Intent Discovery with Frame-guided Semantic Regularization and Augmentation

Yajing Sun and **Rui Zhang** and **Jingyuan Yang** and **Wei Peng** *
Artificial Intelligence Application Research Center, Huawei Technologies

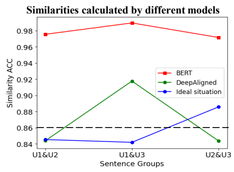
Abstract

Most intent discovery methods leverage representation learning and clustering to transfer the prior knowledge of known intents to unknown ones. The learned representations are limited to the syntactic forms of sentences, therefore, fall short of recognizing adequate variations under the same meaning of unknown intents. This paper proposes an approach utilizing frame knowledge as conceptual semantic guidance to bridge the gap between known intents representation learning and unknown intents clustering. Specifically, we employ semantic regularization to minimize the bidirectional KL divergence between model predictions for frame-based and sentence-based samples. Moreover, we construct a frame-guided data augmenter to capture intent-friendly semantic information and implement contrastive clustering learning for unsupervised sentence embedding. Extensive experiments on two benchmark datasets show that our method achieves substantial improvements in accuracy (5%+) compared to solid baselines.

1 Introduction

Dialogue systems such as chatbots and virtual assistants increasingly rely on their ability to understand and identify user intents, which directly impacts the system's performance. (Tseng et al., 2019; Guo et al., 2021). Real-world dialogue systems have to deal with evolving user needs, resulting in continuously increasing intents. To address these challenges, intent discovery methods that aim to detect unknown user intents from a large number of unlabeled utterances have been developed in recent years (Perkins and Yang, 2019; Min et al., 2020).

Previous intent discovery work focuses on the unsupervised setting with clustering algorithms (Shi et al., 2018; Min et al., 2020). Min et al. (2020) proposed alternating-view k-means for joint multi-view representation learning and clustering analysis. However, the unsupervised approaches often produce unsatisfactory performance because of their low degree of utilization of prior knowledge of known intents. Therefore, recent works (Hsu et al., 2019; Lin et al., 2020; Zhang et al., 2021) invariantly pretrain an intention detection model using limited known intents data, followed by leveraging clustering techniques to infer unknown intention. Zhang et al. (2021) propose an iterative representation learning and clustering method, DeepAligned, for intention detection and discovery.



Utterances	Intention label
U1: I am not seeing recent cash withdrawal on my account.	Pending cash withdrawal
U2: I have a cash withdrawal that I don't recognize.	Cash withdrawal not recognized
U3: Why cash withdrawal on my account that I don't remember?	

Figure 1: Similarities between different sentences calculated by BERT (red) and DeepAligned (green). The black dash line represents the similarity threshold to cluster a same class. U2 and U3 have the same intent, but U1 and U3 have different intentions even with a phrase overlap of "cash withdrawal on my account".

Learning intention-friendly semantic representation is vital for semi-supervised methods to improve the performance of intent detection and discovery. However, it's insufficient to merely rely on the labeled data. Methods pretrained on known intents are limited to the syntactic forms of sentences, therefore, fall short of recognizing adequate variations under the same meaning of unknown intents. Figure 1 displays the semantic similarities between the sentences with the overlapping phrase "cash withdrawal on my account" but a different intention. Red and green polylines represent the sentence similarity produced by the BERT (Devlin et al., 2019) and DeepAligned. It can be observed that both BERT and DeepAligned models ill-group U1 and U3 into the same class because the similarity between U1 and U3 is higher than the decision threshold (i.e., 86% illustrated

* Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2023, pages 14254–14261
July 9-14, 2023

