How to Control Sentiment in Text Generation: A Survey of the State-of-the-Art in Sentiment-Control Techniques

Michela Lorandi and **Anya Belz**
ADAPT Research Centre, Dublin City University
{michela.lorandi, anya.belz}@adaptcentre.ie

Abstract

Recent advances in the development of large Pretrained Language Models, such as GPT, BERT and Bloom, have achieved remarkable performance on a wide range of different NLP tasks. However, when used for text generation tasks, these models still have limitations when it comes to controlling the content and style of the generated text, often producing content that is incorrect, irrelevant, or inappropriate in the context of a given task. In this survey paper, we explore methods for controllable text generation with a focus on sentiment control. We systematically collect papers from the ACL Anthology, create a categorisation scheme based on different control techniques and controlled attributes, and use the scheme to categorise and compare methods. The result is a detailed and comprehensive overview of state-of-the-art techniques for sentiment-controlled text generation categorised on the basis of how the control is implemented and what attributes are controlled and providing a clear idea of their relative strengths and weaknesses. [1]

1 Introduction

In recent years, there has been a surge of interest in developing algorithms and models for Controllable Text Generation (CTG). This research field aims to enable users to generate text with specific attributes, controlling e.g. the text's sentiment, topic, or level of formality. In this survey paper, we focus on state-of-the-art CTG techniques that control sentiment. We provide a comprehensive overview of the existing literature and categorise approaches based on their implementation of control, and which specific attributes they control.

Our main contributions are as follows:

- We propose a categorisation scheme for Sentiment-Controlled Text Generation specifying control attributes and how control is implemented.
- We analyse state-of-the-art techniques for Sentiment Control, and categorise each technique using the proposed categorisation scheme.
- We compare the selected papers in terms of performance, efficiency and generality.

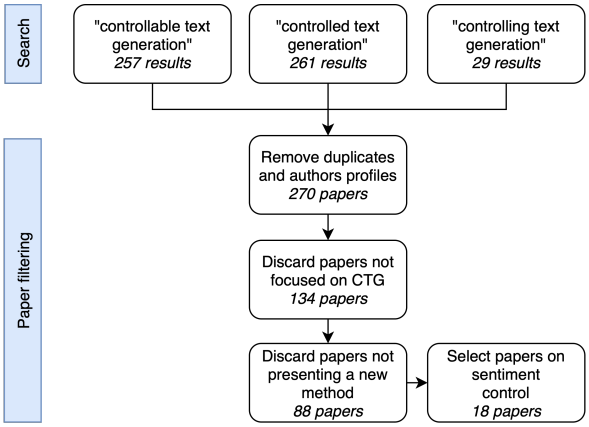

Figure 1: Paper selection process from ACL Anthology to which we added 1 paper (see in text).

The paper is structured as follows. Section 2 summarises two previous related survey papers, while Section 3 outlines the scope of the present survey, the method we used for systematic paper selection, and some high-level statistics for the selected papers. The proposed categorisation scheme is described in Section 4, consisting of (i) different types of controlled attributes (Section 4.1), and (ii) different types of control implementation techniques (Section 4.2). Section 5 describes the CTG techniques from the surveyed papers in terms of the categorisation scheme, including which attributes are controlled. Section 6 compares the different techniques in terms of their generality, performance, and efficiency. We conclude with suggested future directions (Section 7), some discussion (Section 8) and final remarks (Section 9).

[1] The categorised list of papers can be found in our GitHub repository https://github.com/DCU-NLG/sentimentCTG-survey

Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 341–353

2 Related Research

Prabhumoye et al. (2020) propose a schema of the language generation pipeline based on five components that control the generation process: external input, sequential input, generator operations, output, and training objective. They argue that control of specific attributes requires modification of these five components, and present an overview of existing control techniques in terms of which component different techniques use to exert the control. The work focuses on how the proposed schema can be applied to enable control of text generation with a particular focus on autoregressive models.

As part of a general introduction and overview of techniques in pretrained language model (PLM) based CTG and evaluation methods, Zhang et al. (2022) propose a set of control conditions (semantic, structural, lexical), and broadly group together methods for CTG into finetuning, retraining/refactoring, and postprocessing. The work addresses only Transformer-based PLMs, and distinguishes seven subtypes of methods, based on how the control signal works with the PLM.

In this survey, we consider all types of methods that have been used for sentiment-controlled Text Generation, not just Transformer-based PLMs, and we conduct a systematic paper selection process. We provide a categorisation scheme based on Control Attribute Types and Control Implementation Techniques that we use to characterise and compare the selected methods. Finally, we provide comparisons in terms of performance and efficiency.

3 Survey Scope and Paper Selection

This paper aims to fill a gap in the current literature by surveying recent models applied to Controllable Text Generation (CTG) with a specific focus on sentiment control. Furthermore, we propose a categorisation of the selected papers based on controlled attributes, and how the control is implemented.

We conducted an otherwise unrestricted search on the ACL Anthology using the keywords "controllable text generation," "controlled text generation" and "controlling text generation," as shown in Figure 1. After removing duplicates, authors' profiles, and non-paper resources, we obtained 270 papers. From this original pool, we discard papers that are not strictly related to CTG, such as papers that mention CTG but do not explore the task. Next we only retain papers which present a new model or control method, discarding those

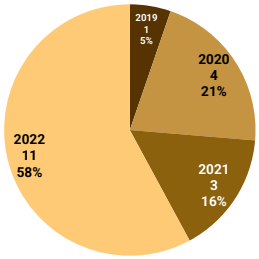

Figure 2: Distribution of selected papers across years.[2]

that e.g. only present a new dataset or perform a comparative study. After filtering, we are left with 88 papers from which for the present survey we select all papers implementing sentiment control, thus giving us 18 papers. We add one more relevant paper (Dathathri et al., 2019) not present in the ACL Anthology that was cited in our pool of papers. Table 1 lists the resulting 19 papers in the left-most column.

