 (0.14)	0.77 (0.22)	0.86 (0.29)	85.9%

Table 4.3: Results of age-controlled dialogue generation: **younger**-targeted models, conditioned on **neutral prompts**. Format: *average metric (standard error)*. **ppl.** is perplexity w.r.t. GPT-1. **Dist-**n (for $n = 1, 2, 3$) is 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$_{FT}$. **Acc.** is BERT$_{FT}$'s accuracy when classifying the row's samples. Values in **bold** are the best in the column; in blue, the second best.

Model	**ppl.** ↓ better	**Dist-1** ↑ better	**Dist-2** ↑ better	**Dist-3** ↑ better	$\bar{P}_O$ ↑ better	**Acc.** ↑ better
G-baseline	27.50 (6.58)	**0.87** (0.09)	**0.94** (0.04)	0.90 (0.06)	0.38 (0.42)	-
G-100MCW	27.56 (6.60)	0.86 (0.10)	0.93 (0.04)	**0.90** (0.05)	0.37 (0.42)	-
G-BoW$_{FB}$	27.58 (7.07)	0.86 (0.10)	0.93 (0.04)	0.90 (0.06)	0.42 (0.42)	43.0%
G-BoW$_{100MIU}$	**27.25** (6.15)	**0.87** (0.09)	0.93 (0.04)	0.90 (0.06)	0.38 (0.42)	37.4%
G-Discrim	47.15 (47.56)	0.73 (0.24)	0.75 (0.28)	0.75 (0.27)	**0.76** (0.36)	**74.3%**
D-baseline	37.52 (12.06)	0.86 (0.13)	0.90 (0.08)	0.85 (0.10)	0.24 (0.37)	-
D-100MCW	37.80 (10.89)	0.85 (0.14)	0.89 (0.10)	0.85 (0.10)	0.18 (0.33)	-
D-BoW$_{FB}$	37.85 (11.17)	0.87 (0.12)	0.90 (0.08)	0.86 (0.09)	0.22 (0.35)	21.5%
D-BoW$_{100MIU}$	37.91 (12.27)	0.87 (0.11)	0.90 (0.07)	0.85 (0.10)	0.22 (0.34)	21.9%
D-Discrim	41.17 (20.72)	0.87 (0.12)	0.89 (0.13)	0.83 (0.16)	0.57 (0.41)	56.7%

Table 4.4: Results of age-controlled dialogue generation: **older**-targeted models, conditioned on **neutral prompts**. Format: *average metric (standard error)*. ppl. is perplexity w.r.t. GPT-1. Dist-n (for $n = 1, 2, 3$) is number of distinct n-grams normalized by text length, as a measure of diversity. $\bar{P}_O$ is the sample's average probability to contain features learned to be older by BERT$_{FT}$. Acc. is BERT$_{FT}$'s accuracy when classifying the row's samples. Values in **bold** are the best in the column; in blue, the second best.

pre-trained on WebText, a corpus of high-quality documents scraped from web pages, which could be over-represented by millennials [2]. Moreover, DialoGPT has very strong bias towards generating younger sounding responses, given a neutral prompt (0.76 average probability to contain detectable younger features). This is most likely due to DialoGPT having been fine-tuned

[2] https://www.statista.com/statistics/272365/age-distribution-of-internet-users-worldwide/

on Reddit threads [Zhang et al., 2020], as the majority of Reddit users are between the ages 20 and 29 [3]. The uncontrolled DialoGPT baseline also produces more perplexing and less diverse text than GPT-2 according to GPT-1 perplexity and the Dist-$n|_{n=1,2,3}$ scores. The higher perplexity is to be expected, as DialoGPT pre-training and fine-tuning method deviates more from GPT-1's than GPT-2, making it likely to produce more unexpected sentences [Zhang et al., 2020].

BoW-based PPLM-setups The G-BoW$_{100MCW}$ setup performs on par with baseline w.r.t. perplexity and diversity (second best perplexity, second best Dist-2, best Dist-3), when compared to the younger generation models. A similar pattern can be seen for older models, where G-100MCW has the second best Dist-2, and best Dist-3 scores. Additionally, the target probabilities seem virtually unaffected by the 100MCW setups, suggesting that perturbing GPT-2's output with an age-agnostic bag-of-words of the most frequently used words in the dialogue dataset does not noticeably shift the writing style towards that of the younger or older age group. This is to be expected, as such a wordlist should be an unbiased representation of the dataset's language style, given similar sample sizes per class. However, when using DialoGPT as an underlying language model, the BoW-based models (including 100MCW) seem to reinforce younger bias for DialoGPT, regardless of age (i.e., for all younger targeted BoW-based models $\bar{P}_Y$ goes up, and $\bar{P}_O$ goes down for all older targeted BoW-based models).

Compared to discriminator-based models, the BoW-based models seem to generate responses that are slightly more likely to contain features of the target age (younger GPT2-BoW$_{FB}$ models result in 0.06 target probability improvement over the baseline, and older GPT2-BoW$_{FB}$ model in 0.04 target probability improvement over baseline). However, these differences could also be due to randomness. For the 50-plus style GPT-2-based response models, the BoW$_{100MIU}$ setup does not manage to generate older sounding language better than the baseline. Furthermore, BoW-based models seem to barely impact the syntactical structure of generated responses, because changes are more local than those made by discriminator-based setups Dathathri et al. [2020]. This is confirmed by the barely altered perplexity and Dist-scores.

Discriminator-based PPLM-setups The GPT-2 discriminator-based older targeted model manages to generate responses that more convincingly resemble the style of the target group (0.38 $\bar{P}_O$ improvement over the GPT-2 baseline). However, this comes at the cost of perplexity

[3] https://www.statista.com/statistics/1125159/reddit-us-app-users-age/

($+19.65$ compared to baseline) and diversity (much lower dist-scores and higher corresponding standard deviations). By contrast, the GPT-2 discriminator-based younger setup does not generate convincingly more younger sounding responses (only 0.04 average target probability improvement over baseline), despite having noticeably worse and less precise perplexity (4.59 increase over baseline) and diversity. And the discriminator-based DialoGPT models, are superior w.r.t. target probability (+0.10 $\bar{P}_Y$ difference compared to DialoGPT baseline for younger models, and +0.34 $\bar{P}_O$ compared to baseline for older models). However, this again comes at the cost of much higher and more volatile perplexity.

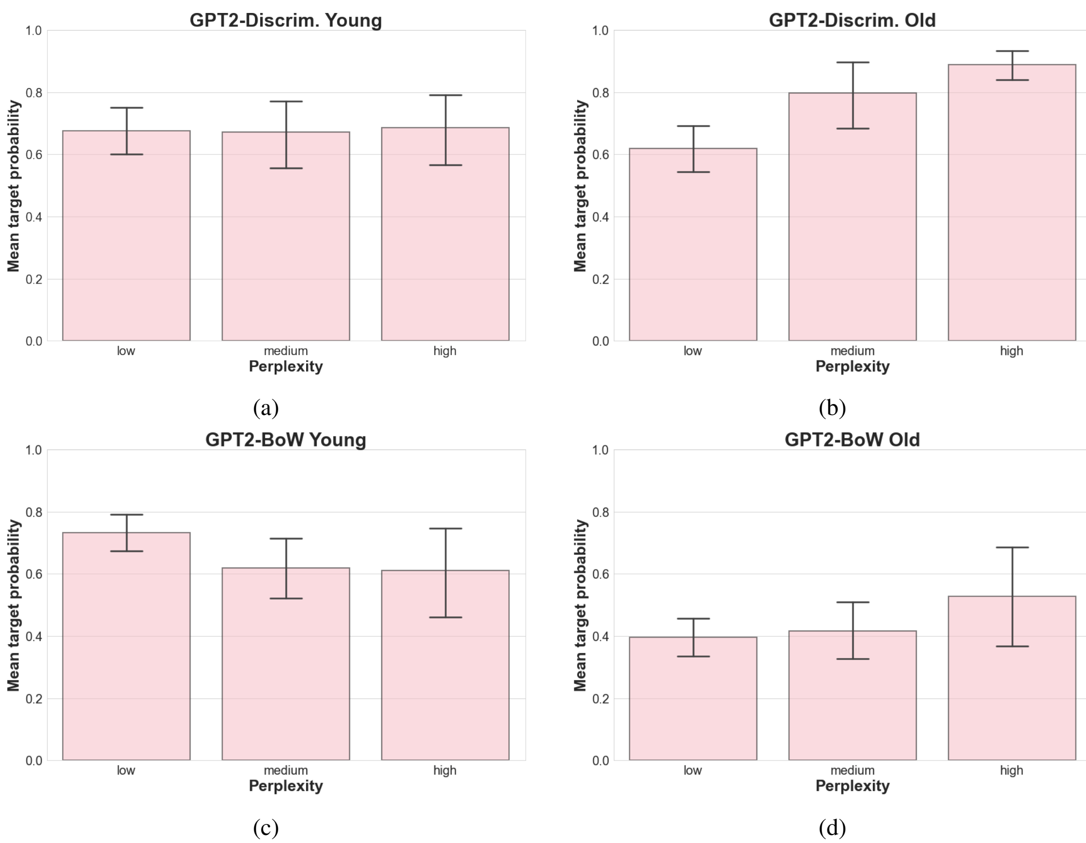
Key Takeaways Overall, according to these results it seems to be possible to control dialogue responses for a certain age group. The used underlying language models are biased (in varying degrees) towards generating younger sounding language. In these tested PPLM-setups, there seems to be a tradeoff between increased control and decreasing perplexity and diversity of generated language. Furthermore, the BoW-based models achieve less detectable levels of control, but preserve the fluency and diversity of generated text. In other words, the discriminator-based models make more invasive changes to the uncontrolled sentences, which can result in less fluent and more repetitive text. However, they do produce more detectably age-appropriate passages, as indicated by BERT$_{FT}$'s assigned target probabilities.

4.4 Controlled Dialogue Generation Analyses

By means of quantitative and qualitative analyses, I seek to study which relationships affect the quality and attribute relevance of the generated responses. The discriminator-based setups, and the BoW-models with the highest average target probabilities (i.e., BoW$_{FB}$ for GPT-2, and BoW$_{100MIU}$ for DialoGPT) are considered for the analyses. The following sections report a series of analyses about the relationship between perplexity and target probability (Section 4.4.1), the effects of response length on generation quality (Section 4.4.2), the impact of the prompt's style on generation style and quality (Section 4.4.3), and qualitatively observable patterns in generated samples 4.4.5.

4.4.1 The Relationship between Perplexity and Target Probability

Figures 4.1 and 4.2 show bar charts depicting the relationship between the average target probability (y-axes) and perplexity (x-axes) assigned to the dialogue responses to neutral prompts (see Table 4.2) generated by various PPLM-model setups. The error bars around the average target probabilities are 95% confidence intervals. Based on the distribution of perplexity observed over all generated responses to neutral prompts, perplexity is binned into three consecutive intervals, bordered by the tertiles of distribution (i.e., the two points that divide the distribution of perplexity into three parts, each containing a third of the distribution). These intervals are named low perplexity (ppl < 27.52), medium perplexity ($27.52 \leq$ ppl < 35.63), and high perplexity (ppl ≥ 35.63).



(a) (b) (c) (d)

Figure 4.1: Mean target probability (x-axes) assigned to GPT-2-based models' samples by BERT_{FT} for increasing ranges of GPT-1 perplexity (y-axes). Error bars are 95% confidence intervals.

