Change or Not: A Simple Approach for Plug and Play Language Models on Sentiment Control

Chen Xu, [1] **Jianyu Zhao,** [2] **Rang Li,** [2] **Changjian Hu,** [2] **Chuangbai Xiao** [1*]
[1]Beijing University of Technology
[2]Lenovo Research
chenxu05037@gmail.com, cbxiao@bjut.edu.cn

Abstract

Text generation with sentiment control is difficult without fine-tuning or modifying the model architecture. Plug and Play Language Model (PPLM) utilizes an external sentiment classifier to update the hidden states of GPT-2 at each time step. It does not change the parameters but achieves competitive performance. However, fluency is impaired due to the instability of the hidden states. Moreover, the classifier is not strong because of the way it is trained with partial texts, hence it is difficult to guide the generation in the process. To solve the above problems, in this paper, we first propose a fixed threshold method based on the Valence-Arousal-Dominance (VAD) lexicon to decide whether to change a word, which keeps the fluency of the original LM to the greatest extent. Furthermore, for the improvement of sentiment alignment, we propose a dynamic threshold method that utilizes VAD-based loss to make the threshold dynamic. Experiments demonstrate that our methods outperform the baseline with a great margin significantly both on fluency and sentiment accuracy.

Introduction

Transformer-based (Vaswani et al. 2017) pre-trained language models have made significant advances in natural language generation (Radford et al. 2019). Most unconditional LMs are trained on a huge text through a log-likelihood objective. Because of their remarkable fluency, there are growing interests in conditional text generation (Keskar et al. 2019). PPLM (Dathathri et al. 2019) solves the conditional text generation problem without changing the architecture or weights of pre-trained LM but utilizing an external sentiment classifier to calculate loss, which is then backpropagated to the original LM's hidden states at each time step. Hence, the word is sampled from the perturbed distribution through the recomputation. However, such excess modification may cause semantic confusion which impairs the fluency a lot. Moreover, in the process of generation, the discriminator may not always provide an accurate loss because of the difficulties in predicting the sentiment only based on partial text generated so far.

To address the aforementioned problems, we first incorporate VAD Lexicon (Mohammad 2018), a list of 20,000 English words and their valence, arousal, and dominance scores ranged between 0 and 1. The valence measures the sentiment direction of a word. We then propose a Fixed Threshold sampling method based on PPLM (PPLM-FT) to decide whether to change a word or not. The threshold ensures that only words making a relatively remarkable effect on the valence can replace the unperturbed words. Therefore, the unperturbed word would not be influenced if the discriminator does not guide well at certain steps. This keeps the fluency of the original LM to the greatest extent. To improve sentiment accuracy, we propose a Dynamic Threshold method (PPLM-DT) that enables VAD-based loss to make the threshold dynamic.

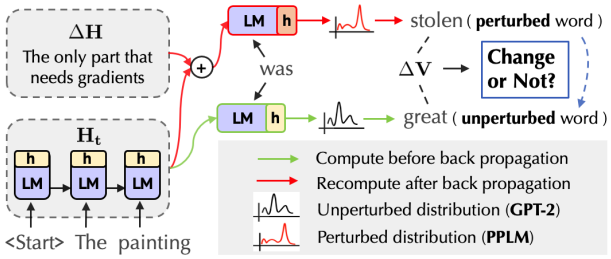

Figure 1: The whole architecture of our sampling method at a certain step. h is the last hidden layer of LM. H_t stands for the historic hidden states to the current time step t. ΔH is the update to H_t, such that generation with $(H_t + \Delta H)$ shifts the distribution towards the desired sentiment. ΔV stands for the valence difference of two candidate words.

Threshold for Sampling

Fixed Threshold

Figure 1 shows the whole architecture of our sampling methods. We could see that the word sampling process happens twice a time step. Therefore, two candidate words are sampled from unperturbed (before backpropagation) and perturbed distribution (after backpropagation) respectively. The difference of valence value between two candidate words is compared with the fixed threshold. Therefore, only the perturbed word that makes a relatively remarkable effect on the valence can replace the unperturbed word. Otherwise, the unperturbed word will not be changed, which keeps the flu-

*Corresponding author

ency the same as the original LM level to the greatest extent.