The 19 papers span the period 2019 (1 paper), 2020 (4 papers), 2021 (3 papers), and 2022 (11 papers), illustrating the rapidly growing interest in this topic, as shown in Figure 2 . Papers report work using Complete Training techniques (3 papers), Model Fine-Tuning (3), Disentanglement (1), Modification of Token Distribution (6), and Hybrid techniques (6). In 10 of the papers, multiple attributes are controlled simultaneously, whereas in 9, single attributes are controlled one at a time. 14 papers are designed for free text generation (rather than a specific task), 2 methods are for Story Generation, 2 for Conversational Agents, and one for Topic to Essay Generation. We return to properties of techniques in more detail in Section 6 and Table 1.

4 Categorisation Scheme

We collect all selected papers and annotate them based on different aspects, such as control attributes addressed and architecture used to solve the control problem with a specific focus on how the control is implemented and embedded in the proposed architecture. Using the collected information, we cluster control attributes and models to create a categorisation scheme for Sentiment-Controlled Text Generation in which we specify types of controlled attributes and types of control implementation. The created categorisation scheme will be used to cate-

[2]The proposed scheme is specific for Sentiment-Controlled Text Generation, but we are currently working on a general scheme for CTG.

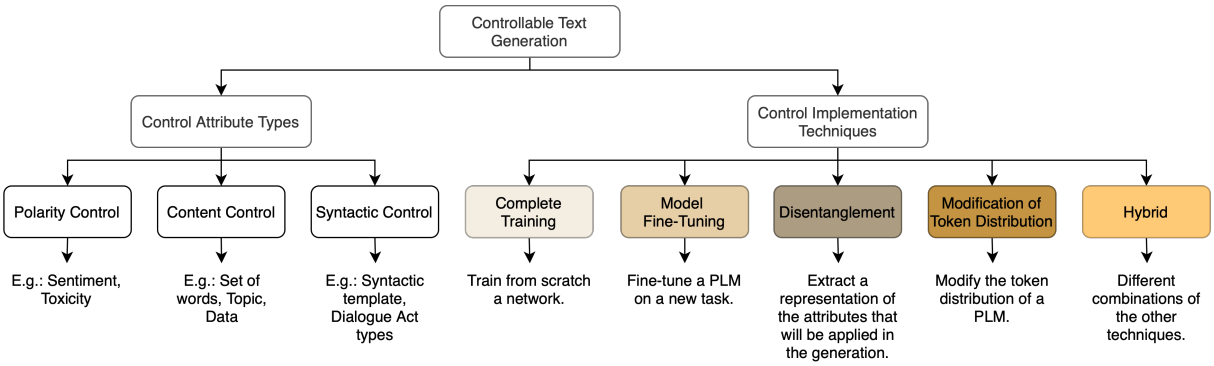

Figure 3: Categorisation of Sentiment-Controlled Text Generation methods[3] considering Control Attributes Types (left) and Control Implementation Techniques (right).[2]

gorise and describe the studied papers.

4.1 Types of controlled attributes

Sentiment-Controlled Text Generation methods control different attributes, such as sentiment, set of words, and topics. In order to help gain insights into different control attributes currently in use and how they are controlled and combined in different methods, we distinguish three Control Attribute Types (Figure 3 left): Polarity Control, Content Control, and Syntactic Control.

Polarity Control covers attributes that control whether otherwise similar content is expressed with positive vs. negative judgment, toxic vs. neutral attitude, different political stance, or one of several competing perspectives e.g. in sport. For example, the negative sentence *The office is poorly maintained and dirty at all times* and the positive sentence *The office is well maintained and clean at all times* have a different polarity, but are otherwise similar in content, i.e. both are about cleaning and maintenance of an office. In the 19 papers in our survey, we encountered two attribute of this type: *Sentiment* and *Toxicity*.

Content Control attributes control the core content of a text. For example, the set of words *{burger, good, fries}* may be used to enforce presence of these words in the output sentence, e.g.: *The **burger** was very **good** and the **fries** are excellent.* In our 19 papers we encountered 4 attriibutes of this type: *Topic*, *Action*, *Character*, *Need*.

Syntactic Control attributes control the structure, syntax, and grammar of the output sentence. For example, we can give the model a syntactic template *(S (S) (,) (CC) (S))* to generate the final text *the film is a visual treat, but almost unsurpassed.* (Yang et al., 2021). In our 19 papers, there was just one attribute of this type: *Tense*.

A system can in principle simultaneously control multiple attributes of the same or different types, thus enabling more fine-grained control. Table 1 lists the seven control attribute types encountered in the 19 papers in our survey, and which paper implements control over which attribute, in column 5.

4.2 Types of control implementation

We categorise controllable text generation techniques based on how they implement the control, as shown in Figure 3, right half, distinguishing four types: Complete Training, Model Fine-Tuning, Disentanglement, Modification of Token Distribution, and Hybrid.

Complete Training techniques train new models from scratch in order to obtain specialised models specifically trained for a Controllable Text Generation task. **Model Fine-Tuning** techniques use a pre-trained model which is fine-tuned to adapt it to the new task including control attributes. **Disentanglement** techniques extract a representation of the control attributes which is applied to steer text generation toward the specified attributes. **Modification of Token Distribution** techniques modify the token distribution of a pre-trained model in order to steer text generation. **Hybrid** techniques include two or more types of the above control implementation techniques.

