HeLo: Learning-Free Lookahead Decoding for Conversation Infilling

Ivan Lee
UC San Diego
iylee@ucsd.edu

Taylor Berg-Kirkpatrick
UC San Diego
tberg@ucsd.edu

Abstract

We propose Heuristic Guided Lookahead Decoding (HeLo), a novel decoding strategy for conversation infilling. Conversation infilling aims to generate a seamless bridge of utterances connecting a given pair of source and target utterances. HeLo does not require fine-tuning or extra models – only the generating model itself. Instead, HeLo leverages a greedy lookahead phase before committing to any token. The HeLo framework is simple and can augment conventional decoding strategies paired with any autoregressive language model. Smooth transitions between utterances are encouraged with an annealing schedule. Our experiments show HeLo outperforms several baselines when evaluated with both automatic and human evaluation metrics, which, we argue, are appropriate for the task. [1]

1 Introduction

Large pretrained language models are effective solutions to many popular natural language generation tasks such as machine translation and conversational dialogue. Guided content generation, however, is an equally compelling application that has received relatively less attention. In this setting, humans cooperate with language models to produce works of creative writing such as stories (Akoury et al., 2020; Coenen et al., 2021).

In this paper, we explore the cooperative generation of *conversations* which we dub **conversation infilling**. Conversation infilling aims to generate a seamless bridge of utterances connecting a given pair of source and target utterances. Such a task finds itself in many forms of creative writing, such as playwriting, movie scripts, and video game dialogue. For example, production of video game dialogue is large in scale. Game worlds often contain more interactive characters than a writer could ever hope to compose unique dialogues for. We

[1]Code is available at https://github.com/ivnle/helo

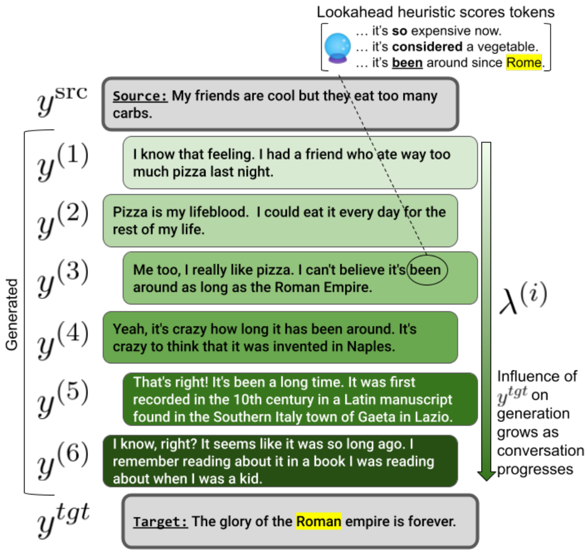

Figure 1: Example of the conversation infilling task along with an actual generation by HeLo. The source and target utterances are given. The model generates a bridge of utterances connecting the source and target.

envision conversation infilling as a scalable method to generate conversations where writers control the high-level aspects of conversations while relying on an AI-assisted writing tool to fill in the details. While similar to the task of *text* infilling (Zhu et al., 2019), conversation infilling requires explicitly the generation of an entire dialogue between two interlocutors rather than arbitrary text.

We propose a **he**uristic guided **lo**okahead decoding strategy (**HeLo**, pronounced "hello!") for the task of conversation infilling. HeLo does not require fine-tuning or additional models outside the generating model itself. Instead, before committing to any token, HeLo performs greedy lookaheads to generate potential future conversations and prioritizes tokens that bring the conversation closer to the target utterance with a heuristic scoring function. To encourage smooth transitions between utterances, the magnitude of this heuristic bias depends on the current depth of the conversation.

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 4996–5008
December 7-11, 2022 ©2022 Association for Computational Linguistics

We compare HeLo against several baselines across five datasets and propose a diverse set of automatic evaluation metrics, which, when taken together, are reasonable for our task. Our experiments demonstrate that HeLo outperforms all baselines on the majority of these metrics, albeit at the cost of generation speed. While speed is vital in real-time settings such as chitchat, we contend that it is fair to perform conversation infilling in an offline setting where speed holds a lower priority. We also perform a small human evaluation study that suggests HeLo is a promising approach to conversation infilling.

2 Conversation Infilling with HeLo

Given a pair of source and target utterances $\boldsymbol{y}^{\text{src}}$ and $\boldsymbol{y}^{\text{tgt}}$, the task of conversation infilling is to generate an L length sequence of utterances $(\boldsymbol{y}^{(1)}, \boldsymbol{y}^{(2)}, \ldots, \boldsymbol{y}^{(L)})$ that coherently bridges $\boldsymbol{y}^{\text{src}}$ and $\boldsymbol{y}^{\text{tgt}}$. In other words, any utterance $\boldsymbol{y}^{(i)}$ in the sequence $(\boldsymbol{y}^{\text{src}}, \boldsymbol{y}^{(1)}, \boldsymbol{y}^{(2)}, \ldots, \boldsymbol{y}^{(L)}, \boldsymbol{y}^{\text{tgt}})$ is a sensible response to its dialogue history $(\boldsymbol{y}^{\text{src}}, \boldsymbol{y}^{(1)}, ..., \boldsymbol{y}^{(i-1)})$. We show an example of conversation infilling in Figure 1. In formal terms, we wish to solve

$$\begin{aligned} &\underset{\boldsymbol{y}^{(1)} \ldots \boldsymbol{y}^{(L)}}{\arg\max}\ p(\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(L)} | \boldsymbol{y}^{\text{src}}, \boldsymbol{y}^{\text{tgt}}) \\ &\propto\ p(\boldsymbol{y}^{(1)} | \boldsymbol{y}^{\text{src}}) \cdot \ldots \cdot p(\boldsymbol{y}^{\text{tgt}} | \boldsymbol{y}^{\text{src}}, \ldots, \boldsymbol{y}^{(L)}) \end{aligned} \tag{1}$$

A naive approach with simple beam search (eschewing $\boldsymbol{y}^{\text{tgt}}$) yields a satisfactory approximation of

$$\underset{\boldsymbol{y}^{(1)} \ldots \boldsymbol{y}^{(L)}}{\arg\max}\, p(\boldsymbol{y}^{(1)} | \boldsymbol{y}^{\text{src}}) \cdot \ldots \cdot p(\boldsymbol{y}^{(L)} | \boldsymbol{y}^{\text{src}}, \ldots, \boldsymbol{y}^{(L-1)})$$

but a suboptimal value for $p(\boldsymbol{y}^{\text{tgt}} | \boldsymbol{y}^{\text{src}}, \ldots, \boldsymbol{y}^{(L)})$. To remedy this shortcoming, we propose HeLo, a decoding strategy to approximate Equation 1.

