Attend, Select and Eliminate: Accelerating Multi-turn Response Selection with Dual-attention-based Content Elimination

Jianxin Liang[1,2], **Chang Liu**[1,3], **Chongyang Tao**[4], **Jiazhan Feng**[1,2], **Dongyan Zhao**[1,3,5,6]*
[1] Wangxuan Institute of Computer Technology, Peking University
[2] School of Intelligence Science and Technology, Peking University
[3] Center for Data Science, Peking University
[4] Microsoft Corporation
[5] Institute for Artificial Intelligence, Peking University
[6] National Key Laboratory of General Artificial Intelligence, Peking University
{liangjx,liuchang97,fengjiazhan,zhaody}@pku.edu.cn
{chotao}@microsoft.com

Abstract

Although the incorporation of pre-trained language models (PLMs) significantly pushes the research frontier of multi-turn response selection, it brings a new issue of heavy computation costs. To alleviate this problem and make the PLM-based response selection model both effective and efficient, we propose an inference framework together with a post-training strategy that builds upon any pre-trained transformer-based response selection models to accelerate inference by progressively selecting and eliminating unimportant content under the guidance of context-response dual-attention. Specifically, at each transformer layer, we first identify the importance of each word based on context-to-response and response-to-context attention, then select a number of unimportant words to be eliminated following a retention configuration derived from evolutionary search while passing the rest of the representations into deeper layers. To mitigate the training-inference gap posed by content elimination, we introduce a post-training strategy where we use knowledge distillation to force the model with progressively eliminated content to mimic the predictions of the original model with no content elimination. Experiments on three benchmarks indicate that our method can effectively speeds-up SOTA models without much performance degradation and shows a better trade-off between speed and performance than previous methods.

1 Introduction

Constructing intelligent dialogue systems has attracted wide attention in the field of natural language processing (NLP) in recent years. There are two approaches widely used for the dialogue system, generation-based and retrieval-based methods. The former views conversation as a generation problem (Vinyals and Le, 2015; Serban et al., 2016; Zhang et al., 2020b), while the latter aims to select the optimal response from candidates given a dialog context (Wu et al., 2017; Tao et al., 2019b; Xu et al., 2021; Han et al., 2021; Feng et al., 2022). Since retrieval-based methods can usually provide fluent and informative responses, they are widely adopted in a variety of industrial applications such as XiaoIce (Shum et al., 2018) from Microsoft and AliMe Assist (Li et al., 2017) from Alibaba.

Context
A: can someone help me with installing drivers? this is the output file.
B: What drivers are you installing
A: I try to install the video card drivers, and it says to check out the log file of it.
B: Give more detail. How do you try to install those drivers? which log file is that.
A: The ones that ship with Ubuntu.
Response
B: This might be heavily connected, so maybe you have another driver manager running other open windows synaptic.

Table 1: A dialogue example from Ubuntu Corpus. The light gray words are eliminated in shadow layers, the light red words are eliminated in mediate layers, and the black words are retained all the time and sent to the deeper layer for the context and response matching.

We focus on multi-turn response selection in retrieval-based dialogue systems in this paper. Recently advances of pre-trained language models (Devlin et al., 2019) further push the research

[*] Corresponding author: Dongyan Zhao.

Findings of the Association for Computational Linguistics: ACL 2023, pages 6758–6770
July 9-14, 2023 ©2023 Association for Computational Linguistics

frontier of this field by providing a much powerful backbone for representation learning (Whang et al., 2020; Gu et al., 2020) and dialogue-oriented self-supervised learning (Xu et al., 2021; Zhang and Zhao, 2021; Han et al., 2021). Although significant performance improvement has been made by these PLM-based response selection models, they usually suffer from substantial computational cost and high inference latency due to the growing model size, presenting challenges for their development in resource-limited real-world applications. Therefore, there is an urgent need to accelerate PLM-based response selection models while maintaining their satisfactory performance.

To accelerate PLM-based multi-turn response selection, one direct idea is to avoid *unnecessary* calculation when joint modeling dialogue context and response. Through empirical observation, we find that there are many unimportant contents that are either redundant (i.e., repeated by many context turns) or less relevant to the topic, especially in the lengthy dialogue context (Zhang et al., 2018). If accurately identified and appropriately eliminated, the removal of the unnecessary calculation on them can bring minimum performance degradation. Drawing inspiration from Goyal et al. (2020), we propose an inference framework together with a post-training strategy customized for PLM-based multi-turn response selection, where unimportant contents are progressively identified and dropped as the calculation goes from shallow layers to deep. In our framework, we seek to answer three research questions (*RQs*): (1) how to accurately identify these unimportant contents, (2) how to properly decide the intensity of elimination for these unimportant contents under various computation demands, and (3) how to eliminate unnecessary calculations on those contents at the minimum cost of performance degradation. As the answer to the above questions, we propose an inference framework together with a post-training strategy customized for PLM-based multi-turn response selection as illustrated in Table 1. For *RQ1*, we propose a dual-attention-based method to measure the relative importance of tokens in context and response as we find this method is in accordance with our empirical observation. For *RQ2*, we adopt evolutionary search (Cai et al., 2019) to build the Pareto Frontier of performance-efficiency map and choose proper retention configurations (i.e., which defines how many tokens are passed to the next layer for each layer) from the frontier. For *RQ3*, we notice the gap between the proposed efficient inference framework and training and employ knowledge distillation (Hinton et al., 2015) to mitigate this gap by forcing the model with progressively eliminated contents to mimic the predictions of the original model with no content elimination.