Table 1 shows the control implementation type(s) addressed by each of the 19 papers in our survey in column 4.

[3]The colours have been tested with Adobe Color Accessibility Tools Color Blind Safe.

Work	Model	Task	Control	Control Attributes							Sentiment Datasets
				S	T	To	A	C	Te	N	
Complete Training											
Qiao et al. (2020)	CVAE	TEG	Multiple	✓	✓						ZHIHU (Feng et al., 2018)
Betti et al. (2020)	GAN	FT	Single	✓	✓						Movie Reviews (Socher et al., 2013), Customer Reviews (Hu and Liu, 2004)
Xie et al. (2022)	Enc-Dec	SG	Multiple	✓				✓		✓	Story Commonsense
Model Fine-Tuning											
Qian et al. (2022)	GPT-2	FT	Multiple	✓	✓	✓					PPLM Prompts
Gu et al. (2022b)	BERT	FT	Multiple	✓	✓	✓					PPLM Prompts
Fang et al. (2022)	MA	FT	Multiple	✓	✓				✓		Yelp
Disentanglement											
Yu et al. (2021)	GPT-2	FT	Multiple	✓	✓						PPLM Prompts
Modification of Token Distribution											
Dathathri et al. (2019)	MA	FT	Multiple	✓	✓						PPLM Prompts
Madotto et al. (2020)	MA	CA	Single	✓	✓						(Adiwardana et al., 2020) prompts
Goswamy et al. (2020)	GPT-2	FT	Multiple	✓	✓						-
Kumar et al. (2022)	MA	FT	Single	✓		✓					PPLM Prompts
Gu et al. (2022a)	MA	FT	Single	✓	✓	✓					PPLM Prompts
Landsman et al. (2022)	MA	FT	Single	✓							OpenWebText Corpus Prompts
Hybrid											
Wang et al. (2022)	BART	SG	Multiple	✓			✓	✓			ROCStory (Rashkin et al., 2018)
Tian et al. (2022)	Enc-Dec	CA	Single	✓							weibo.com (Shang et al., 2015)
Liu et al. (2021)	GPT-2	FT	Single	✓		✓					OpenWebText Corpus Prompts
Zhang and Song (2022)	GPT-2	FT	Single	✓		✓					OpenWebText Corpus Prompts
Krause et al. (2021)	MA	FT	Single	✓	✓	✓					Bookcorpus (Zhu et al., 2015)
Liu et al. (2022)	GPT-2	FT	Multiple	✓	✓						IMDb, OpeNER (Agerri et al., 2013), SenTube (Uryupina et al., 2014)

Table 1: Overview of methods for Sentiment Control in Text Generation. Models: MA=Model Agnostic. Tasks: SG=Story Generation, TEG=Topic to Essay Generation, FT=Free Text, CA=Conversational Agent. Control Attributes: S=Sentiment, T=Topic, To=Toxicity, A=Action, C=Character, Te=Tense, N=Need.

5 Characterisation of CTG Techniques

Table 1 provides an overview of the 19 papers we survey (listed in column 1), in terms of the control implementation technique used (headings inserted into the rows), the type of model used (column 2), the NLP task implemented by the model (column 3), the attributes controlled by the technique (column 5), whether a single or multiple attributes are controlled at a time (column 4), and finally the datasets used in training (column 6).

In the remainder of this section, we summarise each of the 19 papers in our survey in more detail, grouped together in terms of the control implementation technique used.

5.1 Complete Training

Betti et al. (2020) propose a text GAN composed of one generator and two discriminators. The generator is a Relational Memory with self-attention (Santoro et al., 2018) with the objective to generate text consistent with the specified control attribute. The syntax discriminator distinguishes between real and generated sentences, while the semantic discriminator assesses whether the generated sentence expresses the control attribute, e.g. positive sentiment. To solve the well-known problem of differentiation in GANs applied to text, the Gumbel-softmax trick (Jang et al., 2016) is applied. This approach enables control only for one attribute at a time and it has been evaluated on sentiment and topic control.

In order to enable multi-attribute control, Qiao et al. (2020) propose a Sentiment-Controllable topic-to-essay generator that deploys a Conditional Variational Auto-Encoder in adversarial training. The model simultaneously controls the topics of the essay and the sentiment of each sentence composing the essay. The topic control is achieved using a Topic Graph Attention, which includes a topic knowledge graph in the generation process. Sentiment control is achieved by injecting the sentiment representation both in the encoder and the

decoder.

In a different direction, Xie et al. (2022) propose a psychology-guided story generation method that controls storytelling as the protagonist's psychological state changes. This technique enables multi-attribute control considering the protagonist of the story (Character), their chain of emotions (Emotion), and chain of needs (Need) representing the evolution of the psychological state of the protagonist. The model is an encoder-decoder architecture with the addition of psychology controllers designed to integrate the local and global psychological state into the story context representation.

5.2 Model Fine-Tuning

Model Fine-Tuning can be achieved in many ways. One way is to focus on prefix tuning, i.e. fine-tuning a model to extract continuous attribute-specific vectors that will be prepended to the activations of the pre-trained model to steer text generation. E.g., Qian et al. (2022) fine-tune GPT-2 (Radford et al., 2019) to obtain prefixes, but they use the contrast between prefixes, for example, positive vs negative sentiment, to encourage the desired attribute and discourage the opposite attribute. In this method, only the prefixes are trained and GPT-2 weights are kept frozen.