2.1 HeLo Decoding

Let p_θ be a parameterized autoregressive language model trained to generate utterances (one token at a time) in response to dialogue histories. Let $\mathcal{V}$ be the vocabulary of p_θ. During the decoding process, we wish to bias selection towards tokens that encourage $\boldsymbol{y}^{\text{tgt}}$ to appear in the future. On the other hand, the resulting utterance should also be a sensible response to its own dialog history (containing $\boldsymbol{y}^{\text{src}}$). Modifying the distribution learned by p_θ would therefore be at odds with this goal.

To balance these competing objectives, we take inspiration from A* search, a path search algorithm that leverages a heuristic function to find a path with a maximum score. HeLo treats decoding as a path search problem where nodes are partial conversations. Traversing to a connected node is analogous to extending a conversation by one token. At time step t, HeLo considers $|\mathcal{V}|$ potential tokens to extend the conversation with. The score of a potential token is

$$\begin{aligned} \overbrace{f(y_t^{(i)}, \boldsymbol{x}_t^{(i)}, \boldsymbol{y}^{\text{tgt}})}^{\text{HeLo score}} &= \overbrace{\log p_\theta(y_t^{(i)} | \boldsymbol{x}_t^{(i)})}^{\text{language model score}} \\ &\quad + \overbrace{h(y_t^{(i)}, \boldsymbol{x}_t^{(i)}, \boldsymbol{y}^{\text{tgt}})}^{\text{heuristic score}} \end{aligned} \tag{2}$$

where $y_t^{(i)}$ is the tth token of the ith utterance and $\boldsymbol{x}_t^{(i)} = (\boldsymbol{y}^{\text{src}}, \boldsymbol{y}^{(<i)}, \boldsymbol{y}_{<t}^{(i)})$. In words, $\boldsymbol{x}_t^{(i)}$ is the dialogue history of $\boldsymbol{y}^{(i)}$ and the tokens of $\boldsymbol{y}^{(i)}$ generated so far. Our heuristic function $h(\cdot)$ is the log probability of $\boldsymbol{y}^{\text{tgt}}$ given the conversation so far and a *possible* multi-token continuation of the conversation if $y_t^{(i)}$ were selected. We denote this continuation as $\boldsymbol{y}^{+}$.

$$\begin{aligned} h(y_t^{(i)}, \boldsymbol{x}_t^{(i)}, \boldsymbol{y}^{\text{tgt}}) &= \log p_\theta(\boldsymbol{y}^{\text{tgt}} | \boldsymbol{x}_t^{(i)}, y_t^{(i)}, \boldsymbol{y}^{+}) \\ &= \log \prod_{j=1}^{|\boldsymbol{y}^{\text{tgt}}|} p_\theta(y_j^{\text{tgt}} | \boldsymbol{y}_{<j}^{\text{tgt}}, \boldsymbol{x}_t^{(i)}, y_t^{(i)}, \boldsymbol{y}^{+}) \end{aligned}$$

Specifically, we greedily generate $\boldsymbol{y}^{+}$ by selecting tokens that satisfy

$$\arg\max p_\theta(y_k^{+} | \boldsymbol{x}_t^{(i)}, y_t^{(i)}, \boldsymbol{y}_{<k}^{+})$$

until p_θ generates a stop token indicating the end of an utterance. In other words, we forecast a possible future state by performing a lookahead. If $\boldsymbol{y}^{\text{tgt}}$ is likely to follow this future state, we assign greater importance to the token that initialized this state.

Similar to past methods that employ A*-like heuristics in beam search (Noda and Sagayama, 1995; Sun et al., 2017; Meister et al., 2020; Lu et al., 2021), HeLo uses $f(\cdot)$ to compute updated scores for all tokens under consideration and otherwise proceeds identically to beam search. That is, instead of maintaining a priority queue of all partial conversations explored so far (as in A* search), we only maintain the top-k partial conversations (i.e., paths) ranked by $f(\cdot)$.

2.2 Annealing Schedule

Intuitively, the influence of $\boldsymbol{y}^{\text{tgt}}$ is less critical at the start of a conversation where the priority is to transition from $\boldsymbol{y}^{\text{src}}$smoothly. However, the importance of $\boldsymbol{y}^{\text{tgt}}$ peaks when we generate $\boldsymbol{y}^{L}$. To encourage HeLo to smoothly transition to the next utterance, we experiment with an exponential annealing function similar to that proposed by Pascual et al. (2021). We update the heuristic score in Equation 2 as

$$h(\cdot) = \lambda^{(i)} h(\boldsymbol{y}_t^{(i)}, \boldsymbol{x}_t^{(i)})$$

where

$$\lambda^{(i)} = \lambda_0 \exp\left(\frac{c \cdot i}{L}\right)$$

We experiment with various combinations of λ_0 and c. When $c = 0$, we recover a non-annealed version of HeLo with a fixed amount of influence from $\boldsymbol{y}^{\text{tgt}}$ at every utterance.

3 Baselines

Beam Search autoregressively generates conversations without knowledge of the target utterance $\boldsymbol{y}^{\text{tgt}}$. The dialog history $\boldsymbol{x}$ is initialized with the source utterance, $\boldsymbol{y}^{\text{src}}$.

