AN EQUAL-SIZE HARD EM ALGORITHM FOR DIVERSE DIALOGUE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Open-domain dialogue systems aim to interact with humans through natural language texts in an open-ended fashion. However, the widely successful neural networks may not work well for dialogue systems, as they tend to generate generic responses. In this work, we propose an Equal-size Hard Expectation–Maximization (EqHard-EM) algorithm to train a multi-decoder model for diverse dialogue generation. Our algorithm assigns a sample to a decoder in a hard manner and additionally imposes an equal-assignment constraint to ensure that all decoders are well-trained. We provide detailed theoretical analysis to justify our approach. Further, experiments on two large-scale, open-domain dialogue datasets verify that our EqHard-EM algorithm generates high-quality diverse responses[1].

1 INTRODUCTION

Open-domain dialogue systems aim to generate natural language text utterances to hold open-ended conversations with humans (Li et al., 2017a; Wang et al., 2021b). These systems have shown great success, and are seamlessly integrated into our society in the form of chatbots. For example, XiaoIce[2], a chatbot from Microsoft, has gained over 660 million active users from all over the world across more than 40 platforms since its launch in 2014 (Zhou et al., 2020). Recent advances in deep learning allow neural networks to be trained in an end-to-end fashion with large volumes of dialogue data, significantly improving the performance of open-domain dialogue systems (Serban et al., 2016; Wu et al., 2021).

However, neural dialogue systems tend to generate generic responses such as *I don't know* (Li et al., 2016b; Wang et al., 2021b). One of the possible causes is the *one-to-many mapping* phenomenon in the dialogue task, where a dialogue context may correspond to multiple plausible responses (Wei et al., 2019; Bao et al., 2020; Khan et al., 2020). Learning to generate a dialogue utterance is thus analogous to learning a *multimodal* distribution of a continuous variable, where a *mode* refers to a peak in the distribution; in the dialogue task, a mode can be thought of as a set of similar responses. The widely used cross-entropy training is not good at capturing different modes, as it encourages the prediction to cover all plausible responses, forcing the model to learn an overly smoothed distribution. Consequently, these neural dialogue models resort to generating generic responses (Wei et al., 2019).

Previous studies address the generic responses mainly at training time or inference time. Training-time approaches explore different training objectives: Li et al. (2016b) apply reinforcement learning to optimize a customized reward function that discourages generic responses. Khan et al. (2020) train a generative adversarial network (GAN) for dialogue generation, as GAN discourages generic responses because they can be easily identified by the discriminator. However, RL and GAN training is known to be difficult, and a recent study finds that "the generation performance of text GANs is not convincingly better, or even worse, than standard MLE training" (He et al., 2021)

Wang et al. (2021b) apply label smoothing to encourage less probable words to be generated. Among inference approaches, Vijayakumar et al. (2016) encourage diverse beam results by penalizing similar beams. Holtzman et al. (2020) apply nucleus sampling to allow plausible but less probable words to be decoded.

[1]Our code is available at `https://github.com/anonymous-1759/EqHard-EM`.
[2]`http://www.xiaoice.com`

Another way to address the generic responses is to explicitly capture different modes with mixture models. Shen et al. (2019) train a multi-decoder model with the Expectation–Maximization (EM) algorithm (Dempster et al., 1977) to address the one-to-many mapping phenomenon in the diverse machine translation setting. However, their direct application of the standard EM algorithms suffers from the *decoder-collapsing problem*, where the multi-decoder model degenerates to a single-decoder model. The authors attempt to alleviate this problem by disabling the dropout layers during the E-step of the EM algorithm, but their results show that the model is still susceptible to collapses.

To this end, we propose a novel EM variant to multi-decoder diverse dialogue generation. A standard EM algorithm assigns a sample to all decoders by the posterior probability, thus known as Soft-EM; it suffers from *synchronous-training* collapse, where the posterior probabilities tend to be similar and all decoders are also trained similarly. The Hard-EM variant trains the decoder having the highest posterior probability; it suffers from *non-training* collapse due to the rich-gets-richer phenomenon. Our proposed EM algorithm avoids both types of collapses by adopting hard assignments and imposing an equal-assignment constraint. Hence, we call our approach Equal-size Hard EM (EqHard-EM).