Similarly, Gu et al. (2022b) fine-tune BERT (Devlin et al., 2019) to obtain prefixes. The idea is to have an Autoencoder structure, i.e. the encoder-decoder reconstructs the input sentence, to map attribute-relevant sentences to latent representations of attributes. At inference time, the model searches the attribute representation in the attribute space and uses it as a prefix for the decoder. In the case of multiple attributes the intersection of attributes is taken as the prefix, instead of contrastive prefixes (Qian et al., 2022). In this setting, the decoder is fixed, while the encoder is fine-tuned to get the attribute representations.

Fang et al. (2022) further explore the usage of Variational Autoencoders to learn a latent representation of control attributes. The idea is to use contrastive learning to separate the latent space into several parts, thus obtaining learnable vectors associated with a control attribute. At inference time, all the vectors associated with the desired attribute are extracted and combined with a Dirichlet distribution to produce a latent variable, which is fed to the decoder.

All three methods allow the control of multiple attributes (sentiment and topic) at the same time. The last supports control of the tense of the sentences together with the other attributes.

5.3 Disentanglement

Yu et al. (2021) learn an alignment function to transform the control attribute into an aligned attribute representation. The Bayes rule is used to separate attributes encouraging the alignment function to encode different attributes to different representations. The aligned representation is given to a pre-trained LM (PLM) to steer the generation toward the given control attributes. This method enables control of multiple attributes at the same time (sentiment and topic).

5.4 Modification of Token Distribution

Dathathri et al. (2019) propose a Plug and Play Language Model (PPLM) which uses external attribute classifiers to guide text generation without requiring any training of the PLM. The PLM is used to obtain the next token distribution, which is fed to external classifiers, called Attribute Models, to assess whether the token correctly expresses the desired attributes. The internal latent representations of the LM are updated with a backward pass using the gradients of the attribute models to increase the likelihood of the desired attributes. Finally, the next token distribution is recomputed taking into account the updated latent representations. This model allows control of multiple attributes at a time, such as sentiment and topic.

Inspired by this work, Madotto et al. (2020) propose a variation of PPLMs in which the backward pass is executed n times depending on the desired intensity of the control attribute. Furthermore, they add Residual Adapters (Houlsby et al., 2019) on top of each transformer layer to steer the PLM output distribution without changing its parameters.

Goswamy et al. (2020) propose a different variation of PPLMs based on GPT-2, in which a modified loss is considered to take into account the intensity of the controlled sentiment. Furthermore, instead of considering only positive/negative sentiment, control over 8 emotion categories is enabled.

Starting from PPLMs, Gu et al. (2022a) observe that using a controller alone leads to the trade-off problem, i.e. the controller used to modify the token distribution only focuses on how to make the prefix related to the desired attribute without taking into account the original distribution of the LM. In this way, the controller takes over the LM's control for the next token distribution. In order to alleviate

Model	Control Impl	Attribute Relevance %			Fluency ↓	Diversity ↑			
		Overall	Pos	Pos Prob	Ppl	Dist-1	Dist-2	Dist-3	Avg
Yu et al. (2021)	D	-	-	64.49	36.62	0.48	**0.85**	*0.91*	**0.75**
Qian et al. (2022)	MFT	-	83.3	-	-	-	-	-	-
Gu et al. (2022b)	MFT	**86.7**	-	-	**28.4**	-	-	-	0.49
Dathathri et al. (2019)	MTD	78.8	-	-	46.6	0.36	0.77	*0.91*	0.68
Kumar et al. (2022)	MTD	-	**96**	-	28.9	**0.53**	0.77	0.74	0.68
Gu et al. (2022a)	MTD	-	-	**66.58**	48.52	0.40	0.80	*0.91*	0.70

Table 2: Comparison of techniques evaluated using the PPLM prompts. Different models are used to compute Attribute Relevance and Perplexity, making techniques comparison difficult. MTD=Modification of Token Distribution, MFT=Model Fine-Tuning, D=Disentanglement, Pos=Positive, Pos Prob=Positive probability, Ppl=Perplexity.

this problem, they propose a weighted decoding method that adds a regulator module that permits fine-grained adjustment of a bias signal from the controller. At every step, the regulator detects differences between the PLM distribution and the target attribute and it determines whether to suppress or amplify the bias signal. This method is model agnostic and has been evaluated with sentiment, topic, and toxicity attributes.

The last two methods propose sampling procedures that can be applied to any LM. Landsman et al. (2022) propose to modify beam search by reweighing the token candidate likelihoods to control different attributes. Diverse beam search (Vijayakumar et al., 2016) is used to decode k candidates, which are then scored using an attribute model. The obtained scores are used to reweigh the original likelihoods to produce a reweighed candidate distribution that considers both fluency and attribute characteristics. The resulting distribution is used to sample the next token.

Lastly, Kumar et al. (2022) propose a sampling method combining LM log-likelihoods with arbitrary constraints in a single energy function generating samples in a non-autoregressive manner. The idea is to use a PLM without changing its distribution but sampling from it considering different constraints, i.e. control attributes. The constraints are discriminative classifiers trained from scratch or fine-tuned. This method allows multi-attribute control (sentiment and toxicity).

5.5 Hybrid

Hybrid techniques combine two or more Control Implementation techniques. One possibility is to combine Complete Training and Fine-Tuning, for example, designing a model composed of different modules in which some modules are trained from scratch and some are fine-tuned models. In this context, Tian et al. (2022) propose a conversation model that generates empathetic responses and guides the mood of the conversation in a positive direction while acknowledging the user's emotion. The idea is to extract the sentiment from the conversation context using a fine-tuned sentiment evaluator and use both the context and the extracted sentiment to steer the generation of the next response by generating a responding strategy that will be used by the Conditional Conversation model to generate the final response. The proposed method enables only single-attribute control (of sentiment).