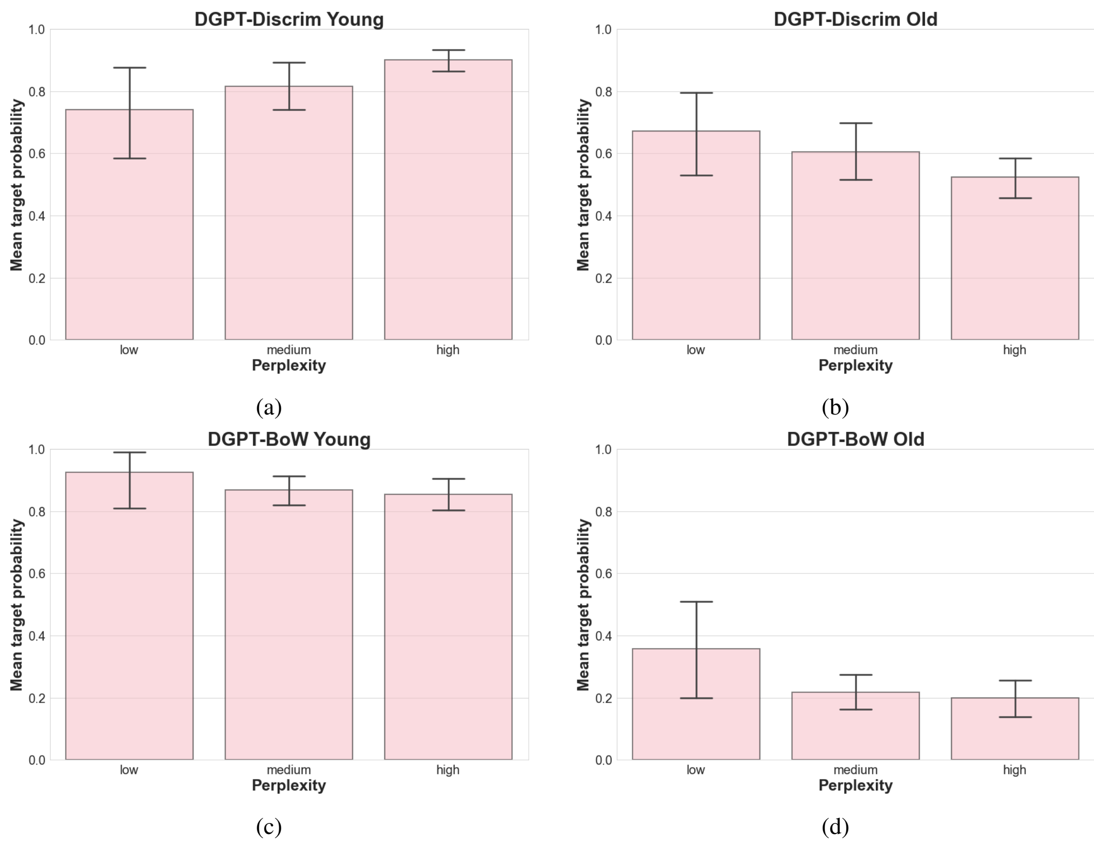


Figure 4.2: Mean target probability (y-axes) assigned to DialoGPT-based models' samples by BERT$_{FT}$ for increasing ranges of GPT-1 perplexity (x-axes). Error bars are 95% confidence intervals.

GPT2-based older targeted PPLM-setups The GPT2-based older models (Figures 4.1b and 4.1d) show a pattern of increasing perplexity coinciding with higher and more precise assigned target probabilities. Especially so for the older targeted discriminator-based GPT-2 setup, whose responses with relatively high perplexity (50+) are (at a 5% level) significantly more likely than those with low perplexity to contain features learned to be older by BERT$_{FT}$. High-perplexity responses of this model are also assigned probabilities with more precision, as indicated by the narrower confidence regions. These observations suggest that, controlling the output of GPT-2 to contain more older related features, the responses become more perplexing (and arguably less fluent). This relationship is more salient for discriminator-based control, because the perturbations are more invasive to the sentence structure, which is in line with earlier observations. It also suggests that PPLM prioritizes enforcing attribute control over promoting fluency, when making perturbations to the underlying language model's output distribution.

GPT2-based younger targeted PPLM-setups By contrast, there does not appear to be a clear pattern for the younger targeted GPT-2-based models' responses (Figures 4.1a and 4.1c). Higher perplexity responses generated by younger targeted BoW-based GPT-2's seem to coincide being less detectably younger sounding (with lower precision) than lower perplexity ones. And the responses generated by the younger targeted discriminator-based GPT-2 PPLM setups appear to receive roughly equal average target probabilities for increasing perplexity, with slightly lower precision. It must be noted that there are no significant differences at the 5% level between the average $\bar{P}_Y$, so conclusions about the relationship between perplexity and target probability should be taken tentatively.

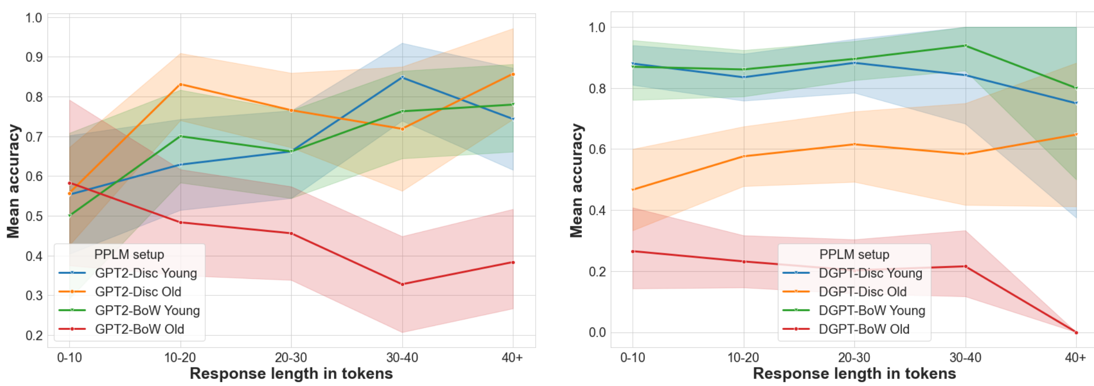
DialoGPT-based older targeted PPLM-setups The DGPT-based (i.e., DialoGPT-based) models targeted towards generating older related features (Figures 4.2b and 4.2d) display patterns of decreasing target probabilities for increasing perplexity. However, it must be emphasized that responses generated by both these models are rarely seen by BERT_{FT} as likely to contain linguistic features associated with their target age (overall $\bar{P}_O$ of 0.22 and 0.57 for DGPT-BoW Old and DPGT-Discrim Old, respectively). This makes it less reliable to draw conclusions about the relationship between the style-adherence and perplexity of their samples, as they are often not style-adherent to begin with.

DialoGPT-based younger targeted PPLM-setups DialoGPT's strong proclivity to generate younger sounding responses is noticeable in Figures 4.2a and 4.2c, as depicted by the high average target probabilities and relatively narrow confidence intervals. Furthermore, Figure 4.2a shows a similar pattern as the ones observed for the older targeted GPT-2-based setups, that is, increasing detectability of target age features for increasing perplexity, with an almost significant difference in average assigned target probabilities between the low and high perplexity responses (as depicted by their slightly overlapping confidence intervals). Additionally, one could argue that the BoW-based DGPT younger targeted setup (Figure 4.2c) shows a similar, yet much less pronounced, pattern to its GPT-2-based counterpart (Figure 4.1c), where the average assigned target probability decreases for increasing perplexity. However, the differences between the average target probabilities are not statistically significant and too small to draw a conclusion from this result alone.

4.4.2 The Relationship between Response Length and Various Evaluation Metrics

The number of tokens in a generated response coincides with noticeable differences in the used automated evaluation metrics. It is therefore important to get a clearer picture of how the various measures for fluency and control change for different sequence lengths. Moreover, properly understanding these relationships can inform developers of adaptive dialogue systems about preserving output quality and adaptation of responses of arbitrary lengths.

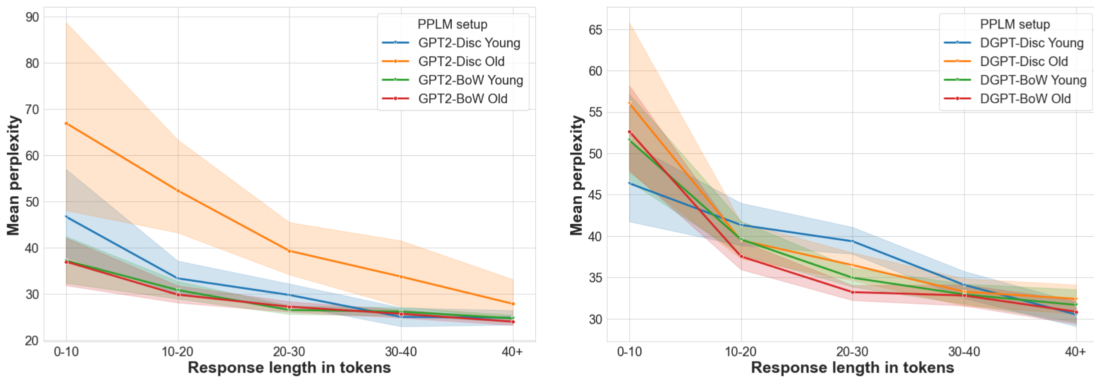
Response length (on the x-axes) is plotted against various evaluation metrics in Figures 4.3 (average BERT$_{FT}$ accuracy), 4.4 (GPT-1 perplexity), 4.5 (normalized number of distinct unigrams), 4.6 (normalized number of distinct bigrams), and 4.7 (normalized number of distinct trigrams).



(a) (b)

Figure 4.3: Mean BERT$_{FT}$ accuracy. GPT-2-based models (left), DialoGPT-based models (right). Translucent error bands represent 95% confidence intervals. Plots best viewed in color.

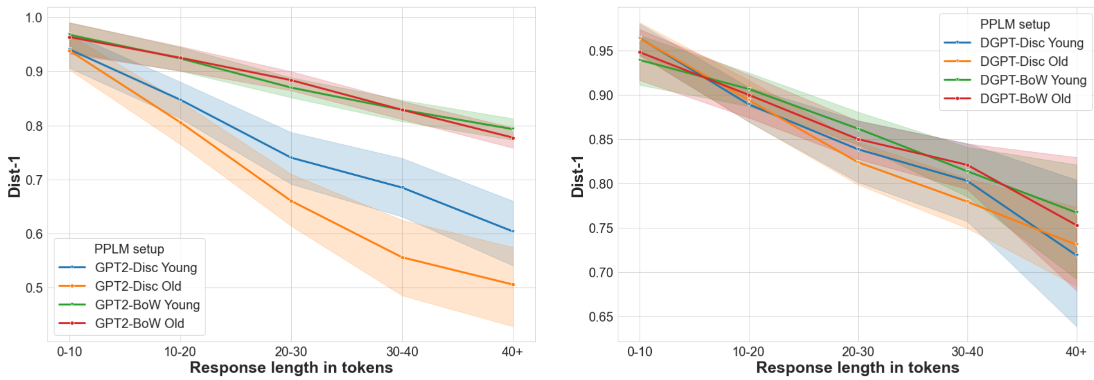
BERT$_{FT}$ Accuracy versus Response Length Figure 4.3a shows a slight upward trend in average accuracy with greater uncertainty for increasing response length for all GPT-2-based models, except for the BoW-based older generation model. That is, longer sequences are, on average, slightly easier to classify, though with less precision. This is probably due to the fact that longer sentences contains more information to base predictions on. By contrast, the DialoGPT-based models in Figure 4.3 do not seem to show a clear general trend that mean accuracy follows for increasing response length. However, it does seem that DialoGPT's strong bias towards generating younger sounding responses causes output from DialoGPT-based younger generation models to be much easier to classify than that from the older generation models. Overall, it can be seen that the BoW-based older targeted models (GPT-2 and DialoGPT) are significantly more challenging to classify at almost every length bracket.



