	ConvAI2				Reddit			
Model	**R@1**	**R@3**	**R@5**	**P@1**	**R@1**	**R@3**	**R@5**	**P@1**
Random	1.03±0.09	2.99±0.12	4.83±0.04	1.18±0.12	0.60±0.06	1.88±0.24	3.35±0.34	0.69±0.04
PMI	16.96	34.15	46.39	19.11	6.90	16.06	22.98	7.79
Neural	17.81±0.35	34.59±0.42	44.88±0.66	19.91±0.57	7.22±0.26	16.81±0.20	23.89±0.21	8.12±0.35
Kernel	16.23±0.50	32.07±0.84	42.62±0.76	17.57±0.87	7.38±0.17	17.10±0.28	24.81±0.70	8.24±0.22
DKRN	18.03±0.15	34.60±0.56	45.06±0.95	20.09±0.38	7.11±0.21	16.47±0.72	23.42±0.98	8.08±0.29
Ours (CKC)	**19.31**±0.44	**36.26**±0.45	**46.32**±0.57	**21.98**±0.66	**8.23**±0.31	**17.83**±0.25	**24.89**±0.12	**9.17**±0.28

Table 2: Test results (in %) for next-turn keyword prediction. Results are averaged over 3 random seeds.

	ConvAI2				Reddit			
Model	**R@1**	**R@3**	**R@5**	**MRR**	**R@1**	**R@3**	**R@5**	**MRR**
PMI	48.67±0.25	75.88±0.49	86.38±0.15	64.74±0.26	45.31±0.70	68.93±0.37	79.75±0.46	60.42±0.50
Neural	47.93±0.47	75.53±0.62	86.36±0.20	64.25±0.38	44.96±0.21	68.75±0.27	79.59±0.23	60.18±0.22
Kernel	48.55±0.51	75.57±0.32	86.04±0.04	64.47±0.37	44.55±0.33	68.47±0.24	79.66±0.38	59.92±0.30
DKRN	48.44±0.34	75.78±0.20	86.83±0.16	64.64±0.17	44.92±0.45	68.84±0.45	79.59±0.65	60.19±0.44
Ours (CKC)	**59.90**±0.41	**83.03**±0.31	**92.15**±0.17	**73.50**±0.26	**50.02**±0.41	**72.94**±0.33	**82.87**±0.22	**64.33**±0.35

Table 3: Test results (in %) for keyword-augmented response retrieval. Results are averaged over 3 random seeds.

	ConvAI2		Reddit	
Model	**Succ. (%)**	**#Turns**	**Succ. (%)**	**#Turns**
PMI	14.6	5.83	5.1	4.88
Neural	18.9	6.07	11.1	5.99
Kernel	20.7	5.89	10.6	5.83
DKRN	25.6	4.54	18.4	4.42
Ours (CKC)	**28.9**	**4.23**	**22.7**	**4.19**

Table 4: Self-play simulation results.

	ConvAI2		Reddit	
Model	**Succ. (%)**	**Smo.**	**Succ. (%)**	**Smo.**
PMI	16.0	3.05	6.3	2.68
Neural	17.3	2.77	11.0	2.85
Kernel	22.3	2.88	12.3	2.57
DKRN	25.0	3.01	17.7	2.81
Ours (CKC)	**29.3**	**3.27**	**22.3**	**3.08**

Table 5: Human evaluation results. Smo. denotes conversation smoothness.

Baselines and Model Settings

We compare our model with the following baselines: PMI (**?**), Neural (**?**), Kernel (**?**) and DKRN (**?**). We follow their released implementations[9]. All baselines are trained and evaluated using the same filtered datasets as our model.

We initialize the embedding layer of all models using GloVe embedding of size 200 (**?**). All hidden sizes in GRU and GGNN are set to 200. We use one layer in GGNN and set $\lambda_k = 0.01$. We optimize our model using Adam (**?**) with batch size of 32, an initial learning rate of 0.001 and a decay rate of 0.9 for every epoch.

Result Analysis

In this section, we present the experimental results, model analysis, case study and limitations.

Next-Turn Keyword Prediction

The results for next-turn keyword prediction are presented in Table 2. Among all baselines except Random, the non-parameterized PMI performs worst, and Neural, Kernel and DKRN performs comparably on both datasets. Our proposed model achieves consistent better performance than all baselines across all metrics and datasets, suggesting that incorporating CKG triplets into keyword prediction helps.

[9]We fixed a bug in DKRN where the keyword transition mask is obtained using train+valid+test datasets.

Next-Turn Keyword Prediction		
Model	**R@1**	
Ours (CKC)	**19.31**±0.44	
- concepts	18.56±0.31	
Keyword-Augmented Response Retrieval		
Model	**R@1**	
Ours (CKC)	**59.90**±0.41	
- concepts	53.11±0.43	
- keywords	52.30±0.54	
Self-Play Simulation		
Model	**Succ.** (%)	**#Turns**
Ours (CKC)	**28.9**	**4.23**
- CKG-based strategy	22.3	4.42

Table 6: Ablation study (in %) on ConvAI2.