Another way to enable controllability using a hybrid technique is to combine Fine-Tuning and Modification of Token Distribution. Wang et al. (2022) propose a technique to control Story Generation by fine-tuning an encoder that learns the representation of new special tokens identifying the control attributes, thus allowing the model to properly include this information in the generation process. The next token distribution is obtained by combining the decoder distribution and the attention distribution, which allows the model to copy important information from the specified control attributes. The model allows fine-grained control taking into account the characters of the story with their actions and emotions.

In contrast to Wang et al. (2022) who learn the representation of special tokens during fine-tuning, Liu et al. (2021) propose to modify an LM's token distribution including two fine-tuned versions of the PLM: an expert, focused on the desired attribute, and an anti-expert, focused on the opposite of the desired attribute. The next token distribution is obtained by subtracting the anti-expert distribution from the expert one and combining the result with the distribution of the frozen PLM to maintain fluency. This method enables the control only of one control attribute at a time and it has been tested on sentiment and toxicity attributes.

Similarly, Krause et al. (2021) propose to con-

Target Sentiment	Model	Control Impl	Positive AR % ↑			Fluency ↓	Diversity ↑		
			Pos	Neutr	Neg	Ppl	Dist-1	Dist-2	Dist-3
Positive	Landsman et al. (2022)	MTD	-	**98.87**	**74.37**	51.4	*0.56*	**0.84**	**0.85**
	Zhang and Song (2022)	H	-	94.98	64.96	48.71	0.14	0.50	0.76
	Liu et al. (2021)	H	-	94.46	36.42	**45.83**	*0.56*	0.83	0.83
Negative	Landsman et al. (2022)	MTD	**28.42**	**1.99**	-	53.29	0.57	**0.85**	**0.85**
	Zhang and Song (2022)	H	31.24	6.36	-	45.60	0.12	0.48	0.77
	Liu et al. (2021)	H	35.99	3.77	-	45.91	**0.60**	0.84	0.83

Table 3: Comparison of techniques evaluated using the OpenWebText prompts. Different models are used to compute Perplexity, making techniques comparison difficult. AR=Attribute Relevance, Ppl=Perplexity, Pos=Positive prompts, Neutr=Neutral prompts, Neg=Negative prompts, MFD= modification of token Distribution, H=Hybrid.

trast the desired control attribute and its opposite. Instead of fine-tuning specialised LMs for each attribute, GPT-2 is fine-tuned with control codes to obtain a Class-Conditional LM (CCLM). At each time step, the generation is guided by computing classification probabilities for all possible next tokens via the Bayes rule by normalizing two class-conditional distributions: conditioned on the desired attribute and conditioned on the undesired attribute. Like the previous method, it allows the control of one attribute at a time and has been evaluated using sentiment, topic, and toxicity attributes.

Liu et al. (2022) also use a CCLM which is fine-tuned using an external discriminator to generate texts with the desired attributes, supporting multi-attribute control. The token distribution is modified based on a contrastive generator that learns effective representations by bringing together positive samples, i.e. samples with desired attributes, and separating negative samples, i.e. samples without desired attributes. The obtained distribution is combined with the distribution of a PLM to maintain the fluency of the generated text. The generated text is fed to the external discriminator to assess whether it contains the desired attributes or not. The model has been tested on the joint control of sentiment and topic.

Zhang and Song (2022) explore the contrast between desired and undesired attributes proposing a fine-tuned LM incorporating the attribute knowledge of a discriminator, similarly to Liu et al. (2022), to optimize continuous virtual tokens called control-prompts. The learned control-prompts are used as prefixes to steer a fixed conditional LM to generate attribute-specific texts. The LM is fine-tuned using (i) likelihood training, encouraging the LM to generate tokens with higher probability as scored by the discriminator assessing the desired attribute, and (ii) unlikelihood training, keeping the generated tokens away from lower-probability candidates.

6 Comparison of Different Techniques

In this Section, we compare the methods from the last section in terms of performance, efficiency, and generality.

6.1 Performance

In the performance comparison below, three quality criteria from the CTG field are used, namely attribute relevance, fluency, and diversity. **Attribute relevance** (AR) (Yu et al., 2021) assesses the proportion of texts correctly generated with the desired sentiment, i.e. the accuracy of the sentiment attribute measured using an external classifier. Details of the external classifier depends on the evaluation procedure, for more details refer to Appendix A. In some cases, instead of reporting the accuracy, the probability of the text being positive is reported (Pos Prob). **Fluency** is calculated as the perplexity of an external LM (Pichel Campos et al., 2018), while **diversity** is measured as the proportion of unique n-grams obtained using the Distinct metric (Dist-n in Table 2) (Li et al., 2016). Since both AR and fluency are calculated using an external component, it is difficult to obtain a fair comparison due to the usage of different models.

We consider the techniques that have been evaluated using the prompts used in the evaluation of PPLM (Dathathri et al., 2019) and the prompts extracted from OpenWebText (Gokaslan and Cohen, 2019), as detailed below.

In Table 2, we compare 6 methods that have been evaluated using the PPLM prompts, i.e. 15 prefixes used to start text generation. Perplexity is calculated using three different models; as regards attribute relevance, all the methods train or fine-tune a different classifier (for details regarding models see Appendix A.1). Performance results

are not directly comparable for these two criteria due to the usage of different models.

