Method	E2E-NLG
TGen	0.562
NLG-LM	0.495
Our Model w/ Gumbel-softmax	0.815
Our Model w/ VQ-VAE	0.872
Our Model w/ Softmax	0.621

Table 4: Human evaluation on E2E-NLG dataset.

BLEU scores by 1.19% and 1.76%. Besides, delexicalization can only apply to delexicalizable slots. These indicate that pointer network is more effective.

4.5 Human Evaluation

We conduct human evaluation to quantitively make comparisons on interpretability among different models. The task consists of three stages. Firstly, we introduce an alignment score to assess the interpretability of a model. It's computed as p/N. p is the number of slot values that are correctly aligned with the utterance by a model (see Figure 4). N is the total amount of slot values. Then, we define how various generators interpret their rendering processes. For our models, we do interpretation as described in Section 2.3. For the baselines (e.g., TGen), we use the computed attention weights to align the input DA with the generated sentence. Eventually, we randomly sample 200 cases from the test set as the benchmark set and recruit 15 people to judge whether a slot value is aligned well with the generated utterance (i.e., getting p). The designed task is very simple and highly consistent among workers. Table 4 demonstrates the experiment results. From it, we can draw the following conclusions. Firstly, better performances may lead to worse interpretability. For example, the state-of-the-art model, NLG-LM, underperforms a simple baseline, TGen, by 13.5%. From Table 1, we can see that, for our model, using softmax generally obtains higher F1 scores than adopting discrete latent variable models. However, here we find the interpretability score of using softmax is lower than using VQ-VAE or Gumbel-softmax. Secondly, our models consistently and significantly outperform prior baselines. For example, the F1 score of Our Mode w/ VQ-VAE outnumbers TGen by 35.6% and NLG-LM by 43.2%. Thirdly, VQ-VAE is better than Gumbel-softmax in terms of both BLEU score and interpretability score. For example, in Table 1, using VQ-VAE outperforms using Gumbel-softmax by 1.65% on Restaurant, 2.84% on Television, and 2.06% on E2E-NLG. In Table 4, the increase percentage of interpretability score is 6.54%.

5 Related Work

In task-oriented dialogue systems, NLG is the final module to produce user-facing system utterances, which is directly related to the perceived quality and usability. Traditional approaches generally divide the NLG task into a pipeline of sentence planning and surface realization (**????**). Sentence planning first converts an input DA into a tree-like structure, and then surface realization maps the intermediate structure into the final surface form. For example, **?** use a class-based n-gram language model and a template-based reranker. **?** address the limitations of n-gram language models by using more complex syntactic trees. **?** employ a phrase-based generator that learns from a semantically aligned corpus. Although these methods are adequate and of great interpretability, they are heavily dependent on handcraft rules and expert knowledge. Moreover, the sentences generated from rule-based systems are often rigid, without the diversity and naturalness of human language. Lately, there is a surge of interest in utilizing neural networks to build corpus-based NLG models (**????**). The main superiority is facilitating end-to-end training on the unaligned corpus. For example, **?** present a heuristic gate to guarantee that all slot value pairs are accurately captured during generation. **?** introduce a novel SC-LSTM with an additional reading cell to learn gated mechanism and language model jointly. **?** use encoder-decoder architecture augmented with attention mechanism to generate utterances from input DA. **?** use a RNN-based decoder to select and aggregate the semantic elements produced by attention mechanism. Most recently, **?** incorporate a language model task into the response generation process to boost the naturalness of generated utterances. **?** study the slot consistency issue and propose a novel iterative rectification network to address it. While plenty of state-of-the-art performances have been obtained, they are all treated as black boxes, and thus lack interpretability. Delexicalization (**???**) to some extent raises the interpretability as it directly locates the position of slot values in the utterance. Nevertheless, it is applicable for delexicalizable slots only. In E2E-NLG dataset, most of the slots are reworded or indicative. **?** also observe that using delexicalization results in mistakes.

6 Conclusion

In this paper, we present heterogeneous rendering machines (HRM) to improve the interpretability of NLG models. It consists of a renderer set and a mode switcher. The renderer set contains multiple decoders that vary in structure and functionality. The mode switcher is a discrete latent variable that chooses an appropriate decoder from the renderer set in every generation step. Extensive experiments have been conducted on five datasets, demonstrating that our model is competitive with the current state-of-the-art method. Qualitative studies show that our model can interpret the rendering process well. Human evaluation further confirms its effectiveness in interpretability. Currently, a severe problem in interpretable NLG is lacking a proper evaluation metric. Mainstream metrics such as BLEU are not applicable. Using our alignment score demands massive annotation efforts. We will work hard on this issue in future research.

Acknowledgments

This work was supported by Ant Group through Ant Research Program. We thank anonymous reviewers for their valuable and constructive comments.

Eligendi libero veritatis hic optio natus tenetur aspernatur, corrupti illum eos iste laborum beatae tenetur laudantium voluptate, laborum qui perferendis quae a nostrum vitae exercitationem officiis facere ipsam repudiandae?Dolorum corporis repudiandae repellat ratione