(a) (b)

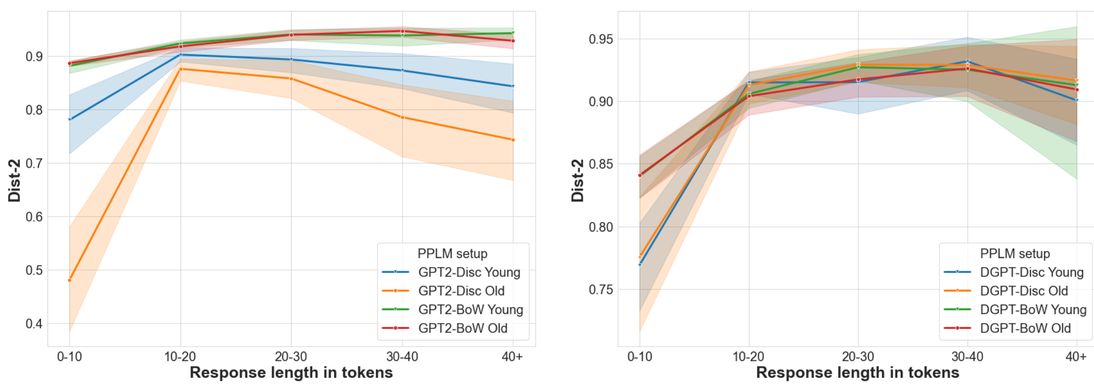
Figure 4.4: Perplexity. GPT-2-based models (left), DialoGPT-based models (right). Translucent error bands represent 95% confidence intervals. Plots best viewed in color.

Perplexity versus Response Length Figure 4.4 shows a clear downward trend for both sets of PPLM-setups. Irrespective of the underlying language model being used, longer responses are deemed less perplexing, with more certainty, by GPT-1 than shorter ones. It is worth emphasizing that the model with the highest target probability improvement over its relevant baseline, G-Discrim$_{Old}$, is found to produce significantly more perplexing responses than its GPT-2-based counterparts at most length brackets (see Figure 4.4a). This finding resonates with Figure 4.1b and the idea, that especially for older targeted generation models, increased levels of attribute relevance coincide with worse perplexity. However, it must be noted that the downward slope of perplexity for increasing response length could be attributable to the nature of calculating perplexity, rather than generation properties of the models. Namely, perplexity essentially averages the sum of the negative exponentiated probabilities $p(\texttt{word}|\texttt{context})$, for every word in a sentence. Because the context increases with every successive word, and larger contexts typically result in less uncertainty, shorter sequences are often given unfairly high perplexities.



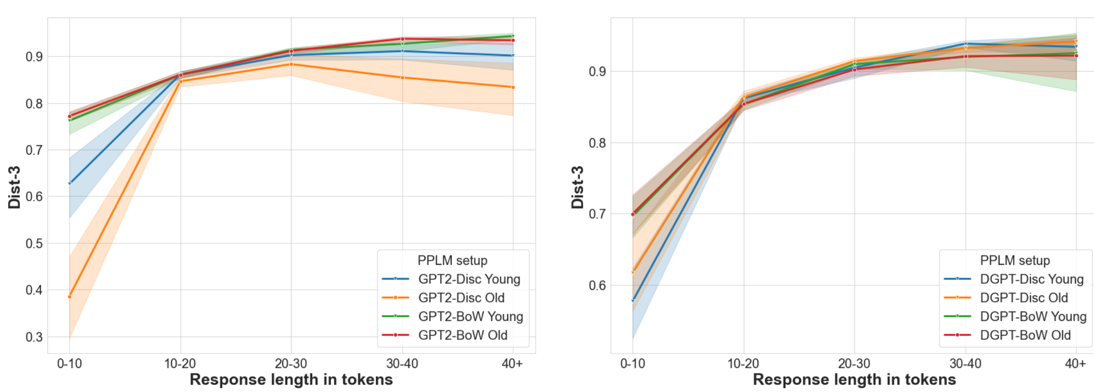
(a) (b)

Figure 4.5: Dist-1. GPT-2-based models (left), DialoGPT-based models (right). Translucent error bands represent 95% confidence intervals. Plots best viewed in color.



(a) (b)

Figure 4.6: Dist-2. GPT-2-based models (left), DialoGPT-based models (right). Translucent error bands represent 95% confidence intervals. Plots best viewed in color.



(a) (b)

Figure 4.7: Dist-3. GPT-2-based models (left), DialoGPT-based models (right). Translucent error bands represent 95% confidence intervals. Plots best viewed in color.

Diversity versus Response Length Figure 4.5 shows that diversity w.r.t. unigrams decreases for longer responses. This is most likely due to the fact that longer sentences have an *a priori* higher probability of containing repeated words. E.g., stopwords like "the" and "of" are likely to appear multiple times in longer sentences. The same figure shows that GPT-2-BoW models generate significantly more diverse responses w.r.t. unigrams for almost every bracket of response length. This could be attributable to BoW-based control altering base-GPT2's generated sentences at the token-level, thus being more likely to preserve the unigram diversity of the uncontrolled baseline (the GPT-2 baseline's Dist-1 is always in the upper register in Tables 4.3 and 4.4).

Figure 4.6 shows that variety w.r.t. bigrams makes an initial upward jump between response length of 1-10 and 10-20. Dist-2 then follows a mild downward trend for both GPT2- and DialoGPT-based models However, detailed inspection of Figure 4.5a shows that only the discriminator-based setups have a negative slope, whereas the BoW-based setups follow a very slight upward trend. Thus, the BoW-based GPT-2 models produce significantly more diverse language w.r.t. bigrams for most response lengths, attributable to the same reason mentioned above.

Figure 4.7 shows similar patterns: an initial upward jump in trigram diversity between shortest and second-to-shortest length brackets. GPT-2 models then show a slight downward trend for the discriminator-based setups from the 20-30 lengths onward, while the BoW-based models become slightly more diverse w.r.t. trigrams.

Overall, it can again be seen that BoW-based models generate detectably more diverse responses (with greater precision), and remain to do so as response length increases. The decreasing diversity of discriminator-based generated responses further confirms that more invasive control during generation impedes textual variety.

4.4.3 The Effects of Prompt Style on PPLM Control

Recall that given a conditioning text, i.e., `prompt`, a predefined style attribute a, and some controlled dialogue generation model parameterized by θ, generating a style-controlled piece of text $\mathbf{x}$ entails modeling $p_\theta(\mathbf{x}|a, \texttt{prompt})$. It is therefore reasonable to expect the output distribution of controlled generation model p_θ to depend (to some extent) on the content and style of the conditioning text, `prompt`. Indeed, the content and style of prompts are found to strongly influence the output of neural text generation models [Fan et al., 2018, Lester et al., 2021]. Thus, studying the effects of a prompt's age-style (i.e., whether a prompt is considered younger, older, or neutral by BERT$_{FT}$) on the style and grammatical quality of PPLM-setups is of great importance, as it could inform developers of adaptive dialogue systems about mitigation of prompt-induced biases.

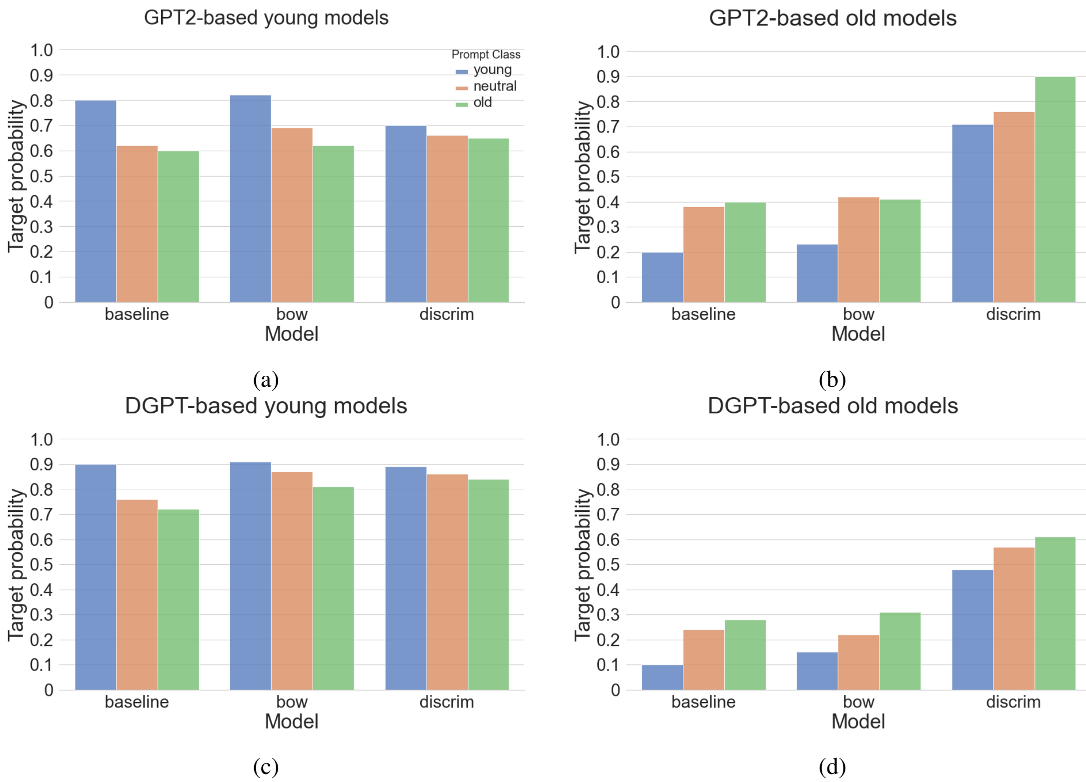
It is worth mentioning that the effects of prompt-style on PPLM-generation are not considered by Dathathri et al. [2020] and Madotto et al. [2020]. Studying these effects is an important extension of their methods, as not quantitatively taking into account the effects prompt-style obfuscates the degree to which one can conclude whether detectable attribute-adherence is the result of controlled generation or prompt-induced bias.

Figures 4.8 and 4.9 depict the average target probability and perplexity over responses generated by the baseline, the best BoW-based model, and the discriminator-based model, when prompted with a prompt of either younger, neutral, or older style. More specifically, each bar represents a metric (target probability or perplexity) averaged over $N = 270$ samples generated by a single model, when presented with five prompts of the same age-style. E.g., the blue bar in Figure 4.8a represents the average probability of samples generated by GPT-2 + frequency-based BoW to contain features learned to be younger by BERT$_{FT}$, when the model was presented younger sounding prompts. The explicit numerical values of Figures 4.8 and 4.9 are found in Tables 4.5, 4.6, 4.7, and 4.8.

Ideally, the neutrally prompted baseline's assigned probability of generating age-specific responses should be around 0.50. Furthermore, a prompt should shift the target probability in the direction of the prompt class, e.g., a younger sounding prompt should shift a younger targeted model's target probability upwards, and an older targeted model's target probability downwards. It is known from previous results in Tables 4.3 and 4.4 that the language model baselines are (in varying degrees) biased towards generating younger sounding responses to neutral prompts. Nevertheless, It can be expected that the impact of prompt-style persists, albeit to differing degrees based on the (dis)similarity in style between the prompt and response, and the type of attribute model being used (BoW or discriminator).

Figure 4.8 shows that the models' target probabilities indeed move accordingly with the prompts' styles. E.g., the younger prompted younger targeted models achieve the highest target probability, then the neutral prompted ones, and then the older prompted ones (Figures 4.8a and 4.8c). The same pattern holds the other way around: an older prompted older targeted model has the highest (older) target probability, then neutrally prompted, and then the younger prompted ones (Figures 4.8b and 4.8d). In these last two sub-figures, it can also be clearly seen that the discriminator-based older targeted generation models achieve substantial target probability improvements over the baseline (and BoW-based models), for every style of prompt. By contrast, the younger generation models (Figures 4.8a and 4.8c) do not show the same pattern: discriminator-based models achieve similarly subtle improvements in target probability over their baselines as the BoW-based models do. Figure 4.8a even shows the discriminator-based models to perform worse than the baseline and BoW-based models.

