Winnie: Task-Oriented Dialog System with Structure-Aware Contrastive Learning and Enhanced Policy Planning

Kaizhi Gao, Tianyu Wang, Zhongjing Ma, Suli Zou*

School of Automation, Beijing Institute of Technology
gaokaizhi_0@126.com, wangtianyu1122@gmail.com, {mazhongjing, sulizou}@bit.edu.cn

Abstract

Pre-trained encoder-decoder models are widely applied in Task-Oriented Dialog (TOD) systems on the session level, mainly focusing on modeling the dialog semantic information. Dialogs imply structural information indicating the interaction among user utterances, belief states, database search results, system acts and responses, which is also crucial for TOD systems. In addition, for the system acts, additional pre-training and datasets are considered to improve their accuracies, undoubtedly introducing a burden. Therefore, a novel end-to-end TOD system named **Winnie** is proposed in this paper to improve the TOD performance. First, to make full use of the intrinsic structural information, **supervised contrastive learning** is adopted to narrow the gap in the representation space between text representations of the same category and enlarge the overall continuous representation margin between text representations of different categories in dialog context. Then, a **system act classification task** is introduced for policy optimization during fine-tuning. Empirical results show that Winnie substantially improves the performance of the TOD system. By introducing the supervised contrastive and system act classification losses, **Winnie** achieves state-of-the-art results on benchmark datasets, including **MultiWOZ2.2**, **In-Car**, and **Camrest676**. Their end-to-end combined scores are improved by **3.2**, **1.9**, and **1.1** points, respectively.

Introduction

The extensive adoption of intelligent customer services and personal assistants has sparked a growing fascination with constructing Task-Oriented Dialog (TOD) systems. These systems assist users in diverse tasks through natural language conversations, including tasks like table reservations, hotel bookings, and more. To effectively serve users, TOD systems need to possess the ability to comprehend user intentions, devise appropriate strategies, and generate responses that resemble human-like interactions.

Conventional TOD systems adopt a modular pipeline approach. The natural language understanding module identifies user intentions and extracts relevant information from utterances. Then, the dialogue state tracking (Zhang et al. 2020a) module monitors slot values and updates the belief state, capturing the entire context information. The belief state is used to query a task-related database to align with the user's objective, e.g., determining matching entities or booking availability. Based on the matched results, the dialogue policy module determines the next system action, guiding the system to generate an appropriate response. Finally, the natural language generation module converts the system action into a natural language response. However, this sequential approach has a significant drawback, as errors can propagate from one module to another, affecting subsequent subtasks.

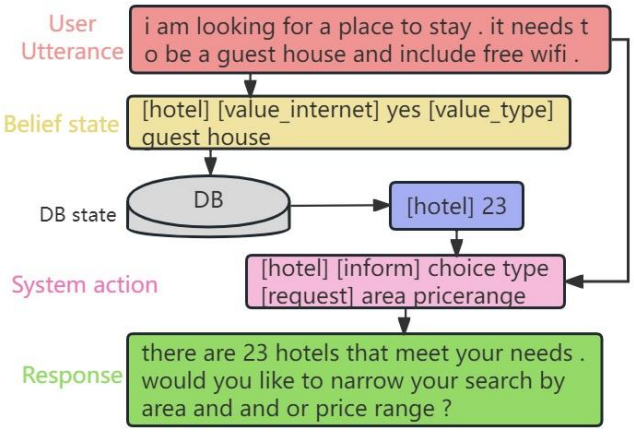

Figure 1: The first turn for Task-Oriented Dialog Systems.

To address this limitation, researchers are exploring module integration into a unified model, establishing an end-to-end neural architecture for TOD systems. Due to the significant improvement of Pre-trained Language Models(PLMs) (Yang et al. 2019) on various downstream tasks, researchers construct the end-to-end TOD system (Hosseini-Asl et al. 2020) by modeling a cascaded generation problem based on PLMs (Su et al. 2022). Through multi-task training, typical generative models, e.g., GPT and T5 (Raffel et al. 2019), can convert the whole dialog context into system responses. Nevertheless, prior research (Zhang et al. 2020c) indicates that inherent disparities in linguistic patterns exist between human conversations and regular texts. Simply fine-tuning PLMs are initially trained on plain texts. For downstream dialogue tasks, they hamper the model's ability to adequately

*Correponding author.

capture conversational patterns and linguistic knowledge, resulting in sub-optimal performance (Wu and Xiong 2020; Zeng and Nie 2021). As a result, to tackle this issue, researchers propose Pre-trained Conversation Models (PCMs) (Wu et al. 2020), which involve pre-trained vanilla PLMs on the extensive dialog corpora.

Notably, the TOD system is different from open-domain dialog generation (Bao et al. 2020), containing intermediate states in addition to user utterances and responses. We define the overall dialog task flow containing intermediate states as dialog structure. Currently, most works have shown that incorporating dialog structure and dialog policy into dialog generation can significantly improve responses quality, e.g., IEHSA (Wang et al. 2023) and GALAXY (He et al. 2021). However, the existing end-to-end TOD system does not take full advantage of the dialog structure and neglects the exploitation of the dialog policy. In addition, most existing datasets do not have intermediate states; the collection of corpus containing intermediate states for the pre-trained process of PCMs is difficult, and the pre-trained process is very time-consuming. Therefore, a novel end-to-end TOD system named **Winnie** is proposed in this paper. We only use the dialogue-Transformer layer to model dialog structural information as a residual to attach to the dialog semantic representation output by the dialog encoder, and we don't use additional training corpora to perform the pre-training process of PCMs.

Firstly, to comprehensively capture the structural information within the dialogue, a dialogue-level Transformer layer is incorporated to handle the long-range context dependencies between utterances. Each utterance's representation is captured by a pre-trained language model. In contrast to previous methods that solely employ PLMs as feature extractors, the dialogue-level Transformer layer is more adept at modeling dialog structural information and leveraging the unique characteristics of the dialog compared to plain text. Consequently, the model can extract more abundant contextual information from the dialog history.