We evaluate our proposed method on three benchmarks for multi-turn response selection: Ubuntu (Lowe et al., 2015), Douban (Wu et al., 2017) and E-commerce (Zhang et al., 2018). Experimental results show that our proposed method can accelerate the inference of PLM-based multi-response selection models with acceptable performance degradation under various computation constraints, while significantly outperforming previous acceleration methods. We also conduct comprehensive analyses to thoroughly investigate the effectiveness of proposed components.

We summarize the contributions of this paper as follows: (1) We propose Attend, Select and Eliminate (ASE), an efficient inference framework customized for PLM-based multi-turn response selection models that identify and progressively eliminate unimportant contents. (2) We propose a knowledge-distillation-based post-training strategy to mitigate the training-inference gap and decrease the performance degradation caused by content elimination. (3) We conduct comprehensive experiments on three benchmarks to verify the effectiveness of our proposed method and prove its superiority over other acceleration methods.

2 Related Work

Recently, methods based on pre-trained models are relatively popular, Whang et al. (2020) introduced the next sentence prediction and mask language model tasks in the PLMs into the conversation corpus, conducted post-domain training, and finally treated the context as a long sequence, and adjusted the model directly by fine-tuning the model. Compute context-response match scores. Xu et al. (2021) tries to introduce self-supervised learning tasks to increase the difficulty of model training, and the results show the effectiveness of these works. From the perspective of data augmentation, BERT-FP (Han et al., 2021) splits the context into multiple sets of short context-response pairs and introduces a conversational relevance task, which achieves state-of-the-art performance.

Although the performance of the pre-training

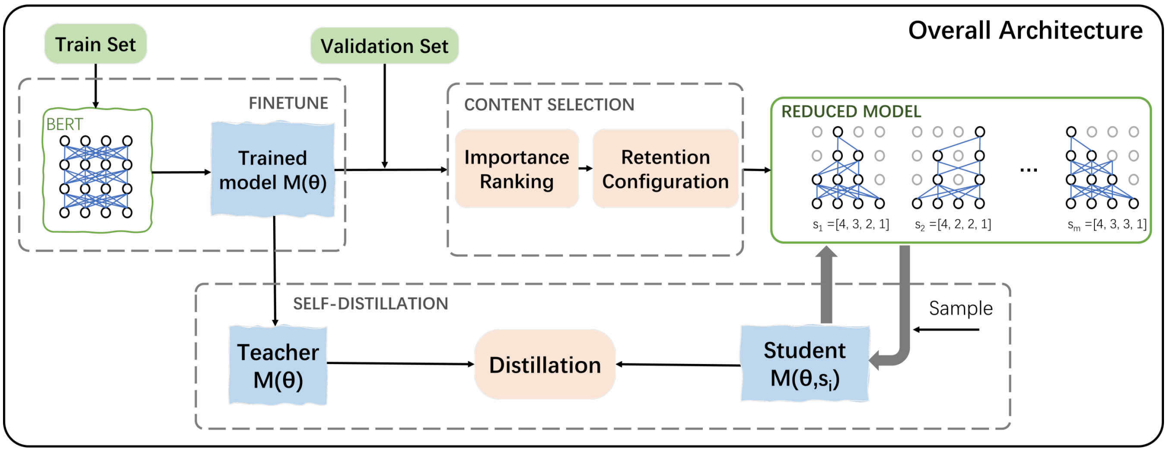


Figure 1: The Overall framework ASE.

model is powerful, it also brings some problems. The expensive computational cost and high inference latency hinder the further implementation of the PLMs to a certain extent. Some works try to alleviate this problem, one of the branches is to reduce the model size, such as distillation (Jiao et al., 2020; Wang et al., 2021; Liu et al., 2022a,b), structural pruning (Michel et al., 2019; Fan et al., 2019; Gordon et al., 2020; Hou et al., 2020) and quantization (Zafrir et al., 2019; Shen et al., 2020; Zhang et al., 2020a; Bai et al., 2021), etc. Goyal et al. (2020) adopts the Attention Strategy to select the important tokens with a fixed length configuration, but its speed ratio cannot be selected as needed and once full training can only get a model with a fixed speedup.

Since existing method Goyal et al. (2020) is mainly evaluated on single-sentence or sentence-pair tasks, it not fully suitable for response selection where the model needs to understand the relationship between all the utterances in a dialogue session and learn the interaction of the utterances closely related to the response. Therefore, we propose to select and eliminate the token representation based on context-to-response and response-to-context attention (i.e., dual-attention, **DualA**), which make good use of the relationship between context-response.

3 Task Formulation

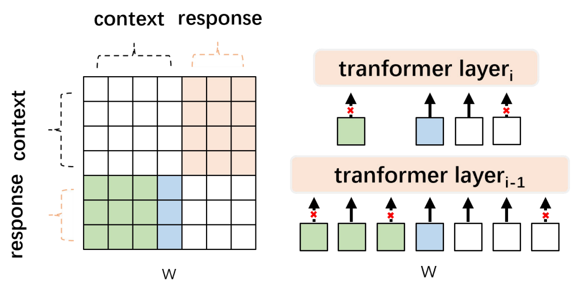
Considering a dialogue system given a dialogue dataset $D = \{(c_i, r_i, y_i)\}_{i=1}^{n}$. Each sample in the dataset is a triple that consists of context c_i, response r_i, and ground truth label y_i. $c_i = \{u_1, u_2, ..., u_l\}$ is dialogue context with l utterances and $\{u_j\}_{j=1}^{l}$ are arranged in a temporal order. r_i is a response candidate and $y_i = 1$ represents r_i is a proper response for the context c_i, otherwise $y_i = 0$. The core problem of this research is to learn a matching model $M(\cdot, \cdot)$ which can measure the matching degree between context and response.