Diversity is the only metric for which we can have a fair comparison and here we can see that the technique proposed by Yu et al. (2021) manages to have good variety in the generated texts, while also maintaining good fluency.

Lastly, we compare three techniques evaluated using the OpenWebText (OWT) prompts, composed of neutral, positive, or negative prompts (Table 3). Two different models have been used to compute perplexity, while for the computation of attribute relevance, all the techniques use the same classifier. More details are in Appendix A.2.

In Table 3, we can see that the technique proposed by Landsman et al. (2022) obtains the highest AR and Diversity in both positive and negative target sentiment. This suggests that the proposed technique is able to generate text in the correct sentiment using diverse tokens. The same classifier is used to compute AR for every technique, allowing a fair comparison between them. On the other hand, different LMs are used to calculate Fluency, so it is more difficult to decide whether the differences are due to the model used during evaluation or due to the proposed technique.

Model	# trainable parameters
Complete Training	
Qiao et al. (2020)	68M
Betti et al. (2020)	1 generator + 2 discriminators*
Xie et al. (2022)	280M + state trackers, planners*
Model Fine-Tuning	
Qian et al. (2022)	491.520K/attribute
Gu et al. (2022b)	110M
Fang et al. (2022)	117M
Disentanglement	
Yu et al. (2021)	2M
modification of token Distribution	
Dathathri et al. (2019)	∼1K/attribute
Madotto et al. (2020)	5.175M
Goswamy et al. (2020)	∼1K/attribute
Kumar et al. (2022)	774M
Gu et al. (2022a)	0
Landsman et al. (2022)	0
Hybrid	
Wang et al. (2022)	407M
Tian et al. (2022)	337M + Enc + Strategy gen*
Liu et al. (2021)	1548M/attribute
Zhang and Song (2022)	117M
Krause et al. (2021)	345M
Liu et al. (2022)	External Discriminator*

Table 4: Comparison of studied techniques in terms of the number of trainable parameters. More details in Appendix B. *=total number of trainable parameters unclear.

6.2 Generality and parameter efficiency

In terms of generality, some of the techniques we have discussed are highly specialised and require many modifications to adapt them to include more or new control attributes. For example, the technique proposed by Xie et al. (2022) is specifically designed to control emotions and needs representing the psychological state of the story's protagonist. Other techniques require the training or fine-tuning of specific models for each control attribute (Liu et al., 2021).

In terms of efficiency, we see some techniques that require the storage and usage of multiple LMs (Liu et al., 2021). On the other hand, many techniques are model agnostic, so they can be applied to any PLM allowing reuse of existing models (Landsman et al., 2022 and Dathathri et al., 2019). In Table 4, we compare the studied techniques in terms of the number of trainable parameters. In Model Agnostic techniques, we consider the number of parameters considering the models used in the reference paper. Unfortunately, it is not possible for all techniques to correctly identify the number of parameters. In general, the modification of token Distribution techniques have fewer trainable parameters than others. The techniques proposed by Gu et al. (2022a) and Landsman et al. (2022) have 0 trainable parameters because they are sampling procedures using a PLM without any training or fine-tuning. More details regarding model parameters in Appendix B.

7 Future direction and work

In this section, we summarise the future direction and work described in the analysed papers. Overall, we can identify two suggested directions: model generalisation and fine-grained control.

Model generalisation. Different works suggest to explore the generalisation of the proposed models to explore their capabilities across domains. This can be achieved by introducing the usage of more controlled attributes, such as writer's style and dialog acts (Betti et al., 2020, Yu et al., 2021 and Liu et al., 2022), and the usage of more tasks, such as poetry generation, machine translation, and intelligent education agents (Xie et al., 2022 and Fang et al., 2022).

Fine-grained control. Some works also suggest

to explore the capabilities of the proposed methods to support fine-grained control. For example, we can extend the methods to include control attributes in the table-to-text scenario (Zhang and Song, 2022) or explore correlation between different attributes combination to enable fine-grained control (Gu et al., 2022b).

8 Discussion

In this Section, we discuss issues and trends observed in the studied techniques, which suggest possible future directions for the field.

Lack of a standard evaluation procedure. We observe that it is difficult to directly compare the performance of models evaluated on the same dataset using the same metrics, due to the usage of different methods for the metrics' evaluation. In fact, considering the six methods evaluated with PPLM prompts (Section 6.1), we observe that each method has been evaluated using a different classifier to calculate the AR metric. The usage of different classifiers affects the final result of the metric, thus requiring that every work recomputes all the evaluations to have a fair comparison with previous work. Furthermore, we observe that the papers use different datasets for the evaluation, making the comparison between papers even more difficult.

Lack of combination of different control attribute types. We observe that it is mainly topic and toxicity that are explored in combination with sentiment. While topic is a Content Control attribute, other content control attributes, such as data or set of words, are not explored in combination with sentiment, suggesting that a possible future direction is to investigate the combination of such control attributes. Furthermore, we see that Syntactic Control is not explored in combination with sentiment. Moreover, there are not many combinations of different attributes in the *same* category. For example, sentiment and toxicity (Polarity Control) are used together to enable multiple control in just two papers (Qian et al., 2022; Gu et al., 2022b)), but not many other attributes are widely explored.

9 Conclusion

We have reported a systematic survey of Sentiment-Control Text Generation techniques spanning the years 2019–2022. We proposed a categorisation scheme to analyse the studied papers based on the control attributes used and on how the control is implemented. We compared the papers based on their performance, generality and efficiency. While analysing the selected papers, we observed some issues and trends, such as the lack of a standard evaluation procedure and the lack of combinations between different control attribute types.