Keyword-Augmented Response Retrieval

The results for keyword-augmented response retrieval are presented in Table 3. The baselines differ in which next-turn keyword prediction model is used. It is surprising that all baselines perform comparably regardless of the next-turn keyword prediction model. This may suggest that the baselines are unable to effectively leverage the predicted keyword information into response retrieval. Our model achieves substantially better performance than all baselines on both datasets. The performance improvement can be primarily attributed to 1) we additionally incorporate utterance-

Target: **music**
A: Hey, how are you doing?
H: I'm well, thanks. Working on a **party** I'm planning.
A: I am sitting here listening to **pearl jam**, my favorite **band**.
H: Super cool! Do you **sing**? I was just **singing** in my **shower**.
A: No, but I was in **jazz band** in hs.
H: Congrats! I love **music** and playing my **guitar** and **violin**.
A: That's awesome! However, my favorite is **country music**.

Table 7: Case study from self-play simulations on ConvAI2. *A* denotes our model and *H* denotes the base model.

related CKG triplets into utterance representation learning; and 2) we propose an additional keyword matching module to match the predicted keywords with candidate keywords, whereas baselines directly match predicted keywords with candidate utterances.

Keyword-Guided Conversation

The self-play simulation results for keyword-guided conversation are presented in Table 4. DKRN performs best among all baselines, which can be primarily attributed to its strategy of selecting keyword-related responses. This strategy requires a pool of confident candidates to select from. A larger pool will lead to higher success rate but lower smoothness because potentially less likely candidates can be selected. In all experiments, we set the pool size to 100. Our model also leverages this strategy but instead use weighted path lengths to measure keyword relatedness. Our model outperforms all baselines in both metrics on both datasets. Note that the success rates on ConvAI2 are consistently larger than that on Reddit across all models, which can be partially due to the higher next-turn keyword prediction accuracy on ConvAI2. The human evaluation results are presented in Table 5. The results for success rate are similar to that in self-play simulations. Among all baselines, DKRN has slightly more robust performance in smoothness on both datasets. Our model obtains consistently better performance in both success rate and smoothness on both datasets, suggesting that our model can select confident candidates that are also related to the target keyword.

Model Analysis

Table 6 presents the ablation study of our model across multiple tasks on the ConvAI2 test set. In both next-turn keyword prediction and keyword-augmented response retrieval, removing concepts representation from our model leads to degraded performance in R@1, suggesting that CKG triplets are helpful in learning the semantic representation of utterances. In keyword-augmented response retrieval, unlike other baselines that do not leverage keywords effectively, our model performs noticeably worse when keywords are removed, showing that our design of matching keywords separately indeed contribute to the overall matching. Finally, we examine the impact of our CKG-guided keyword selection strategy on self-play simulations. The results in Table 6 show that replacing our CKG-based strategy by the embedding-based strategy (**??**) leads to worse performance in both success rate and number of turns.

Case Study

We present a case study from our self-play simulations in Table 7. Our model can lead the conversation from a starting keyword "party" to the target keyword "music" smoothly and fast.

Limitations

One major limitation of existing approaches including ours is the mediocre accuracy of retrieving keyword-related responses (this is different from keyword-augmented response retrieval where the ground-truth responses do not necessarily correlate with the input keywords), which bottlenecks the overall target success rate. In fact, for both DKRN and our model, the target keyword can be successfully selected most of the time during self-play simulations, however, both models can not retrieve the keyword-related responses given the selected target keyword accurately. A potential solution to this problem is to train the keyword-augmented response retrieval model on datasets where input keywords and ground-truth responses are correlated, which is left to future work.

Conclusion

We study the problem of imposing conversational goals/keywords on open-domain conversational agents. The keyword transition module in existing approaches suffer from noisy datasets and unreliable transition strategy. In this paper, we propose to ground keyword transitions on commonsense and propose two GNN-based models for the tasks of next-turn keyword transition and keyword-augmented response retrieval, respectively. Extensive experiments show that our proposed model obtains substantially better performance on these two tasks than competitive baselines. In addition, the model analysis suggests that CKG triplets and our proposed CKG-guided keyword selection strategy are helpful in learning utterance representation and keyword transition, respectively. Finally, both self-play simulations and human evaluations show that our model can achieve better success rate, reach the target keyword faster, and produce smoother conversations than baselines.

Acknowledgments

This research is supported, in part, by Alibaba Group through Alibaba Innovative Research (AIR) Program and Alibaba-NTU Singapore Joint Research Institute (JRI) (Alibaba-NTU-AIR2019B1), Nanyang Technological University, Singapore. This research is also supported, in part, by the National Research Foundation, Prime Minister's Office, Singapore under its AI Singapore Programme (AISG Award No: AISG-GC-2019-003) and under its NRF Investigatorship Programme (NRFI Award No. NRF-NRFI05-2019-0002). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of National Research Foundation, Singapore. This research is also supported, in part, by the Singapore Ministry of Health under its National Innovation Challenge on Active and Confident Ageing (NIC Project No. MOH/NIC/COG04/2017 and MOH/NIC/HAIG03/2017). Assumenda tempora doloremque aliquid, vel