Overall, Figures 4.8 and 4.9 and Tables 4.5, 4.6, 4.7, and 4.8 show that the style of the prompt clearly nudges the assigned probability of containing age-related features in the direction of the prompt's style. Put more concisely, the style of the prompt strongly influences the style of the generated response. Moreover, using strongly younger prompt results in heavily reinforced younger bias for GPT2 and DialoGPT. Similarly, using an older sounding prompt results in a slightly neutralized younger bias for both language models. Additionally, the discriminator-based older targeted models (GPT-2 and DialoGPT) always yield the highest relative improvements over the baselines. However, they fail to do so when attempting to generate younger sounding responses, which could suggest that the stylistic features learned to be younger and older by BERT$_{FT}$ lie at different linguistic levels, i.e., learned distinguishing features for younger language could lie mostly at the lexical level, and older ones at the syntactic level. Moreover, this could imply that they, the linguistic styles representing the younger and older age group, are not equally challenging to control for a PPLM-setup. Finally, the aforementioned tables show that the effects of prompt-style are largely limited to the target probabilities. For perplexity and distinctiveness, the younger and older prompted results show similar patterns to the neutral-prompt setting: BoW-based models achieve smaller increases in control, but maintain relatively desirable perplexity and diversity. Whereas, the discriminator-based models achieve higher levels of control, at the cost of worse perplexity and Dist-scores.



(a) (b) (c) (d)

Figure 4.8: The average target probabilities assigned to responses generated by various PPLM-setups, conditioned on younger, neutral, or older prompts. The plots are best viewed in color.

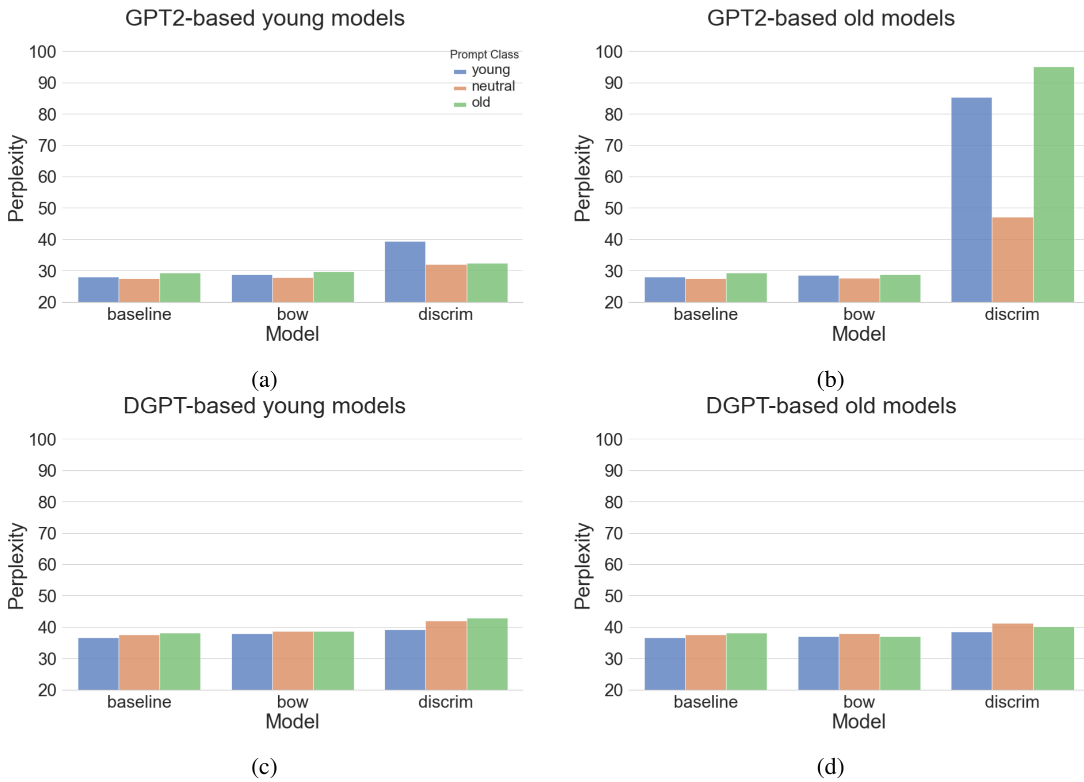


Figure 4.9: The average perplexities of responses generated by various PPLM-setups, conditioned on younger, neutral, or older prompts. The plots are best viewed in color. The plots are best viewed in color.

Model	**ppl.** ↓ better	**Dist-1** ↑ better	**Dist-2** ↑ better	**Dist-3** ↑ better	$\bar{\boldsymbol{P}}_{\boldsymbol{Y}}$ ↑ better	**Acc.** ↑ better
G-baseline	28.05 (±6.12)	0.85 (±0.13)	0.91 (±0.08)	0.88 (±0.08)	0.80 (±0.33)	-
G-100MCW	**27.71** (±6.20)	0.85 (±0.12)	0.91 (±0.09)	0.88 (±0.09)	0.75 (±0.37)	-
G-B$_{FB,Y}$	28.81 (±7.09)	0.86 (±0.12)	**0.92** (±0.08)	**0.89** (±0.08)	0.82 (±0.32)	83.3%
G-B$_{100MIU,Y}$	28.49 (±6.49)	0.86 (±0.12)	0.91 (±0.08)	0.88 (±0.08)	0.83 (±0.32)	83.0%
G-D$_Y$	39.32 (±37.49)	0.84 (±0.21)	0.61 (±0.40)	0.57 (±0.40)	0.70 (±0.40)	70.7%
D-baseline	36.69 (±9.11)	0.87 (±0.10)	0.91 (±0.06)	0.87 (±0.08)	0.90 (±0.24)	-
D-100MCW	36.93 (±9.18)	0.86 (±0.11)	0.91 (±0.06)	0.88 (±0.07)	0.90 (±0.25)	-
D-B$_{FB,Y}$	37.35 (±8.60)	0.88 (±0.10)	0.91 (±0.06)	0.87 (±0.08)	0.90 (±0.26)	90.0%
D-B$_{100MIU,Y}$	37.87 (±8.32)	0.88 (±0.10)	0.91 (±0.07)	0.87 (±0.09)	**0.91** (±0.24)	**92.6%**
D-D$_Y$	39.22 (±14.96)	**0.89** (±0.12)	0.86 (±0.19)	0.79 (±0.23)	0.89 (±0.25)	91.1%

Table 4.5: Results of age-controlled dialogue generation: **younger**-targeted models, conditioned on **young prompts**. Format: *average metric (standard error)*. **ppl.** is perplexity w.r.t. GPT-1. **Dist-**n (for $n = 1, 2, 3$) is number of distinct n-grams normalized by text length, as a measure of diversity. $\bar{\boldsymbol{P}}_{\boldsymbol{Y}}$ is the sample's average probability to contain features learned to be younger by BERT$_{FT}$. **Acc.** is BERT$_{FT}$'s accuracy when classifying the row's samples. Values in **bold** are the best in the column; in blue, the second best.

Model	ppl. ↓ better	Dist-1 ↑ better	Dist-2 ↑ better	Dist-3 ↑ better	$\bar{P}_O$ ↑ better	Acc. ↑ better
G-baseline	28.05 (±6.12)	0.85 (±0.13)	0.91 (±0.08)	0.88 (±0.08)	0.20 (±0.33)	-
G-100MCW	**27.71** (±6.20)	0.85 (±0.12)	0.91 (±0.09)	0.88 (±0.09)	0.25 (±0.37)	-
G-B$_{FB,O}$	28.54 (±6.45)	0.86 (±0.12)	**0.92** (±0.08)	**0.89** (±0.08)	0.23 (±0.36)	22.6%
G-B$_{100MIU,O}$	28.18 (±5.70)	0.87 (±0.11)	**0.92** (±0.08)	0.89 (±0.09)	0.21 (±0.34)	21.5%
G-D$_O$	85.40 (±150.28)	0.67 (±0.30)	0.62 (±0.31)	0.62 (±0.32)	**0.71** (±0.40)	**70.5%**
D-baseline	36.69 (±9.11)	0.87 (±0.10)	0.91 (±0.06)	0.87 (±0.08)	0.10 (±0.24)	-
D-100MCW	36.93 (±9.18)	0.86 (±0.11)	0.91 (±0.06)	0.88 (±0.07)	0.10 (±0.25)	-
D-B$_{FB,O}$	37.25 (±9.45)	0.87 (±0.11)	0.91 (±0.06)	0.87 (±0.08)	0.12 (±0.29)	11.1%
D-B$_{100MIU,O}$	37.04 (±8.78)	**0.88** (±0.10)	0.91 (±0.05)	0.88 (±0.07)	0.15 (±0.32)	15.2%
D-D$_O$	38.46 (±14.91)	0.82 (±0.15)	0.87 (±0.15)	0.83 (±0.17)	0.48 (±0.44)	47.4%

Table 4.6: Results of age-controlled dialogue generation: **older** 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 number of distinct n-grams normalized by text length, as a measure of diversity. $\bar{\boldsymbol{P}}_{\boldsymbol{O}}$ is the sample's average probability to contain features learned to be older by BERT$_{FT}$. **Acc.** is BERT$_{FT}$'s accuracy when classifying the row's samples. Values in **bold** are the best in the column; in blue, the second best.

Model	ppl. ↓ better	Dist-1 ↑ better	Dist-2 ↑ better	Dist-3 ↑ better	$\bar{P}_Y$ ↑ better	Acc. ↑ better
G-baseline	29.34 (±10.30)	0.86 (±0.09)	**0.94** (±0.04)	0.90 (±0.06)	0.60 (±0.43)	-
G-100MCW	**29.14** (±10.11)	0.86 (±0.10)	0.93 (±0.04)	0.90 (±0.06)	0.60 (±0.44)	-
G-B$_{Y,FB}$	29.61 (±10.28)	0.86 (±0.10)	0.93 (±0.04)	**0.91** (±0.06)	0.62 (±0.43)	61.1%
G-B$_{Y,100MIU}$	29.51 (±0.09)	0.87 (±0.09)	0.93 (±0.05)	0.90 (±0.06)	0.68 (±0.42)	68.5%
G-D$_Y$	32.34 (±19.88)	0.77 (±0.20)	0.84 (±0.19)	0.80 (±0.23)	0.65 (±0.43)	65.4%
D-baseline	38.18 (±12.03)	0.86 (±0.12)	0.90 (±0.08)	0.86 (±0.09)	0.72 (±0.38)	-
D-100MCW	37.73 (±11.88)	0.85 (±0.13)	0.90 (±0.08)	0.86 (±0.09)	0.73 (±0.39)	-
D-B$_{Y,FB}$	38.24 (±11.53)	0.86 (±0.12)	0.90 (±0.08)	0.86 (±0.10)	0.81 (±0.34)	82.6%
D-B$_{Y,100MIU}$	38.66 (±11.57)	0.85 (±0.12)	0.90 (±0.07)	0.86 (±0.09)	0.81 (±0.33)	80.7%
D-D$_Y$	42.93 (±20.18)	**0.90** (±0.14)	0.79 (±0.22)	0.68 (±0.28)	**0.84** (±0.30)	**85.2%**

Table 4.7: Results of age-controlled dialogue generation: **younger**-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 number of distinct n-grams normalized by text length, as a measure of diversity. $\bar{\boldsymbol{P}}_{\boldsymbol{Y}}$ is the sample's average probability to contain features learned to be younger by BERT$_{FT}$. **Acc.** is BERT$_{FT}$'s accuracy when classifying the row's samples. Values in **bold** are the best in the column; in blue, the second best.