Secondly, we incorporate supervised contrastive learning (SCL) (Khosla et al. 2020) to differentiate representations of distinct category utterances in the representation space. This approach fosters cohesion among utterances of the same category and ensures mutual exclusivity among those of different categories by fully leveraging dialog structural information. Compared to cross-entropy loss, the supervised contrastive loss improves training stability and enhances the model's generalization (Gunel et al. 2020). Additionally, for effective integration of the PLM and the dialogue-level Transformer layer, we introduce both the residual connection and max pooling. These measures contribute to optimizing the overall performance of the model.

Thirdly, an auxiliary system act classification task is introduced to boost the performance of policy planning for TOD. Compared with the end-to-end generation task, using a classification task for explicit policy injection can enable the model to capture more abundant system act annotation information and accelerate the model convergence.

Finally, we employ T5 (Raffel et al. 2019), a pre-trained Transformer with an encoder-decoder structure, as the backbone model and augment it using contrastive and classification losss. The proposed model, Winnie, achieves state-of-the-art results on three TOD datasets. Moreover, ablation experiments and case studies further validate the efficiency of the contrastive and classification losss on the TOD task.

In conclusion, the main contributions of our work can be summarized as follows:

- To the best of our knowledge, we are the first to incorporate SCL into the TOD system, resulting in a substantial enhancement of the model's capability to generate natural language responses.
- By incorporating system act classification as an auxiliary task, the model captures more abundant system act annotation information and further improves the performance of policy planning.
- Our model is straightforward to implement as it does not rely on any additional annotations or extensive datasets for pre-training.
- Empirical results show Winnie achieves state-of-the-art performance on the MultiWOZ 2.2, CamRest676, and In-car datasets.

Related Work

Pre-trained Language Models (PLMs) are trained with Transformer on extensive textual data, leading to a significant enhancement in dialog systems' performance. PLMs can be classified into two types based on the attention mechanism used during pre-training: uni-directional and bi-directional. For natural language understanding tasks such as classification or machine reading comprehension, BERT (Devlin et al. 2019) and Roberta (Liu et al. 2019) are pre-trained using a bi-directional transformer to capture rich contextual semantic meanings. In contrast, for natural language generation tasks, GPT and T5 (Raffel et al. 2019) utilize the uni-directional transformer decoder to optimize the likelihood of left-to-right generation. In recent times, the introduction of the unified language model UniLM (Dong et al. 2019) has allowed for both bi-directional and uni-directional attention with versatile self-attention mask schemes. In the context of TOD systems, all three aforementioned models are employed. TOD-BERT (Wu et al. 2020) adopts BERT as the backbone model, DialoGPT (Zhang et al. 2020c) utilizes GPT as the backbone model, and GALAXY (He et al. 2021) employs UniLM as the backbone model.

Pre-trained Conversation Models (PCMs) extend the pre-training of PLMs on dialogue corpora (Henderson et al. 2020; Mehri, Eric, and Hakkani-Tür 2020; Zhang et al. 2020c), aiming to narrow the disparity between regular texts and human dialogues. These models can be broadly categorized into two types: open-domain and task-oriented. The former entails training PLMs on conversational data from diverse sources collected from platforms such as Reddit or Twitter to facilitate dialog response generation. Notably, DialogGPT (Zhang et al. 2020c), an extension of GPT2, is pre-trained on 147M conversations from Reddit. As the same time, Blender (Roller et al. 2021) underwents pre-training on

1.5B dialogs in open-domain settings, showcasing impressive dialog response generation capabilities. Additionally, to address the challenge of generating multiple diverse responses for a single input in an open-domain dialog, PLATO (Bao et al. 2020) incorporats discrete latent variables, and the second line of PCM designs tailors models specifically for TOD tasks. For instance, in SC-GPT (Peng et al. 2020), natural language responses are generated with the assumption that both dialog acts and slot-tagging results are available. SOLOIST (Peng et al. 2021) utilizes a Transformer-based auto-regressive language model to create a task bot, integrating diverse dialog modules into a unified neural model, and performed pre-training on two TOD datasets. GALAXY (He et al. 2021) utilized a consistency regularization loss in a semi-supervised approach to training the dialog policy using both labeled and unlabeled dialog corpora. PPTOD (Su et al. 2022) utilized task-specific prompts on the T5 model (Raffel et al. 2019) to transform various TOD tasks into text-to-text generation tasks, enabling the incorporation of more diverse TOD corpora. In contrast to these methods, our proposed approach, Winnie, does not employ a separate pre-training phase. Instead, it solely adopts supervised contrastive loss and system act classification loss as additional optimization goals during fine-tuning.

Contrastive Learning seeks to minimize the difference between representations of two semantically similar utterances while maximizing the distinction between representations of dissimilar utterances. This approach can be classified into two groups depending on the label requirement. The first approach is un-supervised contrastive learning (un-SCL). For example, SimCLR (Chen et al. 2020) employs pairs of augmented images from the same original image as positive samples and images from different sources as negative samples, effectively optimizing the contrastive loss. Similarly, ConSERT (Yan et al. 2021) incorporates self-supervised contrast loss into the fine-tuning process of BERT. The second approach is SCL, which maximizes the utilization of supervised signals. Khosla et al. extends SCL with a self-supervised training approach, leading to the grouping of samples with the same label in the embedding space while pushing samples from different categories away from each other (Khosla et al. 2020). Due to the potential issues with model training instability and convergence to local optima when using cross-entropy loss, incorporating supervised contrastive loss during the fine-tuning stage leads to remarkable performance gains for the model in few-shot learning scenarios, as demonstrated by SCL (Gunel et al. 2020). SimCSE (Gao, Yao, and Chen 2021) uses entailment and contradicting pairs from the annotated NLI dataset as positive and negative samples in SCL. ConvFiit (Vulić et al. 2021) applies SCL for intent recognition tasks during fine-tuning, treating all samples within the same class as positive instances. In contrast, our approach involves adopting SCL during fine-tuning to model the dialog structural information comprehensively.

