ECO Decoding: Entropy-Based Control for Controllability and Fluency in Controllable Dialogue Generation

Anonymous ACL submission

Abstract

Controllable Dialogue Generation (CDG) enables chatbots to generate responses with desired attributes, and weighted decoding methods have achieved significant success in the CDG task. However, using a fixed constant value to manage the bias of attribute probabilities makes it challenging to find an ideal control strength that satisfies both controllability and fluency. To address this issue, we propose a novel dynamic control strength method that considers the uncertainty of the model's generation and classification probabilities. Specifically, we dynamically adjust the control strength at each generation step based on the entropy of the language model's next token probabilities and the entropy of the attribute classifier's probability estimates. Experimental results on various existing models demonstrate that our decoding method achieves high control performance while maintaining fluency compared to existing decoding strategies across all models. Additionally, our approach alleviates the probability interpolation issue in multi-attribute controlled generation, yielding superior performance.

1 Introduction

Recently, Controllable Dialogue Generation (CDG) (Zhang et al., 2023; Zeng et al., 2023) has been proposed to enhance the realism and accuracy of responses generated by conversational models, improving the user experience. CDG enables chatbots to generate responses tailored to desired attributes like emotion and dialog-act. Among studies on controllable generation, **weighted decoding** methods (Yang and Klein, 2021; Arora et al., 2022) have achieved significant success.

In the field of controllable generation, training-based methods such as alignment tuning and **weighted decoding** approaches (Yang and Klein, 2021; Arora et al., 2022) have achieved notable success. While alignment tuning suffers from the disadvantage of requiring the entire model to be retrained, weighted decoding can be easily applied during the inference stage and enables the generation of controlled responses by training an attribute classifier with relatively little data. Consequently, we focused on this weighted decoding strategy to effectively generate controllable responses.

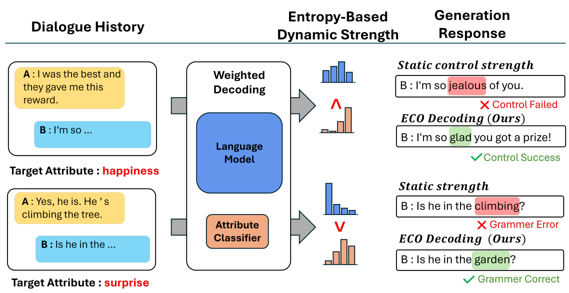

Figure 1: Controllable Dialog Generation method based on dynamic weighting with Eco Decoding. By dynamically determining the weights between the language model probability distribution and the attribute control probability distribution, it is possible to perform attribute control while maintaining fluency.

In weight decoding methods, generating responses controlled by desired attributes involves the adjustment of the next token probability distribution modeled by the language model. This is achieved by multiplying the attribute probability of the generated response obtained from the attribute classifier and the next token probability. In this process, the control strength is used as the exponent of the attribute probability to control attribute bias. As the control strength increases, the generated tokens become more dependent on the token rank of attribute probability.

Multiplying the attribute probability alters the probability distribution of the language model, which can affect language modeling performance. When static control strength is used, the same control probability is continuously reflected in the generated sentence, even if the sentence has already received sufficient attribute control or if specific words need to be generated for fluency. This can lead to a trade-off between controllability and fluency. Furthermore, the fact that the appropriate control strength varies depending on the situation is an important issue. If this is not properly accounted for, it can lead to decreased efficiency. Figure 1 shows an example of a failed response generation with these fixed static control strength.

In this paper, we propose the **E**ntropy-based **CO**ntrol strength decoding method, named **ECO decoding**, to resolve the aforementioned problem with static control strength. Our method can generate controlled responses that achieve high controllability as well as maintain text fluency. During the decoding process, the entropy (Shannon, 2001) of each probability distribution from language model and the attribute classification model is calculated at every generation step, and calculated entropy is used as a dynamic factor to adjust the control strength on the response probabilities at each generation step.

Specifically, with respect to the language model's probability distribution, if a particular token is assigned a high probability, which implies low entropy, it is considered contextually and syntactically appropriate. To maintain fluency, the language model's prediction is given priority. In contrast, when token probabilities are uniformly distributed, implying high entropy, the model is considered less confident, and to achieve a higher degree of control, the bias toward the attribute probabilities is increased.