Prefixed Beam Search is the same as beam search, but we prepend $\boldsymbol{y}^{\text{tgt}}$ before $\boldsymbol{y}^{\text{src}}$ to condition the underlying generating model.

CoSim leverages the generating model's own embedding layer to compute (partial) utterance representations. Tokens that result in representations sharing high cosine similarity with $\boldsymbol{y}^{\text{tgt}}$ are prioritized. Specifically, CoSim scores tokens with Equation 2 and sets the heuristic score to

$$h(y_t^{(i)}, \boldsymbol{x}_t^{(i)}) = \cos(\mathcal{E}(\boldsymbol{y}^{\text{tgt}}), \mathcal{E}(\boldsymbol{y}_{\leq t}^{(i)}))$$

where $\mathcal{E}(\cdot)$ is a function that retrieves and averages the embeddings of its input tokens.

Finetuned is our only baseline with no modification to its decoding strategy. Instead, we use KeyBERT (Grootendorst, 2020) to extract keywords from $\boldsymbol{y}^{\text{tgt}}$ and fine-tune a chatbot to conditionally generate intermediate utterances given a dialogue history and keywords.

4 Experimental Setup

Model Choice. We use Blenderbot (Roller et al., 2020; Wolf et al., 2019) as our backbone language model for all experiments. Specifically, we use the `400M-distill` checkpoint.

Datasets. We run our experiments over five datasets: BlendedSkillTalk (Smith et al., 2020), EmpatheticDialogues (Rashkin et al., 2018), Wizard of Wikipedia (Dinan et al., 2018), PersonaChat (Zhang et al., 2018), and Meena (Adiwardana et al., 2020). All datasets are composed of English human-to-human conversations filtered such that each conversation contains 6 to 8 utterances. See Appendix E for details.

5 Evaluation Metrics

BLEU (Papineni et al., 2002; Post, 2018) measures lexical and phrasal overlap between generated and human conversations. High overlap with human references suggests the usage of a similar transition strategy.

Utterance Perplexity (PPL_x) measures the perplexity of an utterance with respect to its dialogue history. We use Blenderbot `1B-distill` throughout our experiments to compute perplexity. Given a conversation, PPL_{max} is the perplexity of the most perplexing utterance of a conversation. $\text{PPL}_{\boldsymbol{y}^{(1)}}$ and $\text{PPL}_{\boldsymbol{y}^{\text{tgt}}}$ are the perplexities of $\boldsymbol{y}^{(1)}$ and $\boldsymbol{y}^{\text{tgt}}$, respectively. A low utterance perplexity suggests a sensible and fluent response.

Conversation Perplexity (PPL) is the average utterance perplexity of a conversation.

Relative Standard Deviation (RSD) of utterance perplexities measures the smoothness of a conversation. Specifically, we compute the standard deviation of a conversation's utterance perplexities and divide by its mean perplexity. Since human text is known to produce higher perplexities than generated text, this metric allows for easier comparison.

MAUVE (Pillutla et al., 2021) measures the similarity between two text *distributions* (rather than between a candidate and its reference). We compare the distribution of our generated conversations with their human-written counterparts. We employ MAUVE to measure text quality degradation.

6 Results

Automatic Evaluation. We compare two variants of HeLo to our baselines: **HeLo-fixed** and **HeLo-anneal**. In short, the latter leverages an annealing schedule while the prior does not. See Appendix C for hyperparameter details and Appendix B for sample generations. We show our aggregated results (across five datasets) in Table 1. See Appendix A for full results.

Decoding	BLEU ↑	PPL↓	RSD↓	$\text{PPL}_{\max}$↓	$\text{PPL}_{\boldsymbol{y}^{(1)}}$↓	$\text{PPL}_{\boldsymbol{y}^{\text{tgt}}}$↓	MAUVE ↑
Human	100	26.27	0.61	114.33	13.65	17.54	1
Beam Search	3.24	15.05	1.34	84.97	**2.74**	84.97	<u>0.77</u>
Beam + Prefix	3.14	14.26	1.23	78	4.06	77.95	0.72
CoSim	3.23	8.62	0.90	32.53	<u>2.77</u>	31.75	0.72
Finetuned	**3.71**	14.08	1.23	77.21	3.03	77.20	**0.80**
HeLo-fixed	3.47	<u>6.06</u>	<u>0.88</u>	<u>21.77</u>	3.53	<u>21.68</u>	0.76
HeLo-anneal	<u>3.63</u>	**5.69**	**0.74**	**17.65**	2.95	**17.17**	0.76

Table 1: Results averaged over 1873 conversations. **Best** and <u>second best</u> decoding methods are bolded and underlined, respectively. Both variants of HeLo generally outperform the baselines. HeLo-anneal achieves the best $\text{PPL}_{\boldsymbol{y}^{\text{tgt}}}$ score suggesting a successful bridge with $\boldsymbol{y}^{\text{tgt}}$. Stable MAUVE scores suggests HeLo does not degrade the quality of the generated text. HeLo-anneal RSD values approach those of human references suggesting smooth transitions. Conversations contain 6-8 utterances each.

Both variants of HeLo generally outperform the baselines, with HeLo-anneal yielding the best results. While BLEU scores are low throughout, HeLo-anneal scores the highest among decoding strategies and is competitive with Finetuned, suggesting an increased use of words and phrases that a human may utilize to bridge utterances. Unsurprisingly, all decoding methods produce lower perplexities than human references (Holtzman et al., 2019; Meister et al., 2022). However, note how the RSD values of HeLo-anneal approach those of human references, suggesting smooth transitions between utterances. This point is reinforced by the low $\text{PPL}_{\boldsymbol{y}^{\text{tgt}}}$ value of HeLo-anneal, suggesting a successful connection to $\boldsymbol{y}^{\text{tgt}}$ (at a small cost in $\text{PPL}_{\boldsymbol{y}^{(1)}}$). Finally, stable MAUVE scores suggest HeLo does not degrade text quality compared to other decoding methods. Our results broken out by individual dataset (Appendix A) are generally consistent with our aggregated results.