We conducted experiments on the Weibo (Gao et al., 2019) and OpenSubtitles (Lison et al., 2018) datasets. Results show that our EqHard-EM algorithm alleviates the decoder-collapsing problem that occurs in both Soft-EM and Hard-EM, allowing the multi-decoder model to specialize and generate high-quality and diverse responses. In addition, we provide in-depth theoretical and empirical analyses to better understand our EqHard-EM algorithm.

To sum up, our contributions are three-fold: 1) We propose the EqHard-EM algorithm to alleviate the collapse issues in soft and hard EM variants; 2) We provide detailed theoretical analysis to justify our equal assignment; and 3) We conduct extensive empirical experiments to show the effectiveness of our approach.

2 Related Work

Open-domain dialogue generation is a challenging task as neural dialogue models tend to generate generic responses such as "I don't know" (Li et al., 2016b; Wang et al., 2021b). In addition to the training-time and inference-time approaches mentioned in Section 1, another line of work addresses the problem by using cue words to guide the response generation. Yao et al. (2017) enhance the gated recurrent unit to fuse cue words that are heuristically selected based on point-wise mutual information. Gao et al. (2019) jointly train a cue word inference model along with the decoder network to further improve performance. Zou et al. (2021) apply non-autoregressive generation to incorporate cue words in the generated response. Incorporating commonsense knowledge is also investigated to tackle the generic responses. Zhou et al. (2018) use a large-scale knowledge graph to ground the generation process. Wu et al. (2020) make further improvements by retrieving and fusing relevant knowledge facts. These approaches do not directly address the one-to-many phenomenon in open-domain dialogue generation, but rely on heuristics to make the responses more meaningful.

Expectation–maximization (EM, Dempster et al., 1977) is a principled approach to training mixture models. The algorithm has been successfully applied in various fields (Dayan & Hinton, 1997; Sigg & Buhmann, 2008; Li et al., 2019). Specifically in natural language processing, EM is applied to unsupervised or semi-supervised training of hidden Markov models for sequential labeling (Rabiner & Juang, 1986), known as the classic Baum–Welch algorithm (Baum et al., 1970). Hofmann (1999) apply the EM algorithm for probabilistic latent semantic analysis (pLSA), where the topic is an unobserved variable. In the neural regime, the most related work to ours is Shen et al. (2019), as they apply EM to diverse machine translation. By contrast, we propose a theoretically justified EqHard-EM variant for diverse dialogue generation.

Another related (but different) research topic is mixture of experts (MoE, Jacobs et al., 1991; Hampshire & Waibel, 1992; Shazeer et al., 2017), where an expert k predicts a real-valued feature vector $\boldsymbol{e}_k$, mixed by predicted probabilities $\boldsymbol{p}$ as $\sum_k p_k \boldsymbol{e}_k$. The main difference between MoE and mixture models is that we mix decoders' probabilities rather than features, as $P(\mathbf{r}|\mathbf{c})$ can be thought of as the marginal probability $\sum_k P(z=k|\mathbf{c})P(\mathbf{r}|z=k,\mathbf{c})$. However, we do not adopt such factorization because a uniform prior of $P(z|\mathbf{c})$ makes the problem degrade to the optimization of $P(\mathbf{r}|z=k,\mathbf{c})$ individually for different ks. Admittedly, we might train a model to learn the mixing probability

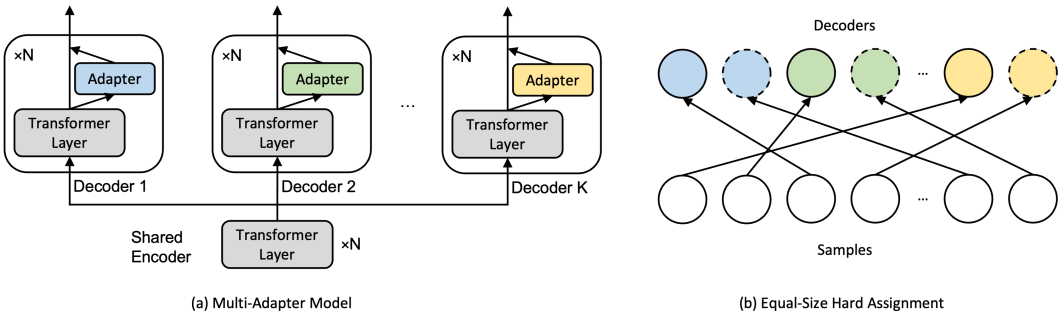

Figure 1: (a) Our multi-adapter neural architecture. (b) The equal-size hard assignment scheme. Dashed circles: decoders are conceptually duplicated when we solve the assignment problem.