Method

In this section, a novel end-to-end TOD system, Winnie, is proposed. It consists of the dialog encoder, dialogue-level Transformer layer, belief state decoder, and system act/response decoder. Firstly, we introduce the data transmission process after the user utterances are input into the dialogue system. Then, we explain the two novel methods used in this paper to enhance the system's performance.

Dialog Model Based on Transformer

The proposed model adopts a sequence-to-sequence architecture, illustrated in Figure 2. In the initial turn, the encoder processes the user utterance. As the dialogue progresses to turn t, the encoder takes all previously generated outputs and the user utterances $\{\mathrm{turn}_0, ..., \mathrm{turn}_{\mathrm{t}-1}, \mathrm{U_t}\}$, where turn is $\{U, B, D, A, R\}$, U is user utterance, B is belief state, D is database search result, A is system act, and R is response. Here, we use the second turn as an example. After $\{\mathrm{turn}_0, \mathrm{U}_1\}$ is fed into the T5, the representation of the current context H_1 is acquired:

$$H_1 = \mathrm{T5Encoder}(\mathrm{turn}_0, \mathrm{U}_1) \tag{1}$$

where $H_1 \in \mathcal{R}^{s\times d}$, s represents the length of the sequence, and d indicates the hidden dimension.

The T5Encoder output, represented as H_1, is forwarded into the max-pooling ($\mathrm{MaxPooling}$) layer to obtain the condensed representation of the utterances after aggregation, as follows:

$$u_0, b_0, d_0, a_0, r_0, u_1 = \mathrm{MaxPooling}(H_1) \tag{2}$$

To acquire the context information of the historical dialog, we employ a Transformer-based dialogue-level encoder. The multi-head attention mechanism adeptly apprehends the interplay among diverse conversations and consolidates diverse features to yield the ultimate hidden representation. This approach thoroughly models the complex interdependence among various utterances and contextual connections. The multi-head attention score for all utterances in a context, specifically between two different utterances in a conversation represented as h_j, h_k, can be calculated using the subsequent equations:

$$\mathrm{Atten}(Q, K, V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{3}$$

$$\mathrm{head}_i = \mathrm{Atten}(h_j W_i^Q, h_k W_i^K, h_k W_i^V) \tag{4}$$

$$\mathrm{MultiHead}(Q, K, V) = [\mathrm{head}_1; ...; \mathrm{head}_n]W^O \tag{5}$$

where $W_i^Q \in \mathcal{R}^{d\times d_q}, W_i^K \in \mathcal{R}^{d\times d_k}, W_i^V \in \mathcal{R}^{d\times d_v}$, and the parameters $Q^O \in \mathcal{R}^{d\times d}$ can be optimized, where d_q, d_k, and d_v represent the dimensions of query, key, and value vectors, respectively. The variable n refers to the number of attention heads.

Hence, contextual utterance representation that captures contextual dependencies can be acquired using the previously mentioned dialog-level Transformer layer:

$$H_{\mathrm{context}} = [u_0; b_0; d_0; a_0; r_0; u_1] \tag{6}$$

$$H_{\mathrm{dcontext}} = \mathrm{DialogueTransformer}(H_{\mathrm{context}}) \tag{7}$$

where $H_{\mathrm{context}} \in \mathcal{R}^{cs\times d}$ indicates utterances in a conversation within the dialog context size cs, and $H_{\mathrm{dcontext}} \in \mathcal{R}^{cs\times d}$ denotes the utterances after context modeling.

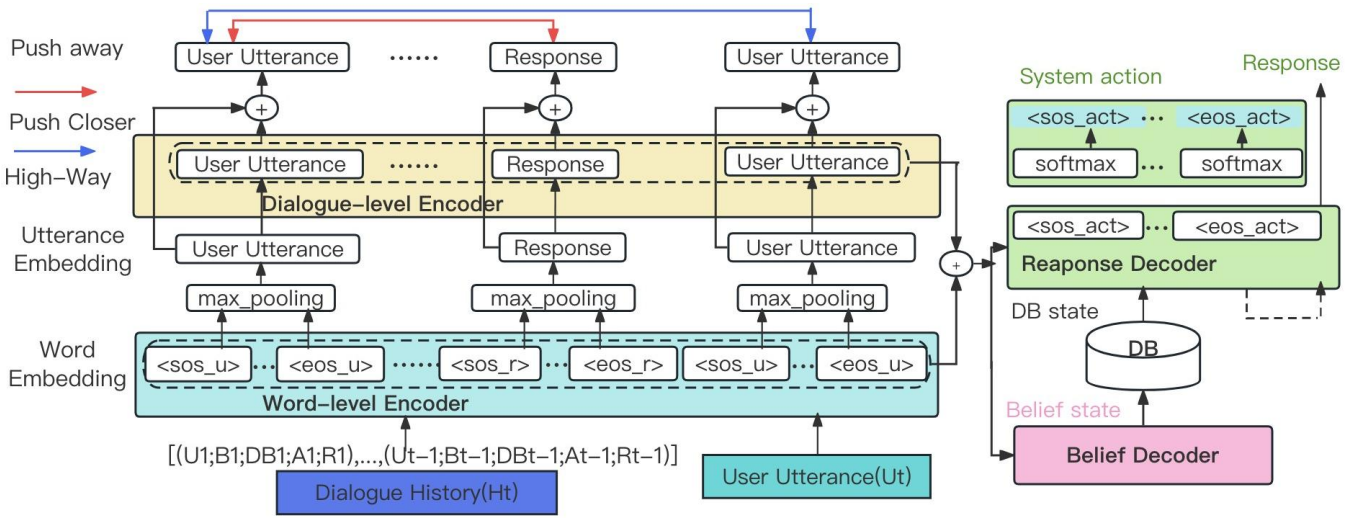

Figure 2: The proposed Winnie model consists of a dialog encoder, a dialogue-level Transformer layer, a belief decoder, and a response decoder. The representation of each utterance, acquired through max-pooling the hidden state, serves as input to the upper-level dialogue-level Transformer, capturing dialog structure information. These utterance representations, containing semantic and structural information, are subsequently used to decode the belief state, system act, and response.