This dynamic control method effectively balances the language model's fluency with the attribute classifier's controllability, thereby achieving an optimal trade-off between naturalness and the desired attribute expression in the final generated sentences. To validate our intuition, we experiment with three existing controllable generation models using the DailyDialog (Li et al., 2017) dataset. Experimental results demonstrate that ECO decoding achieves high controllability while maintaining text fluency across all models.

Our main contributions are as follows:

1. We raise the issue of static control strength in existing weighted decoding methods and propose a dynamic control strength approach to generate responses with high controllability as well as maintain fluency.

2. We show that the ECO decoding methodology enables multi attribute control over single attribute based weighted decoding methodologies.

3. Experimental results show that the ECO decoding method outperforms the existing weighted decoding methods for all existing controllable generation models.

2 Related Work

2.1 Weighted Decoding

Controllable dialogue generation aims to generate a response, $R = \{r_1, r_2, ..., r_N\}$, with desired attributes, given dialogue history h and attribute c, using a pre-trained auto-regressive model (e.g. GPT2, (Radford et al., 2019), DialoGPT (Zhang et al., 2020)). Emotion and dialog-act can be attributes for controllable dialogue generation.

To condition on attribute c, the response generation given a dialogue can be formulated as follows:

$$P(R|h, c) = \prod_{i=1}^{N} P(r_i|r_{<i}, h, c) \quad (1)$$

Using Bayesian factorization, $P(r_i|r_{<i}, h, c)$ can be converted into the following equation.

$$P(r_i|r_{<i}, h, c) \propto P(r_i|r_{<i}, h)P(c|r_{\leq i}, h)^{\lambda} \quad (2)$$

where the first term $P(r_i|r_{<i}, h)$ represents the next token probability modeled by a language model, and the second term $P(c|r_{\leq i}, h)$ represents the attribute probability of the generated response obtained from the attribute classifier. In addition, control strength λ is added to the exponential term of the attribute probabilities to control attribute bias.

When dealing with multi-attribute control, Equation 3 can be extended by introducing the product of multiple attribute classifiers, assuming that the attributes are conditionally independent:

$$P(r_i|r_{<i}, h, C) \propto P(r_i|r_{<i}, h) \prod_{c_j \in C} P(c_j|r_{\leq i}, h)^{\lambda} \quad (3)$$

where C denotes the set of target attributes. The product of probabilities is typically implemented as the sum of logits.

2.2 Weighted Decoding Models

FUDGE Yang and Klein, 2021 trained a classification model for partial sequences through an

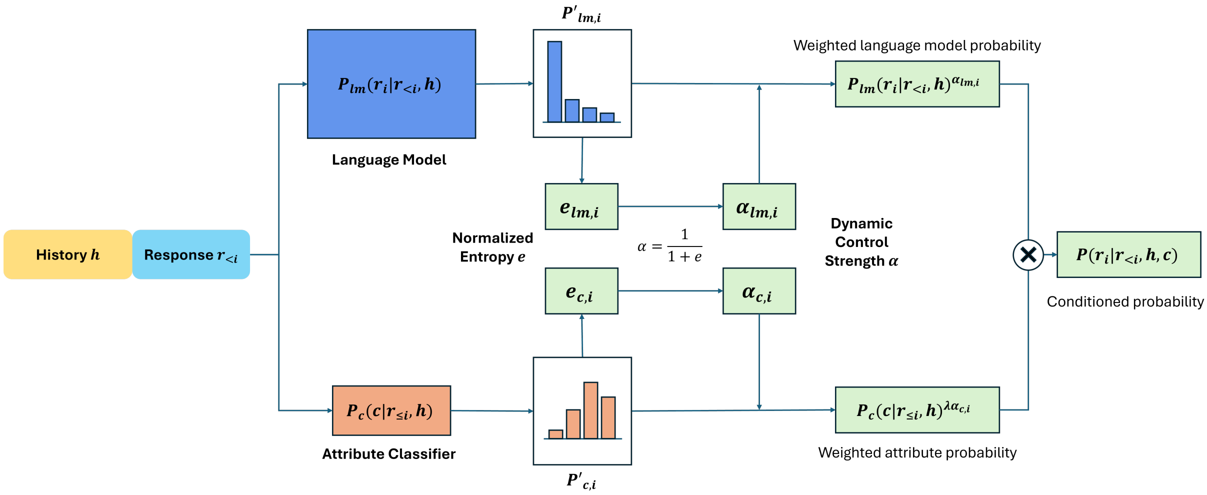

Figure 2: An illustration of controllable dialogue generation using the weighted decoding method.

external attribute classifier. Specifically, for each training example $\{(x,c)\}$, where x is sentence and c is class label, the classifier is trained on all partial sequences $\{(x_{1:i},c)\}$ at each step. During inference, at a given time step i, the classifier predicts the probability that appending the top k candidate tokens to the generated text will satisfy the attribute c in future generations.