Acknowledgements

This work was conducted with the financial support of the Science Foundation Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224 and the Science Foundation Ireland under Grant Agreement No. 13/RC/2106_P2 at the ADAPT SFI Research Centre at Dublin City University. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Rodrigo Agerri, Montse Cuadros, Sean Gaines, and German Rigau. 2013. Opener: Open polarity enhanced named entity recognition. *Procesamiento del Lenguaje Natural*, (51):215–218.

Federico Betti, Giorgia Ramponi, and Massimo Piccardi. 2020. Controlled text generation with adversarial learning. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 29–34, Dublin, Ireland. Association for Computational Linguistics.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*.

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

Xianghong Fang, Jian Li, Lifeng Shang, Xin Jiang, Qun Liu, and Dit-Yan Yeung. 2022. Controlled text generation using dictionary prior in variational autoencoders. In *Findings of the Association for Computa-*

tional Linguistics: ACL 2022, pages 97–111, Dublin, Ireland. Association for Computational Linguistics.

Xiaocheng Feng, Ming Liu, Jiahao Liu, Bing Qin, Yibo Sun, and Ting Liu. 2018. Topic-to-essay generation with neural networks. In *IJCAI*, pages 4078–4084.

Aaron Gokaslan and Vanya Cohen. 2019. Openwebtext corpus.

Tushar Goswamy, Ishika Singh, Ahsan Barkati, and Ashutosh Modi. 2020. Adapting a language model for controlled affective text generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2787–2801, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yuxuan Gu, Xiaocheng Feng, Sicheng Ma, Jiaming Wu, Heng Gong, and Bing Qin. 2022a. Improving controllable text generation with position-aware weighted decoding. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3449–3467, Dublin, Ireland. Association for Computational Linguistics.

Yuxuan Gu, Xiaocheng Feng, Sicheng Ma, Lingyuan Zhang, Heng Gong, and Bing Qin. 2022b. A distributional lens for multi-aspect controllable text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1023–1043, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Mark Heitmann, Christian Siebert, Jochen Hartmann, and Christina Schamp. 2020. *More than a feeling: Benchmarks for sentiment analysis accuracy*. Ssrn.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

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

Sachin Kumar, Biswajit Paria, and Yulia Tsvetkov. 2022. Gradient-based constrained sampling from language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2251–2277, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

David Landsman, Jerry Zikun Chen, and Hussain Zaidi. 2022. BeamR: Beam reweighing with attribute discriminators for controllable text generation. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 422–437, Online only. Association for Computational Linguistics.

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

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

Guisheng Liu, Yi Li, Yanqing Guo, Xiangyang Luo, and Bo Wang. 2022. Multi-attribute controlled text generation with contrastive-generator and external-discriminator. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5904–5913, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150.

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

Jose Ramom Pichel Campos, Pablo Gamallo, and Iñaki Alegria. 2018. Measuring language distance among historical varieties using perplexity. application to European Portuguese. In *Proceedings of the Fifth*

Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018), pages 145–155, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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

Jing Qian, Li Dong, Yelong Shen, Furu Wei, and Weizhu Chen. 2022. Controllable natural language generation with contrastive prefixes. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2912–2924, Dublin, Ireland. Association for Computational Linguistics.

Lin Qiao, Jianhao Yan, Fandong Meng, Zhendong Yang, and Jie Zhou. 2020. A sentiment-controllable topic-to-essay generator with topic knowledge graph. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3336–3344, Online. Association for Computational Linguistics.

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

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

Hannah Rashkin, Antoine Bosselut, Maarten Sap, Kevin Knight, and Yejin Choi. 2018. Modeling naive psychology of characters in simple commonsense stories. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2289–2299.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Adam Santoro, Ryan Faulkner, David Raposo, Jack Rae, Mike Chrzanowski, Theophane Weber, Daan Wierstra, Oriol Vinyals, Razvan Pascanu, and Timothy Lillicrap. 2018. Relational recurrent neural networks. *Advances in neural information processing systems*, 31.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1577–1586.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Zhiliang Tian, Yinliang Wang, Yiping Song, Chi Zhang, Dongkyu Lee, Yingxiu Zhao, Dongsheng Li, and Nevin L. Zhang. 2022. Empathetic and emotionally positive conversation systems with an emotion-specific query-response memory. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6364–6376, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Olga Uryupina, Barbara Plank, Aliaksei Severyn, Agata Rotondi, and Alessandro Moschitti. 2014. Sentube: A corpus for sentiment analysis on youtube social media. In *LREC*, pages 4244–4249.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.

Xinpeng Wang, Han Jiang, Zhihua Wei, and Shanlin Zhou. 2022. CHAE: Fine-grained controllable story generation with characters, actions and emotions. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6426–6435, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Yuqiang Xie, Yue Hu, Yunpeng Li, Guanqun Bi, Luxi Xing, and Wei Peng. 2022. Psychology-guided controllable story generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6480–6492, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Erguang Yang, Mingtong Liu, Deyi Xiong, Yujie Zhang, Yao Meng, Changjian Hu, Jinan Xu, and Yufeng Chen. 2021. Syntactically-informed unsupervised paraphrasing with non-parallel data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2594–2604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dian Yu, Zhou Yu, and Kenji Sagae. 2021. Attribute alignment: Controlling text generation from pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2251–2268, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hanqing Zhang and Dawei Song. 2022. DisCup: Discriminator cooperative unlikelihood prompt-tuning for controllable text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3392–3406, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2022. A survey of controllable text generation using transformer-based pre-trained language models. *arXiv preprint arXiv:2201.05337*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

A Performance Comparison of Different Techniques

In this Section, we provide further details about the automatic evaluation reported in Section 6.1. Furthermore, we report all the models that have been used in the different techniques to calculate the evaluation metrics explained.

