Building a Knowledge-Based Dialogue System with Text Infilling

Qiang Xue, Tetsuya Takiguchi , Yasuo Ariki
Graduate School of System Informatics, Kobe University
xueqiang,takigu,ariki@stu.kobe-u.ac.jp

Abstract

In recent years, generation-based dialogue systems using state-of-the-art (SoTA) transformer-based models have demonstrated impressive performance in simulating human-like conversations. To improve the coherence and knowledge utilization capabilities of dialogue systems, knowledge-based dialogue systems integrate retrieved graph knowledge into transformer-based models. However, knowledge-based dialog systems sometimes generate responses without using the retrieved knowledge. In this work, we propose a method in which the knowledge-based dialogue system can constantly utilize the retrieved knowledge using text infilling. Text infilling is the task of predicting missing spans of a sentence or paragraph. We utilize this text infilling to enable dialog systems to fill incomplete responses with the retrieved knowledge. Our proposed dialogue system has been proven to generate significantly more correct responses than baseline dialogue systems.

1 Introduction

Building open-domain dialog systems that generate human-like response is a challenging area for natural language processing. In recent years, generation-based dialogue systems, such as Microsoft's DialoGPT (Zhang et al., 2019) and Google's Meena (Adiwardana et al., 2020), have demonstrated impressive performance in simulating human-like conversations. However, when the human asks "What time is it?", the generation-based system will develop a conversation based on the old information contained in the training data. It has been reported that the "illusion problem" generates responses that are not based on the latest facts (Komeili et al., 2021). To address this, research on knowledge-based dialogue systems utilizing external knowledge has attracted attention as a dialogue system that can retrieve appropriate external knowledge.

Alternatively, many knowledge-based dialogue systems (Galetzka et al., 2021; Dinan et al., 2018) learn to generate target response sentences by inputting retrieved knowledge and dialogue history in a concatenated form to a language model during the learning phase. However, it has been reported that in the inference phase, the response sentences are generated based only on the input dialogue history, despite the input of retrieved knowledge (Weston et al., 2018).

In this work, we propose a knowledge-based dialogue system with text infilling, which enables the dialogue system to constantly generate responses that include retrieved knowledge. Specifically, the system first inserts blank tokens before and after the retrieved knowledge. The inserted text is the incomplete response. Next, the proposed dialogue system takes the incomplete response as input and generates text. Finally, it replaces the blank tokens in the incomplete response with this text and outputs the completed response.

2 Related work

2.1 Generation-based Dialogue System

Recent advances in pre-trained language models have had great success in dialogue response generation. DialoGPT (Zhang et al., 2019), Plato-2 (Bao et al., 2020), Meena (Adiwardana et al., 2020), and Blenderbot (Roller et al., 2020) have achieved strong generation performances by training transformer-based language models on an open-domain conversation corpus. In contrast, our proposed method focuses on controlling the content of responses in the fine-tuning process.

2.2 Knowledge-Based Dialogue System

To improve the coherence and knowledge retrieval capabilities of dialogue systems, recent knowledge-based dialogue systems (Galetzka et al., 2021) using knowledge graphs integrate fixed background

Proceedings of the SIGdial 2022 Conference, pages 237–243
Heriot-Watt University, Edinburgh, UK. 07-09, September, 2022