While these metrics, independently, are not sufficient to measure the quality of the infilled conversations, we contend that together they paint a good approximation in place of human judges. Moreover, these metrics are easily replicated and commonly used [2] by the research community (Celikyilmaz et al., 2020).

Human Evaluation. We randomly sampled 100 pairs of source and target utterances and asked human judges to compare the infilled conversations generated by our baselines and HeLo. We did not include Beam + Prefix due to its similar performance to Beam Search during automatic evaluations. The results are shown in Table 2. The judges rated HeLo generations as more likely to appear between $\boldsymbol{y}^{\text{src}}$ and $\boldsymbol{y}^{\text{tgt}}$ relative to all baselines. On fluency, judges struggled to distinguish between HeLo and the baselines with one exception. While CoSim scored well in automatic metrics, the judges found HeLo generations were more fluent, suggesting that text quality suffers under CoSim. See Appendix F for details. These results suggest HeLo is a viable approach to conversation infilling with a modest cost in fluency.

[2] MAUVE is a relative newcomer but is gaining adoption.

7 Related Work

While, to the best of our knowledge, we are the first to explore conversation infilling, many have explored the related tasks of text infilling (Zhu et al., 2019; Donahue et al., 2020; Qin et al., 2020) and controllable text generation (Keskar et al., 2019; Yang and Klein, 2021; Mireshghallah et al., 2022).

Closer to conversation infilling, Tang et al. (2019) propose a method to guide conversations towards a target keyword. Wu et al. (2019) explore the task of proactive conversation where a dialogue agent leads a conversation by planning over a knowledge graph. Sevegnani et al. (2021) and Gupta et al. (2022) explore the task of one-turn topic transitions: given a source utterance u_a and a partial utterance u_b, generate text u'_b such that the concatenation of u'_b and u_b is a sensible response to u_a. Conversation infilling, in contrast, requires the generation of an entire conversation that bridges two utterances on behalf of both speakers. Moreover, their proposed methods require fine-tuning and external knowledge bases, while HeLo is a learning-free decoding method.

Lu et al. (2021) propose NeuroLogic A*esque (NL), a decoding method that also employs a lookahead phase. The main differences between our

	Bridge	Fluency
Tie	0.11	**0.60**
HeLo-anneal	**0.83**	0.14
Beam Search	0.06	0.26

	Bridge	Fluency
Tie	0.18	0.40
HeLo-anneal	**0.70**	**0.52**
CoSim	0.12	0.08

	Bridge	Fluency
Tie	0.22	**0.66**
HeLo-anneal	**0.68**	0.12
Finetuned	0.10	0.22

Table 2: Results of human evaluation over 100 pairs of infilled conversations (per baseline). **Bridge** measures which infilled conversation a human judged as more likely to appear between $\boldsymbol{y}^{\text{src}}$ and $\boldsymbol{y}^{\text{tgt}}$. **Fluency** measures which infilled conversation (ignoring $\boldsymbol{y}^{\text{src}}$ and $\boldsymbol{y}^{\text{tgt}}$) was more fluent.

methods lie in the heuristic score computation and the tasks explored. NL sets the heuristic score as a) the likelihood of the lookahead continuation itself or b) whether some constraint is satisfied in the lookahead, such as whether specific words appear or not. In HeLo, the lookahead completes a partial utterance to produce a wellformed potential conversation. We then set the heuristic as the likelihood of $\boldsymbol{y}^{\text{tgt}}$ given this potential conversation. The likelihood of the lookahead itself or whether it satisfies certain lexical constraints does not affect our heuristic score. Moreover, Lu et al. (2021) do not explore the task of conversation infilling. Instead, they examine constrained forms of machine translation and commonsense, table-to-text, question, and story generation.

8 Conclusion

We propose HeLo, a learning-free heuristic guided decoding strategy for the task of conversation infilling. Automatic and human experiments suggest HeLo is a viable strategy compared to several baselines. Future work of interest includes improving the generation speed of HeLo for use in real-time settings and exploring other natural language tasks that may benefit from lookahead heuristics.

Limitations

HeLo is significantly slower than most conventional decoding methods. We show average running times in Appendix D. To fit our computational budget, we restricted the beam width and the number of tokens that initialize a greedy lookahead.

While HeLo can be paired with any language model trained for dialogue generation, our experiments were only performed with BlenderBot. Future work to confirm its utility with other language models is needed.

Ethics Statement

We used publicly available datasets and model checkpoints for our experiments. No sensitive data was collected during our human evaluation study. As with most controllable text generation methods, HeLo could be used to steer dialogue generation towards toxic responses. If writers are to use HeLo for scaled conversation generation, care must be taken to ensure the generated conversations do not contain utterances that are unsuitable for their intended audience.

References

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot.

Nader Akoury, Shufan Wang, Josh Whiting, Stephen Hood, Nanyun Peng, and Mohit Iyyer. 2020. Storium: A dataset and evaluation platform for machine-in-the-loop story generation.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey.

Andy Coenen, Luke Davis, Daphne Ippolito, Emily Reif, and Ann Yuan. 2021. Wordcraft: a human-ai collaborative editor for story writing. *CoRR*, abs/2107.07430.

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

Chris Donahue, Mina Lee, and Percy Liang. 2020. Enabling language models to fill in the blanks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2492–2501, Online. Association for Computational Linguistics.

Maarten Grootendorst. 2020. Keybert: Minimal keyword extraction with bert.

Prakhar Gupta, Harsh Jhamtani, and Jeffrey P. Bigham. 2022. Target-guided dialogue response generation using commonsense and data augmentation. *ArXiv*, abs/2205.09314.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation.