4 Methodology

We aim to accelerate the inference of PLM-based multi-turn response selection models by proposing Attend, Select and Eliminate (ASE) that progressively identifies and eliminates unimportant contents to avoid unnecessary calculations. The overall framework is illustrated in Figure 1. There are three crucial questions that need to be answered: (1) how to accurately identify the unimportant contents, (2) how to properly decide the intensity of content elimination, and (3) how to effectively mitigate the training-inference gap in our framework and decrease the performance degradation. In the following part of this section, we elaborate on our method by answering the above three research questions.

4.1 Content Selection

In the specific scenario of multi-turn dialogue, there is a lengthy context with multiple turns and a single sentence of candidate response and the model aims to measure their semantic similarity. To achieve this goal, existing PLM-based methods calculate the interaction of all contents without distinction, regardless of the various importance of contents where many of them are redundant or topic-irrelevant. In order to eliminate them for inference acceleration, we need to accurately identify them first during encoder flow as in Figure 2(b).



(a) self-attention matrix (b) encoder information flow

Figure 2: (a) The averaged attention weights of posted by the blue response part as the token w's mutual-importance. (b) between the encoders, tokens are eliminated and selected to be sent to the next layer.

4.1.1 Empirical Methods

The multi-turn context accounts for a large proportion of the input pair (c_i, r_i), making it a good choice to start our content selection. For multi-turn context, the easiest way is to conduct content selection in sentence-level. Empirically, the last few utterances in the dialogue context are more close to the response in the dialogue flow, so they might be more important than the utterances in the beginning. Hereby, we can also simply select the last k utterances in the original context as the new context (i.e., $c_i = \{u_j\}_{j=n+1-k}^{n}$) and concatenate them with the candidate response, resulting in the setting that we denote as Last$_k$. Similarly, we can select other context utterances, such as the first k utterances and randomly selected k utterances which are denoted as First$_k$ and Rand$_k$, respectively.

4.1.2 Dual-attention-based Content Selection

Although simply adopting empirical methods (i.e., Last$_k$) yields plausible results as will be shown in our experiments later, this approach takes all the last k utterances without distinction, regardless of the various importance of utterances and tokens. A reasonable way is to conduct content selection in a more fine-grained manner (i.e., token-level). Recent works have shown that the importance of a token can be measured by the total attention weights it receives from other tokens (Goyal et al., 2020; Kim and Cho, 2021), denoted as **AM**. However, AM treats all tokens in the input sequence equally without distinction, neglecting the imbalanced relationships between tokens in context and response. Intuitively, for a token in the context, the attention it receives from other context tokens reflects its importance in the context, which we call self-importance, and the attention obtained from response tokens reflects its importance for semantic matching, which we call mutual-importance. Therefore, we propose to disentangle the attention received by a token into two parts: (1) the self-attention within a context or response and (2) the mutual-attention between a context and a response, and jointly consider them when measuring the importance of a token, and we call it **DualA**. Specifically, take a token w in the context for example in Figure 2(a), we use the averaged attention weights posed by the response tokens on it as its mutual-importance score, formulated as:

$$g_{\text{c,mutual}}(w) = \frac{1}{H \cdot |T_{res}|} \cdot \sum_{h=1}^{H} \sum_{w' \in T_{res}} A_h[w', w], \tag{1}$$

where T_{res} means the set of tokens belonging to the response, A_h represents the attention received by token w from w' on head h, and H denotes the number of attention heads. While for the self-importance of w, we adopt the averaged attention weights posed by other context tokens on it:

$$g_{\text{c,self}}(w) = \frac{1}{H \cdot |T_{con}|} \cdot \sum_{h=1}^{H} \sum_{\substack{w' \in T_{con} \\ w' \neq w}} A_h[w', w], \tag{2}$$

where T_{res} means the set of context tokens. We then jointly consider the self-importance and the mutual-importance of w by a weighted sum of $g_{\text{c,self}}(w)$ and $g_{\text{c,mutual}}(w)$:

$$g_{\text{c}}(w) = \alpha_c \cdot g_{\text{c,self}}(w) + \beta_c \cdot g_{\text{c,mutual}}(w), \tag{3}$$

where α_c, β_c that satisfy $0 \le \alpha_c, \beta_c \le 1$ and $\alpha_c + \beta_c = 1$ are weights for calculating the overall importance score for context tokens. Similarly, we can calculate the overall importance score for the tokens in the response with the only difference lying in the weights for response tokens α_r, β_r:

$$g_{\text{r}}(w) = \alpha_r \cdot g_{\text{r,self}}(w) + \beta_r \cdot g_{\text{r,mutual}}(w). \tag{4}$$

It should be noted that our method can be viewed as a generalization of typical attention-based importance measurement (Goyal et al., 2020), and can flexibly balance the influence of self-attention and dual-attention parts.

4.2 Retention Configuration Search

After having the basis for evaluating the importance of the token, the model needs to determine *retention configuration*, i.e., how to properly decide the intensity of content elimination and how many tokens to keep and pass to deeper encoder layers.

Given a PLM-based model $M(\theta)$ with m encoder layers, and θ is the parameter of model M. $S = \{s_1, s_2, \cdots, s_n\}$ is a set called *retention configurations* where $s_i = [l_1,\ l_2,\ l_3,\ \cdots,\ l_m]$ is a monotonically non-increasing sequence and l_j indicates that l_j tokens are kept from the output of the l_{j-1}-th encoder layer and passed to the l_j-th encoder layer. According to s, the model $M(\theta)$ keeps and eliminates the corresponding number of tokens in each encoder, $M(\theta)$ can get faster inference, but the performance may degrade.