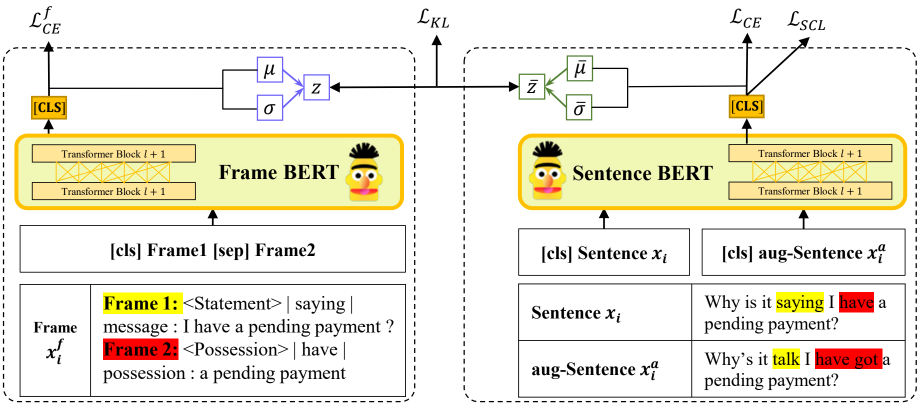


Figure 2: Intention Encoder Module. Given a labeled intent sample x_i, the frame annotation x_i^f and the augmented sample x_i^a, the module fine-tunes Sentence-BERT and Frame-BERT with four objectives. In representation learning, we use the labeled utterance set $\mathcal{D}_{\text{labeled}}$, and unlabeled utterance set $\mathcal{D}_{\text{unlabeled}}$ is used in the clustering stage.

as a black dash line in Figure 1). But ideally, the model should group U2 and U3 into the same intent class (depicted as the blue line in Figure 1). It is important to enhance the semantic representation learning ability of intent discovery methods. In this paper, we propose an approach utilizing frame knowledge as conceptual semantic guidance to bridge the gap between known intents representation learning and unknown intents clustering. Frame knowledge is defined as the conceptual structure based on FrameNet (Baker et al., 1998; Das et al., 2014), capturing the background knowledge necessary to understand a situation. Specifically, we introduce a novel frame-guided semantic regulation and augmentation method to enhance intention discovery. Semantic regularization minimizes the bidirectional KL divergence between model predictions made on the frame and sentence-based samples to learn conceptual semantic information. In the meantime, we construct a frame-guided data augmenter to capture intent-friendly semantic information and implement contrastive learning for semi-supervised sentence embedding learning. Finally, we perform an iterative representation learning and clustering method for intention detection and discovery. Extensive experiments on two benchmark datasets show that our method achieves substantial improvements, over 5% in accuracy compared to solid baselines.

2 Model

2.1 Task Formulation

Let $I_{\text{known}} = \{I_{\text{known}}^1, I_{\text{known}}^2, \ldots, I_{\text{known}}^n\}$ denote the set of n known intents and $I_{\text{unknown}} = \{I_{\text{unknown}}^1, I_{\text{unknown}}^2, \ldots, I_{\text{unknown}}^m\}$ denote the set of m unknown intents. Given a labeled utterance set $\mathcal{D}_{\text{labeled}} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l) \mid y_i \in I_{\text{known}}\}$ which contains n known intents and an unlabeled utterance set $\mathcal{D}_{\text{unlabeled}} = \{(\tilde{x}_1, \tilde{y}_1), (\tilde{x}_2, \tilde{y}_2), \ldots, (\tilde{x}_r, \tilde{y}_r) \mid \tilde{y}_i \in I_{\text{known}} \cup I_{\text{unknown}}\}$ containing n known intents and m unknown intents, intent discovery aims to group utterances into different clusters using the PLMs trained by labeled utterance. In our work, we exploit KAF-SPA (Zhang et al., 2023) to extract the frame annotation x_i^f, as shown in the Figure 2 and produce augmented sentence x_i^a for the utterance x_i.

2.2 Model Architecture

We adopt an iterative representation learning and clustering for intention discovery. In the representation learning stage, the parameters of the intention encoder module, as shown in Figure 2, are updated based on labeled data from known intents. The module proposes frame-guided semantic regularization and augmentation with four objectives to capture the conceptual semantic information of utterances. In the clustering stage, inspired by DeepAligned, we calculate pseudo-labels for unlabeled data and fine-tune Sentence-BERT for self-supervised learning. Specifically, we freeze the Frame-BERT parameters and minimize the KL divergence of Sentence-BERT and Frame-BERT models to retain the conceptual semantic representation of known intents.

2.3 Intention Encoder Module

The module is responsible for learning intention-friendly semantic representation of utterance and performing intention detection. Specifically, given an labeled intent samples x_i, the corresponding frame annotation x_i^f and the augmented sample x_i^a, we first obtain their intent representation z_i and z_i^a using Sentence-BERT with a pooling layer, and frame-based intention representation z_i^f with Frame-BERT encoder with a pooling layer. Then we employ four objectives in intention pretraining process: instance/frame-wise intent detection $\mathcal{L}_{\mathrm{CE}}(\theta)$ and $\mathcal{L}_{\mathrm{CE}}^f(\theta)$, semantic regularization $\mathcal{L}_{\mathrm{KL}}(\theta)$ and unsupervised semantic augmentation $\mathcal{L}_{\mathrm{SCL}}(\theta)$ via clustering contrastive learning.