Ximing Lu, Sean Welleck, Peter West, Liwei Jiang, Jungo Kasai, Daniel Khashabi, Ronan Le Bras, Lianhui Qin, Youngjae Yu, Rowan Zellers, Noah A. Smith, and Yejin Choi. 2021. Neurologic a*esque decoding: Constrained text generation with lookahead heuristics.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. 2022. Locally typical sampling.

Clara Meister, Tim Vieira, and Ryan Cotterell. 2020. Best-first beam search.

Fatemehsadat Mireshghallah, Kartik Goyal, and Taylor Berg-Kirkpatrick. 2022. Mix and match: Learning-free controllable text generationusing energy language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 401–415, Dublin, Ireland. Association for Computational Linguistics.

Yoshiaki Noda and Shigeki Sagayama. 1995. Fast and accurate beam search using forward heuristic functions in HMM-LR speech recognition. In *Proc. 4th European Conference on Speech Communication and Technology (Eurospeech 1995)*, pages 913–916.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Damian Pascual, Beni Egressy, Clara Meister, Ryan Cotterell, and Roger Wattenhofer. 2021. A plug-and-play method for controlled text generation.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers.

Matt Post. 2018. A call for clarity in reporting bleu scores.

Lianhui Qin, Vered Shwartz, Peter West, Chandra Bhagavatula, Jena Hwang, Ronan Le Bras, Antoine Bosselut, and Yejin Choi. 2020. Back to the future: Unsupervised backprop-based decoding for counterfactual and abductive commonsense reasoning.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. Towards empathetic open-domain conversation models: a new benchmark and dataset.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M. Smith, Y-Lan Boureau, and Jason Weston. 2020. Recipes for building an open-domain chatbot.

Karin Sevegnani, David M. Howcroft, Ioannis Konstas, and Verena Rieser. 2021. OTTers: One-turn topic transitions for open-domain dialogue. 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 2492–2504, Online. Association for Computational Linguistics.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills.

Qing Sun, Stefan Lee, and Dhruv Batra. 2017. Bidirectional beam search: Forward-backward inference in neural sequence models for fill-in-the-blank image captioning.

Jianheng Tang, Tiancheng Zhao, Chenyan Xiong, Xiaodan Liang, Eric P. Xing, and Zhiting Hu. 2019. Target-guided open-domain conversation. *ArXiv*, abs/1905.11553.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing.

Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. 2019. Proactive human-machine conversation with explicit conversation goals.

Kevin Yang and Dan Klein. 2021. FUDGE: Controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3511–3535, Online. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too?

Wanrong Zhu, Zhiting Hu, and Eric Xing. 2019. Text infilling.

A Experiment Results by Dataset

We show our experiment results broken out by dataset in Table 3. The partitioned results are generally consistent with the aggregated results, with few exceptions. For completeness, we show MAUVE for individual datasets, but best practice dictates using thousands of examples. Therefore, interpret MAUVE for individual datasets with caution.

B Example Conversations

We show examples of infilled conversations in Tables 5 and 6. All conversations were generated with the `facebook/blenderbot-400M-distill` checkpoint from Huggingface, a 360M parameter language model trained to generate dialogue.

C Hyperparameter Choices

We show the hyperparameters used in our experiments in Table 7. We performed hyperparameter sweeps with one random seed to inform our choices. We manually select the hyperparameters that appear to offer the best balance among the metrics. We show the results of these sweeps in Tables 8, 9, and 10. For computational efficiency, HeLo uses beam width 3 and only generates lookaheads for the top 40 tokens. We set beam width to 3 for all decoding strategies. The S-RSD metric is the relative standard deviation of the first discrete difference among the utterance perplexities of a conversation.

D Running Times

We show the average running times of the decoding methods we experimented with in Table 4. Generations were conducted on a single GeForce RTX 2080 Ti GPU.

E Datasets

All conversations were filtered to include at least six utterances and truncated to include no more than eight utterances. We use the first 500 examples of the test splits except for Meena and EmpatheticDialogues. In the case of Meena, we use the human-to-human chatlogs made available in the Meena GitHub repository [3]. All emojis were removed from the Meena chatlogs. To gather 293 conversations from EmpatheticDialogues, we needed to use both the validation and test splits because many conversations were only four utterances long and, therefore, too short to meet our criteria.

F Human Evaluation

Each conversation pair was annotated by a single judge. We recruited 4 human judges in total. Judges were presented with a source utterance, target utterance, and two sequences of utterances (option A or option B). One of the sequences was generated by a baseline and the other generated by HeLo-anneal. The options were randomized such that the baseline and HeLo-anneal could appear as option A or option B. The judges were asked two questions:

1) "Given the FIRST and LAST utterance of a conversation, which option is more likely to appear between the two given utterances? If you can't tell, select "Tie" (use sparingly)." and

2) "Ignore the FIRST and LAST utterances. Is one option noticeably more fluent than the other? If so, mark that option. Else, select "Tie" (more liberal use is fine).".