Dynamic Threshold

In order to improve the sentiment control and the balance with fluency, we design a VAD-based loss used to influence the threshold at each time step. This sentiment loss is defined as the difference between the valence of generated words so far and the valence of target sentiment. For example, if the task is positive control and the generated words so far have shown enough positivity, the sentiment loss will be small, so there is a higher probability of keeping the same word as the original LM, and vice versa. Our VAD-based loss for threshold is defined as:

$$\mathcal{L}_t = \sum_{i=1}^{k} p(w_i) \left| (\sum_{j=1}^{t-1} V(w'_j) + V(w_i))/t - V(tgt) \right| \quad (1)$$

where $\mathcal{L}_t$ represents the sentiment loss for the current time; p represents the softmax probability; w_i represents the i-th word within top-k probabilities; w' means the already generated word; V means the valence score; t means the current time step; tgt means the target sentiment: positive or negative.

By adding this simple VAD-based loss to the original PPLM loss, we realize that our VAD-based loss, in essence, makes the threshold dynamic based on whether the words have expressed enough target sentiment so far.

Experiments

We experiment to study the sentiment control and fluency over the generated texts given different prompts. We set the fixed threshold to 0.01, $V(positive)$ to 0.6, and $V(negative)$ to 0.4 after analyzing the lexicon. Besides, the k is set to 10 and for words not in the lexicon, their valence scores are set to 0.5 (neutral). For the evaluation process and other hyper-parameters, we keep the same with PPLM.

Automatic Evaluation

Following PPLM, the sentiment accuracy (ACC), perplexity (PPL), and distinct n-grams (Dist-n) are reported. PPL measures the fluency and Dist-n for the diversity.

Human Evaluation

Three external occupational annotators participate in the evaluation. For fluency, annotators are asked to give each individual sample a score on a scale of 1-5 and the average is used. For A/B testing on sentiment accuracy, the majority-voting is used on each pair of all 3 combinations of methods.

Experimental Results

Note that to reduce the randomness, we enlarge the number of samples for automatic evaluation from 45 (15 prompts $\times$ 3 samples) to 500 (50 prompts $\times$ 10 samples) for each class. Moreover, the statistic test is performed in both automatic and human evaluation. Table 1 reports the automatic evaluation performance. We could see that both PPLM-FT and PPLM-DT methods significantly outperform PPLM in all metrics. The latter shows stronger controllability. Table 2 shows the human evaluation results. We use 45 samples (same as PPLM) for each class. The results exhibit the same trend with automatic evaluation.

Methods	ACC	PPL	Dist-1	Dist-2	Dist-3
PPLM	59.60	48.42	0.205	0.583	0.806
PPLM-FT	61.41†	**44.14**†	**0.219**†	**0.635**†	**0.851**†
PPLM-DT	**63.73**$^{\dagger *}$	45.24†	0.216†	0.634†	0.850†

† $p < 0.001$, comparison with PPLM
* $p < 0.001$, comparison with PPLM-FT

Table 1: Automatic evaluation of methods on the sentiment control task. Statistical significance is computed with Wilcoxon signed-rank test.

Methods	Sentiment Accuracy	Fluency
PPLM	0.38	2.95
PPLM-FT	0.43	**3.70**†
PPLM-DT	**0.5***	3.65†

† $p < 0.001$, comparison with PPLM
* $p < 0.05$, comparison with PPLM

Table 2: Human evaluation of methods on the sentiment control task. Statistical significance is computed with one-tailed binomial test for the sentiment accuracy and two-tailed T-test for the fluency.

Conclusion

In this paper, we address the non-fluency of PPLM by proposing a method PPLM-FT, in order to decide whether to change a word. For the improvement of sentiment alignment and the balance with fluency, we further propose a method PPLM-DT that makes the threshold dynamic. Both automatic metrics and human assessment demonstrate that our methods significantly outperform the baseline both on fluency and sentiment accuracy.

References

Dathathri, S.; Madotto, A.; Lan, J.; Hung, J.; Frank, E.; Molino, P.; Yosinski, J.; and Liu, R. 2019. Plug and play language models: a simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164* .

Keskar, N. S.; McCann, B.; Varshney, L. R.; Xiong, C.; and Socher, R. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858* .

Mohammad, S. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *ACL (Volume 1: Long Papers)*, 174–184.

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

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NIPS*, 5998–6008.