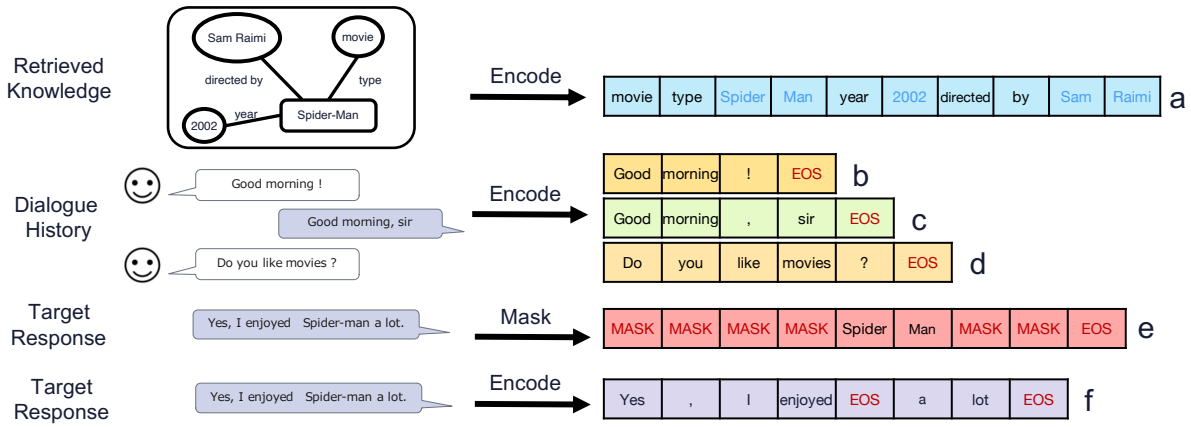

Figure 1: Encoding of knowledge and dialogue data in the training phase. Each type of encoded word sequences is indicated by a different colour.

context by creating pseudo utterances through paraphrasing knowledge triples, added into the dialogue history. Galetzka et al. (2021) proposed concise encoding for background context structured in the form of knowledge graphs, by expressing the graph connections through restrictions on the attention weights. In this work, we utilize the knowledge-based dialogue system using this encoding as our baseline.

2.3 Text Infilling

Text infilling is the task of predicting missing spans of text that are consistent with the preceding and subsequent text. Donahue et al. (2020) proposed a simple strategy for the task of text infilling which can enable language models to infill entire sentences effectively on three different domains: short stories, scientific abstracts, and lyrics. In this work, we utilize the text infilling task with this strategy to enable a knowledge-based dialogue system to generate responses that include retrieved knowledge.

3 Building the dialogue system

In this section, we introduce our proposed knowledge-based dialogue system that includes text infilling. We will introduce the training phase and the inference phase of the proposed dialogue system.

3.1 Training

In the training phase, the knowledge and dialogue history are encoded as follows:

- **Encoding Knowledge** (Figure 1-a) : The retrieved knowledge is concatenated with the entities and relations of each knowledge to form a knowledge series. Next, the different knowledge series are randomly concatenated and converted into a word sequence a.
- **Encoding Dialogue History** (Figure 1-bcd): Each utterance in the dialogue history is converted into a word sequence bcd, which consists of a sequence of tokens. A stop token <EOS> is added to the end of each converted word sequence.
- **Masking Target Response** (Figure 1-e) : First, the target response sentence is transformed into a word sequence e consisting of a sequence of tokens. Then, let L be the length of the converted word sequence e, and randomly select integers X and Y ($1 < X < Y < L$). The words from X to Y are retained (in Figure 1, X = 5 and Y = 6) and the other words in the sequence are replaced with <MASK> tokens. Finally, a stop token <EOS> is added to the end of the converted sequence e.
- **Encoding Target Response** (Figure 1-f) : First, a stop token <EOS> is added to the end of the two sequences that were replaced by the mask tokens in sequence e. Next, the two sequences are concatenated into sequence f.

The sequences encoded as described above are concatenated in the order of $abcdef$ and used as input to the language model. The training task is to maximize the probability of generating the target word sequence f.

3.2 Inference

The flow of the dialog system during the inference phase is as follows:

- **Encoding Knowledge and Dialogue History** (Figure 1-$abcd$) : The input data in the inference

phase are converted into word sequence $abcd$, as in the training phase in section 3.1.