4.4.4 Visual Analysis of BERT$_{FT}$ Attention Patterns

I use visualizations of attention mechanisms [Vig, 2019] in BERT$_{FT}$'s attention heads to analyze recurring patterns when assigning target probabilities to generated prompt-response pairs. An attention weight can be interpreted as an indication of how important a particular token is when producing the next representation of the current token [Bahdanau et al., 2015, Clark et al., 2019]. Please note that the PPLM-method does not change the attention weights of the underlying

Model	ppl. ↓ better	Dist-1 ↑ better	Dist-2 ↑ better	Dist-3 ↑ better	$\bar{P}_O$ ↑ better	Acc. ↑ better
G-baseline	29.34 (±10.30)	0.86 (±0.09)	**0.94** (±0.04)	**0.90** (±0.06)	0.40 (±0.43)	-
G-100MCW	29.14 (±10.11)	0.86 (±0.10)	0.93 (±0.04)	**0.90** (±0.06)	0.40 (±0.44)	-
G-B$_{O,FB}$	**28.81** (±10.10)	0.86 (±0.10)	0.93 (±0.05)	**0.90** (±0.06)	0.41 (±0.43)	41.1%
G-B$_{O,100MIU}$	29.05 (±9.80)	0.86 (±0.09)	0.93 (±0.04)	**0.90** (±0.06)	0.40 (±0.43)	39.6%
G-D$_O$	95.21 (±174.42)	0.65 (±0.27)	0.78 (±0.18)	0.78 (±0.18)	**0.90** (±0.25)	**90.3%**
D-baseline	38.18 (±12.03)	0.86 (±0.12)	0.90 (±0.08)	0.86 (±0.09)	0.28 (±0.38)	-
D-100MCW	37.73 (±11.88)	0.85 (±0.13)	0.90 (±0.08)	0.86 (±0.09)	0.27 (±0.39)	-
D-B$_{O,FB}$	37.80 (±11.74)	0.86 (±0.12)	0.90 (±0.07)	0.87 (±0.08)	0.28 (±0.39)	29.3%
D-B$_{O,100MIU}$	36.93 (±11.68)	**0.87** (±0.12)	0.90 (±0.09)	0.86 (±0.09)	0.31 (±0.41)	29.6%
D-D$_O$	40.08 (±16.77)	0.85 (±0.14)	0.88 (±0.10)	0.83 (±0.14)	0.61 (±0.42)	61.1%

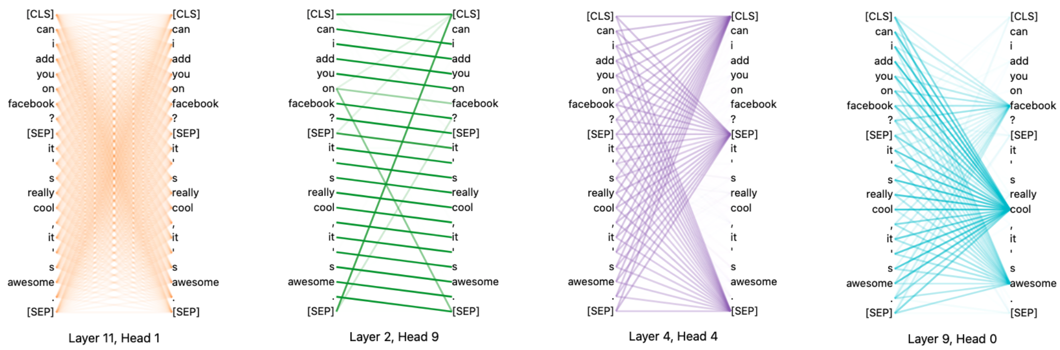
Table 4.8: Results of age-controlled dialogue generation: **older** targeted models, conditioned on **older prompts**. Format: *average metric (standard error)*. **ppl.** is perplexity w.r.t. GPT-1. **Dist-n** (for $n = 1, 2, 3$) is number of distinct n-grams normalized by text length, as a measure of diversity. $\bar{P}_O$ is the sample's average probability to contain features learned to be older by BERT$_{FT}$. **Acc.** is BERT$_{FT}$'s accuracy when classifying the row's samples. Values in **bold** are the best in the column; in blue, the second best.

language models, so visualizing the attention weights of the generation models would not reveal any age-related attention patterns. By contrast, BERT$_{FT}$ is fine-tuned to detect age-related patterns, so visualizing its attention maps can inform decisions about which features are important when assigning a target-age probability to a generated sentence. Despite this seemingly natural interpretation, there is much debate about the validity of using attention mechanisms (as opposed to, e.g., saliency methods) as explanations for model output [Jain and Wallace, 2019, Wiegreffe and Pinter, 2019, Bastings and Filippova, 2020]. However, as Vig [2019] and Clark et al. [2019] suggest, attention weight visualization can be used tentatively as a complementary analysis tool to add to sets of different analysis methods to inform researchers about, e.g., possible linguistic patterns that may be attended to by attention-based models.

To provide a clear reading experience of the analyses presented below, I recapitulate a few important concepts about BERT and how to read attention weight visualizations. During pre-processing, BERT tokenizes the input text and adds a special token `[CLS]` to the beginning of the text, and another special token `[SEP]` is appended to the text. If an input sequence consists of multiple sentences (e.g., a question-answer, or prompt-response input), `[SEP]` tokens are also used to separate the sentences. BERT$_{FT}$ is a fine-tuned version of BERT-base-uncased [Devlin et al., 2019], which consists of 12 layers of 12 attention heads. A specific attention head is referred to as Head `layer_number-head_number`, where `layer_number` and `head_number` both range from 0 to 11. E.g., Head 3-1 refers to Head 1 in Layer 3. Furthermore, attention weights are visualized as colored lines between tokens of the input sequence, where a thicker line corresponds to a greater attention weight, and the color represents the layer in which the head is present. The input text is displayed twice in parallel columns, to make visualizations of self-attention possible (visualized as lines between identical tokens in the same positions). Because BERT is designed to be deeply bidirectional, tokens can also attend to tokens in previous positions in the input text.

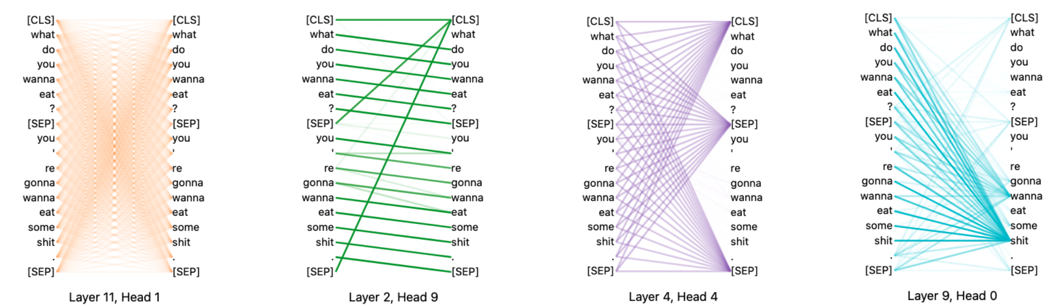
Figures 4.10, 4.11, 4.12, and 4.13 show attention visualizations of BERT_{FT} when processing prompt-response pairs that are cherry-picked from the age-targeted prompted results, presented in Tables 4.5 and 4.8. The generated sequences are chosen to display more pronounced examples of recurring attention patterns. Two younger targeted generated responses to younger sounding prompts are shown in Figures 4.10 and 4.11, and two older targeted generated responses to older sounding prompts are shown in Figures 4.12 and 4.13. All prompts and responses received target probabilities from BERT_{FT} of at least 95%.

Heads in the same layer have the tendency to attend to similar patterns and linguistic phenomena Clark et al. [2019]. The results of the analyses seem to confirm this behavior, as recurring patterns are observed for heads in the same layers. For instance, it can be seen (in Figures 4.10a, 4.11a, 4.12a, and 4.13a) that heads in the last layer (11) tend to broadly disperse attention among all tokens. Recurring patterns are also observed in earlier layers, such as Head 2-9 attending to the next token in the sequence (Figures 4.10b, 4.11b, 4.12b, and 4.13b), and Head 4-4 attending to the special tokens (Figures 4.10c, 4.11c, 4.12c, and 4.13c). Certain heads also seem to pay special attention to age-related linguistic features, specifically certain tokens associated with an age group (mentioned in Section 4.4.5). This is most noticeable in Head 9-0 (Figures 4.10d, 4.11d, 4.12d, and 4.13d), which seems to consistently devote the majority of its attention to tokens that are found to be indicative of age. For instance, Head 9-0 attends strongly to younger sounding tokens like *facebook*, *awesome*, *cool*, and slang and swear words in Figures 4.10d and 4.11d. And the same attention head then focuses strongly on tokens associated with older age in Figures 4.12d and 4.13d: e.g., *workers*, *union*, *greetings*, or *fellow*.



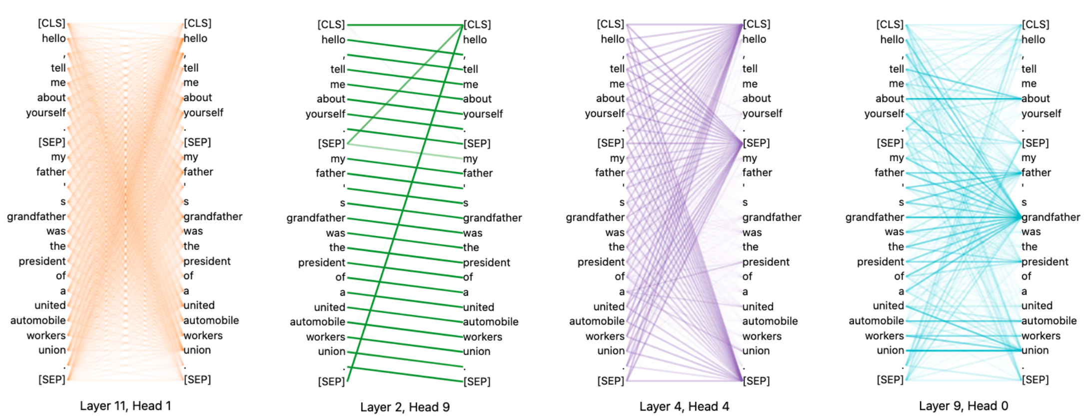
(a) Attends broadly. (b) Attends to next token. (c) Attends to special tokens. (d) Age-related attention.

Figure 4.10: Attention weights visualizations of four of BERT$_{FT}$'s attentions heads and the patterns to which they presumably attend when processing representations for a cherry-picked prompt-response pair generated by **younger** targeted GPT2-Discrim.



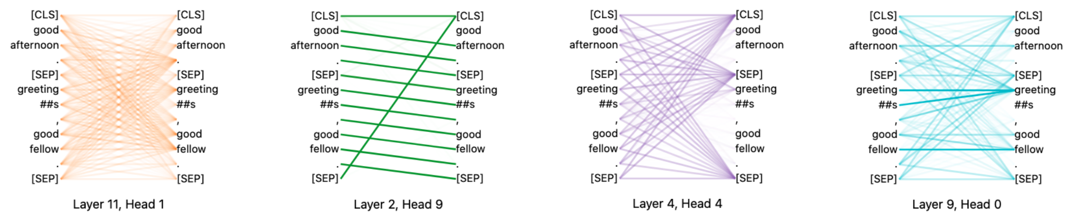
(a) Attends broadly. (b) Attends to next token. (c) Attends to special tokens. (d) Age-related attention.

Figure 4.11: Attention weights visualizations of four of BERT$_{FT}$'s attentions heads and the patterns to which they presumably attend when processing representations for a cherry-picked prompt-response pair generated by **younger** targeted GPT2-BoW$_{100MIU}$.



(a) Attends broadly. (b) Attends to next token. (c) Attends to special tokens. (d) Age-related attention.

Figure 4.12: Attention weights visualizations of four of BERT$_{FT}$'s attentions heads and the patterns to which they presumably attend when processing representations for a cherry-picked prompt-response pair generated by **older** targeted GPT2-Discrim.



(a) Attends broadly. (b) Attends to next token. (c) Attends to special tokens. (d) Age-related attention.

Figure 4.13: Attention weights visualizations of four of BERT$_{FT}$'s attentions heads and the patterns to which they presumably attend when processing representations for a cherry-picked prompt-response pair generated by **older** targeted DGPT-Discrim.