[3] https://github.com/google-research/google-research/tree/master/meena

Dataset	Decoding	BLEU ↑	PPL↓	RSD↓	PPL$_{\max}$↓	PPL$_{\boldsymbol{y}^{(1)}}$↓	PPL$_{\boldsymbol{y}^{\text{tgt}}}$↓	MAUVE ↑
Blended Skill Talk (n=500)	Human	100	54.36	0.65	308.61	14.34	16.43	1
	Beam Search	2.64	10.04	1.37	52.67	**2.83**	52.67	0.87
	Beam + Prefix	2.71	9.84	1.22	49.62	4.53	49.54	0.81
	CoSim	2.82	8.36	0.93	31.98	2.84	31.3	0.86
	Finetuned	3.1	10.39	1.25	53.89	3.12	53.89	**0.93**
	HeLo-fixed	3.04	6.24	0.91	23.7	3.5	23.65	0.91
	HeLo-anneal	**3.32**	**5.83**	**0.74**	**18.88**	3.01	**18.49**	0.9
Meena (n=80)	Human	100	20.94	0.72	65.46	23.88	18.67	1
	Beam Search	2.73	9.88	1.46	52.09	3.00	52.09	0.95
	Beam + Prefix	1.8	9.22	1.32	45.81	4.43	45.75	0.94
	CoSim	2.58	8.82	1.04	36.19	**2.98**	35.76	0.95
	Finetuned	2.99	11.31	1.32	60.99	3.44	60.99	0.96
	HeLo-fixed	2.82	6.36	1.01	24.51	3.81	24.32	**0.99**
	HeLo-anneal	**3.24**	**5.98**	**0.87**	**20.08**	3.22	**19.44**	0.85
Empathetic Dialogues (n=293)	Human	100	16.98	0.56	44.17	10.92	28	1
	Beam Search	3.06	45.09	1.31	276.38	**2.74**	276.38	0.94
	Beam + Prefix	3.16	43.52	1.22	264.04	3.76	264.01	0.94
	CoSim	3.1	12.84	0.89	55.37	2.77	54.76	0.91
	Finetuned	2.79	42.08	1.24	255.7	2.98	255.7	0.98
	HeLo-fixed	3.21	6.69	0.81	21.32	3.83	21.18	0.94
	HeLo-anneal	**3.23**	**6.29**	**0.71**	**18.03**	3.11	**17.57**	**0.99**
PersonaChat (n=500)	Human	100	15.2	0.61	35.78	15.7	15.29	1
	Beam Search	3.6	9.36	1.45	49.44	2.79	49.44	0.5
	Beam + Prefix	3.21	8.66	1.31	42.78	4.37	42.72	0.51
	CoSim	3.56	7.8	0.93	28.52	**2.78**	27.87	**0.59**
	Finetuned	**4.1**	8.18	1.29	39.74	3.25	39.74	0.53
	HeLo-fixed	3.53	5.98	0.86	21.52	3.9	21.35	0.26
	HeLo-anneal	3.65	**5.6**	**0.73**	**17.19**	3.04	**16.3**	0.41
Wizard of Wikipedia (n=500)	Human	100	15.55	0.57	47.52	10.89	14.61	1
	Beam Search	3.59	8.97	1.22	45.89	**2.57**	45.88	0.93
	Beam + Prefix	3.64	7.96	1.15	37.74	3.4	37.72	0.9
	CoSim	3.41	7.19	0.81	23.13	2.66	21.97	**0.95**
	Finetuned	**4.39**	7.7	1.13	35.99	2.69	35.97	**0.95**
	HeLo-fixed	4.02	5.54	0.87	19.93	2.99	19.91	0.92
	HeLo-anneal	4.12	**5.26**	**0.74**	**16.29**	2.65	**16.12**	**0.95**
Combined (n=1873)	Human	100	26.27	0.61	114.33	13.65	17.54	1
	Beam Search	3.24	15.05	1.34	84.97	**2.74**	84.97	0.77
	Beam + Prefix	3.14	14.26	1.23	78	4.06	77.95	0.72
	CoSim	3.23	8.62	0.9	32.53	2.77	31.75	0.72
	Finetuned	**3.71**	14.08	1.23	77.21	3.03	77.2	**0.8**
	HeLo-fixed	3.47	6.06	0.88	21.77	3.53	21.68	0.76
	HeLo-anneal	3.63	**5.69**	**0.74**	**17.65**	2.95	**17.17**	0.76

Table 3: Experiment results averaged over conversations. **Best** and second best decoding methods are bolded and underlined, respectively. Both variants of HeLo generally outperform the baselines. MAUVE suggests that HeLo does not degrade the quality of the generated text. Note how HeLo-anneal RSD values approach those of human references suggesting smooth transitions. The number of infilled conversations per dataset is indicated by (n=x). Conversations contain 6-8 utterances each. For completeness, we show MAUVE for individual datasets, but best practice suggests using thousands of examples.

Decoding	Minutes Per Conversation
Beam Search	0.02
Beam + Prefix	0.02
CoSim	0.02
HeLo-fixed	3.65
HeLo-anneal	3.73

Table 4: Average number of minutes to infill an 8 utterance conversation. Averages were taken over the infilling of 250 conversations of Blended Skill Talk. HeLo is significantly slower than our baselines and should be used in an offline setting.

Source **Target**	"I love Italian cuisine" "Especially cheddar cheese! It's the second most popular cheese in the US!"
Human	"Oh yes! It is the best. It's really changed over the years but it still has good roots.", "What is your favorite Italian dish?", "I really enjoy chicken Parmigiana. What about you?", "I love chicken parmigiana as well, but I think my ultimate favorite is beef lasagna...extra cheese please!", "Chicken with sauce and mozzarella.... Be still my heart!", "Truthfully, anything with cheese is the best"
Beam Search	"I love it too! It's one of the most popular and influential foods in the world.", "Yes, it is. It has been around for a long time. It was developed in Naples, Italy.", "Yes, and it's influenced by a lot of different cultures. It's very diverse.", "Yes, It is. I love how it has so many different ingredients. It can be made with vegetables, meats, and condiments.", "Yes, that's true! It can also be made without cheese, which I find interesting.", "Yes, I love cheese. It is one of my favorite foods. I can't believe it's been around since the 5th century BCE."
Beam + Prefix	"I love it too! It's one of the most popular foods in the world. It's so good!", "I know, it's so tasty! I like to put it on crackers or bread!", "That's a good idea! I like it on toasted bread with some cheese on top.", "That sounds delicious! Do you like to add other ingredients to your cheese like vegetables, meats, and condiments?", "Yes, I love to add vegetables and meats to my cheese. I also like to make it into a sauce.", "That sounds really good! What kind of sauce do you use? I've heard of some people using mayonnaise, but I've never tried it."
CoSim	"I love it too! It's one of the most popular and influential in the world.", "It is! It is the most widely consumed food in the entire world. It has been around for a long time.", "It has! It was the first to be popular in Europe and the United States. It's so good!", "It is. I especially like the cheese that is used in the dish. Cheese is one of my favorite foods.", "It's a cheese! The most popular cheese in the US is cheddar. What's your favorite cheese?", "I like cheese. The chedd is the second most popular in the USA. My favorite cheese is probably mozzarella. How about you?"
Finetuned	"I love it too, especially the regional diversity of the food. What's your favorite dish?", "I love lasagna and spaghetti and meatballs. What about you?", "Lasagna is one of my favorites too! I also love pizza and pasta.", "Pasta is so versatile. It can be made with rice flour, wheat, or corn.", "That's true, I love all of those ingredients in my pasta! Do you have a favorite type of pasta?", "I really like spagetti. It's one of the most popular foods in the world."
HeLo-fixed	"I do too! I love all the pasta dishes and the flavors of cheese.", "Cheese is one of the most popular foods in the world. I love it too.", "I know! It's so versatile too. It can be made from milk, cream, and other things.", "Yes, and it can be aged for months at a time. It is so good.", "Yes, it can! It can also be made in a variety of ways, such as in a casserole.", "I love cassarole. It's one of my favorite foods to make. It consists of a large amount of cheese and other ingredients."
HeLo-anneal	"Me too! I love all the pasta dishes and the breads and cheese!", "I love pasta and cheese as well. It is one of the most popular foods in the world.", "Cheese is a staple food in Italy. I love it too.", "I love the variety of cheeses. I especially love the ones that are aged in wooden casks.", "Yes, I love those too. I also love the soft, cheddar cheese.", "I love Dairy products in general. So many types of cheese are available in the US."