In theory, there can be $\binom{l_0}{l_1} \times \binom{l_1}{l_2} \times \cdots \times \binom{l_{m-1}}{l_m}$ possible combinations for each s. By using evolutionary algorithms (Cai et al., 2019), we search for the Pareto Frontier to make the optimal tradeoffs between performance and efficiency which can satisfy various given computation constraints.

4.3 Training Framework

In the aforementioned sections, we have introduced our accelerated inference framework for PLM-based multi-turn response selection models. Here, we present our training framework.

Given a pre-trained language model such as BERT (Devlin et al., 2019), we first adapt it to the task of multi-turn response selection by using the SOTA method (i.e., BERT-FP (Han et al., 2021)) on some multi-turn response selection dataset, obtaining the model $M(\theta)$. Then we conduct *retention configuration search* (described in Sec. 4.2) based on our proposed method DualA to obtain a set of optimal retention configurations S^*.

Algorithm 1: Model Training Steps

```
Input: PLM (i.e.,BERT$_{base}$) ;
       Datasets $D_{train}$ and $D_{dev}$;
Initialize retention set $S$;
Training BERT$_{base}$ on $D_{train}$ to get $M(\theta)$
  using BERT-FP (Han et al., 2021);
repeat
    Sort the tokens based on the importance
      through Eq.(3) and Eq.(4) ;
    Generate new $s'$ by evolutionary
      algorithms (Cai et al., 2019);
    Update $S$ based on the efficiency and
      performance on $D_{dev}$ of $M(\theta, s')$;
until $S$ converges to get $S^*$;
repeat
    Randomly sample a configuration $s_j$
      from $S^*$;
    Optimize $M(\theta, s_j)$ by minimizing K-L
      divergence through Eq.(5);
until convergence;
Output: $M(\theta^*)$ and $S^*$
```

Now with the trained model $M(\theta)$ and S^* with n retention configurations, we can get n acceleration settings for model inference with various speedup ratio, denoted as $G = \{M(\theta, s_1), \cdots,\ M(\theta, s_n)\}$. Although one can directly utilize $M(\theta, s_j)$ for faster inference, we argue that there is a gap between the training and our proposed accelerated inference framework. The previously trained model $M(\theta)$ didn't encounter the situation where the input sequence of tokens is progressively eliminated from shallow layers to deep layers. Therefore, we propose to mitigate this training-inference gap with once-for-all self-distillation. Specifically, we fix $M(\theta)$ as the teacher and make a copy of it as the student. During self-distillation, the teacher receives the complete inputs without content elimination and produces a probability distribution $p_{M(\theta)}(c_i, r_i)$ of whether the response is appropriate to the context or not. While for the student, in order to ensure it can be customized to all retention configurations S^* simultaneously with the same parameters θ^*, we randomly sample the configuration s_j and compute its output distribution under content elimination setting as $p_{M(\theta', s_j)}(c_i, r_i)$, which is used to compute the KL-divergence with the teacher's outputs following Hinton et al. (2015):

$$\mathcal{L}_{\theta'} = D_{\mathrm{KL}}(p_{M(\theta)}(c_i, r_i) \| p_{M(\theta', s_j)}(c_i, r_i)). \quad (5)$$

After self-distillation, we obtain the adapted model $M(\theta^*)$ customized for all the searched optimal retention configurations S^*, making our final inference acceleration settings $G^* = \{M(\theta^*, s_1),\ \cdots,\ M(\theta^*, s_n)\}$ efficient at the minimum cost of performance degradation.

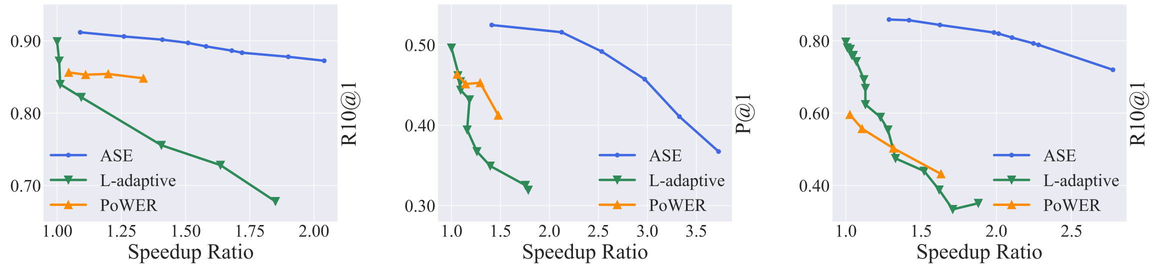
5 Experiments

5.1 Dataset

We evaluate our framework on three widely used multi-turn response selection benchmarks: the Ubuntu Corpus (Lowe et al., 2015), the Douban Corpus (Wu et al., 2017)and the E-commerce Corpus (Zhang et al., 2018).