4.4.5 Qualitative Analysis of Generated Dialogue Responses

Similar to the previous manual inspection of dialogue utterances in Section 3.4.2 and most informative n-grams in Section 3.4.3, what follows now is a qualitative analysis of the generated dialogue responses for various cases. This section reports a quick overview of how the responses generated by the best performing models are distributed among various categories relating to their level of perplexity (low, medium, or high) and classification status (correctly or incorrectly classified). Then, there follows a manual inspection of the most distinct sets of cases (also referred to as sub-samples) for observable age-related linguistic patterns, and discuss how they relate to my expectations, the previous qualitative analyses, and earlier work on the relationship between age and language.

The generated responses to neutral prompts are split by (1) whether or not BERT$_{FT}$ correctly classified a response as its target class, and (2) the level of perplexity: low (ppl. < 27.52), medium ($27.52 \leq$ ppl. < 35.63), or high (ppl. ≥ 35.63). See Section 4.4.1 for an explanation of the rationale behind these intervals. Table 4.10 shows how the generated responses are distributed among these cases for the best performing BoW-based models (among all BoW-based models), and discriminator-based models for both underlying language models and target age groups. As can be seen, the majority of the responses generated by GPT-2 BoW-based models lie in the low perplexity range, whereas the discriminator-based GPT-2 models also generate considerable amounts of high-perplexity responses (e.g., 37.2% high-perplexity correctly classified for G-Discrim$_{Old}$). By contrast, the DialoGPT-based models all show the majority of their distributions lying in the medium-to-high perplexity range.

To narrow down the comparison, the qualitative inspection of samples is limited to responses generated by the discriminator-based models of both language models and age groups (i.e., samples generated by G-Discrim$_{Young}$, G-Discrim$_{Old}$, D-Discrim$_{Young}$, and D-Discrim$_{Old}$). Moreover, these models are all among those with the highest target probability improvements over their respective baselines in Tables 4.3 and 4.4, so the differences in language style should be most pronounced between these setups. Furthermore, to emphasize the differences between perplexity, I only consider low versus high perplexity samples. Finally, the samples used for qualitative inspection are also split by whether or not they were correctly classified by BERT$_{FT}$. To summarize, the qualitative inspections are performed on a total of 16 splits: by model (G-Discrim$_{Young}$, G-Discrim$_{Old}$, D-Discrim$_{Young}$, or D-Discrim$_{Old}$), perplexity (low or high), and classification outcome (correct or incorrect). Table 4.9 shows examples of generated responses containing the patterns and observations discussed in the remainder of this section.

PROMPT	Good weather we're having.
GPT-2$_{Young}$	It's awesome weather.
GPT-2$_{Old}$	It's not been quite a good weather for long.
DialoGPT$_{Young}$	Gonna make the weather better.
DialoGPT$_{Old}$	It's not the best weather in the country.
PROMPT	Hi, how's it going?
GPT-2$_{Young}$	It's been almost 2 weeks now since my boyfriend has been gone.
GPT-2$_{Old}$	I've had my first surgery recently.
DialoGPT$_{Old}$	You have you're first born daughter going on with the surgery????
PROMPT	Can we talk?
GPT-2$_{Young}$	We have to stop talking about this is all going so stupid lol lol, sh*t just got weird lol LOL.
GPT-2$_{Old}$	The question was raised in a response sent by an official.
DialoGPT$_{Young}$	You guys are doing a lot.
DialoGPT$_{Old}$	We have talked about your past your previous history.

(a) Low perplexity & correctly classified.

PROMPT	Good weather we're having.
GPT-2$_{Young}$	The Weather is nice, the temperature cool, the traffic good.
GPT-2$_{Old}$	The sun was setting when we were getting up with a huge rain.
DialoGPT$_{Young}$	Cold weather we can't handle it
PROMPT	Hello, tell me about your latest holiday.
GPT-2$_{Young}$	I'm trying hard to stay away.
GPT-2$_{Old}$	It's not my favorite holiday.
DialoGPT$_{Old}$	It was a great weekend!
PROMPT	Can we talk?
GPT-2$_{Young}$	Can we talk with some sense?
GPT-2$_{Old}$	No, not about your favourite club.
DialoGPT$_{Young}$	We have a meeting in a bit.
DialoGPT$_{Old}$	We have a couple more people interested!

(b) Low perplexity & incorrectly classified.

Table 4.9: Cherry-picked examples of low-perplexity young-targeted (red rows) and old-targeted (blue rows) responses to neutral prompts (yellow rows) generated by various PPLM-setups, and grouped by whether or not they were correctly classified by BERT$_{FT}$. High-perplexity samples are omitted due to excessive amounts of gibberish and nonsensical strings present in the samples. Note that some combinations of cases and prompts do not show responses generated by all four models, because these responses are either all nonsensical strings, or simply do not exist. The shown responses are chosen because they most clearly display the manually observed patterns described in this section.

Manual inspection of the sub-samples (i.e., high or low perplexity, and correctly or incorrectly classified) shows that correctly classified low-perplexity samples from models targeted towards the older age group use more formal and complex words than their younger targeted counterparts, e.g., *quite*, *significant*, *powerful*, or *institutions*. This observation is in line with the findings of Pennebaker and Stone [2003], which suggest that verbal indicators of cognitive abilities are expected to increase until mid-adulthood (see Section 2.1.1). Samples generated by these older targeted models also show recurring topics that are typically associated with older age, e.g., children (*my son* or *your daughter*), history, or politics. Responses about healthcare related subjects are also more common among the samples generated by these models, as indicated by the use of words like *surgery*. This pattern is also observed in the BoWs that are empirically extracted from older dialogues of the BNC dialogue dataset, reported in Section A.1. When viewing high-perplexity samples from the same set of models (i.e., those that are correctly classified and target towards the older age group), there appears to be a substantial increase in the amount of gibberish and nonsensical sequences (space-less sequences of words, repetitions of the same words, or sequences of punctuation marks). Inspection of the incorrectly classified samples from models targeted towards the older class, there is also a considerable increase in the amount of nonsensical strings. Remarkably, there also seem to be more linguistic patterns associated with younger age in this sub-sample, e.g., words of excitement like *favourite* or *best*, informal vocabulary (*pretty much*).

The low-perplexity correctly classified samples generated by models targeted towards the younger age group also contain clear indications of their target age. Similar to the observations from the qualitative analysis of dialogue utterances in Section 3.4.2 and the findings of Schler et al. [2006] relating to similar age groups, for the younger targeted models, there appears to be more usage of slang words, neologisms, swear words, and informal language, like *yeah*, *dude*, *cool*, *kinda*, and *lol*. Furthermore, these sub-samples are also characterized by increased use of words of excitement and exclamation: e.g., *awesome*, *really*, *love*, *fun*, or *amazing*. Especially in the correctly classified low-perplexity responses generated by G-Discrim$_{Young}$, it can be seen that there is a strong presence of topics such as dating (indicated by words such as *girlfriend* or *boyfriend)*, having parents (*my dad*), parties, and student life (*roommate*). An interesting pattern observed in this sub-sample is one often associated with millennial or social networking language: the tendency to end a (serious-sounding) statement with *lol* or *haha* as a means of softening the perceived severity of the statement or as a signal of interlocutor involvement [Newitz, 2019, Tagliamonte and Denis, 2008]. For example, *We have to stop talking about this all going so stupid lol lol*. When inspecting the high-perplexity and/or incorrectly classified samples generated by the younger targeted models, similar patterns are observed. Namely, a substantial increase in non-alphabetical strings, and gibberish.

When comparing the samples from GPT-2 and DialoGPT-based models, it can be seen that the responses generated by DialoGPT are more dialogic, whereas GPT-2 sometimes generates sequences that look more like sentence completions. This is to be expected based on the differences in pre-training methods between the two language models [Zhang et al., 2020]. Furthermore, there appear to be a lot more nonsensical low-perplexity responses generated by DialoGPT-based models. Perplexity remains a rough proxy for fluency, and observations like these confirm this problem. That is, low perplexity often does not imply lack of gibberish or nonsense. DialoGPT's strong bias towards generating younger sounding language is also noticeable in its generated samples. For example, DialoGPT-based models targeted towards the older group still produce lots of words of excitement. However, this older targeted DialoGPT-based model does succeed in producing significantly fewer usages of slang or swear words, when compared to its younger targeted counterpart.

model	**low ppl. | ✓**	**low ppl. | ✗**	**med ppl. | ✓**	**med ppl. | ✗**	**high ppl. | ✓**	**high ppl. | ✗**
G-BoW$_{Young}$	48.0%	15.6%	15.6%	10.4%	6.7%	3.7%
G-BoW$_{Old}$	25.6%	37.7%	11.9%	14.8%	5.6%	4.4%
G-Discrim$_{Young}$	33.3%	15.9%	17.4%	8.5%	17.0%	7.9%
G-Discrim$_{Old}$	23.8%	18.2%	13.4%	3.3%	37.2%	4.1%
D-BoW$_{Young}$	5.2%	0.4%	40.0%	5.2%	43.3%	5.9%
D-BoW$_{Old}$	3.0%	5.9%	10.4%	35.2%	8.5%	37.0%
D-Discrim$_{Young}$	5.9%	2.7%	23.0%	5.6%	56.9%	5.9%
D-Discrim$_{Old}$	7.0%	4.1%	19.3%	11.5%	30.3%	27.8%

Table 4.10: Distribution of dialogue responses generated by various PPLM-setups (under the **model** column) into categories of low, medium, or high perplexity, and whether or not they are correctly classified as their target age group by BERT$_{FT}$ (✓: correctly classified, ✗: incorrectly classified). Alternating row colors are for readability and bear no meaning.

Chapter 5

Discussion

5.1 Summary of Key Findings and Interpretations

In this thesis, the problems of automated age detection from dialogue and age-adaptive controlled dialogue generation are investigated. First, I studied the extent to which purely text-based NLP models can detect age-related linguistic features in dialogue data, and which features drive their predictions. Subsequently, I studied the extent to which age-adaptive dialogue generation is possible using an approach based on Plug-and-Play language models (PPLM) [Dathathri et al., 2020].

The results of the age detection experiments indicated that a fine-tuned version of BERT, BERT$_{FT}$, is capable of detecting age-related linguistic features in dialogue utterances with reasonable accuracy. BERT$_{FT}$ was particularly useful for age detection when the dialogue fragment is long enough to contain discriminative signal. At the same time, it is observed that much simpler models based on n-grams achieve comparable performance, which suggests that, in dialogue, 'local' features can be indicative of the language of speakers from different age groups. This is showed to be the case, with both lexical and stylistic cues being informative to these models in this task.

Furthermore, the age detection results informed the development of controllable dialogue generation systems using PPLM. On the one hand, the presumed locality of age-indicative features, as suggested by the comparable performance achieved by the n-gram models, motivates the use of unigram-based bag-of-words attribute models in PPLM-setups. On the other hand, the use of more sophisticated neural discriminator attribute models in PPLM-setups is motivated by the superior performance of neural discriminators in Experiment 1, and the notion that abstract linguistic styles, like those of specific age groups, are not easily represented as a bag-of-words. This difficulty of expressing an age group's linguistic style as a manually curated bag-of-words motivates the use of empirically generated bag-of-words attribute models (see Section 4.2.1). Specifically, the empirically generated lists of unigrams per age group used as BoW attribute models are (1) the lists of unigrams deemed most informative by Experiment 1's best-performing unigram-based classifier, and (2) the lists of the most frequently used words by a specific age group, after omitting the overlapping frequently used words between both age groups. Overall, the analyses of Experiment 1 predominantly compare the classification performance and feature importance of n-gram and Transformer-based neural models. This comparison is continued in the controlled dialogue generation experiments, in the form of comparing BoWs versus neural discriminators as attribute models in PPLM-setups. Also, the best performing classifier from the age detection experiments, BERT$_{FT}$, is used in the second phase of experiments to evaluate the prompts and generated responses for their resemblance to the linguistic style of the younger or older age group.