Director Arora et al., 2022 addressed the inefficiency issue of requiring a external model during inference. It integrates the language model and attribute classification functionality into a single model, overcoming the inefficiency of the external classifier evaluating the attribute for every candidate token. To address this issue, an additional classification head is introduced, which takes the last hidden state as input and computes the probability that each token in the vocabulary satisfies the specified attribute. This allows for the effective incorporation of attribute information without the need for a external classifier.

DASC Zhang et al., 2023 addressed the computational inefficiency issues arising from dual-head architectures. DASC introduces Attribute Token Embedding and Attribute Semantic Embedding concepts, employing a semantic space-based weighted decoding mechanism to reduce the number of parameters while improving computational efficiency. Each token is associated with an embedding that captures its attribute semantics, and these embeddings are projected into an attribute semantic space via attribute-specific linear layers. This design facilitates smooth control over multiple attributes and enables effective interpolation among attribute embeddings, allowing more diverse range of attribute combinations.

3 Methodology

3.1 Entropy-based Control Strength

The existing weighted decoding methods apply a fixed control strength and they are not flexible enough to handle situations where stronger or no more control is needed. In such cases, they may fail to control attribute, or even if they succeed, the fluency and grammar may degraded. To solve this problem, we propose the ECO decoding method that utilizes the entropy of the probability distribution to dynamically adjust the control strength. Dynamic control strength allows to achieve higher attribute control rates, while maintaining generation quality, including context and grammar.

Entropy is a measure of the uncertainty of a probability distribution, which is lower when the probability distribution is focused on a specific value and higher when it is more evenly distributed. Given this property, the higher the entropy of the next token probability distribution is, the more likely it is to contain a variety of plausible candidates. This is an advantageous property for exploring plausible options that satisfy desired attribute. Based on this insight, a novel mechanism of dynamically controlling strength is developed by weighting probability distributions from language models and it controls each property inversely to their entropy score. That is, distributions with lower uncertainty are more strongly reflected. Figure 2 shows how ECO decoding is working by using dynamic control strength based on both of the language model entropy and

Model	Accuracy	Rouge-1	Rouge-L	Dist-1	Dist-2	Grammar
			Emotion			
DialoGPT	-	9.00	8.53	0.58	0.76	90.21
FUDGE	76.98	9.06	8.60	0.60	0.75	90.30
+ ECO decoding	81.03 (+4.05)	9.13 (+0.07)	8.64 (+0.04)	0.62 (+0.02)	0.75 (-)	90.34 (+0.04)
Director	79.94	8.83	8.37	0.59	0.70	90.23
+ ECO decoding	82.82 (+2.88)	8.82 (-0.01)	8.34 (-0.03)	0.59 (-)	0.71 (+0.01)	90.30 (+0.07)
DASC	74.65	8.25	7.87	0.58	0.70	90.30
+ ECO decoding	75.74 (+1.09)	8.22 (-0.03)	7.79 (-0.08)	0.58 (-)	0.71 (+0.01)	90.39 (+0.09)
			Dialog-act			
DialoGPT	-	9.14	8.66	0.57	0.78	91.24
FUDGE	41.07	9.21	8.75	0.59	0.78	90.98
+ ECO decoding	46.42 (+5.35)	9.21 (-)	8.79 (+0.04)	0.62 (+0.03)	0.79 (+0.01)	91.00 (+0.02)
Director	70.96	10.43	9.94	0.62	0.78	91.18
+ ECO decoding	71.56 (+0.60)	10.46 (+0.03)	9.96 (+0.02)	0.63 (+0.01)	0.79 (+0.01)	91.15 (-0.03)
DASC	42.59	9.53	9.03	0.59	0.75	91.13
+ ECO decoding	47.17 (+4.58)	9.52 (-0.01)	9.05 (+0.02)	0.60 (+0.01)	0.76 (+0.01)	91.13 (-)

Table 1: Evaluation results for a single attribute of emotion or dialog-act on the DailyDialog test set. The scores in brackets indicate the performance gap between static control and dynamic control settings.

the attribute entropy. ECO decoding can be applied to the existing weighted decoding methods and requires no additional modules or training.