Instance/frame-wise Intent Detection The objective aims to achieve intention detection based on the labeled utterances. We fine-tune sentence-BERT and frame-BERT with transformation heads respectively by minimizing the cross-entropy loss $\mathcal{L}_{\mathrm{CE}}$ and $\mathcal{L}_{\mathrm{CE}}^f$ respectively, where

$$\begin{aligned} p(I_{\text{known}} \mid x_i) &= \operatorname{softmax}(z_i * W + b)) \\ \mathcal{L}_{\mathrm{CE}} &= -\frac{1}{N}\sum_j y \cdot \log\left(p(I_{\text{known}}^j \mid x_i)\right) \end{aligned} \tag{1}$$

and $\mathcal{L}_{\mathrm{CE}}^f$ is similarly defined by replacing x_i, z_i with x_i^f, z_i^f in Equation (1).

In Equation (1), $p(I_{\text{known}} \mid x_i)$ denotes the probability of assigning to x_i or x_i^f to the known intents. After fine-tuning two BERT models based on labeled utterances, the transformation heads are discarded. In the second clustering stage, BERT with a pooling layer is continuously fine-tuned while Frame-BERT with a pooling layer is frozen, which is regarded as conceptual semantic guidance.

Unsupervised Semantic Regularization To guide the Sentence-BERT concentrating on the conceptual semantic representation, we apply semantic regularization to minimize the bi-directional KL-divergence between model predictions made by sentence-BERT and frame-BERT modules. Given the intent representation z_i and z_i^f, we assume $\mathcal{P}_\phi(z|z_i)$ and $\mathcal{P}_\psi(\overline{z}|z_i^f)$ take the following forms:

$$\mathcal{P}_\phi(z|z_i) \sim \mathcal{N}(\mu, \sigma^2\mathbf{I}),\;\; \mathcal{P}_\psi(\overline{z}|z_i^f) \sim \mathcal{N}(\overline{\mu}, \overline{\sigma}^2\mathbf{I}) \tag{2}$$

The μ and σ are calculated based on the z_i,

$$\begin{aligned} \mu &= z_i \cdot \mathbf{W}_\mu + b_\mu \\ \log \sigma^2 &= z_i \cdot \mathbf{W}_\sigma + b_\sigma \\ z &= \mu + \sigma \odot \epsilon \end{aligned} \tag{3}$$

and $\overline{\mu}, \overline{\sigma}$ are calculated in a similar way by replacing μ, σ, z, z_i with $\overline{\mu}, \overline{\sigma}, \overline{z}, z_i^f$ with learnable parameters $\overline{W}_\mu$,$\overline{W}_\sigma$, $\overline{b}_\mu$ and $\overline{b}_\sigma$ in Equation (3). Then we formulate the negative KL divergence:

$$\mathcal{L}_{\mathrm{KL}}(\theta) = \sum_i \mathcal{P}_\phi(z|z_i) \log\left(\frac{\mathcal{P}_\phi\left(z|z_i\right)}{\mathcal{P}_\psi(\overline{z}|z_i^f)}\right) \tag{4}$$

Semi-supervised Semantic Augmentation As shown in the Figure 2, the positive samples x_i^a are generated by replacing the target words in the original samples x_i with another word that belongs to the same frame as the target words. Then we employ contrastive learning to model instance-wise semantic similarities by pulling together intents belonging to the same class while pushing apart samples from different classes in one batch $\mathcal{B}$. The equation is as follows:

$$\mathcal{L}_{\mathrm{SCL}}(\theta) = -\log \frac{\exp(\operatorname{sim}(z_i, z_i^a))}{\sum_{j=1}^{\mathcal{B}} \mathbb{1}_{[j\neq i]} \exp(\operatorname{sim}(z_i, z_j))} \tag{5}$$

where $\mathcal{B}$ denotes the batch size. Finally, The intention encoder module is trained with the loss $\mathcal{L}$.