The controlled dialogue generation results indicate that it is possible to use PPLM to generate dialogue responses that have been adapted to the styles of different age groups, to an extent that is identifiable by Experiment 1's best classifier. It can be seen that the discriminator-based PPLM-setups typically achieve higher levels of detectable control (i.e., statistical resemblance to a specific writing style) than the BoW-based setups, but generate significantly more perplexing and repetitive responses. This could be attributable to the fact that BoW-based control is more local (i.e., token-level) and less invasive than discriminator-based control, which can operate at the structural level. The results also indicate that the underlying language models used in the PPLM-setups, GPT-2 and DialoGPT, are both biased towards generating younger sounding language (according to the best performing classifier from Experiment 1). This is most likely due to both models having been pre-trained on texts scraped from web pages, which are typically over-represented by millennials [Radford et al., 2019, Zhang et al., 2020]. Quantitative analyses of the dialogue generation results revealed that the outputs of some PPLM-setups show a correlation between improved target probability and worsened perplexity and diversity. This could be interpreted as a tradeoff between control and quality. Also, the style of the prompt (i.e., whether it is classified by BERT$_{FT}$ as younger, neutral, or older) steers the style of the generated response in the direction of the prompt's style. This observation indicates the importance of taking prompt-induced biases into account when developing controlled generation models. It also appears that BoW-based models are significantly worse at controlling for older sounding language than they are for younger sounding language. This could be attributable to the the possibility that overcoming the younger biases of the underlying language models (GPT-2 and DialoGPT), to produce older sounding responses, needs more invasive perturbations than BoW-based attribute models can bring about, given the current parameter-settings. Finally, qualitative inspection of generated responses reveal that differences in style are noticeable by use of age-indicative words, formality of language, and the prevalence of certain topics.

5.2 Limitations

There are several limitations to take into account when interpreting the results presented in this thesis. The generalizability of the age detection results are limited by the coarse granularity with which the age groups are defined. One could imagine that if one aims to leverage the predicted age signal in real-world conversational systems, a fine-grained grouping is more reasonable. Furthermore, age group information remains difficult to identify in short utterances. A non-negligible proportion of utterances seem too short for reliable classification, and it remains to be investigated if NLP models or even humans could identify the targets. Acoustic signals are not considered in this work, but they are important characteristics of speaking style, and could also be taken into account when developing audio-based conversational systems. On the one hand, this richer signal (i.e., the combination of acoustic and textual information) could lead to better classification performance. On the other hand, the use of acoustic signals could lead to ethical privacy-related issues, that aggregated text-based approaches might have to a lesser extent. This is discussed in more detail in Section 5.4.

With respect to the evaluation of the generated responses (see Section 4.2.4), the reliability of the controlled dialogue generation results is impacted by perplexity (measured by GPT-1) being a crude proxy for fluency. Perplexity also lacks generalizability as an evaluation measure, because it only measures the uncertainty assigned by one language model. Automated evaluation of fluency could be improved by considering an aggregated perplexity measure (i.e., perplexity averaged over an ensemble of language models). Alternatively, indications of a generated response's fluency could be made more reliable by having humans participate in an evaluation study to rate the responses, and taking their opinions into account.

It is also important to realize that the learned representations for younger and older style used in this research are specific to the BNC [Love et al., 2017], and should not be interpreted as general representations of the speaking style of people of ages 19-29 and 50-plus. Even within the context of the BNC, the representations used for classification and generation correspond to textual features which have been learned by Experiment 1's models to have a high likelihood of coming from a younger or older speaker, and are therefore approximations of the age groups' speaking styles.

Finally, due to the lack of age-labeled transcribed spontaneous dialogue data, I was constrained to using the same dialogue dataset to develop the attribute models (i.e., the BoWs or neural discriminators used in PPLM-setups), and to train evaluation model (i.e., BERT$_{FT}$). It must be noted that this is still sound practice, because the attribute and evaluation models are not trained jointly, and do not share representations. Moreover, the approach of using an external classifier for evaluation is based on the work of Dathathri et al. [2020] and Madotto et al. [2020]. At the same time, the assigned target probabilities indicating attribute adherence would be more generalizable if the evaluation model were trained on a different dataset of spontaneous dialogue utterances between speakers of the same age.

5.3 Future Research Directions

While the classification task in Experiment 1 was performed at the level of single dialogue utterances, future work may take into account larger dialogue fragments, such as the entire dialogue or a fixed number of turns. This would make the setup more comparable to discourse, but but would require making experimental choices and dealing with extra computational challenges. Moreover, it could be tested whether the language used by a speaker is equally discriminative when talking to a same-age (this work) or a different-age interlocutor. Also, future work might involve leveraging acoustic signals for automated age-detection and controlled generation. Acoustic signals can contain important features indicative of speaking style, and could therefore substantially improve classification and controlled generation performance.

Developing interactive age-adaptive conversational systems was beyond the scope of this thesis. Therefore, an interesting future research objective would be to use Plug-and-Play Conversational Models (PPCM) for age-adaptive dialogue generation [Madotto et al., 2020]. PPCM solves a latency problem of PPLM, making it usable in interactive settings. However, the use of PPCM in combination with more abstract speaking styles and BoW-based attribute models has not been researched (to the best of my knowledge). Using PPCM for age-adaptive dialogue generation would require developing and training residual adapter modules, and generating style-specific dialogue datasets. Furthermore, the age detection results suggest the importance of n-gram features for age-identification. Therefore, an interesting future research direction would be to adapt PPLM and PPCM to be compatible with n-gram lists (for $n > 1$) as attribute models. This would require finding a way to by-pass the need to re-train large underlying language models like GPT-2 and DialoGPT for arbitrary n-grams. Finally, the current set of tools to analyze which features drive controlled generation in PPLM-setups is limited. It would be valuable to develop probing methods for PPLM-setups. Namely, visualizations or saliency methods targeted towards understanding the activation space perturbations made by PPLM could provide important insights about the effects of attribute models on controlled generated output.

5.4 Ethical and Environmental Considerations

There is growing concern around the ethics, environmental costs, and societal dangers associated with powerful large-scale language models [Brown et al., 2020, Bender et al., 2021]. There is a non-negligible societal risk accompanying the recent trend of deploying ever larger language models that are more capable of producing texts that are indistinguishable from human-written ones. Namely, such language models can be used to spread misinformation (e.g., fake news), or to perpetuate harmful social biases. Plug-and-Play language models carry the same risks, if not to a greater extent, as they can be used to produce personalized, and ultimately more convincing text. As seen in the analysis of generated responses in Section 4.4.5, some examples of the models' outputs can be seen as exacerbating biases present in the used dialogue dataset (and society) (e.g., older speakers talking about healthcare, younger speakers talking about partying). This bias-amplification could be attributable to the models also to picking up and generating topic-related features, aside from purely stylistic ones. Future considerations for work on controlled generation should therefore also include methods to ensure adaptation happens at the stylistic level, and avoid amplification of (topical) biases in the data being used.

Furthermore, as previously mentioned in Section 5.2, style-detection performance of adaptive dialogue systems could be improved by the use of acoustic signals, in addition to purely text-based information. However, privacy concerns should be taken into account when augmenting signals for dialogue systems, as this leads to more user-information being processed, making users easier to identify. Researchers should therefore consider adaptation methods that identify users with as sparse a signal as possible.

However, Dathathri et al. [2020] also demonstrate the benevolent applicability of PPLM for language detoxification, which can help to reduce the perpetuation of harmful biases. Furthermore, Plug-and-Play methods, which avoid substantial re-training costs of massive language models, can help to reduce the energy consumption and dampen the environmental impact of developing modern deep learning models [Strubell et al., 2019]. Additionally, the same methods that generate convincing texts (like GPT-2) can be used to build more sophisticated tools for detecting artificially generated texts Gehrmann et al. [2019]. Overall, I believe that the positive applications of PPLM for, e.g, personalized dialogue generation, outweigh the potential risks of misuse.

Chapter 6

Conclusion

In this thesis, I first investigate the extent to which text-based NLP-models can detect age-related linguistic features in dialogue data, and which features drive their predictions. Then, I study the extent to which age-adaptive dialogue generation is possible using Plug-and-Play language models (PPLM) [Dathathri et al., 2020].

The results of the age detection experiments show that a fine-tuned BERT model is capable of detecting age-related linguistic features in dialogue utterances with reasonable accuracy, especially when the dialogue fragment is long enough to contain discriminative signal. However, simpler n-gram-based models achieve comparable performance, suggesting that, in dialogue, 'local' features can be indicative of the language of speakers from different age groups. This is shown to be the case, with both lexical and stylistic cues being informative to these models in this task. Furthermore, the age detection results informed the subsequent experiments about controlled dialogue generation using PPLM.

The controlled dialogue generation results show that it is possible to use PPLM to generate dialogue responses that possess detectable linguistic features associated with specific age groups. Discriminator-based PPLM-setups typically achieve higher levels of detectable control than bag-of-words (BoW) based setups, but generate significantly more perplexing and repetitive responses. This could be attributable to the fact that BoW-based control is more local (i.e., at the token-level) and less invasive than discriminator-based control, which can operate at the structural level.

Overall, I believe that the research presented in this thesis is a promising step towards the development of adaptive conversational systems. In particular, the development of age-adaptive conversational systems can benefit from these results. Because consistent language style differences were found between age groups, systems whose language generation capabilities aim to be consistent with a given age group should reproduce these patterns. This could be achieved, as shown in this work, by embedding Plug-and-Play modules that control the generation of a system's output, which could lead to more natural interactions between human speakers and a dialogue system. In an increasingly automated society, where dialogue systems often takeover simple tasks related to, e.g., customer service, perceived natural interaction between users and these systems is crucial to their optimal functioning. And finally, the development of dialogue systems with which users feel they can naturally interact brings us closer to achieving the longstanding goal of AI-powered human-like conversation.

Bibliography

Central Limit Theorem, pages 66–68. Springer New York, New York, NY, 2008. ISBN 978-0-387-32833-1. doi: 10.1007/978-0-387-32833-1_50. URL https://doi.org/10.1007/978-0-387-32833-1_50.

E. E. Abdallah, J. R. Alzghoul, and M. Alzghool. Age and gender prediction in open domain text. *Procedia Computer Science*, 170:563–570, 2020.

L. J. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *CoRR*, abs/1607.06450, 2016. URL http://arxiv.org/abs/1607.06450.

D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1409.0473.

A. Bapna and O. Firat. Simple, scalable adaptation for neural machine translation. 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 1538–1548, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1165. URL https://www.aclweb.org/anthology/D19-1165.

J. Bastings and K. Filippova. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 149–155, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.blackboxnlp-1.14. URL https://aclanthology.org/2020.blackboxnlp-1.14.

E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. 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, 2021.

C. M. Bishop. *Pattern recognition and machine learning*. springer, 2006.

D. Bohm and L. Nichol. *On dialogue*. Routledge, 2013.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

M. Burtsev, A. Seliverstov, R. Airapetyan, M. Arkhipov, D. Baymurzina, N. Bushkov, O. Gureenkova, T. Khakhulin, Y. Kuratov, D. Kuznetsov, A. Litinsky, V. Logacheva, A. Lymar, V. Malykh, M. Petrov, V. Polulyakh, L. Pugachev, A. Sorokin, M. Vikhreva, and M. Zaynutdinov. DeepPavlov: Open-source library for dialogue systems. In *Proceedings of ACL 2018, System Demonstrations*, pages 122–127, Melbourne, Australia, July

2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-4021. URL https://aclanthology.org/P18-4021.

Y.-N. Chen, A. Celikyilmaz, and D. Hakkani-Tür. Deep learning for dialogue systems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 8–14, Vancouver, Canada, July 2017. Association for Computational Linguistics. URL https://aclanthology.org/P17-5004.

K. Clark, U. Khandelwal, O. Levy, and C. D. Manning. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy, Aug. 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4828. URL https://aclanthology.org/W19-4828.