- **Masking Knowledge** (Figure 2-e) : First, for the retrieved knowledge, we randomly select one entity of retrieved knowledge and transform it into a word sequence e. Next, integers X and Y ($0 < X, Y \leq MaskLen$, where $Masklen$ is a hyperparameter of mask tokens' number.) are randomly selected. X and Y <MASK> tokens are added in the left and right side of word sequence e. Finally, a stop token <EOS> is added to the end of the word sequence e.
- **Text Infilling** (Figure 2-f) : The word sequences encoded as described above are concatenated in the order abcde and input to the language model. The language model generates word sequence f sequentially by using a decoding strategy. Text Infilling is stopped when the second stop token <EOS> is generated.
- **Output** (Figure 2-g) : The stop token <EOS> splits the word sequence f into two word sequences, which are converted into word sequence g by replacing the left and right parts of the mask tokens <MASK> in e. The word sequence g is the output of the inference phase.

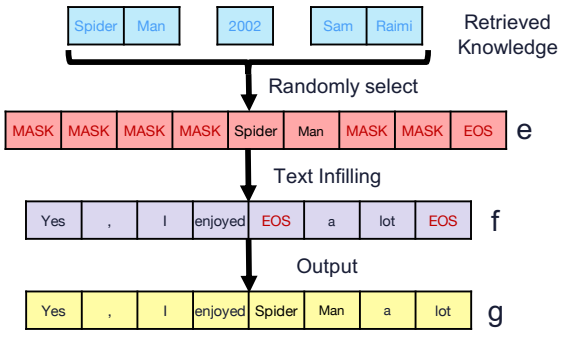

Figure 2: Encoding of knowledge and the output in the inference phase.

4 Experiments

We conducted experiments on the OpenDialKG dataset (Moon et al., 2019) which contains 15,000 dialogues. The dataset was collected in a Wizard-of-Oz setup, by connecting two human participants who were tasked to have an engaging dialogue about a given topic.

4.1 Experimental Details

Following the Zhang et al. (2019); Galetzka et al. (2021) work and section 3, we built 3 different types of the dialogue systems: a dialogue system without knowledge (generation-based dialogue system), a dialogue system with knowledge (knowledge-based dialogue system), and dialogue system with knowledge and text infilling (the proposed dialogue system).

We utilized DialoGPT-small (Zhang et al., 2019) as language model of 3 different dialogue systems. Table 1 shows the hyperparameters of the language models.

Table 1: Hyperparameters of language models

Total parameters	117M
Optimizer	AdamW
Max dialogue history	3
Decoding strategie	Greedy
Epochs	10
Batch size	4
MaskLen	10
Learning rate	6.0e-5

4.2 Evaluation metric

Automatic In the experimental evaluation, the quality of the response sentences is evaluated from two angles: diversity and correctness. DIST-n (Li et al., 2015), which represents the number of types of n-grams in the response sentences, is used as the evaluation index for diversity. BLEU-n (Papineni et al., 2002) and NIST-n (Doddington, 2002), which represent the degree of similarity between the response and the correct response, are used as evaluation indices for correctness. NIST-n is a variant of BLEU-n that weights n-gram matches by their information gain, i.e., it indirectly penalizes uninformative n-grams.

Ent-Res, which we employ to calculate the proportion of responses containing at least one entity of retrieved knowledge to all responses, and AvgLen, which represents the average number of words in the response sentences, are also used as evaluation indices. Furthermore, in order to compare the proposed method with previous models, we have listed the results achieved by previous models. The proposed method and conventional methods were compared using the same metrics, including the faithfulness metric FeQA (Durmus et al., 2020) and correctness metrics Rouge-L and BLEU-4.

Human In human evaluation, we use an evaluation technique called Best-Worst Scaling (BWS) (Flynn and Marley, 2014), which can handle a long list of options and always generates discriminating results. We employ three metrics at the utterance-level and dialogue-level: naturalness, informative-

Table 2: Results of the response sentences generated by each dialogue system. Higher is better.