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \mathcal{L}_{\mathrm{CE}}^f + \mathcal{L}_{\mathrm{KL}} + \mathcal{L}_{\mathrm{SCL}} \tag{6}$$

3 Experiments

3.1 Set up

The experiments are conducted on CLINC (Larson et al., 2019), and Banking77 (Casanueva et al., 2020) datasets. CLINC contains $22,500$ utterances with 150 intents, and Banking77 consists of $13,083$ utterances with 77 intents. Followed DeepAligned (Zhang et al., 2021) model, we randomly select 10% of training data as labeled and 75% of all intents as known intents. The model is compared with strong baselines: BERT-MCL (Hsu et al., 2019), CDAC+ (Lin et al., 2020) and DeepAligned (Zhang et al., 2021). Three widely used metrics are adopted to evaluate the clustering results: Accuracy (ACC), Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI).

3.2 Training details

For a fair comparison with previous work, hyperparameters are adopted as suggested in Zhang et al. (2021). The parameters of Frame-BERT and $\overline{\mu}$, $\overline{\sigma}$ are frozen during clustering. The unlabeled data are aligned per 10 epochs in the clustering stage.

Method	CLINC								
	25%			50%			75%		
	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI
BERT-MCL(Hsu et al., 2019)	24.35	16.82	65.06	47.21	36.72	78.39	69.66	59.92	87.72
CDAC+ (Lin et al., 2020)	64.64	50.35	84.25	69.02	54.15	86.18	69.89	54.33	86.65
DeepAligned* (Zhang et al., 2021)	-	-	-	66.31	60.68	88.22	70.76	67.28	91.28
Our model	**65.47**	**57.64**	**87.47**	**72.84**	**65.24**	**90.44**	**77.07**	**72.61**	**92.3**
	Banking77								
	25%			50%			75%		
BERT-MCL(Hsu et al., 2019)	24.53	15.51	49.46	42.28	29.80	62.50	61.14	47.43	75.68
CDAC+ (Lin et al., 2020)	48.79	34.88	67.65	51.97	38.61	70.62	53.83	40.97	72.25
DeepAligned (Zhang et al., 2021)	48.88	36.81	70.45	59.23	47.82	76.52	64.90	53.64	79.56
Our model	**59.25**	**49.32**	**68.33**	**62.37**	**52.10**	**79.49**	**70.78**	**61.75**	**84.82**

Table 1: The results on CLINC150 and Banking77 datasets in the semi-scenario for 25%, 50%, and 75% known intent ratios (KIR). * represents the results implemented by us.

3.3 Result and Discussion

3.3.1 Main results

We report the results across all models on the CLINC and Banking77 with different known intents ratios in Table 1. Our model achieves the best results on all metrics and datasets compared to the baselines, demonstrating the effectiveness of the proposed approach. Another observation is that the smaller ratios of known intents are included in the fine-tuning, the more significant improvements are achieved compared to baselines. The contribution can be explained in two aspects. Firstly, frame regularization and augmentation alleviate the overfitting problem when fine-tuning. Secondly, freezing Frame-BERT in the clustering stage guides the Sentence-BERT to avoid noisy pseudo-labeled data distribution.

3.3.2 Ablation study

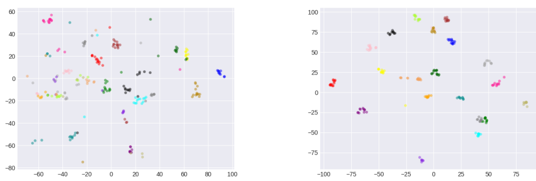
To understand the effect of the proposed method, we perform an ablation study (in Table 2) by removing frame regularization and augmentation, respectively. There is a significant drop in performance on all datasets with all metrics, especially for the Banking77 datasets when frame regularization and augmentation are absent. The dropped performance verifies the effectiveness of frame-based semantic guidance in bridging the gap between known intents representation learning and unknown intent clustering. The degradation is more obvious in Banking77 datasets because the intentions is fine-grained and closed compared to CLINC datasets.

3.3.3 Case study

To further verify the model's effectiveness, we visualize the representation in the clustering stage of DeepAligned and our model (in Figure 3). We can find the distribution obtained by DeepAligned are sparse, with more utterance clusters indicating higher errors in the output. In comparison, the representation learned from our model appears aggregated to fewer groups, potentially ascribed to the effect of the conceptual semantic learning ability of the proposed method. The results are consistent with the case study shown in Figure 1. Our model performs closer to an ideal model in alleviating the ill influence the syntactic forms bring.

	CLINC			Banking77		
	ACC	ARI	NMI	ACC	ARI	NMI
our model	77.07	72.61	92.3	70.78	61.75	84.82
w/o regularization	72.36	67.87	91.54	63.05	56.13	82.51
w/o augmentation	71.16	67.73	91.35	66.53	57.95	82.77

Table 2: Ablation study with KIR(75%) on CLINC and Banking77 datasets. "w/o" stands for "without".