Language Model Entropy Dynamic control strength $\alpha_{x,i}$ is separately calculated for i-th generation step, and it can have different values while a sentence is generated. To calculate control strength, we select the top-k candidate tokens. From the probability distribution $P_{lm,i}$ of the language model, we construct the set S, which consists of the k tokens with the highest probabilities. Let $P'_{lm,i}$ denote the partial probability distribution of top-k tokens in S.

$$P'_{lm,i} = \{P_{lm}(t|r_{<i}, h)|t \in S\} \tag{4}$$

To convert the partial probability distribution $P'_{lm,i}$ into a probability distribution, we recompute the probability distribution of the top-k tokens using a softmax function with temperature τ_{lm}.

$$e_{lm,i} = Entropy(Softmax(P'_{lm,i}/\tau_{lm})) \tag{5}$$

Attribute Entropy Weighted decoding methodologies for CDG utilize attribute classifier P_c to reflect attributes. For each candidate token t in the top-k token set S, concatenates the current sequence $r_{<i}$ with t and computes the probability $P_{c,i}([r_{<i}; t], h)$ which represents the probability of token t being part of the generated response while aligning with the target attribute to be controlled. The set $P'_{c,i}$ is the probabilities of the target attribute for all candidate tokens in top-k token set S. The attribute entropy $e_{c,i}$ is computed based on a probability distribution normalized by softmax the set of attribute probabilities $P'_{c,i}$ over τ_c, where τ_c is the attribute temperature for softmax.

$$P'_{c,i} = \{P_c([r_{<i}; t], h)|t \in S\} \tag{6}$$

$$e_{c,i} = Entropy(Softmax(P'_{c,i}/\tau_c)) \tag{7}$$

Entropy Based Control Strength To assign higher weights to probability distributions with higher entropy, we utilize a control strength formula with an inverse function structure, as shown in Equation 8. The control strength $\alpha_{x,i}$ is applied to both the language model probability distribution P_{lm} and the attribute probability distribution P_c. The language model probability distribution and the attribute probability distribution are reflected by a power of their respective weight $\alpha_{x,i}$. The attribute probability distribution additionally reflects the strength scale factor λ. The value of λ allows to adjust whether to focus more on attribute control or language modeling performance. The final probability distribution for generating the next token $P(r_i|r_{<i}, h, c)$ is computed by multiplying the two weighted probability distributions as shown in Equation 9. If each of the control strength $alpha$ values were fixed at 1, the same result would be

Model	Accuracy(Emo)	Accuracy(Act)	Rouge-1	Rouge-L	Dist-1	Dist-2	Grammar
DialoGPT	-	-	9.00	8.53	0.58	0.76	90.21
FUDGE	66.17	44.17	8.21	7.82	0.57	0.74	90.20
+ ECO decoding	66.41 (+0.24)	45.57 (+1.40)	8.20 (-0.01)	7.81 (-0.01)	0.58 (+0.01)	0.74 (-)	90.21 (+0.01)
Director	80.48	60.65	9.41	8.99	0.58	0.73	90.22
+ ECO decoding	81.18 (+0.7)	61.20 (+0.65)	9.49 (+0.08)	8.97 (-0.02)	0.58 (-)	0.74 (+0.01)	90.23 (+0.01)
DASC	75.19	51.17	8.22	7.67	0.60	0.77	90.05
+ ECO decoding	77.22 (+2.03)	54.12 (+2.95)	7.60 (-0.62)	7.15 (-0.52)	0.61 (+0.01)	0.78 (+0.01)	90.19 (+0.14)

Table 2: Evaluation results for multiple attributes setting on the DailyDialog test set. The scores in brackets indicate the performance gap between static control and dynamic control settings.

obtained as with the traditional weighted decoding methodologies.

$$\alpha_{x,i} = 1 + \left(\frac{1}{1 + e_{x,i}}\right) \tag{8}$$

$$\begin{aligned} P(r_i|r_{<i}, h, c) \propto\ & P_{lm}(r_i|r_{<i}, h)^{\alpha_{lm,i}} \\ & \times P_c(c|r_{\leq i}, h)^{\lambda * \alpha_{c,i}} \end{aligned} \tag{9}$$