Dialogue System	DIST-1	DIST-2	BLEU-1	BLEU-2	NIST-2	NIST-4	Ent-Res	Avg Len
Generation-Based	**11.93**	**36.79**	15.74	8.71	1.39	1.43	20%	10.86
Knowledge-Based	10.77	31.84	17.77	10.47	1.62	**1.69**	42%	10.45
Ours	9.09	32.18	**18.79**	**10.64**	**1.64**	**1.69**	**100%**	13.11

Table 3: Results of human evaluation using Best Worst Scaling (BWS).

Systems		Generation-Based	Knowledge-Based	Ours
naturalness	Best	30%	**40%**	30%
	Worst	35%	**21%**	44%
informativeness	Best	19%	33%	**48%**
	Worst	55%	27%	**18%**
coherence	Best	36%	**38%**	26%
	Worst	33%	**22%**	44%

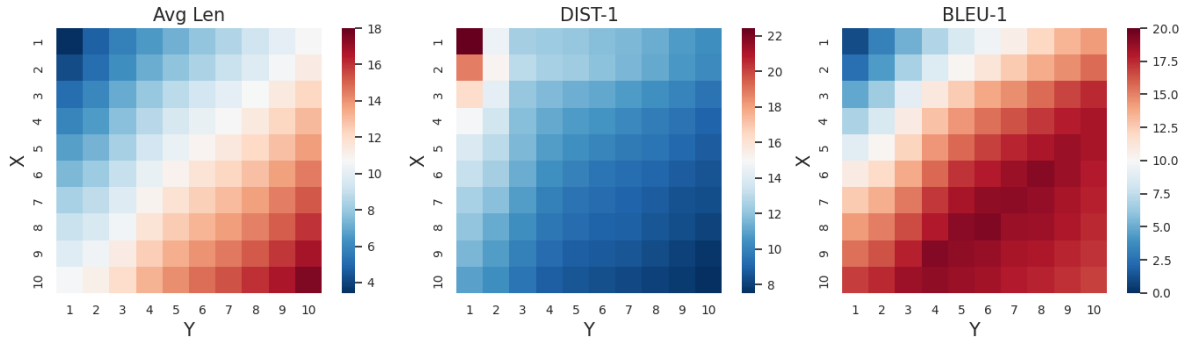

Figure 3: Heat map comparing the different X and Y impacts of the proposed dialogue system on three metrics. The blue shades denote lower values, white middle and black higher, with dark blue representing the lowest and dark black the highest values. X and Y denote the number of <MASK> tokens added to the left and right side of the word sequence e in Figure 2.

Table 4: Results of other dialogue system on OpenDialKG test data. Higher is better.

Dialogue System	FeQA	Rouge-L	BLEU-4
AdptBot	23.1	**31.0**	10.1
GPT-2+KE	19.5	19.0	5.5
GPT-2+KB	26.54	30.0	11.1
GPT-2+NPH	**28.9**	**31.0**	**11.3**
FSB	25.3	29.17	6.08
Ours	22.7	23.97	4.0

ness, and coherence. We randomly select 33 generated response examples. Three workers are asked to choose the best one and the worst one for three metrics in terms of response quality of each dialogue system with respect to the dialogue history.

- Naturalness is an utterance-level metric, judging whether the response is natural or not.
- Informativeness is also an utterance-level metric, evaluating whether the response is informative or not.
- Coherence is a dialogue-level metric, measuring whether the response is relevant and consistent with the context.

4.3 Results and Discussion

Table 2 shows the results of the response sentences generated by each dialogue system. The table shows that the proposed dialogue system reached the highest scores in the correctness evaluation index. This confirms the effectiveness of the text infilling task. On the other hand, the proposed method reached the highest value in the Ent-Res score, but the lowest value in the DIST-1 score. The improvement of the diversity in response sentences by the proposed method is a topic for future work.

Table 3 shows the results of human evaluation using Best Worst Scaling (BWS). We observed that the proposed method achieves lower naturalness and coherence scores compared to other models. Since the generated responses of the proposed method always contain entities as a result of the text infilling, the responses of the proposed method are rendered unnatural and incoherent. In future work, we will develop a module capable of determining whether knowledge should be embedded in the responses.