Model	Ubuntu				Douban							E-commerce			
	R_{10}@1	R_{10}@2	R_{10}@5	Speed	MAP	MRR	P@1	R_{10}@1	R_{10}@2	R_{10}@5	Speed	R_{10}@1	R_{10}@2	R_{10}@5	Speed
SMN	0.726	0.847	0.961	-	0.529	0.569	0.397	0.233	0.396	0.724	-	0.453	0.654	0.886	-
DAM	0.767	0.874	0.969	-	0.550	0.601	0.427	0.254	0.410	0.757	-	0.526	0.727	0.933	-
MRFN	0.786	0.886	0.976	-	0.571	0.617	0.448	0.276	0.435	0.783	-	-	-	-	-
IOI	0.796	0.894	0.974	-	0.573	0.621	0.444	0.269	0.451	0.786	-	0.563	0.768	0.950	-
MSN	0.800	0.899	0.978	-	0.587	0.632	0.470	0.295	0.452	0.788	-	0.606	0.770	0.937	-
BERT	0.808	0.897	0.975	1x	0.591	0.633	0.454	0.280	0.470	0.828	1x	0.610	0.814	0.973	1x
BERT-DPT	0.851	0.924	0.984	1x	-	-	-	-	-	-	-	-	-	-	-
BERT-SL	0.884	0.946	0.990	1x	-	-	-	-	-	-	-	0.776	0.919	0.991	1x
BERT-FP	0.911	0.962	0.994	1x	0.644	0.680	0.512	0.324	**0.542**	**0.870**	1x	0.870	**0.956**	0.993	1x
BERT+**ASE***	0.813	0.902	0.976	2.0x	0.591	0.639	0.462	0.283	0.475	0.814	2x	0.664	0.837	0.973	2.3x
BERT+**ASE**†	0.828	0.910	0.979	1.1x	0.602	0.646	0.469	0.290	0.489	0.837	1.3x	0.700	0.852	0.971	1.4x
BERT-FP+**ASE***	0.897	0.955	0.991	1.5x	0.633	0.678	0.511	0.323	0.525	0.844	2x	0.843	0.941	0.993	1.4x
BERT-FP+**ASE**†	**0.914**	**0.964**	**0.994**	1.1x	**0.650**	**0.691**	**0.532**	**0.343**	0.536	0.856	1.4x	**0.872**	0.954	**0.996**	1.1x

Table 2: Model comparison on three benchmarks. BERT-FP is the previous SOTA model. **ASE*** and **ASE**† are two representative points of the models with a different speedup ratio.



(a) Ubuntu

(b) Douban

(c) E-commerce

Figure 3: Model performance-efficiency comparison of BERT-FP equipped with different accelerating methods.

5.2 Experimental Settings

We use BERT-FP's trained model to search on the validation set and get k (k<20) different length configurations. We adopt the weighted sum of the distillation loss and the cross-entropy loss, as the training objective function running 5 to 8 epochs. We employ recall rate $R_n@k$ as the evaluation metric. Especially for some samples in the Douban corpus having more than one true candidate response, we use MAP, MRR, and P@1 same as Tao et al. (2019b) and Yuan et al. (2019). For inference efficiency, we employ FLOPs (floating-point operations) speedup ratio compared to the BERT model as the measure, as it is agnostic to the choice of the underlying hardware. To avoid the pseudo improvement by pruning padding, we evaluate all models with input sequences without padding to the maximum length such as to pad length to 256.

5.3 Comparison Methods

We compare our method with these baselines: **(1)Interaction-based Models** where the context and response candidate interact with each other at the beginning stage. SMN (Wu et al., 2017), DAM (Zhou et al., 2018), IOI (Tao et al., 2019b), MSN (Yuan et al., 2019), MRFN (Tao et al., 2019a). **(2)BERT-based Models** where the context and response are concatenated together and feed into BERT-based models to BERT (Devlin et al., 2019), BERT-DPT (Whang et al., 2020), BERT-SL (Xu et al., 2021), BERT-FP (Han et al., 2021). **(3)Inference Accelerated Models** PoWER-BERT (Goyal et al., 2020), L-Adaptive (Kim and Cho, 2021).

5.4 Overall Performance

Table 2 and Figure 3 shows the overall comparison results with baselines. We can see that with ASE, the performance and efficiency of the BERT and BERT-FP are greatly improved. Specifically, BERT-FP+ASE† performs slightly better than the model BERT-FP on Ubuntu and E-commerce and achieves a significant improvement by 2.0% in $P@1$ and by 1.9% in $R_{10}@1$ on Douban. BERT-FP+ASE* achieves comparable performance with a double speed on Douban. The ASE also gives the vanilla BERT significant performance improvement: 9.0% in $R_{10}@1$ at 1.4x speed, 5.4% in $R_{10}@1$ at 2.3x on E-commerce, and slightly better performance with a double speed on Ubuntu and Douban. The detail of the BERT with ASE is shown in Appendix. Figure 3 compares the effect of combining BERT-FP with three different accelerating methods: ASE, PoWER-BERT, and L-adaptive. It can be seen that with ASE, BERT-FP achieves better results than with other method by a large margin, which demonstrates that extracting important tokens based on dual attention is feasible for accelerating the inference of multi-turn response selection. In contrast, both baselines have shown a large decline due to the incomplete adaptation of the task.

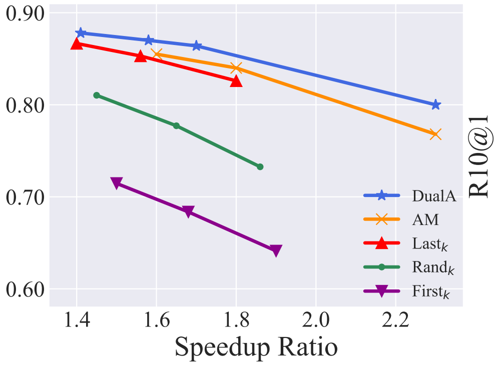


Figure 4: Comparison between different content selection strategies without self-distillation on Ubuntu.