3.2 Multiple Attribute Control Strength

Existing weighted decoding methodologies struggle to control multiple attributes simultaneously due to their fixed control strength. When using a fixed control strength for each attribute, the search space of attribute control strengths grows exponentially. Furthermore, even when control strength is applied, effectively incorporating more than two attributes remains a main challenge. In contrast, our proposed ECO-decoding method enables CDG to control generation by reformulating the final probability distribution based on multiple attributes. Dynamic control strength $\alpha_{x,i}$ adjusts the weight of probability distributions at each generation step based on the entropy of the language model and the entropy of each attribute, allowing more flexible and adaptive multi-attribute control. When C is the set of controlling attributes, the multiple attribute control formula for ECO-decoding is as follows:

$$\begin{aligned} P(r_i|r_{<i}, h, C) \propto\ & P_{lm}(r_i|r_{<i}, h)^{\alpha_{lm,i}} \\ & \times \prod_{c_j \in C} P_{c_j}(c_j|r_{\leq i}, h)^{\lambda * \alpha_{c_j,i}} \end{aligned} \tag{10}$$

4 Experiments

4.1 Datasets

Daily Dialog (Li et al., 2017) is an English open-domain dialogue dataset with two controllable attributes: emotion and dialog-act. For dialogues, each utterance is regarded as the response, and all preceding utterances are used as the corresponding dialogue history. For the dialog-act attribute, it consists of training (75,957), validation (7,059), and test (6,740) dialogues, and it is composed of four classes (inform, question, directive, and commissive). For the emotion attribute, experiments were conducted with 6 classes (anger, disgust, fear, happiness, sadness, and surprise), excluding the "no emotion" attribute value. The dataset consisted of training (13,681), validation (882), and test (1,286) dialogues.

4.2 Experimental Settings

Language Model We use DialoGPT (Zhang et al., 2020), pre-trained on a large-scale dialogue corpus, as the baseline model. For most of our experiments, the DialoGPT-small (176M) is used, but the DialoGPT-large (1.1B) is used for comparison across different sizes. In addition, we also use a Llama2-7B (Touvron et al., 2023) to evaluate the applicability of ECO decoding in LLM.

Weighted Decoding Methods We evaluate and compare the performances of ECO decoding with those of various controllable generation models with weighted decoding methods, including FUDGE (Yang and Klein, 2021), Director (Arora et al., 2022), and DASC (Zhang et al., 2023), and all of them use DialoGPT as the backbone model for comparison.

Implementation Details For the three weighted decoding method, the language model is frozen and each attribute classifier is trained on the training dataset. FUDGE is trained for 30 epochs with a batch size of 8 and a learning rate of 2e-5 for each attribute. For the Director, each attribute is fine-tuned for 20 epochs with a batch size of 32 and a learning rate of 1e-5. For DASC, each attribute is fine-tuned for 30 epochs with a batch size of 4 and a learning rate of 1e-5. All methods use

Model	Accuracy	Rouge-1	Rouge-L	Dist-1	Dist-2	Grammar
Small Model (176M)						
DialoGPT	-	9.00	8.53	0.58	0.76	90.21
Director	79.94	8.83	8.37	0.59	0.70	90.23
+ ECO decoding	82.82 (+2.88)	8.82 (-0.01)	8.34 (-0.03)	0.59 (-)	0.71 (+0.01)	90.30 (+0.07)
Large Model(1.1B)						
DialoGPT	-	11.54	10.89	0.75	0.73	87.28
Director	75.66	11.76	11.15	0.74	0.73	87.18
+ ECO decoding	76.05 (+0.39)	11.82 (+0.06)	11.23 (+0.08)	0.74 (-)	0.73 (-)	87.25 (+0.07)
Large Language Model (7B)						
Llama2	-	15.23	12.99	0.35	0.08	90.60
Director	75.43	15.95	13.74	0.35	0.80	90.51
+ ECO decoding	75.66 (+0.23)	15.88 (-0.07)	13.63 (-0.03)	0.35(-)	0.80 (-)	90.55 (+0.04)

Table 3: Evaluation results for attributes of emotion on the DailyDialog test set with various size of model. The scores in brackets indicate the performance gap between static control and dynamic control settings.

Model	Accuracy	Interest	Sensible
Emotion			
Director	2.82	2.96	2.75
+ ECO decoding	3.19 (+0.37)	3.16 (+0.20)	3.15 (+0.40)
Dialog-act			
Director	3.04	2.93	2.78
+ ECO decoding	3.42 (+0.38)	3.41 (+0.48)	3.36 (+0.58)

Table 4: Human Evaluation on DailyDailog test set (single attribute)

greedy search (Li et al., 2016b), and the maximum sequence length is set to 128. All experiments are run on a single NVIDIA GeForce RTX 3090.