A.1 Comparison using PPLM prompts

Perplexity has been calculated using three different models. Dathathri et al. (2019) and Gu et al. (2022a) used GPT (Radford et al., 2018), Qian et al. (2022), Gu et al. (2022b), and Yu et al. (2021) used GPT-2 Large, Kumar et al. (2022) used while GPT-2 XL. Regarding attribute relevance, all the methods trained or fine-tuned a different classifier. Dathathri et al. (2019) trained a classifier on SST-5 (Socher et al., 2013), while Gu et al. (2022a) trained a classifier on IMDb movie reviews (Maas et al., 2011). Qian et al. (2022) and Gu et al. (2022b) fine-tuned RoBERTa (Liu et al., 2019) and DeBERTa (He et al., 2020), respectively, on the Yelp Review dataset (Zhang et al., 2015). Finally, Yu et al. (2021) fine-tune BERT with IMDb movie reviews dataset, while Kumar et al. (2022) fine-tuned SieBERT (Heitmann et al., 2020) on 15 different polarity datasets.

A.2 Comparison using OWT prompts

The techniques in Table 3 have been evaluated using the OpenWebText prompts, which are prompts randomly selected from OpenWebText dataset. For each selected prompt 25 completions are generated using a base LM. Based on the average sentiment of these completions, each prompt is labeled as neutral, positive, or negative resulting in 3 datasets of prompts containing respectively 5K, 2,5K, and 2.5K prompts

In order to compute perplexity, Zhang and Song (2022) used GPT-2 Large, and Landsman et al., 2022 and Liu et al., 2021 used GPT-2 XL. While for the computation of attribute relevance, all the techniques used the same classifier, i.e. a DistilBERT (Sanh et al., 2019) sentiment classifier provided by Huggingface and fine-tuned on SST-2 (Socher et al., 2013).

B Parameters Comparison of Different Techniques

In Table 5, we show the number of parameters in each studied technique. We report all the components of the techniques with the respective parameters (Column 2), the number of trainable parameters (Column 3), and the total number of parameters (Column 4).

In some of the techniques, we can see a huge difference between the number of trainable parameters and the total number of parameters, for example, (Zhang and Song, 2022) and (Landsman et al., 2022).

Model	Model components	# trainable parameters	# parameters
	Complete Training		
Qiao et al. (2020)	Encoder (biGRU) + Recognition network (MLP) + Prior network (MLP) + Sentence decoder (GRU) + Discriminator (CNN)	68M	68M
Betti et al. (2020)	Generator (Relational Memory with self-attention) + Syntax Discriminator (Conv net) + Semantic Discriminator (Conv net)	1 generator + 2 discriminators*	1 generator + 2 discriminators*
Xie et al. (2022)	Encoder (BART) 140M + State trackers, planners (BiGRU) + Decoder (BART) 140M	280 + state trackers, planners*	280 + state trackers, planners*
	Model Fine-Tuning		
Qian et al. (2022)	PLM (GPT-2 medium) 345M + 491.520K/attribute	491.520K/attribute	345.491M
Gu et al. (2022b)	Encoder (BERT-base-uncased) 110M + Decoder (GPT-2 medium) 345M	110M	455M
Fang et al. (2022)	Encoder (BERT-base-uncased) 110M + Decoder (GPT-2) 117M + Deep Dual function network 1K	117M	227M
	Disentanglement		
Yu et al. (2021)	PLM (GPT-2 medium) 345M + Attribute Alignment function (MLP) 2M	2M	347M
	modification of token Distribution		
Dathathri et al. (2019)	PLM (GPT-2 medium) 345M + PPLM Discriminator ∼1K/attribute	∼1K/attribute	345M
Madotto et al. (2020)	PLM (DialoGPT medium) 345M parameters + Residual Adapters 5.175M parameters + Discriminator ∼1K/attribute	5.175M	350.175M
Goswamy et al. (2020)	PLM (GPT-2 medium) 345M + PPLM Discriminator 1K/attribute	1K/attribute	345M
Kumar et al. (2022)	PLM (GPT-2 Large) 774M + Discriminative Classifier (GPT-2 Large) 774M	774M	1548M
Gu et al. (2022a)	PPLM 345M + Trainable Regulator (TF-IDF)	0	345M
Landsman et al. (2022)	PLM (DExperts expert) 774M	0	774M
	Hybrid		
Wang et al. (2022)	PLM (BART-large-cnn) 407M	407M	407M
Tian et al. (2022)	Encoder + Emotion detector (BERT) 110M + Responding Strategy Generator + GPT 117M + BERT 110M	337M + Encoder + Strategy generator*	337M + Encoder + Strategy generator*
Liu et al. (2021)	PLM (GPT-2 Large) 774M + Expert (GPT-2 Large) 774M and anti-Expert (GPT-2 Large) 774M /attribute	1548M/attribute	1548M/attribute + 774M
Zhang and Song (2022)	CLM (GPT-2 large) 774M + Attribute Discriminator (GPT-2 small) 117M	117M	891M
Krause et al. (2021)	PLM (GPT-2 medium) 345M	345M	345M
Liu et al. (2022)	PLM (GPT-2 medium) 345M + External Discriminator (biGRU)	External Discriminator*	345M + External Discriminator*

Table 5: Comparison of studied techniques in terms of the number of parameters. In Model Agnostic techniques, we consider the number of parameters considering the models used in the reference paper. * the total number of trainable parameters is unclear.