5.5 Discussions

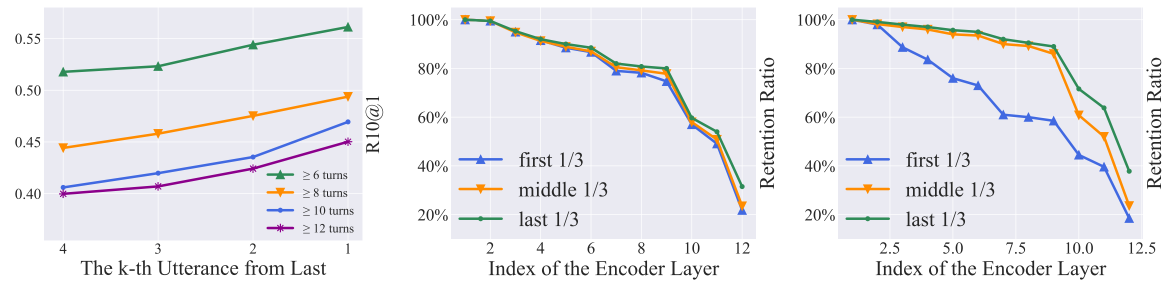
Comparison between different content selection strategies. Intuitively, the latter utterances may be helpful for the multi-turn response selection. We compare several different strategies, including empirical methods (i.e., Last_k, First_k, and Rand_k), the attention-based method AM and dual-attention-based method DualA.

Figure 4 shows the results of these strategies with k=3, 4, and 5 on Ubuntu. It can be seen that based on the three simple empirical strategies, Last_k, First_k, and Rand_k, the model can also achieve good performance with a certain inference speed. Strategy Last_k performs much better than strategy First_k and Rand_k, which validates our hypothesis that latter utterances in context may be more helpful and more important for selecting appropriate responses. Most importantly, the performance-efficiency tradeoffs of our proposed strategy based on dual attention are completely better than the other strategies. This result shows that to achieve the effect of faster inference, DualA, a fine-grained strategy of selecting token, is more effective than the utterance-level selection method for the response selection.

The effects of using only the k-th utterance from last as the context. To understand the effect of utterances in different positions on the task of response selection, we test the performance using only the k-th from last utterance as context. From the validation set, we first filter out examples where the context is too short and keep the examples where the context consists of more than 6, 8, 10, and 12 utterances on Ubuntu. Then, the k-th utterance from last of the context and the candidate response are concatenated, being fed to a trained model for classification. As experimental results in Figure 5(a) show, the overall performance of the model is relatively low. Even for the last utterance of the context, also the previous turn of the response, the performance is still not high. However, model performance increases rapidly as the utterance position moves forward under these four settings, which means that the closer the utterance to the candidate response, the better the performance for the response selection. This is also in line with the actual chat scene of human beings, where both parties usually respond to each other's current utterance.

The distribution of the selected token representations. Under the same retention configuration, the token selected by different strategies will be different. To better observe which tokens are selected by strategies, we divide the dialogue context into three parts, the first third, middle third, and last third of the context. On the Ubuntu IRC V1 corpus, we set the same retention configuration for both strategies, then as the encoder layer deepens, we count the distribution of token in the context part that is selected using AM and DualA.

In Figure 5(b), under the same retention configuration, it can be seen that under the method AM which uses the total attention weights it receives from other tokens to evaluate the token's importance, as the encoder layer deepens, the proportion of token selected in the last third part is slightly higher, while the first third and the middle third are basically the same. However, there is almost no difference in the distribution of the three parts. While in Figure 5(c), under the method DualA based on the dual-attention of the context and response, it can be seen that as the encoder layer deepens, the percentage of token selected in the first third of the context drops sharply. The middle and last third parts still retain a large part. Until after the ninth encoder layer, the middle and last parts begin to



(a) (b) AM (c) DualA

Figure 5: (a) Effect of using only single utterance for response selection. The distribution of selected tokens as the encoder layer deepens based on (b) AM and (c) DualA. Selection strategies are at the same configuration on Ubuntu.

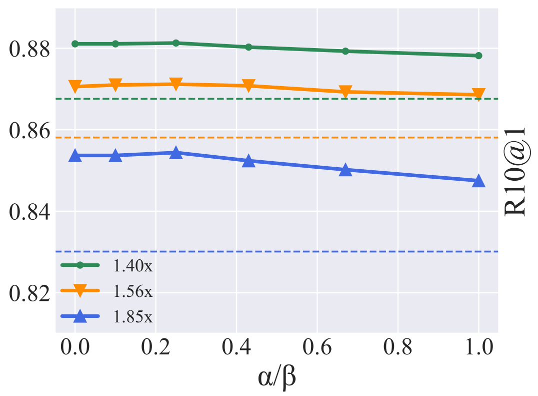


Figure 6: Hyper-parameter tuning for α and β at different Speedratio without dynamic self-distillation on Ubuntu. The dashed and solid lines represent the performance of AM and our method DualA, respectively.

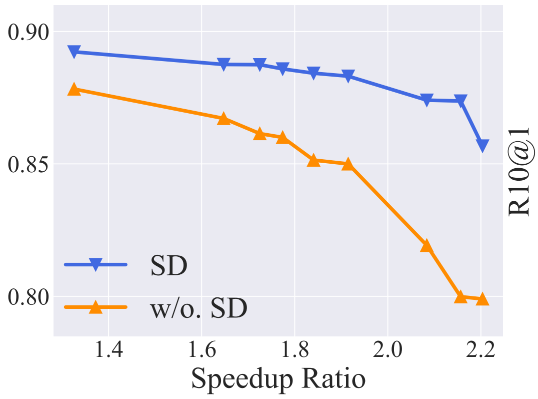
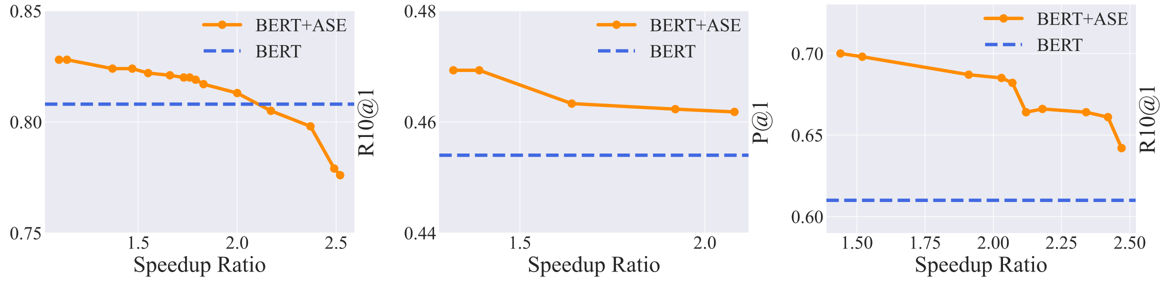