4.3 Evaluation Metrics

Automatic Evaluation To evaluate the controllability, we train RoBERTa (Liu et al., 2019) based evaluators using the DailyDialog training data to estimate attribute accuracy. The two evaluators achieved accuracies of 89.66% and 80.60% on the test set for emotion and dialog-act, respectively. Note that the attribute evaluator is independent from the attribute classifier used in the weight decoding model. For the generation quality, we use ROUGE-1 and ROUGE-L (Lin, 2004) scores to evaluate match scores between generated responses and ground-truth references. We evaluate the diversity of generated responses using 1-gram and 2-gram distinctness(Li et al., 2016a), referred to as Dist-1 and Dist-2. For grammar checking, We utilize the probability of grammaticality given by a RoBERTa-based CoLA grammaticality model (Liu et al., 2019; Warstadt et al., 2019; Morris et al., 2020).

Human Evaluation Experiments on the Director model, which showed the best performance in emotion and dialog-act attributes, conducted human evaluation based on sampling 10 contexts for each attribute value from the test set. We evaluate our generated responses based on three aspects: (1) Accuracy: Response is generated according to the desired attribute. (2) Interest: Response is specific and novel, and it can lead to more engaging conversation. (3) Sensible: Response is grammatically correct and contextually coherent. We asked three expert evaluators to rate each metric on a scale of 1 to 5, with higher scores being better.

4.4 Experimental Results

Single Attribute Control Table 1 presents the results of evaluation on how effectively the proposed ECO decoding, which leverages dynamic weighting, enhances the controllability of dialogue generation models while maintaining fluency. We first determine a baseline grammar score using outputs generated by the Backbone Model without any attribute control. For each controllable generation model (FUDGE, Director, and DASC), we then search for the optimal control strength λ that achieves a grammar score comparable to this baseline. Based on the selected λ, we evaluate Accuracy, Dist, and ROUGE.

The results indicate that applying ECO decoding consistently improves the accuracy for emotion and dialog-act attributes, outperforming methods that rely on static control strengths. Notably, unlike existing decoding approaches, ECO decoding achieves these improvements without degrading the grammar score. Moreover, while maintaining a grammar score close to the baseline, ECO decoding preserves or even slightly improves the Dist and ROUGE metrics. The experimental results demonstrate that ECO decoding leverages dynamic weighting to more actively reflect specific attributes in the generated responses while simultaneously maintaining grammatical fluency and overall response quality.

Multi Attribute Control Multi attribute control typically involves combining attribute probabilities via multiplication. Consequently, when interpolating across multiple distributions (Lin and Riedl, 2021; Kumar et al., 2021), differences in scale and calibration can make it difficult to maintain a proper balance, often leading to a decline in overall controllability compared to single-attribute control.

In Table 2, multi attribute control for the Emotion attribute achieves grammar performance on par with single-attribute control, yet exhibits a decrease in overall controllability. Conversely, multi-attribute control for the Dialo-act attribute appears to yield higher controllability relative to single attribute control. However, this does not necessarily indicate an actual improvement in controllability; rather, it likely reflects the selection of a relatively lower grammar score baseline due to differences in the experimental data.

Compared to methods using a fixed control strength, experimental results show that applying ECO decoding can alleviate the interpolation problem, thereby improving controllability for both Emotion and Dialo-act attributes. Furthermore, similar to single-attribute control, experimental results demonstrate that dynamic weighting helps consistently maintain and enhance grammatical fluency and overall response quality in multi-attribute generation.

Language Model Scaling Table 3 shows the results of evaluating the performance of the Director method and ECO decoding on models of different sizes.The results show that in all cases, from the smallest 176M model to the 7B model, the ECO method achieves higher grammar scores while maintaining good control over attributes. This suggests that applying ECO decoding to traditional weighted decoding methods can achieve better performance regardless of model size.