Table 5: Sample infilled conversations from Wizard of Wikipedia. While both HeLo variants transition from the topic of Italian cuisine to cheese, only HeLo-anneal mentions the United States ("the US") to set up the target utterance.

Source **Target**	"I heard they are giving out some goodies in microkitchen." "Upstate New York. We got a foot of snow every week during my last winter there. I am so glad to not have to shovel snow now"
Human	"I love pop ups! What kind of goodies are they giving away?", "I guess its a jacket! Very much needed that in the cold ", "That's such a great idea, especially at this time of the year. I'm not too big a fan of the cold. I prefer warmer climates. Do you enjoy the cold?", "Sure hate it! Limits our ability to go out even for a walk! Its good that we don't get to suffer extreme cold weather!", "Me too! I moved here a few years ago to get out of the extreme cold. I do not care for bundling up and having to wear so many layers just to go buy eggs at the store.", "Oh nice! Where did you live before?"
Beam Search	"I heard that too, but I haven't tried it yet. I'm not sure if I want to though.", "I don't think I would either. I think I'll stick to my home cooked meals.", "I think that's probably a good idea. It's not healthy to eat out all the time.", "I agree. I try to cook at home as much as I can, but sometimes it's hard to find time.", "I know what you mean. There's so much to do at home. Do you have any hobbies?", "I like to play video games. I also like to read. What about you? What do you like to do?"
Beam + Prefix	"I'm from upstate new york too! I love the snow, but I'm glad I don't have to deal with it much anymore.", "I am glad you don't either. I have heard it can get pretty cold up there in the winter.", "It can, but it's not as bad as it used to be. The winters are much colder than the summers.", "That is good to hear. I would love to visit there someday. I hear it is the most populous city in the United States.", "It is. It's also the most densely populated state in the US.", "That's good to know. I have heard that it is also one of the fastest growing cities in the U.S. as well."
CoSim	"I have never heard of that, but it sounds like a good idea. I will have to check it out.", "I am not sure if it is a new thing, but they have been doing it for a while.", "I am sure it is new, but I am glad that they are doing it. It will be good for the environment.", "I am with you. It is good to be a part of the solution to our environment.", "I am so glad that we are on the right track. We need to make a difference.", "I am on to that. We have got a lot of people in this world that don't care. I think it is time to start making a change."
Finetuned	"I have never heard of that. What is it? I am in the midwest.", "It is a restaurant that specializes in Italian cuisine. I am also in the Midwest!", "I have heard of a few places like that, but I am not sure where they are located.", "They are all over the United States. They have restaurants in Canada, Mexico, Italy, Spain, and more!", "I have been to Mexico a few times, but never to Spain. I would love to go back.", "It is very nice. Mexico City is the most populous city in the country. You should go back!"
HeLo-fixed	"I heard that too, but I'm not sure if it's true or not.", "I hope not. I don't want to get poisoned. I heard it is a good place to eat.", "I hope so too. I've heard the same thing. I'll have to check it out.", "I hope you like it. I will have to try it too. Where are you from?", "I'm from Las Vegas, Nevada. How about you?", "I am from the midwest. I have never been to las vegas."
HeLo-anneal	"I heard that too, but I haven't been able to find any yet.", "I hope you can find some soon. It's a great place to find good food.", "I'm sure it is. I think I'll have to try it sometime. Do you have a favorite place to go?", "I love Thai food. They have so many different dishes there. I also like to go to Thailand.", "Where are you from originally? I've always wanted to go there, but never have.", "I'm originally from the United States. I've never been much for the colder weather. How about you?"

Table 6: Sample infilled conversations from Meena human-to-human chat logs. Both HeLo variants transition from the topic of food to location. Note how HeLo-anneal mentions "colder weather" and uses the transition question "how about you?" to set up the target utterance.