Figure 7: The effect of once-for-all self-distillation. **SD** and **w/o. SD** mean with and without self-distillation, respectively.

decrease drastically but are still more than the first third part of the context. This is consistent with the results in Figure 5(a). To a certain extent, this result shows that when the attention of response-to-context is used as the query, the response prefers to focus on the middle and last parts of the context, that is, the tokens that are closer to the response will provide more help in response selection, but are never the same.

Hyper-parameter tuning. According to Equation 4, the self-importance $g_{\text{r,self}}$ and the mutual-importance $g_{\text{r,mutual}}$ have different contributions to selecting tokens. We experiment with the effects on the performance with different $g_{\text{r,self}}$ and $g_{\text{r,mutual}}$ weights. As shown in Figure 6, the horizontal axis is α/β, which represents the weight coefficient of the $g_{\text{r,self}}$ to $g_{\text{r,mutual}}$ during the model selecting tokens belonging to the context. It can be seen that as the α/β increases, the tokens selected in the context change, and the performance also gradually improves, reaching the maximum at $\alpha/\beta = 0.25$. Consistent with our finds in Figure 4, method DualA is consistently performant than AM by a large margin. These results under different speedup ratios show consistent trends, i.e., the method of selecting tokens based on dual-attention is more effective for the response selection task.

The effects of the once-for-all self-distillation. After token selection, we compare model performance on Ubuntu with or without self-distillation. Different from the traditional distillation method, we adopt the once-for-all self-distillation method to distill the teacher's knowledge to the student by sampling different retention configurations during the training. Figure 7 is a comparison of the performance with and without self-distillation. It can be seen that with self-distillation, the performance is significantly improved for the model under all retention configurations, especially at large speedup ratio. As the speedup ratio of the model increases, that is, more tokens are eliminated during



(a) Ubuntu (b) Douban (c) E-commerce

Figure 8: Comparison between BERT+ASE and BERT on (a) Ubuntu, (b) Douban and (c) E-commerce.

inference, and the performance of the model starts to degrade, but the performance improvement of self-distillation is also enhanced. This way of optimizing all the retention in the training once avoids the problem of re-distilling if configuration various during the actual deployment process.

The flexibility of ASE. We demonstrate the flexibility of ASE by applying it on top of vanilla BERT. ASE can be easily integrated with any BERT-like model. We use the bert-base model from Huggingface[1] and finetune it on three benchmarks: Ubuntu, Douban and E-commerce. Then we apply the Dual-attention-based Content Selection method in Section 4.1.2 to search for the optimal retention and perform self-distillation. Figure 8 shows that ASE can boost BERT performance by 2.0% at 1.1x on Ubuntu and 9.0% at 1.4x on E-commerce.

6 Conclusion

In this paper, we propose a new framework of progressively extracting important tokens and eliminating redundant tokens to accelerate inference for multi-turn response selection, which identifies important tokens based on dual-attention of the context and response. The experimental results empirically verify the effectiveness of this method. In the future, we plan to accelerate inference further by combining it with the layer-wise reduction.

Limitations

During the configuration search stage, because this is a multi-objective optimization problem involving performance and efficiency, we use the evolutionary algorithm to search here. Designing a robust and efficient optimization objective is not simple and it will affect the convergence of search results. Limited by hardware, and in order to speed up the search, we use a small subset of the validation set to search retention configuration, which is bound to have a certain impact on the overall search results.

Ethical Statement

In this paper, we propose ASE, an algorithm to accelerate multi-turn response selection by prograssively selecting and eliminating unimportant tokens. The training corpora including the Ubuntu Corpus, the Douban Corpus and the E-commerce Corpus used for evaluating our framework are publicly available and don't pose privacy issues. The algorithm that we propose does not introduce ethical or social bias.

Acknowledgements

We would like to thank the anonymous reviewers for their constructive comments. This work was supported by the National Key Research and Development Program of China (No. 2020AAA0106600).

References

Haoli Bai, Wei Zhang, Lu Hou, Lifeng Shang, Jin Jin, Xin Jiang, Qun Liu, Michael Lyu, and Irwin King. 2021. BinaryBERT: Pushing the limit of BERT quantization. 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 4334–4348, Online. Association for Computational Linguistics.

Han Cai, Chuang Gan, Tianzhe Wang, Zhekai Zhang, and Song Han. 2019. Once-for-all: Train one network and specialize it for efficient deployment. *arXiv preprint arXiv:1908.09791*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of

[1] https://huggingface.co/bert-base-uncased, https://huggingface.co/bert-base-chinese

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.

Angela Fan, Edouard Grave, and Armand Joulin. 2019. Reducing transformer depth on demand with structured dropout. *arXiv preprint arXiv:1909.11556*.