$P(z = k|\mathbf{c})$, predicting which decoder is responsible for a context $\mathbf{c}$. This, however, violates our one-to-many assumption, which suggests one context requires multiple decoders. Therefore, our EM formulation does not predict the decoder z by merely the context $\mathbf{c}$, but determines which decoder fits a given training sample comprising both the context $\mathbf{c}$ and response $\mathbf{r}$.

3 Approach

In this section, we first present a multi-adapter architecture for the dialogue task (Subsection 3.1). Then, we propose EqHard-EM, a novel EM variant, for training such mixture models (Subsection 3.2).

3.1 Neural Architecture

We propose to address the one-to-many mapping phenomenon in the dialogue task with a mixture model. Given a dialogue context, a shared encoder builds an input utterance's hidden representations, based on which mutiple decoders generate a set of output responses. Specifically, we adopt the adapter architecture (Houlsby et al., 2019), where all decoders share Transformer parameters but are inserted with thin adapter layers serving as a mixture component.

For multiple decoders, it is possible to instantiate each with a full-Transformer model (Shen et al., 2019), but this leads to a large number of parameters that may not fit into the memory of an ordinary GPU. To this end, we propose a multi-adapter architecture (Figure 1a), where different decoders share most Transformer parameters, but differ by a few inserted, thin adapter layers (Houlsby et al., 2019). In fact, the adapter model is widely used for parameter-efficient domain adaptation (Artetxe et al., 2020; Wang et al., 2021a; Ngo Trung et al., 2021). We propose to apply this architecture to mixture models. Our preliminary experiments show that the multi-adaptor model results in less than one point degradation in BLEU scores, but reduces 76% parameters compared with full-Transformer decoders. Therefore, we use the multi-adaptor model in our main experiments.

Specifically, an adaptor layer applies two projections of linear transformation and non-linear activation. Further, a residual connection is included to ensure that the Transformer layers are not significantly disturbed. This process can be formulated as

$$\text{AdapterLayer}(\mathbf{x}) = \mathbf{x} + \mathbf{W}_1(\phi(\mathbf{W}_2\mathbf{x})) \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^d$ is the input to the adapter layer; $\mathbf{W}_1 \in \mathbb{R}^{d\times d'}$ and $\mathbf{W}_2 \in \mathbb{R}^{d'\times d}$ are the projection matrices with $d' < d$; and ϕ is a non-linear activation unit such as ReLU (Agarap, 2018).

3.2 Training Methods

We propose to train our multi-adapter architecture in two stages: 1) Pretraining of the shared Transformer-layer parameters, and 2) EM training of adapter-layer parameters as the mixture model.

In the first stage, we pretrain a T5-small[3] encoder–decoder model (Raffel et al., 2020) with a standard cross-entropy loss since the Transformer parameters are shared among different mixture components; in the second stage, we adopt the standard treatment of adapter modules (Houlsby et al., 2019) and freeze the shared Transformer parameters to minimize the number of trainable parameters for effective adaptation. In the rest of this subsection, we focus on the second stage, i.e., training the adapter parameters of the mixture model.

Consider a multi-decoder model with K decoders. Given a dialogue context $\mathbf{c}$ and response $\mathbf{r}$, our training objective is to maximize the marginal log-likelihood $\log \sum_k P(\mathbf{r}, z = k|\mathbf{c}; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the model parameters and the variable $z \in \{1, \cdots, K\}$ is the choice of the decoder. Since z is unobserved in the training set, it can be viewed as a latent variable, and we apply the expectation–maximization (EM) framework (Dempster et al., 1977) as a principled approach to training latent-variable models. Specifically, the EM algorithm alternates between the E-step and the M-step:

E-step: Estimate the posterior distribution of the latent variable by applying Bayes' rule:

$$Q_k(\mathbf{c}, \mathbf{r}) = P(z = k|\mathbf{c}, \mathbf{r}; \boldsymbol{\theta}) = \frac{P(\mathbf{r}|z = k, \mathbf{c}; \boldsymbol{\theta})P(z = k|\mathbf{c}; \boldsymbol{\theta})}{\sum_{k'=1}^{K} P(\mathbf{r}|z = k', \mathbf{c}; \boldsymbol{\theta})P(z = k'|\mathbf{c}; \boldsymbol{\theta})} \tag{2}$$