Table 5: Samples of responses generated by each dialog system. The retrieved knowledge entities are shown in blue.

Input 1	User A	what about tonto fistfight?
	User B	it was written by sherman alexie
	User A	was he a poet?
	Knowledge 1	poet: sherman alexie film producer: sherman alexie written by: sherman alexie
Generation-Based		yes he was a poet
Knowledge-Based		yes, he was a poet
Ours		yes, he was a poet. the sherman alexie wrote the poems and drawings of a rose.
Input 2	User A	they just got lebron james so that is a big benefit.
	User B	yes i was so happy that he signed. he is one of the greatest, right?
	User A	i'd say so. plus the lakes already have a good record, several nba finals championships under their belt.
	Knowledge 2	team: utah jazz
Generation-Based		i am sure the fans would love it. since lebron james is your favorite player, you must like him?
Knowledge-Based		that's awesome! i think he's a great player too.
Ours		i agree. i like the utah jazz. do you know who won that year?

Table 4 shows the comparison with previous models GPT-2+KB, AdapterBot (Madotto et al., 2020b), GPT-2+KE (Madotto et al., 2020a) and GPT-2+KB with Neural Path Hunter (NPH) (Dziri et al., 2021) and Few-Shot Bot (FSB) (Madotto et al., 2021). Due to the differences in the test dataset and the model sizes, the scores of the proposed method are just reference values. Nevertheless, the proposed method achieves lower FeQA, Rouge-L and BLEU-4 scores compared to previous models. Overall, NPH achieves the best performance, but it can also be applied to the proposed method; we leave this exploration to future work.

Table 5 shows samples of responses generated by each dialogue system. From the table, it can be confirmed that the proposed method can accurately use the retrieved knowledge and generate natural response sentences. On the other hand, the knowledge-based dialogue system is not able to use the knowledge. Despite this, it can be considered that knowledge 1 is not necessary to generate natural response sentences to the dialogue history of input 1. The development of a module that can determine the necessity of knowledge is a subject for future work.

4.4 Impact of <MASK> tokens

The results of the proposed dialogue system with different X and Y are compared using heat maps for various metrics, where X and Y denote the number of <MASK> tokens added to the left side of the word sequence e in Figure 2. Here, the heat map indicates a two-dimensional matrix with scores of the metrics such as BLEU-1 and DIST-1, computed by changing the length X and Y. We show the heat maps for three metrics in Figure 3. The heat maps for other metrics are shown in the Appendix A.

As can be observed in the heat map of Avg Len scores in Figure 3, the larger the sum of X and Y, the higher the value. It can be confirmed that the proposed dialogue system can correctly generate responses of the corresponding length based on X and Y. On the other hand, we can control the length of the generated responses by modifying X and Y. Due to the possibility of duplicate words in longer responses, the values in heat maps of Avg Len and DIST-1 show the opposite trend.

As can be observed in the heat map of BLEU-1 scores in Figure 3, the scores are similar when the sums of X and Y are equal, and the score is highest when the sum of X and Y is around 13. It can be confirmed that the scores are relevant to the sum of X and Y, not X nor Y. If the appropriate X and Y can be determined, the proposed dialogue system will have better performance. However, the appropriate sums of X and Y may be different in various datasets. Developing a method for finding the best combination on X and Y will be the subject of future work.

5 Conclusion

We proposed a knowledge-based dialogue system based on the text infilling method, aiming to improve the problem that the knowledge-based dialogue system generates responses without using retrieved knowledge. The proposed dialogue system can constantly incorporate external knowledge. In our experiments, the proposed dialogue system generated significantly more correct responses than baseline approaches.

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhen Guo, Zhibin Liu, and Xinchao Xu. 2020. Plato-2: Towards building an open-domain chatbot via curriculum learning. *arXiv preprint arXiv:2006.16779*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145.

Chris Donahue, Mina Lee, and Percy Liang. 2020. Enabling language models to fill in the blanks. *arXiv preprint arXiv:2005.05339*.