Jiazhan Feng, Chongyang Tao, Chang Liu, Rui Yan, and Dongyan Zhao. 2022. How to represent context better? an empirical study on context modeling for multi-turn response selection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7285–7298, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mitchell Gordon, Kevin Duh, and Nicholas Andrews. 2020. Compressing BERT: Studying the effects of weight pruning on transfer learning. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 143–155, Online. Association for Computational Linguistics.

Saurabh Goyal, Anamitra Roy Choudhury, Saurabh Raje, Venkatesan Chakaravarthy, Yogish Sabharwal, and Ashish Verma. 2020. Power-bert: Accelerating bert inference via progressive word-vector elimination. In *International Conference on Machine Learning*, pages 3690–3699. PMLR.

Jia-Chen Gu, Tianda Li, Quan Liu, Zhen-Hua Ling, Zhiming Su, Si Wei, and Xiaodan Zhu. 2020. Speaker-aware bert for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2041–2044.

Janghoon Han, Taesuk Hong, Byoungjae Kim, Youngjoong Ko, and Jungyun Seo. 2021. Fine-grained post-training for improving retrieval-based dialogue systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1549–1558, Online. Association for Computational Linguistics.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Lu Hou, Zhiqi Huang, Lifeng Shang, Xin Jiang, Xiao Chen, and Qun Liu. 2020. Dynabert: Dynamic bert with adaptive width and depth. *Advances in Neural Information Processing Systems*, 33:9782–9793.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online. Association for Computational Linguistics.

Gyuwan Kim and Kyunghyun Cho. 2021. Length-adaptive transformer: Train once with length drop, use anytime with search. 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 6501–6511, Online. Association for Computational Linguistics.

Feng-Lin Li, Minghui Qiu, Haiqing Chen, Xiongwei Wang, Xing Gao, Jun Huang, Juwei Ren, Zhongzhou Zhao, Weipeng Zhao, Lei Wang, et al. 2017. Alime assist: An intelligent assistant for creating an innovative e-commerce experience. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2495–2498.

Chang Liu, Chongyang Tao, Jiazhan Feng, and Dongyan Zhao. 2022a. Multi-granularity structural knowledge distillation for language model compression. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1001–1011, Dublin, Ireland. Association for Computational Linguistics.

Chang Liu, Chongyang Tao, Jianxin Liang, Tao Shen, Jiazhan Feng, Quzhe Huang, and Dongyan Zhao. 2022b. Rethinking task-specific knowledge distillation: Contextualized corpus as better textbook. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10652–10658, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909*.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? *arXiv preprint arXiv:1905.10650*.

Iulian Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Sheng Shen, Zhen Dong, Jiayu Ye, Linjian Ma, Zhewei Yao, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. 2020. Q-bert: Hessian based ultra low precision quantization of bert. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8815–8821.

Heung-Yeung Shum, Xiaodong He, and Di Li. 2018. From eliza to xiaoice: challenges and opportunities with social chatbots. *arXiv preprint arXiv:1801.01957*.

Chongyang Tao, Wei Wu, Can Xu, Wenpeng Hu, Dongyan Zhao, and Rui Yan. 2019a. Multi-representation fusion network for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pages 267–275.

Chongyang Tao, Wei Wu, Can Xu, Wenpeng Hu, Dongyan Zhao, and Rui Yan. 2019b. One time of interaction may not be enough: Go deep with an interaction-over-interaction network for response selection in dialogues. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1–11, Florence, Italy. Association for Computational Linguistics.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. 2021. MiniLMv2: Multi-head self-attention relation distillation for compressing pretrained transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2140–2151, Online. Association for Computational Linguistics.

Taesun Whang, Dongyub Lee, Chanhee Lee, Kisu Yang, Dongsuk Oh, and Heuiseok Lim. 2020. An effective domain adaptive post-training method for bert in response selection. In *INTERSPEECH*, pages 1585–1589.

Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2017. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 496–505, Vancouver, Canada. Association for Computational Linguistics.

Ruijian Xu, Chongyang Tao, Daxin Jiang, Xueliang Zhao, Dongyan Zhao, and Rui Yan. 2021. Learning an effective context-response matching model with self-supervised tasks for retrieval-based dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14158–14166.

Chunyuan Yuan, Wei Zhou, Mingming Li, Shangwen Lv, Fuqing Zhu, Jizhong Han, and Songlin Hu. 2019. Multi-hop selector network for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 111–120, Hong Kong, China. Association for Computational Linguistics.

Ofir Zafrir, Guy Boudoukh, Peter Izsak, and Moshe Wasserblat. 2019. Q8bert: Quantized 8bit bert. In *2019 Fifth Workshop on Energy Efficient Machine Learning and Cognitive Computing-NeurIPS Edition (EMC2-NIPS)*, pages 36–39. IEEE.

Wei Zhang, Lu Hou, Yichun Yin, Lifeng Shang, Xiao Chen, Xin Jiang, and Qun Liu. 2020a. TernaryBERT: Distillation-aware ultra-low bit BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 509–521, Online. Association for Computational Linguistics.

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

Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, Hai Zhao, and Gongshen Liu. 2018. Modeling multi-turn conversation with deep utterance aggregation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3740–3752, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Zhuosheng Zhang and Hai Zhao. 2021. Structural pre-training for dialogue comprehension. 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 5134–5145, Online. Association for Computational Linguistics.

Xiangyang Zhou, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, and Hua Wu. 2018. Multi-turn response selection for chatbots with deep attention matching network. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1118–1127, Melbourne, Australia. Association for Computational Linguistics.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitiations,7

□ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract ,1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

□ B1. Did you cite the creators of artifacts you used?
No response.

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

□ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☑ Did you run computational experiments?

5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
5

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
5

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
5

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
5

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?
5

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.