M-step: Maximize the expected log-likelihood with respect to $\boldsymbol{\theta}$:

$$\mathbb{E}_{z\sim Q_k(\mathbf{c},\mathbf{r})}[\log P(\mathbf{r}, z|\mathbf{c}; \boldsymbol{\theta})] = \sum\nolimits_{k=1}^{K} Q_k(\mathbf{c}, \mathbf{r}) \log P(\mathbf{r}, z|\mathbf{c}; \boldsymbol{\theta}) \tag{3}$$

We additionally assume a uniform prior for every context $\mathbf{c}$ as $P(z = k|\mathbf{c}; \boldsymbol{\theta}) = 1/K$, following the one-to-many assumption that each context has multiple plausible responses. This assumption simplifies Eqn. (2) as

$$P(z = k|\mathbf{c}, \mathbf{r}; \boldsymbol{\theta}) = \frac{P(\mathbf{r}|z = k, \mathbf{c}; \boldsymbol{\theta})}{\sum_{k'=1}^{K} P(\mathbf{r}|z = k', \mathbf{c}; \boldsymbol{\theta})} \tag{4}$$

suggesting that the posterior is proportional to the likelihood under the uniform prior assumption.

Soft-EM. The classic EM algorithm, as explained above, assigns a sample to a mixture component by the posterior distribution (2) in a soft manner, and thus is also known as Soft-EM. However, we observe that Soft-EM performs poorly in our experiments due to **synchronous-training collapse**[4], where all decoders are trained almost synchronously and generate similar responses. In particular, the Transformer-layer parameters are pretrained. Even with the inserted adapter layers, all decoders still behave similarly, yielding a similar posterior distribution (2). This in turn yields a similar gradient for each decoder, and as a result, multiple decoders are virtually collapsed to one decoder.

Hard-EM. Alternatively, the assignment in the E-step can be done in a hard manner by choosing the best-fit mixture component

$$k^* = \text{argmax}_k \, P(z = k|\mathbf{c}, \mathbf{r}; \boldsymbol{\theta}) \tag{5}$$

Then, the M-step is approximated as follows

$$\mathbb{E}_{z\sim P(z|\mathbf{c},\mathbf{r};\boldsymbol{\theta})}[\log P(\mathbf{r}, z|\mathbf{c}; \boldsymbol{\theta})] \approx \log P(\mathbf{r}, z = k^*|\mathbf{c}; \boldsymbol{\theta}) \tag{6}$$

Unfortunately, the Hard-EM algorithm suffers from the **non-training collapse**, where only one decoder is trained due to the rich-gets-richer phenomenon. In particular, an arbitrary decoder will be selected more often than the others by Eqn. (5) due to the random initialization of adaptor layers. Then, that decoder will be trained with more samples and perform better than the other decoders, making it even more likely to be selected by Eqn. (5). In practice, we observe that only one decoder is properly trained by the Hard-EM algorithm, making the mixture model ineffective.

[3] In principle, other pretrained language models may also be used. In our development, we have tried finetuning GPT-2 on the OpenSubtitles dataset. It gave a similar, but slightly lower, result: 3.26 BLEU2-F vs T5-small's 3.46 BLEU2-F. Therefore, we chose T5-small as the base model, because it achieves higher performance with fewer parameters.

[4] Shen et al. (2019) refer to this phenomenon as the "D2 Degeneracy" and interpret it as the bypassing of latent variables, similar to the posterior collapse of variational auto-encoders (VAE, Bowman et al., 2016). We find the connection with VAE weak. Instead, we present our own interpretation of synchronous training, which is justified by empirical evidence.

Our proposed EqHard-EM. To this end, we propose an Equal-size Hard Expectation–Maximization (EqHard-EM) algorithm to address the collapse issues. Specifically, we adopt the hard assignments from Hard-EM to avoid synchronous-training collapse, whereas we impose an equal-assignment constraint to ensure that all decoders are adequately trained to avoid non-training collapse.