Human Evaluation Table 4 presents the human evaluation results. Similar to the automatic evaluation results, the ECO decoding approach con-

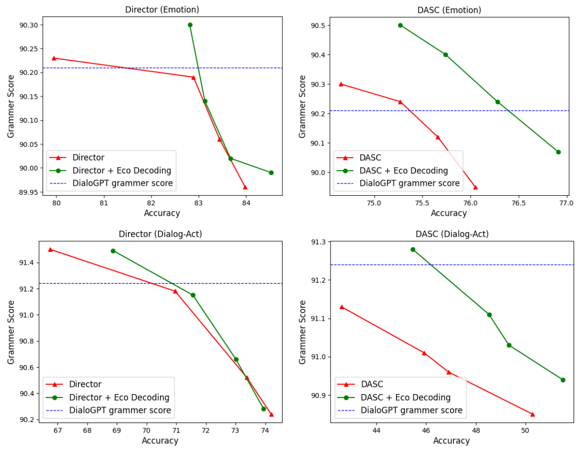

Figure 3: The single attribute control performance of the existing weighted decoding method (red) and ECO decoding (green) with respect to changes in the control strength λ. The y-axis represents grammar, and the x-axis represents accuracy. The blue dot line represents uncontrolled dialogpt's grammar score.

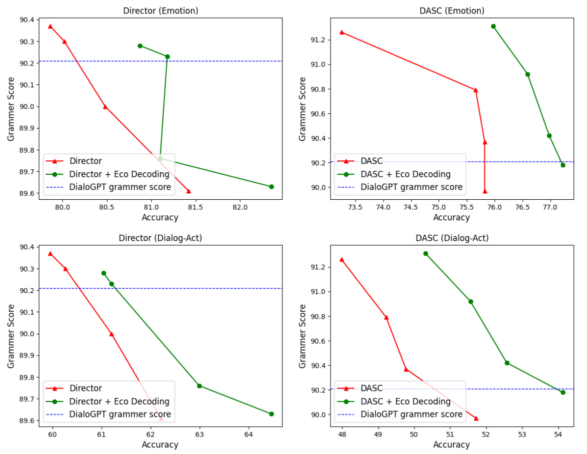

Figure 4: The multi attribute control performance of the existing weighted decoding method (red) and ECO decoding (green) with respect to changes in the control strength λ. The y-axis represents grammar, and the x-axis represents accuracy. The blue dot line represents uncontrolled dialogpt's grammar score.

sistently outperformed traditional methods in generating consistent and controlled responses. For both emotion and dailog-act, accuracy increased by at least 12.5%, and for sensible, performance improved by more than 14%. In addition, evaluation of interest metrics confirmed that when ECO decoding is applied to traditional weighted decoding methods, it can generate engaging responses that drive conversations.

Robustness Test The control strength coefficient *lambda* determines the proportion of weights in the probability distribution of the attribute classifier.

Therefore, a larger lambda will tend to generate tokens that are more attribute-specific, resulting in a trade-off between increased attribute accuracy and decreased grammar score.

Figure 3 shows the results of applying traditional weighted decoding methodology and ECO-decoding for varying the control strength coefficient $lambda$ in a single attribute control setting. The experiments were conducted based on the Director and DASC method for two attributes, emotion and dialog-act, and the red line is for the Director and DASC. The green line is for the application of ECO decoding with each method and the blue bashed line is the grammar score of vanila DialoGPT without attribute control according to each dataset. In all experiments, we observed a trade-off between grammar and accuracy, and showed that for the same grammar score, ECO decoding achieves higher attribute accuracy by dynamically applying weights. In other words, for the same control degree, ECO decoding produces higher quality responses. This demonstrates that our approach has a strong capability in controllable generation to maintain fluency while enhancing controllability, regardless of the λ values.

Figure 4 shows the performance by lambda in the multi attribute control setting. The results are the same for multi attribute as for single, with ECO decoding for each methodology resulting in higher controllability and grammar scores. The interesting thing is that the Director model is not a structured model for multi attribute control, and because of this, there is some performance variation by lambda.

In almost all cases using the dialoGPT-small model, we observed a trend of grammar score increase and then decrease as the lambda value increases. Due to the increase in grammaticality at low λ ranges, some experimental results showed higher grammar score with attribute control than dialoGPT without attribute control. This was not observed when using larger models such as Llama2, which could be thought of as a slight inconsistency in the performance of small language models. For fairness, all experimental results measured grammar score and attribute accuracy after the lambda value where the trade-off occurs.

5 Conclusions

In this paper, we introduce an entropy-based dynamic control strength approach, ECO decoding, to overcome the limitation of static control strength in the previous weighted decoding methods. ECO decoding can be easily adapted to any weighted decoding methodology, and the advantage is that no additional training is required for ECO decoding. Our approach enables the generation of responses with enhanced controllability while maintaining fluency, and extends to a wide variety of multi-attribute environments. Both automated and human evaluations have shown that ECO decoding performs consistently well when applied across a range of controllable generation methods, and has been shown to perform well enough on large models such as Llama2-7B.