Finally, the encoder hidden states with both dialog structural information and semantic information for the decoder can be obtained by the following formula:

$$I_{\text{dec}} = H_1 + \text{explode}(H_{\text{dcontext}}) \tag{8}$$

where explode denotes each utterance representation extended to the length of its sequence and aims to facilitate the addition of dialog semantic representation obtained by the dialog encoder. The loss functions, defined as Eqs (9) and (10), are applied for the belief and system action/response generation, respectively.

$$\mathcal{L}_{\text{belief}} = -logp(B_1|I_{\text{dec}}) \tag{9}$$

$$\mathcal{L}_{\text{resp}} = -logp(A_1, R_1|I_{\text{dec}}, DB_1) \tag{10}$$

Supervised Contrastive Learning

SCL assumes critical aspects receive attention and enhances the stability of few-shot learning when fine-tuning pre-trained models (Gunel et al. 2020). Conventional contrastive learning uses one pair of positive examples, contrasting them against all other samples as negatives. In contrast, supervised contrastive learning regards all examples with the same label in the batch as positive, effectively maximizing the utilization of supervisory signals.

Since we need to capture dialog structural information with SCL, and the model cannot forget the dialog semantic information obtained by the dialog encoder. Therefore, we introduce the residual connection to ensure that the dialogue-level Transformer layer only learns the residual information representing the dialog structure, which is a supplement to dialog semantic information obtained by the dialog encoder. The input I_{SCL} of SCL can be calculated by the following formula:

$$I_{\text{SCL}} = H_{\text{context}} + H_{\text{dcontext}} \tag{11}$$

For TOD, each dialog turn contains five kinds of information, i.e., user utterance, belief state, database search result, system act, and response. The number of samples in each category is highly balanced. Therefore, SCL is suitable for capturing dialog structural information. The SCL for all samples in a batch is represented by the following equations:

$$\mathcal{L}_{\text{SCL}} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \text{SIM}(p, i) \tag{12}$$

$$\text{SIM}(p, i) = log \frac{exp(X_i \cdot X_p / \tau)}{\sum_{a \in A(i)} exp(X_i \cdot X_a / \tau)} \tag{13}$$

where $X \in R^{\text{uttenum} \times d}, i \in I = \{1, 2, ..., \text{uttenum}\}$ indicates the index of the utterances in a batch ($\text{uttenum} = bs * cs$), and bs indicates batch size. $\tau \in R^+$ represents the temperature coefficient, controlling the distance between utterances, $P(i) = I_{j=i} - \{i\}$ corresponds to utterances with the same category as i excluding itself, and $A(i) = I - \{i\}$ denotes all utterances in the batch except itself.

System Act Classification

The response decoder initially generates the system action A_t, comprising domain, action type, and slot triples, followed by the natural language response R_t. It's important to note that the natural language responses are also conditioned on the generated system action, as the decoder generates tokens in an auto-regressive manner. The quality of system action has a significant impact on the result of natural language response. However, most previous methods focus on improving the quality of generating belief states and ignore the quality of system acts. We use the system act classification as an auxiliary task to explicitly inject dialog policy annotation information during the decoding process of the

response decoder. The probability distribution over all possible tokens for the i_{th} system act input token is as follows:

$$p_i = \text{softmax}(W \cdot s_i + b) \tag{14}$$

where s_i is the response decoder hidden state of the i_{th} system act input token, W and b are trainable weights and bias, respectively. We use the cross-entropy loss function for the system act classification, and the formula is as follows:

$$\mathcal{L}_{\text{SAC}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c} \cdot log\hat{y}_{i,c} \tag{15}$$

where N is the token number of tokenized system act, C is the size of vocab corresponding to the system act, $y_{i,c}$ represents the label of token i corresponding to category c, and $\hat{y}_{i,c}$ indicates the probability of token i corresponding to category c output by the dense layer.

Model Training

The model training loss comprises four components: belief state generation, response generation, supervised contrastive loss, and system act classification. The loss is a weighted sum of these four parts, as shown in the following formula:

$$\mathcal{L} = \mathcal{L}_{\text{belief}} + \mathcal{L}_{\text{resp}} + \alpha\mathcal{L}_{\text{SCL}} + \beta\mathcal{L}_{\text{SAC}} \tag{16}$$

In the experiments, we use α to represent the weight for supervised contrastive loss and β to denote the weight for system act classification loss. Specifically, we set α to 0.025 and β to 1.5.

Experimental Settings

This section provides details on the datasets, baseline, evaluation metrics, and parameter settings in the experiments.

Experimental Setup

Our experiments emphasize the end-to-end dialog modeling (E2E) setting, where the model does not receive ground-truth immediate labels. Winnie is initialized with T5-base (Raffel et al. 2019), following the settings of (Lee 2021). The optimizer applied for model training is AdamW with linear-scheduled warm-up strategy. For all datasets, the batch size is 8, the initial learning rate is 5e-4, the warm up ratio is 0.2, and the temperature of SCL is 15.0, 10.0, and 9.0 for MultiWOZ2.2, CamRest676, and In-car, respectively. We train all dialog systems on a single NVIDIA Geforce RTX 3090 and choose the checkpoint model with the best performance on the validation dataset. For MultiWOZ2.2 and In-car, we train 10 epochs. However, there is too few training data in CamRest676, we use the Hot-Start model trained on the MultiWOZ2.2 to further fine-tune 20 epochs on the training data of Camrest676.