We formulate the E-step as a constrained optimization problem over an assignment matrix $\mathbf{A} \in \{0,1\}^{N\times K}$, where N is the number of samples and K is the number of decoders. The element A_{nk} indicates whether the nth sample is assigned to the kth decoder. We define the cost matrix $\mathbf{C} \in \mathbb{R}^{N\times K}$ by the posterior probability, i.e., $C_{nk} = -Q_k(\mathbf{r}^{(n)}, \mathbf{c}^{(n)})$ as in Eqn. (2). Our equal-size hard E-step aims to

$$\begin{aligned}
\text{minimize}_{\mathbf{A}} \quad & \textstyle\sum_{k=1}^{K}\sum_{n=1}^{N} A_{nk}C_{nk} && (7)\\
\text{subject to} \quad & \textstyle\sum_{k=1}^{K} A_{nk} = 1, \quad \text{for } n = 1,\cdots,N && (8)\\
& \textstyle\sum_{n=1}^{N} A_{nk} = N/K, \quad \text{for } k = 1,\cdots,K && (9)\\
& A_{nk} \in \{0,1\} && (10)
\end{aligned}$$

Compared with Hard-EM, we impose the constraint (9) to ensure that every decoder is assigned with an equal number of samples.

In practice, this optimization problem can be transformed to a balanced assignment problem, and in turn solved by the classic Hungarian algorithm (Kuhn, 1955). The algorithm was later improved by Jonker & Volgenant (1987) to achieve a computational complexity of $O(N^3)$, which is efficient in our batch implementation and contributes to less than 1% of the total training time. As shown by Figure 1b, the reformulation is accomplished by duplicating each decoder and its associated costs by a factor of $\frac{N}{K}$ to match the number of samples. Here, we enforce the batch size N to be divisible by K. Note that the decoders are only conceptually duplicated when we apply the Hungarian algorithm. The actual decoder network is not duplicated, and it does not add to computational cost.

For inference, we follow existing literature (Zhao et al., 2017; Gu et al., 2019; Shen et al., 2019) and consider the multi-response generation setting, where the goal is to generate a set of diverse responses given a dialogue context. We collect the outputs of different decoders by greedy decoding.

Theoretical analysis. We provide theoretical justifications for our EqHard-EM algorithm. Our equal-assignment constraint makes sense intuitively, because the aggregated posterior should be the same as the prior distribution, suppose the posterior distribution is correctly estimated and we have enough samples. This is given by $\sum_{\mathbf{c}}\sum_{\mathbf{r}} P(z|\mathbf{r},\mathbf{c})P(\mathbf{r},\mathbf{c}) = \sum_{\mathbf{c}} P(z|\mathbf{c})P(\mathbf{c}) = \sum_{\mathbf{c}}(1/K)P(\mathbf{c}) = 1/K$ for $z = 1,\cdots,K$. In this part, we develop a theorem on how equal the assignments would be in the imperfect-posterior and finite-sample cases.

Theorem 1. *With probability at least $1-\delta$, we have*

$$\left|\frac{1}{|\mathcal{D}|}\sum_{(\mathbf{r},\mathbf{c})\in\mathcal{D}} Q(z|\mathbf{r},\mathbf{c}) - \frac{1}{K}\right| \le \sqrt{\frac{1}{2|\mathcal{D}|}\log\left(\frac{2K}{\delta}\right)} + \mathbb{E}_{\mathbf{c},\mathbf{r}}\Big[\big|Q(z|\mathbf{c},\mathbf{r}) - p(z|\mathbf{c},\mathbf{r})\big|\Big]$$

for all $z \in \{1,\cdots,K\}$. Here, $Q(z|\mathbf{c},\mathbf{r})$ is an estimated posterior and $\mathcal{D}$ is a dataset with finite samples.

The proof sketch is to consider finite samples (Lemma 1) and the estimated posterior (Lemma 2) separately, which can then be combined in a straightforward manner. See Appendix A for the detailed proof.

Theorem 1 justifies the equal-assignment constraint, because with a high probability, the assignments will converge to uniform if we have a large dataset and a correct posterior. More importantly, the difference is at most linear in the expected error of the estimated posterior $Q(z|\mathbf{c},\mathbf{r})$, such as the one in the E-step, and is sub-linear in $1/|\mathcal{D}|$, supporting the use of our approach in practice. While the term $\sum_{(\mathbf{r},\mathbf{c})\in\mathcal{D}} Q(z|\mathbf{r},\mathbf{c})$ mainly works for soft assignment, we could not adopt soft assignment because it leads to synchronous-training collapse. Nevertheless, the theorem is illuminating to our algorithm as well, as it is a hard-assignment approximation similar to Hard-EM.