(a) DeepAligned. (b) Our model.

Figure 3: TSNE plots for Banking77 with 75% KIR.

4 Conclusion and Future Work

This paper introduces frame knowledge as conceptual semantic guidance and proposes a frame-guided semantic regularization and augmentation method. Experimental results on two benchmarks demonstrate that our method outperforms strong baselines by a wide margin. In future work, we will explore practical ways to reason latent intentions based on the frame ontology knowledge.

Limitations

There are two limitations to this work. (1) the total number of known and unknown intents are predefined, requiring an extension in real-world scenarios; (2) The frame knowledge is predefined and, therefore, inflexible to address complex intents. In addition, some user queries have no frame in FrameNet matching.

There are additional computation costs for frame knowledge learning. The model fine-tunes two BERT(bert-base-uncased, 340M) models in the training stage and runs sentence-BERT in the evaluation stage. The pre-training stage of our model lasts about 10 minutes, and clustering runs for 90 minutes on CLINC with a 75% known intents ratio, both using a single NVIDIA Tesla V100 GPU(32 GB of memory).

References

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45, Online. Association for Computational Linguistics.

Dipanjan Das, Desai Chen, André FT Martins, Nathan Schneider, and Noah A Smith. 2014. Frame-semantic parsing. *Computational linguistics*, 40(1):9–56.

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

Jinyu Guo, Kai Shuang, Jijie Li, and Zihan Wang. 2021. Dual slot selector via local reliability verification for dialogue state tracking. 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 139–151, Online. Association for Computational Linguistics.

Yen-Chang Hsu, Zhaoyang Lv, Joel Schlosser, Phillip Odom, and Zsolt Kira. 2019. Multi-class classification without multi-class labels. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. An evaluation dataset for intent classification and out-of-scope prediction. 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 1311–1316, Hong Kong, China. Association for Computational Linguistics.

Ting-En Lin, Hua Xu, and Hanlei Zhang. 2020. Discovering new intents via constrained deep adaptive clustering with cluster refinement. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8360–8367. AAAI Press.

Qingkai Min, Libo Qin, Zhiyang Teng, Xiao Liu, and Yue Zhang. 2020. Dialogue state induction using neural latent variable models. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3845–3852. ijcai.org.

Hugh Perkins and Yi Yang. 2019. Dialog intent induction with deep multi-view clustering. 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 4016–4025, Hong Kong, China. Association for Computational Linguistics.

Chen Shi, Qi Chen, Lei Sha, Sujian Li, Xu Sun, Houfeng Wang, and Lintao Zhang. 2018. Auto-dialabel: Labeling dialogue data with unsupervised learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 684–689, Brussels, Belgium. Association for Computational Linguistics.

Bo-Hsiang Tseng, Marek Rei, Paweł Budzianowski, Richard Turner, Bill Byrne, and Anna Korhonen. 2019. Semi-supervised bootstrapping of dialogue state trackers for task-oriented modelling. 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 1273–1278, Hong Kong, China. Association for Computational Linguistics.

Hanlei Zhang, Hua Xu, Ting-En Lin, and Rui Lyu. 2021. Discovering new intents with deep aligned clustering. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence,* *IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14365–14373. AAAI Press.

Rui Zhang, Yajing Sun, Jingyuan Yang, and Wei Peng. 2023. Knowledge-augmented frame semantic parsing with hybrid prompt-tuning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section Limitation behind conclusion

☑ A2. Did you discuss any potential risks of your work?
in the section Limitation

☑ A3. Do the abstract and introduction summarize the paper's main claims?
see in the section abstract and introduction

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

B ☑ Did you use or create scientific artifacts?

In the Section Experiments

☑ B1. Did you cite the creators of artifacts you used?
In the Section Experiments, set up subsection

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
In the Section Experiments, set up subsection

☑ 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)?
In the section experiments and section Limitation

☐ 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?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
In the Section Experiments, set up subsection

☑ 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.
In the Section Experiments, set up subsection

C ☑ Did you run computational experiments?

In the Section Experiments, training details subsection

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
In the Section Experiments, training details subsection

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?
In the Section Experiments, training details subsection

□ 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?
Not applicable. Left blank.

☑ 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.)?
In the Model Architecture Section

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

Left blank.

□ 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.