Baselines

We have conducted a performance comparison between Winnie and all previous end-to-end TOD systems. Further, we used T5-base as another baseline to clearly show performance improvements brought by our proposed strategies.

Datasets

We assess the performance of end-to-end Winnie on three well-established task-oriented dialog benchmarks: MultiWOZ2.2 (Zang et al. 2020), CamRest676, and Stanford In-Car Assistant. MultiWOZ is a challenging dataset spanning seven domains with complex ontology and diverse language styles. For MultiWOZ2.2, we follow the data processing in (Yang, Li, and Quan 2020) and use 8438, 1000, and 1000 dialogs for training, validation, and testing, respectively. CamRest676 is a smaller English restaurant-domain dataset and split 406, 135, and 135 for training, validation, and testing. The In-Car dataset consists of conversations between users and an in-car assistant system, encompassing tasks like calendar scheduling, weather information retrieval, and point-of-interest navigation. To process the data, we follow the approach in (Zhang et al. 2020b), dividing it into training, validation, and testing sets with 2425, 302, and 304 dialogs, respectively. For a task-oriented generation, we use delexicalized responses to enable the model to learn value-independent parameters (Zhang, Ou, and Yu 2019).

Evaluation Metrics

For MultiWOZ2.2 (Zang et al. 2020), we employ automatic evaluation metrics to assess response quality and task completion:

- **Inform** evaluates the accuracy of entity information provided by the system.
- **Success** assesses the system's accuracy in providing a correct entity.
- **BLEU** (Papineni et al. 2002) is employed to assess the coherence of the generated replies.
- **Combined** is used as a comprehensive quality measure, as suggested in (Mehri, Srinivasan, and Eskenazi 2019), $Combined = (Inform + Success) \times 0.5 + BLUE$.

To compare with the existing methods, the **Success** is replaced by **SuccessF1** in CamRest676 and Stanford In-Car Assistant. The **Success** rate denotes whether the system answered all requested information or not to assess recall, while **SuccessF1** balances recall and precision.

Results and Analysis

Main Results

Table 1 contains the comprehensive inform rates, success rates, BLEU scores, and combined scores of end-to-end TOD models evaluated on the MultiWOZ2.2 benchmark. In Table 1, Baseline indicates the fine-tuning T5-base using only Language model loss, i.e., $\mathcal{L}_{belief}$ and $\mathcal{L}_{resp}$, which is comparable with BORT (Sun et al. 2022). Further, our proposed Winnie outperforms the baseline system by 3.5 combined scores (from 99.9 to 103.4). For end-to-end TOD modeling, MinTL (Lin et al. 2020), PPTOD (Su et al. 2022), BORT (Sun et al. 2022), and MTTOD (Lee 2021) also use T5 as their backbone models. Winnie outperforms the best performance of them, MTTOD (Lee 2021), with a 3.2 combined score (from 100.2 to 103.4). Furthermore, Winnie, without relying on additional corpora, significantly surpasses the previous state-of-the-art model, GALAXY (He

Model	Pre-trained	Extra corpora	Inform	Success	BLEU	Combined
DAMD (Zhang, Ou, and Yu 2019)	-	no	57.9	47.6	16.4	69.2
LABES (Zhang et al. 2020b)	-	no	68.5	58.1	18.9	82.2
AuGPT (Kulhánek et al. 2021)	GPT-2	yes	76.6	60.5	16.8	85.4
MinTL (Lin et al. 2020)	T5-small	no	73.7	65.4	19.4	89.0
SOLOIST (Peng et al. 2021)	GPT-2	yes	82.3	72.4	13.6	91.0
DoTS (Jeon and Lee 2021)	BERT-base	no	80.4	68.7	16.8	91.4
UBAR (Yang, Li, and Quan 2020)	DistilGPT2	no	83.4	70.3	17.6	94.5
PPTOD (Su et al. 2022)	T5-base	yes	83.1	72.7	18.2	96.1
BORT (Sun et al. 2022)	T5-small	no	85.5	77.4	17.9	99.4
MTTOD (Lee 2021)	T5-base	no	85.9	76.5	19.0	100.2
GALAXY (He et al. 2021)	UniLM-base	yes	85.4	75.7	**19.6**	100.2
Baseline	T5-base	no	85.7	75.5	19.3	99.9
Winnie	T5-base	no	**89.7**	**78.3**	19.4	**103.4**

Table 1: Comparison of end-to-end models evaluated on MultiWOZ 2.2. Previous work results are reported on the official MultiWOZ leaderboard.

Model	Match	F1	BLUE	Comb
Sequicity (Lei et al. 2018)	92.7	85.4	25.3	114.4
LABES (Zhang et al. 2020b)	96.4	82.3	**25.6**	115.0
SOLOIST (Peng et al. 2021)	94.7	87.1	25.5	116.4
GALAXY (He et al. 2021)	98.5	87.7	24.2	117.3
Winnie	**100.0**	**87.8**	24.4	**118.3**

Table 2: Comparison of end-to-end task-oriented dialog systems on Camrest676. The results of all previous works are from original papers. F1 is the abbreviation of Success F1.

Model	Match	F1	BLUE	Comb
SEDST (Jin et al. 2018)	84.5	82.9	19.3	103.0
Sequicity (Lei et al. 2018)	84.5	81.1	21.9	104.7
LABES (Zhang et al. 2020b)	85.8	77.0	22.8	92.8
FSDM (Shu et al. 2019)	84.8	82.1	21.5	105.0
GALAXY (He et al. 2021)	85.3	83.6	23.0	107.5
Winnie	**85.7**	**84.0**	**24.6**	**109.5**

Table 3: Comparison of end-to-end task-oriented dialog systems on In-car. The results of all previous works are from original papers. F1 is the abbreviation of Success F1.