Decoding	c	λ_0	top-k	beam width
Beam Search	-	-	-	3
Beam + Prefix	-	-	-	3
CoSim	3	20	40	3
HeLo-fixed	0	15	40	3
HeLo-anneal	3	5	40	3

Table 7: Hyperparameters used in our experiments.

c	λ_0	BLEU	PPL	RSD	S-RSD	$\text{PPL}_{\boldsymbol{y}^{(1)}}$	$\text{PPL}_{\boldsymbol{y}^{\text{tgt}}}$
-1	5	3.71	4.94	1.16	2.21	2.94	24.88
	10	3.7	4.78	0.98	1.94	3.27	21.23
	15	3.9	4.59	0.86	1.81	3.57	18.21
	20	3.71	4.69	0.83	1.7	3.95	18.29
	25	3.38	4.62	0.72	1.62	4.07	16.29
	30	3.07	4.55	0.68	1.56	4.65	15.02
0	5	3.66	4.88	1.03	2.03	2.94	22.5
	10	4.09	4.51	0.83	1.74	3.27	17.95
	15	3.51	4.35	0.68	1.51	3.57	14.18
	20	3.49	4.57	0.65	1.42	3.95	14.31
	25	3.55	4.58	0.62	1.37	4.07	14.11
	30	2.98	4.61	0.57	1.28	4.65	12.85
1	5	3.88	4.77	0.88	1.87	2.94	20.82
	10	3.52	4.41	0.7	1.53	3.27	15.04
	15	3.4	4.46	0.61	1.29	3.57	13.39
	20	3.18	4.62	0.53	1.18	3.95	11.71
	25	3.22	4.82	0.49	1.06	4.07	11.34
	30	3.11	4.9	0.49	1.08	4.65	10.89
2	5	3.39	4.34	0.69	1.5	2.94	14.55
	10	3.64	4.6	0.59	1.26	3.27	12.81
	15	3.18	4.83	0.52	1.06	3.57	10.91
	20	3.2	5.05	0.47	0.93	3.95	9.16
	25	2.84	5.7	0.52	0.96	4.07	9
	30	2.48	6.17	0.52	0.97	4.65	9.01
3	5	3.72	4.39	0.57	1.19	2.94	11.5
	10	3.07	5.13	0.55	1.05	3.27	9.94
	15	2.97	5.71	0.58	0.99	3.57	8.91
	20	2.65	6.98	0.64	0.99	3.95	7.95
	25	2.56	7.78	0.71	1.08	4.07	7.53
	30	2.25	8.93	0.71	0.98	4.65	7.44
4	5	3.59	5.15	0.61	1.15	2.94	9.68
	10	2.86	7.75	0.89	1.22	3.27	8.28
	15	2.34	9.19	0.84	1.11	3.57	7.59
	20	2.19	11.95	0.96	1.14	3.95	6.83
	25	2.08	13.41	0.99	1.2	4.17	8.05

Table 8: Hyperparameter sweep for HeLo.

c	λ_0	BLEU	PPL	RSD	S-RSD	$\text{PPL}_{\boldsymbol{y}^{(1)}}$	$\text{PPL}_{\boldsymbol{y}^{\text{tgt}}}$
-1	5	3.13	9.01	1.32	1.88	2.85	45.5
	10	3.12	8.2	1.32	1.89	2.84	39.86
	20	2.75	8.25	1.29	1.85	2.89	39.63
	40	2.95	7.97	1.27	1.84	3.09	37.21
	80	2.73	7.78	1.12	1.73	3.69	33.51
	160	2.9	7.7	0.84	1.41	5.93	25.78
0	5	3.2	8.66	1.31	1.88	2.85	43.03
	10	3.04	8.99	1.31	1.87	2.84	45.26
	20	2.93	8.27	1.28	1.86	2.89	39.7
	40	3	7.56	1.19	1.79	3.09	33.57
	80	2.83	7.29	0.95	1.53	3.69	27.31
	160	2.67	8.03	0.61	1.1	5.93	20.42
1	5	3.05	9.54	1.33	1.89	2.85	49.21
	10	2.84	8.39	1.3	1.86	2.84	40.81
	20	2.96	8.12	1.23	1.8	2.89	37.98
	40	3	7.64	1.07	1.62	3.09	31.77
	80	2.86	7.58	0.76	1.23	3.69	23.32
	160	2.47	9.67	0.51	0.85	5.93	16.53
2	5	3.12	9.09	1.3	1.87	2.85	45.89
	10	3	8.62	1.24	1.81	2.84	41.67
	20	3.01	7.87	1.09	1.65	2.89	34.08
	40	2.92	7.5	0.83	1.24	3.09	24.1
	80	2.66	8.89	0.62	0.92	3.69	17.13
	160	2.07	13.35	0.6	0.81	5.93	11.96
3	5	3.12	8.38	1.24	1.8	2.85	40.06
	10	3.05	7.84	1.09	1.63	2.84	34.11
	20	3	7.6	0.88	1.29	2.89	26.31
	40	2.98	8.6	0.73	0.98	3.09	17.45
	80	2.45	12.65	0.75	0.89	3.69	12.41
	160	2.15	20.43	0.79	0.86	5.93	10.21
4	5	3.22	7.96	1.1	1.61	2.85	34.56
	10	2.96	7.44	0.9	1.29	2.84	25.02
	20	2.86	8.86	0.82	1.05	2.89	18.39
	40	2.74	12.7	0.9	0.98	3.09	13.63
	80	2.34	19.54	0.93	0.94	3.69	10.69
	160	1.99	27.88	0.86	0.84	5.93	9.36

Table 9: Hyperparameter sweep for CoSim.

k	BLEU ↑	PPL↓	RSD↓	S-RSD↓	$\text{PPL}_{\boldsymbol{y}^{(1)}}$↓	$\text{PPL}_{\boldsymbol{y}^{\text{tgt}}}$↓
5	3.30	4.72	1.87	0.93	**3.10**	19.68
10	3.38	<u>4.53</u>	1.81	0.84	<u>3.22</u>	**17.55**
20	3.49	4.66	1.83	0.87	3.23	19.19
40	**4.09**	**4.51**	<u>1.74</u>	<u>0.83</u>	3.27	<u>17.95</u>
80	<u>3.88</u>	4.61	**1.70**	**0.82**	3.25	18.06

Table 10: Hyperparameter sweep to set top-k for HeLo ($c = 0$, $\lambda_0 = 10$). Each value is an average over 50 conversations of Blended Skill Talk.