4 EXPERIMENTS

4.1 SETUP

Datasets. We evaluated our approach on two large-scale open-domain dialogue datasets: Weibo (Gao et al., 2019) and OpenSubtitles (Lison et al., 2018). The Weibo dataset is a compilation of posts and replies, crawled from a Chinese micro-blogging website. The dataset features explicit one-to-many mapping because in most cases, a post corresponds to multiple replies. On the other hand, the OpenSubtitles dataset is a collection of movie dialogues, which are parsed from online subtitle files. The dataset is known to be noisy (Vinyals & Le, 2015; Feng et al., 2020; Nakazawa et al., 2021), and therefore it demonstrates the implicit one-to-many phenomenon: a movie utterance may lead to many plausible responses even though the dataset only contains one reference response. Admittedly, there exist other manually constructed, high-quality datasets such as DailyDialog (Li et al., 2017b), but they are usually much smaller and the one-to-many phenomenon is less severe.

Recently, Wen et al. (2022) identify significant overlapping between training, validation, and/or test splits in various open-domain dialogue datasets (including OpenSubtitles) due to the overlights of data collection. We use their cleaned OpenSubtitles dataset, which contains 1M, 12K, and 12K samples for training, validation, and test, respectively. We investigated the Weibo dataset and found significant overlapping between the validation and test sets as well, so we followed their cleaning strategy and obtained 4M, 10K, and 10K training, validation, and test samples. Overall, Weibo and OpenSubtitles allow us to conduct comprehensive analyses in both explicit and implicit one-to-many mapping scenarios and in different languages.

Metrics. Open-domain dialogue evaluation is currently an active area of research due to its one-to-many nature (Tao et al., 2018; Sinha et al., 2020). In particular, the standard BLEU metric, which measures the n-gram similarity between generated and reference responses, has low correlation with human evaluation because a plausible response is not necessarily similar to the reference response (Liu et al., 2016). Therefore, we followed previous work in open-domain dialogue (Zhao et al., 2017; Gu et al., 2019; Li et al., 2020) and generated 10 responses for each dialogue context to compute BLEU-P, BLEU-R, and BLEU-F. Roughly, BLEU-P measures the average quality of the generated responses, whereas BLEU-R measures how well the references are covered. BLEU-F is the harmonic mean.[5] We adopted uni-gram and bi-gram BLEU scores in our experiments, because Liu et al. (2016) show that higher-order BLEU scores are often zero, making the metric rely mostly on smoothing techniques.

For diversity, we adopt the Dist (Li et al., 2016a) and Pairwise-BLEU (Shen et al., 2019) scores. Dist-n is a widely used metric that measures the percentage of distinct n-grams among utterances. We report Dist-1 and Dist-2 between the generated responses for each dialogue context. In addition, we follow Shen et al. (2019) and report Pairwise-BLEU, which measures the pairwise similarity (measured in BLEU) among the responses, i.e., a low Pairwise-BLEU indicates high diversity.

Implementation details. We considered 10 decoders due to the setup of diverse generation evaluation (Zhao et al., 2017; Gu et al., 2019; Li et al., 2020). Each E-step involved 640 samples, assigning 64 samples to each decoder. More implementation details are presented in Appendix B.

4.2 RESULTS AND ANALYSES

Main results. Table 1 shows our main results on Weibo and OpenSubtitles. The T5-small model is a pretrained encoder–decoder Transformer (Raffel et al., 2020), where we experimented both beam search and sampling. For beam search, we set the beam size to be 10, i.e., the number of samples needed for evaluation. As known, beam search does not generate diverse responses (Vijayakumar et al., 2016; Holtzman et al., 2020), shown by low BLEU-R, low Dist scores, but high Pairwise-BLEU. The performance of nucleus sampling is inconsistent in the two experiments: on the Weibo dataset, nucleus sampling improves the diversity to a certain extent, but its BLEU-F and Dist scores are inadmissibly worse than our approach; on OpenSubtitles, nucleus sampling gener-

[5]The terminologies of BLEU-P, BLEU-R, and BLEU-F are inspired by, but different from, the precision and recall in information retrieval (Manning et al., 2008). Interested readers are referred to Zhao et al. (2017) for the details of BLEU-P, BLEU-R, and BLEU-F.