Limitations

Our method improves control performance while maintaining fluency using entropy-based control strength in single attribute conditions. However, we need to validate the effectiveness of the proposed method in multi-attribute conditions. In a multi-attribute setting, it is necessary to consider not only the uncertainty of the language model's generation probabilities but also the degree of influence of each attribute's probability in adjusting the control strength. These topics should be addressed as future work.

Ethics Statement

The proposed method aims to enhance the interest and accuracy of responses generated by chatbots to improve user experience. However, this method could be potentially used for malicious purposes. In our experiments, we focus on attributes like emotion and dialog-act, but if malicious desired attributes such as bias are used, the model could be induced to generate inappropriate responses. Therefore, generating controlled responses using malicious attributes should be restricted.

References

Kushal Arora, Kurt Shuster, Sainbayar Sukhbaatar, and Jason Weston. 2022. Director: Generator-classifiers for supervised language modeling. *Preprint*, arXiv:2206.07694.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. *Preprint*, arXiv:1510.03055.

Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. 2016b. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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. *Preprint*, arXiv:1907.11692.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126.

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.

Claude Elwood Shannon. 2001. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, 5(1):3–55.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Preprint*, arXiv:1805.12471.

Kevin Yang and Dan Klein. 2021. Fudge: Controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Weihao Zeng, Lulu Zhao, Keqing He, Ruotong Geng, Jingang Wang, Wei Wu, and Weiran Xu. 2023. Seen to unseen: Exploring compositional generalization of multi-attribute controllable dialogue generation. *Preprint*, arXiv:2306.10317.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. Dialogpt: Large-scale generative pre-training for conversational response generation. *Preprint*, arXiv:1911.00536.

Zhiling Zhang, Mengyue Wu, and Kenny Q. Zhu. 2023. Semantic space grounded weighted decoding for multi-attribute controllable dialogue generation. *Preprint*, arXiv:2305.02820.

A Cases of ECO decoding

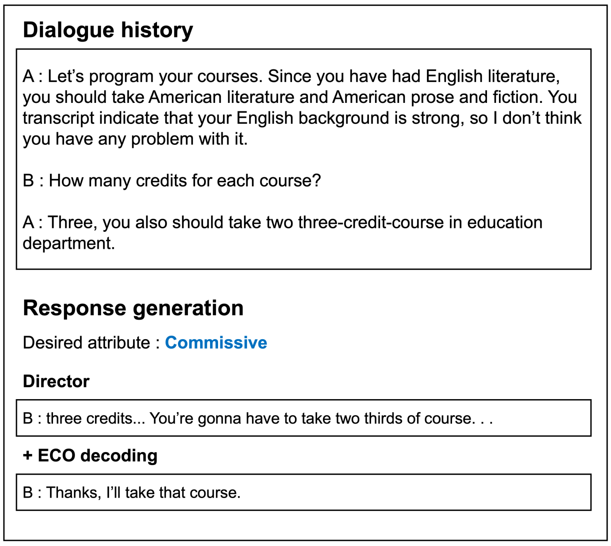

Figure 5: In case where the response fails to satisfy the desired attribute with the existing method but satisfies the desired attribute using ECO decoding.

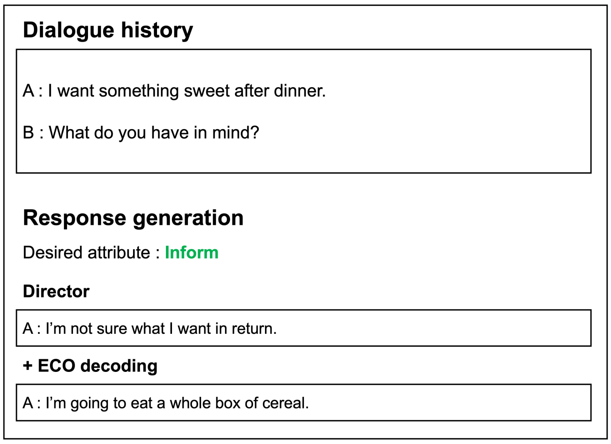

Figure 6: In case where it fails to generate a context-consistent response with the existing method but generates a context-consistent response using ECO decoding.

B Licenses

The DailyDialog dataset is licensed under CC BY-NC-SA 4.0 License. The DialoGPT model is licensed under Contributor License Agreement (CLA) and Llama2 model is licensed under Meta Llama 2 Community License Agreement. The RoBERTa-based CoLA grammaticality model is licenced under MIT License.