T5-model	SAC	SCL	Match	Success	BLUE	Comb
w/	-	-	85.7	75.5	19.3	99.9
w/	w/	-	87.7	76.0	19.2	101.1
w/	-	w/	87.1	76.4	19.1	100.9
w/	w/	w/	**89.7**	**78.3**	**19.4**	**103.4**

Table 4: The performance of the various components in our proposed methods on MultiWOZ is evaluated. SCL represents the Supervised Contrastive Learning, and SAC represents the System Act Classification.

et al. 2021), which utilizes a large external corpus for pre-training, by 3.2 combined scores (from 100.2 to 103.4). This achievement establishes a new state-of-the-art performance in terms of inform rate, success rate, and combined score. Note that Winnie outperforms GLAXY by 4.3 and 2.6 on inform rate (from 85.4 to 89.7) and success rate (from 75.7 to 78.3), respectively, which means that Winnie exhibits superior performance in comprehending dialog context and strategizing dialog policy compared to other models, leading to enhanced task completion. Meanwhile, the results also prove that our proposed SCL and system act classification task are very effective for TOD tasks.

Table 2 and Table 3 present the detailed match rates, success F1s, BLEU scores, and combined scores for end-to-end TOD models on the CamRest676 and the Stanford In-Car Assistant benchmarks, respectively. Winnie outperforms the previous state-of-the-art model GALAXY (He et al. 2021) by 1.0 (from 117.3 to 118.3) and 2.0 (from 107.5 to 109.5) combined scores, establishing a new state-of-the-art performance in terms of match rates, success F1s, and combined scores on both datasets. Note that in Table 2, Winnie achieves 100.0 points on the match rate, which proves that our proposed SCL can effectively model the dialog structural information and further improve the understanding of user intentions. In Table 3, Winnie is the sole model capable of simultaneously achieving state-of-the-art performance in inform rate, success rate, BLEU score, and combined score.

Ablation Study

In this section, we present the ablation results of different Winnie components on MultiWOZ2.2 to explore the impacts of individual modules and combinations on the overall

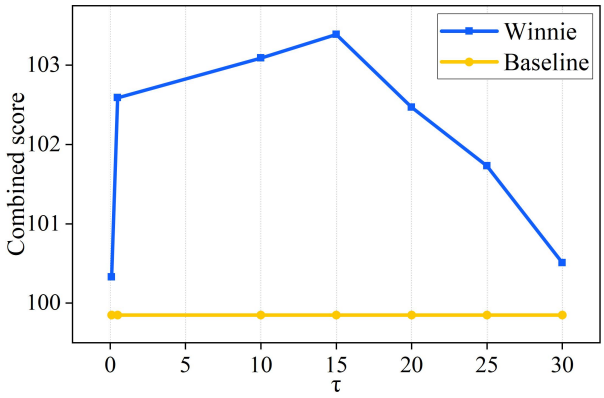

Figure 3: The our model performance with different levels of hyper-parameter τ on the MultiWOZ 2.2.

model performance. As illustrated in Table 4, where "w/" indicates the addition of the single method, "SCL" denotes the auxiliary task of supervised contrastive learning, and "SAC" means the auxiliary task of system act classification. Compared with the baseline, adding any of the two modules makes the overall performance better. SCL and SAC outperform the baseline by 1.0 and 1.2 on the combined score, respectively. In comparison to SCL, SAC exhibits superior performance enhancement capabilities. This superiority arises from the heightened influence of system act quality on response generation quality. Additionally, SAC tasks directly contribute to the enhancement of action state generation quality. Furthermore, integrating these two modules synergistically enhances the end-to-end dialog modeling performance.

Hyper-Parameter Analysis

We empirically scrutinize how the hyper-parameter τ for SCL of Winnie affects the performance of TOD on the MultiWOZ2.2. The value of τ controls the distance between two instances. With a larger τ, the margin between instances of different categories decreases. Conversely, a smaller τ widens the margin between these instances, indicating less similarity in the representation space across various categories. As shown in Figure 3, as the τ value increases from 0.01, the Combined score gradually increases. Obviously, when the τ is 15.0, the Combined score achieves the optimal peak. In Figure 3, with the decreasing of the τ, the instance representation becomes more and more dissimilar, which is of great help modeling the dialogue structural information at first. However, when the value is less than 15.0, the dialog structural representations obtained by the dialogue-level Transformer layer damage the dialog semantic representation obtained by the dialog encoder, leading to performance degradation. Therefore, in the experiment, we chose the optimal value of 15.0.

Case Study

To further analyze **Winnie**'s quality, two cases are presented in Figure 4 and Figure 5. In Figure 4, the baseline obtains

Dialog id: mul0374.json
System action (Sa):[restaurant][request] bookday Sa_Baseline: [restaurant] [inform] postcode phone [general] [reqmore] Sa_Winnie: [restaurant][request] day time
Response (Res):what day will you be dining ? Res_Baseline:the phone number is [value_phone] and the postcode is [value_postcode] . is there anything else i can help you with ? Res_Winnie:what day and time would you like the reservation for

Figure 4: Case1: comparison of baseline and Winnie response generation process, with the differences in red.