N. Dai, J. Liang, X. Qiu, and X. Huang. Style transformer: Unpaired text style transfer without disentangled latent representation. *arXiv preprint arXiv:1905.05621*, 2019.

S. Dathathri, A. Madotto, J. Lan, J. Hung, E. Frank, P. Molino, J. Yosinski, and R. Liu. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1edEyBKDS.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

A. Fan, M. Lewis, and Y. Dauphin. 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, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1082. URL https://aclanthology.org/P18-1082.

J. Ficler and Y. Goldberg. Controlling linguistic style aspects in neural language generation. In *Proceedings of the Workshop on Stylistic Variation*, pages 94–104, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4912. URL https://aclanthology.org/W17-4912.

S. Gehrmann, H. Strobelt, and A. Rush. GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-3019. URL https://aclanthology.org/P19-3019.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

S. Ivanov. The impact of automation on tourism and hospitality jobs. *Information Technology & Tourism*, 22(2):205–215, 2020.

S. Jain and B. C. Wallace. Attention is not Explanation. 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 3543–3556, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1357. URL https://aclanthology.org/N19-1357.

N. S. Keskar, B. McCann, L. Varshney, C. Xiong, and R. Socher. CTRL - A Conditional Transformer Language Model for Controllable Generation. *arXiv preprint arXiv:1909.05858*, 2019.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015. URL http://arxiv.org/abs/1412.6980.

D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

S. Kottur, X. Wang, and V. Carvalho. Exploring personalized neural conversational models. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3728–3734, 2017. doi: 10.24963/ijcai.2017/521. URL https://doi.org/10.24963/ijcai.2017/521.

S. Kullback and R. A. Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

P. Kushneryk, Y. P. Kondratenko, and I. V. Sidenko. Intelligent dialogue system based on deep learning technology. In *ICTERI PhD Symposium*, 2019.

G. Lample, S. Subramanian, E. Smith, L. Denoyer, M. Ranzato, and Y.-L. Boureau. Multiple-attribute text rewriting. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=H1g2NhC5KQ.

B. Lester, R. Al-Rfou, and N. Constant. The power of scale for parameter-efficient prompt tuning, 2021.

C. Li, X. Gao, Y. Li, B. Peng, X. Li, Y. Zhang, and J. Gao. Optimus: Organizing sentences via pre-trained modeling of a latent space. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4678–4699, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.378. URL https://www.aclweb.org/anthology/2020.emnlp-main.378.

J. Li, M. Galley, C. Brockett, G. Spithourakis, J. Gao, and B. Dolan. 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, Aug. 2016a. Association for Computational Linguistics. doi: 10.18653/v1/P16-1094. URL https://aclanthology.org/P16-1094.

J. Li, W. Monroe, A. Ritter, D. Jurafsky, M. Galley, and J. Gao. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Austin, Texas, Nov. 2016b. Association for Computational Linguistics. doi: 10.18653/v1/D16-1127. URL https://aclanthology.org/D16-1127.

J. Li, R. Jia, H. He, and P. Liang. 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, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1169. URL https://aclanthology.org/N18-1169.

M. Li, K. J. Han, and S. Narayanan. Automatic speaker age and gender recognition using acoustic and prosodic level information fusion. *Computer Speech & Language*, 27(1):151–167, 2013.

D. C. Liu and J. Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1):503–528, 1989.

Q. Liu, Y. Chen, B. Chen, J.-G. Lou, Z. Chen, B. Zhou, and D. Zhang. You impress me: Dialogue generation via mutual persona perception. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1417–1427, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.131. URL https://aclanthology.org/2020.acl-main.131.

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

A. Madotto, E. Ishii, Z. Lin, S. Dathathri, and P. Fung. Plug-and-play conversational models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2422–2433, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.219. URL `https://www.aclweb.org/anthology/2020.findings-emnlp.219`.

M. McTear. Conversational ai: Dialogue systems, conversational agents, and chatbots. *Synthesis Lectures on Human Language Technologies*, 13(3):1–251, 2020.

Merriam-Webster. Dialogue. In *Merriam-Webster.com dictionary*. n.d. URL `https://www.merriam-webster.com/dictionary/dialogue`. Retrieved November 20, 2021.

K. Mo, Y. Zhang, S. Li, J. Li, and Q. Yang. Personalizing a dialogue system with transfer reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

A. Newitz. You'e been monetized lol. *New Scientist*, 243(3240):24, 2019.

A. Nguyen, J. Yosinski, and J. Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. IEEE, 2015.

A. Nguyen, J. Clune, Y. Bengio, A. Dosovitskiy, and J. Yosinski. Plug & play generative networks: Conditional iterative generation of images in latent space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017.

D. Nguyen, N. A. Smith, and C. P. Rosé. 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, June 2011. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/W11-1515`.

D. Nguyen, D. Trieschnigg, A. S. Doğruöz, R. Gravel, M. Theune, T. Meder, and F. De Jong. Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1950–1961, 2014.

G. Oppy and D. Dowe. The turing test. 2003.

J. W. Pennebaker and L. D. Stone. Words of wisdom: Language use over the life span. *Journal of Personality and Social Psychology*, 85(2):291–301, 2003. URL `https://doi.org/10.1037/0022-3514.85.2.291`.

S. Prabhumoye, A. W. Black, and R. Salakhutdinov. Exploring controllable text generation techniques. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1–14, Barcelona, Spain (Online), Dec. 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.1. URL `https://aclanthology.org/2020.coling-main.1`.

Q. Qian, M. Huang, H. Zhao, J. Xu, and X. Zhu. 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, 7 2018. doi: 10.24963/ijcai.2018/595. URL `https://doi.org/10.24963/ijcai.2018/595`.

A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving language understanding by generative pre-training. 2018.

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

Y.-P. Ruan, Z.-H. Ling, Q. Liu, Z. Chen, and N. Indurkhya. Condition-transforming variational autoencoder for conversation response generation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7215–7219. IEEE, 2019.

A. Rush. The annotated transformer. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 52–60, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-2509. URL https://aclanthology.org/W18-2509.

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

M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.

R. Sennrich, B. Haddow, and A. Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, Aug. 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL https://www.aclweb.org/anthology/P16-1162.

K. Shuster, E. M. Smith, D. Ju, and J. Weston. Multi-modal open-domain dialogue. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4863–4883, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.emnlp-main.398.

F. Stahlberg, J. Cross, and V. Stoyanov. Simple fusion: Return of the language model. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 204–211, Brussels, Belgium, Oct. 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6321. URL https://www.aclweb.org/anthology/W18-6321.

E. Strubell, A. Ganesh, and A. McCallum. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1355. URL https://aclanthology.org/P19-1355.

S. A. Tagliamonte and D. Denis. Linguistic ruin? lol! instant messaging and teen language. *American speech*, 83(1):3–34, 2008.

D. R. Traum and P. A. Heeman. Utterance units in spoken dialogue. In *Workshop on Dialogue Processing in Spoken Language Systems*, pages 125–140. Springer, 1996.

M. J. van der Goot and T. Pilgrim. 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, 2019.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

J. Vig. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-3007. URL https://www.aclweb.org/anthology/P19-3007.

S. Welleck, J. Weston, A. Szlam, and K. Cho. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1363. URL https://aclanthology.org/P19-1363.

S. Wiegreffe and Y. Pinter. Attention is not not explanation. 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 11–20, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1002. URL https://aclanthology.org/D19-1002.

M. Wolters, R. Vipperla, and S. Renals. Age recognition for spoken dialogue systems: Do we need it? In *Tenth Annual Conference of the International Speech Communication Association (Interspeech)*, 2009.

G. Zeng, W. Yang, Z. Ju, Y. Yang, S. Wang, R. Zhang, M. Zhou, J. Zeng, X. Dong, R. Zhang, H. Fang, P. Zhu, S. Chen, and P. Xie. MedDialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.743. URL `https://www.aclweb.org/anthology/2020.emnlp-main.743`.

S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston. 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, July 2018a. Association for Computational Linguistics. doi: 10.18653/v1/P18-1205. URL `https://aclanthology.org/P18-1205`.

Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and B. Dolan. Dialogpt: Large-scale generative pre-training for conversational response generation. In *ACL, system demonstration*, 2020.

Z. Zhang, S. Ren, S. Liu, J. Wang, P. Chen, M. Li, M. Zhou, and E. Chen. Style transfer as unsupervised machine translation. *CoRR*, abs/1808.07894, 2018b. URL `http://arxiv.org/abs/1808.07894`.

Y. Zheng, G. Chen, M. Huang, S. Liu, and X. Zhu. Personalized dialogue generation with diversified traits. *arXiv preprint arXiv:1901.09672*, 2019.

Appendix A

Supplementary material

A.1 Wordlists used as for BoW Attribute Models

What follows are the explicit wordlists used as attribute models for BoW-based PPLM setups. The empirical approaches used to extract the wordlists are described in Section 4.2.1. "*" indicates masking of foul language.

100 Most Informative Unigrams - Young (19-29) um, cool, sh*t, hmm, uni, cute, tut, massive, awesome, gym, b*tch, lol, grand, pizza, like, excited, yawn, Korea, cigarette, f*ck, fairness, Jesus, annoying, Facebook, quicker, definitely, guess, Sunderland, oo, wanna, mountain, scared, piss, love, miss, Middlesbrough, mhm, specifically, ooh, website, roundabout, photo, nope, blanket, management, ridiculous, mental, pregnant, beers, hate, log, f*cking, cry, cheaper, skinny, plural, burger, hilarious, hint, drunk, fridge, cousin, coke, genuinely, James, mates, smaller, option, balance, saving, basically, leather, nev, shut, frig, mate, yay, invite, maid, nickname, badly, garlic, CD, jokes, Uzbekistan, boyfriend, date, added, Manchester, blah, sh*tty, lang, tempted, stadium, wee, eh, baking, city, honestly, exam

100 Most Informative Unigrams - Old (50 plus) ordinary, Chinese, wonderful, yes, tend, father, photographs, vegetables, hospice, operation, shed, pension, areas, mother, hanging, hospices, glasses, chap, anyhow, tank, surgery, container, cheers, born, church, pain, several, workshop, right, horses, building, extraordinary, vegetarian, biscuit, americano, engine, luck, paint, emperor, lipsy, trombone, occasional, supper, lord, architect, council, roast, schools, bath, asbestos, endometrial, concrete, poodle, recall, diabetes, misty, report, heavens, enormous, lawn, potatoes, email, junk, scabies, mousse, Ebola, churches, sewing, plants, rackets, marmalade, engineering, furniture, photograph, sandwiches, unemployment, xylophone, Piccadilly, flu, claim, arab, nineteen, forgotten, sensible, blancmange, spencer, yards, emails, yellow, scruffy, fungi, garden, boiler, lodge, mostly, Robson, tricky, shark, robin, contracture

Frequency-based Young (19-29) um, sh*t, cool, f*cking, definitely, guess, friends, everyone, literally, dad, sounds, weekend, loads, watch, fair, f*ck, amazing, friend, ha, huh, hate, fun, stay, girl, holiday, blah, hours, uni, month, horrible, massive, Friday, stupid, film, parents, thirty, spend, mate, honest, change, hope, yourself, annoying, wear, wait, ridiculous, anyone, Saturday, tea, dinner, sit, crazy, hell, pound, nine, expensive

Frequency-based Old (50 plus) building, may, water, mother, perhaps, door, lots, business, cancer, area, although, worked, open, cut, number, under, young, nineteen, everybody, garden, church, case, shop, children, certainly, set, coffee, email, gave, white, along, doctor, hear, often, possibly, group, father, outside, wonderful, taken, seem, places, green, given, hand, early, women, space, front, language, dear, light, huge, supposed, country, hospital, otherwise, asked, putting, bits, gosh, wall, woman, almost, particularly, across, word, age, rest, flat, turned, decided, finished, needed, red, bin, hospice, running, slightly, its, middle, local, percent, Chinese, paper, check, high, milk, piece, near, nobody, usually