Esin Durmus, He He, and Mona Diab. 2020. Feqa: A question answering evaluation framework for faithfulness assessment in abstractive summarization. *arXiv preprint arXiv:2005.03754*.

Nouha Dziri, Andrea Madotto, Osmar Zaiane, and Avishek Joey Bose. 2021. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. *arXiv preprint arXiv:2104.08455*.

Terry N Flynn and Anthony AJ Marley. 2014. Best-worst scaling: theory and methods. In *Handbook of choice modelling*, pages 178–201. Edward Elgar Publishing.

Fabian Galetzka, Jewgeni Rose, David Schlangen, and Jens Lehmann. 2021. Space efficient context encoding for non-task-oriented dialogue generation with graph attention transformer. 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 7028–7041.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2021. Internet-augmented dialogue generation. *arXiv preprint arXiv:2107.07566*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Andrea Madotto, Samuel Cahyawijaya, Genta Indra Winata, Yan Xu, Zihan Liu, Zhaojiang Lin, and Pascale Fung. 2020a. Learning knowledge bases with parameters for task-oriented dialogue systems. *arXiv preprint arXiv:2009.13656*.

Andrea Madotto, Zhaojiang Lin, Yejin Bang, and Pascale Fung. 2020b. The adapter-bot: All-in-one controllable conversational model. *arXiv preprint arXiv:2008.12579*.

Andrea Madotto, Zhaojiang Lin, Genta Indra Winata, and Pascale Fung. 2021. Few-shot bot: Prompt-based learning for dialogue systems. *arXiv preprint arXiv:2110.08118*.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

Jason Weston, Emily Dinan, and Alexander H Miller. 2018. Retrieve and refine: Improved sequence generation models for dialogue. *arXiv preprint arXiv:1808.04776*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

A Heat map of metrics

There are the heat maps of the DIST-2, BLEU-2, NIST-2 and NIST-4 scores of the proposed dialogue system with different X and Y on the test set in Figure 4.

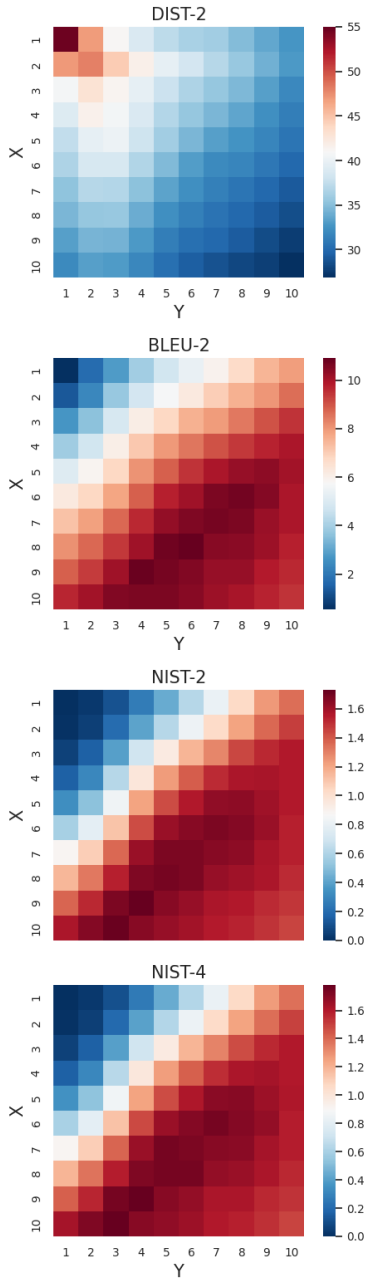

Figure 4: Heat map comparing the different X and Y impacts of the proposed dialogue system on three metrics. The blue shades denote lower values, white middle and black higher, with dark blue representing the lowest and dark black the highest values. X and Y denote the number of <MASK> tokens added to the left and right side of the word sequence e in Figure 2.