Dialog id: sng892.json
Belief State: [hotel] [value_type] guest house [value_pricerange] moderate [value_internet] yes [value_stars] 3[value_name] hamilton lodge Baseline: [hotel] [value_type] guest house [value_pricerange] moderate [value_internet] yes [value_stars] 3 Winnie: [hotel] [value_type] guest house [value_pricerange] moderate [value_internet] yes [value_stars] 3[value_name] hamilton lodge
DB state (DB):[db_1] DB_Baseline:[db_2] DB_Winnie:[db_1]

Figure 5: Case2: comparison of baseline and Winnie response generation process, with the differences in red.

incorrect database search results because the generated belief state lacks certain slot values. On the contrary, Winnie can generate the belief state correctly and get the correct database search results. In Figure 5, the baseline generates the wrong response due to the wrong system act, but Winnie can generate the system act correctly and finally generate correct responses. This further proves that by incorporating SCL and system act classification tasks, the model can understand the dialog context better, capture more abundant system action annotation information, and improve the quality of response generation in the end.

Conclusion

In this paper, we propose innovative additions to TOD systems called Winnie. It incorporates supervised contrastive learning to model dialog structural information, diminishing the distance within the same class and augmenting the gap between different classes. This enables the model to pay more attention to knowledge-dense information types from redundant contexts. Additionally, the system act classification task is used as an auxiliary task to ameliorate the performance of generated system acts by capturing more abundant system act annotation information. We conduct extensive experiments on three freely accessible datasets, showing that our model outperforms other methods with significant improvements. Ablation studies further show that supervised contrastive learning significantly enhances the model's natural language understanding and dialogue state tracking performance, consequently improving response generation performance. Also, system act classification as an auxiliary task can help the model generate system act more accurately.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant U22A2048 and Grant 62373051; in part by the Beijing Natural Science Foundation under Grant IS23063.

References

Bao, S.; He, H.; Wang, F.; Wu, H.; and Wang, H. 2020. PLATO: Pre-trained Dialogue Generation Model with Discrete Latent Variable. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 85–96. Online: Association for Computational Linguistics.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. E. 2020. A Simple Framework for Contrastive Learning of Visual Representations. *ArXiv*, abs/2002.05709.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *North American Chapter of the Association for Computational Linguistics*.

Dong, L.; Yang, N.; Wang, W.; Wei, F.; Liu, X.; Wang, Y.; Gao, J.; Zhou, M.; and Hon, H.-W. 2019. *Unified Language Model Pre-Training for Natural Language Understanding and Generation*. Red Hook, NY, USA: Curran Associates Inc.

Gao, T.; Yao, X.; and Chen, D. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 6894–6910. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Gunel, B.; Du, J.; Conneau, A.; and Stoyanov, V. 2020. Supervised Contrastive Learning for Pre-trained Language Model Fine-tuning. *ArXiv*, abs/2011.01403.

He, W.; Dai, Y.; Zheng, Y.; Wu, Y.; Cao, Z.; Liu, D.; Jiang, P.; Yang, M.; Huang, F.; Si, L.; Sun, J.; and Li, Y. 2021. GALAXY: A Generative Pre-trained Model for Task-Oriented Dialog with Semi-Supervised Learning and Explicit Policy Injection. In *AAAI Conference on Artificial Intelligence*.

Henderson, M.; Casanueva, I.; Mrkšić, N.; Su, P.-H.; Wen, T.-H.; and Vulić, I. 2020. ConveRT: Efficient and Accurate Conversational Representations from Transformers. In Cohn, T.; He, Y.; and Liu, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2161–2174. Online: Association for Computational Linguistics.

Hosseini-Asl, E.; McCann, B.; Wu, C.-S.; Yavuz, S.; and Socher, R. 2020. A Simple Language Model for Task-Oriented Dialogue. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 20179–20191. Curran Associates, Inc.

Jeon, H.; and Lee, G. G. 2021. Domain State Tracking for a Simplified Dialogue System. *ArXiv*, abs/2103.06648.

Jin, X.; Lei, W.; Ren, Z.; Chen, H.; Liang, S.; Zhao, Y. E.; and Yin, D. 2018. Explicit State Tracking with Semi-Supervisionfor Neural Dialogue Generation. *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised Contrastive Learning. *ArXiv*, abs/2004.11362.

Kulhánek, J.; Hudeček, V.; Nekvinda, T.; and Dušek, O. 2021. AuGPT: Auxiliary Tasks and Data Augmentation for End-To-End Dialogue with Pre-Trained Language Models. In Papangelis, A.; Budzianowski, P.; Liu, B.; Nouri, E.; Rastogi, A.; and Chen, Y.-N., eds., *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, 198–210. Online: Association for Computational Linguistics.

Lee, Y. 2021. Improving End-to-End Task-Oriented Dialog System with A Simple Auxiliary Task. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Findings of the Association for Computational Linguistics: EMNLP 2021*, 1296–1303. Punta Cana, Dominican Republic: Association for Computational Linguistics.

Lei, W.; Jin, X.; Kan, M.-Y.; Ren, Z.; He, X.; and Yin, D. 2018. Sequicity: Simplifying Task-oriented Dialogue Systems with Single Sequence-to-Sequence Architectures. In Gurevych, I.; and Miyao, Y., eds., *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1437–1447. Melbourne, Australia: Association for Computational Linguistics.

Lin, Z.; Madotto, A.; Winata, G. I.; and Fung, P. 2020. MinTL: Minimalist Transfer Learning for Task-Oriented Dialogue Systems. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3391–3405. Online: Association for Computational Linguistics.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv*, abs/1907.11692.

Mehri, S.; Eric, M.; and Hakkani-Tür, D. Z. 2020. DialoGLUE: A Natural Language Understanding Benchmark for Task-Oriented Dialogue. *ArXiv*, abs/2009.13570.

Mehri, S.; Srinivasan, T.; and Eskenazi, M. 2019. Structured Fusion Networks for Dialog. In Nakamura, S.; Gasic, M.; Zuckerman, I.; Skantze, G.; Nakano, M.; Papangelis, A.; Ultes, S.; and Yoshino, K., eds., *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, 165–177. Stockholm, Sweden: Association for Computational Linguistics.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Annual Meeting of the Association for Computational Linguistics*.

Peng, B.; Li, C.; Li, J.; Shayandeh, S.; Liden, L.; and Gao, J. 2021. Soloist: Building Task Bots at Scale with Transfer Learning and Machine Teaching. *Transactions of the Association for Computational Linguistics*, 9: 807–824.

Peng, B.; Zhu, C.; Li, C.; Li, X.; Li, J.; Zeng, M.; and Gao, J. 2020. Few-shot Natural Language Generation for Task-Oriented Dialog. *ArXiv*, abs/2002.12328.

Raffel, C.; Shazeer, N. M.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.*, 21: 140:1–140:67.

Roller, S.; Dinan, E.; Goyal, N.; Ju, D.; Williamson, M.; Liu, Y.; Xu, J.; Ott, M.; Smith, E. M.; Boureau, Y.-L.; and Weston, J. 2021. Recipes for Building an Open-Domain Chatbot. In Merlo, P.; Tiedemann, J.; and Tsarfaty, R., eds., *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 300–325. Online: Association for Computational Linguistics.

Shu, L.; Molino, P.; Namazifar, M.; Xu, H.; Liu, B.; Zheng, H.; and Tur, G. 2019. Flexibly-Structured Model for Task-Oriented Dialogues. In Nakamura, S.; Gasic, M.; Zuckerman, I.; Skantze, G.; Nakano, M.; Papangelis, A.; Ultes, S.; and Yoshino, K., eds., *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, 178–187. Stockholm, Sweden: Association for Computational Linguistics.

Su, Y.; Shu, L.; Mansimov, E.; Gupta, A.; Cai, D.; Lai, Y.-A.; and Zhang, Y. 2022. Multi-Task Pre-Training for Plug-and-Play Task-Oriented Dialogue System. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4661–4676. Dublin, Ireland: Association for Computational Linguistics.

Sun, H.; Bao, J.; Wu, Y.; and He, X. 2022. BORT: Back and Denoising Reconstruction for End-to-End Task-Oriented Dialog. In Carpuat, M.; de Marneffe, M.-C.; and Meza Ruiz, I. V., eds., *Findings of the Association for Computational Linguistics: NAACL 2022*, 2156–2170. Seattle, United States: Association for Computational Linguistics.

Vulić, I.; Su, P.-H.; Coope, S.; Gerz, D.; Budzianowski, P.; Casanueva, I.; Mrkšić, N.; and Wen, T.-H. 2021. ConvFiT: Conversational Fine-Tuning of Pretrained Language Models. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 1151–1168. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Wang, J.; Sun, X.; Chen, Q.; and Wang, M. 2023. Information-Enhanced Hierarchical Self-Attention Network for Multiturn Dialog Generation. *IEEE Transactions on Computational Social Systems*, 10(5): 2686–2697.

Wu, C.-S.; Hoi, S. C.; Socher, R.; and Xiong, C. 2020. TOD-BERT: Pre-trained Natural Language Understanding for Task-Oriented Dialogue. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 917–929. Online: Association for Computational Linguistics.

Wu, C.-S.; and Xiong, C. 2020. Probing Task-Oriented Dialogue Representation from Language Models. In *Conference on Empirical Methods in Natural Language Processing*.

Yan, Y.; Li, R.; Wang, S.; Zhang, F.; Wu, W.; and Xu, W. 2021. ConSERT: A Contrastive Framework for Self-Supervised Sentence Representation Transfer. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *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)*, 5065–5075. Online: Association for Computational Linguistics.

Yang, Y.; Li, Y.; and Quan, X. 2020. UBAR: Towards Fully End-to-End Task-Oriented Dialog Systems with GPT-2. *ArXiv*, abs/2012.03539.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.; and Le, Q. V. 2019. *XLNet: Generalized Autoregressive Pretraining for Language Understanding*. Red Hook, NY, USA: Curran Associates Inc.

Zang, X.; Rastogi, A.; Sunkara, S.; Gupta, R.; Zhang, J.; and Chen, J. 2020. MultiWOZ 2.2 : A Dialogue Dataset with Additional Annotation Corrections and State Tracking Baselines. In Wen, T.-H.; Celikyilmaz, A.; Yu, Z.; Papangelis, A.; Eric, M.; Kumar, A.; Casanueva, I.; and Shah, R., eds., *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, 109–117. Online: Association for Computational Linguistics.

Zeng, Y.; and Nie, J.-Y. 2021. An Investigation of Suitability of Pre-Trained Language Models for Dialogue Generation – Avoiding Discrepancies. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 4481–4494. Online: Association for Computational Linguistics.

Zhang, J.; Hashimoto, K.; Wu, C.-S.; Wang, Y.; Yu, P.; Socher, R.; and Xiong, C. 2020a. Find or Classify? Dual Strategy for Slot-Value Predictions on Multi-Domain Dialog State Tracking. In Gurevych, I.; Apidianaki, M.; and Faruqui, M., eds., *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, 154–167. Barcelona, Spain (Online): Association for Computational Linguistics.

Zhang, Y.; Ou, Z.; Hu, M.; and Feng, J. 2020b. A Probabilistic End-To-End Task-Oriented Dialog Model with Latent Belief States towards Semi-Supervised Learning. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 9207–9219. Online: Association for Computational Linguistics.

Zhang, Y.; Ou, Z.; and Yu, Z. 2019. Task-Oriented Dialog Systems that Consider Multiple Appropriate Responses under the Same Context. *ArXiv*, abs/1911.10484.

Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.-C.; Brockett, C.; Gao, X.; Gao, J.; Liu, J.; and Dolan, B. 2020c. DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation. In Celikyilmaz, A.; and Wen, T.-H., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 270–278. Online: Association for Computational Linguistics.