Dataset	Model	BLEU-1†			BLEU-2†			Distinct n-gram†		Pairwise
		F	P	R	F	P	R	Dist-1	Dist-2	BLEU$^{\downarrow}$
Weibo	T5-small beam search (Raffel et al., 2020)	18.59	25.59	14.60	8.51	**17.57**	5.62	10.41	14.17	65.46
	T5-small nucleus sampling (Holtzman et al., 2020)†	20.76	26.13	17.23	10.03	17.44	7.04	16.90	23.79	44.55
	AdaLabel (Wang et al., 2021b)	14.80	21.63	11.25	5.86	13.72	3.72	10.41	14.71	62.46
	DialogBERT (Gu et al., 2021)†	15.99	15.88	16.10	3.55	3.95	3.23	68.95	96.76	0.01
	Multi-adapter + EqHard-EM (ours)	**22.65**	**26.17**	**19.96**	**10.13**	14.55	**7.77**	27.03	42.02	22.32
Open-Subtitles	T5-small beam search (Raffel et al., 2020)	14.39	11.23	20.01	3.46	**2.40**	6.22	13.38	20.71	42.76
	T5-small nucleus sampling (Holtzman et al., 2020)†	16.03	11.21	28.13	2.51	1.47	8.48	60.50	91.26	0.29
	AdaLabel (Wang et al., 2021b)	15.80	11.94	23.38	3.20	2.19	5.94	26.47	43.57	17.21
	DialogBERT (Gu et al., 2021)†	12.86	8.71	24.50	1.20	0.68	4.94	69.70	96.49	0.02
	Multi-adapter + EqHard-EM (ours)	**17.86**	**12.98**	**28.60**	**3.76**	2.36	**9.27**	34.09	50.41	13.15

Table 1: Results on Weibo and OpenSubtitles. † indicates sampling-based approaches; DialogBERT performs poorly with beam search or greedy decoding, also reported by Wen et al. (2022). $^{\uparrow/\downarrow}$The higher/lower, the better. Underlined numbers are the best diversity scores (Dist and Pairwise-BLEU) among non-sampling methods. Note that comparing diversity is only meaningful when BLEU scores are controlled.

ates low-quality texts (very low BLEU-F), which are diverse among themselves.[6] We examined the samples and find two plausible reasons for the inconsistency: 1) The Weibo dataset is in Chinese and BLEU scores measure the overlap of Chinese characters. This is more lenient than English words, making it possible to retain relatively high BLEU-F scores. 2) Nucleus sampling highly depends on the distribution temperature and the nucleus size for sampling. If the sampling scope is too small, then the diversity would not be improved much; if otherwise the sampling scope is too large, the quality would be deteriorated.

By contrast, our EqHard-EM approach is able to perform consistently well on both datasets by simple greedy decoding. Our BLEU-R is much higher than most competing methods. This shows that our EqHard-EM has a high coverage of potential responses, largely alleviating the one-to-many issue in dialogue generation. Overall, we achieve the state-of-the-art BLEU-F scores, while at the same time largely improving diversity. In fact, we outperform all non-sampling methods in terms of Dist and Pairwise-BLEU scores.

We also experimented with recently proposed models: AdaLabel (Wang et al., 2021b) is an encoder–decoder Transformer, trained with adaptive label smoothing to improve diversity. DialogBERT (Gu et al., 2021) features the use of a BERT encoder (Devlin et al., 2019) to capture contextual information. They were alleged to be the state-of-the-art models when the overlapping problem was not realized. However, they perform poorly on the deduplicated datasets; our results in the diverse generation setting are consistent with single-response generation (Wen et al., 2022).

Comparing EM variants. We conducted controlled experiments to compare EM-like algorithms (Table 2). We do not adopt the recurrent dropout trick (Shen et al., 2019) because we find them ineffective in our dialogue setting (see Appendix E). We find Soft-EM suffers from synchronous-training collapse, as it achieves very high Pairwise-BLEU scores. It achieves even lower diversity than beam search because of the synchronous-training collapse, which is characterized by its Pairwise-BLEU scores. This can be best seen by decoder assignment statistics (Figure 2a), where every decoder is assigned with an equal fraction of samples with near-zero standard deviation.

On the other hand, the Hard-EM algorithm also performs poorly. It suffers from the non-training collapse, due to the rich-gets-richer phenomenon. As seen in Figure 2b, only Decoder 6 is trained in the experiment. Hard-EM achieves near-zero Pairwise-BLEU and the lowest Dist scores because the untrained decoders tend to repeat random words, which can be verified in our case studies (Appendix D). As a result, the generated sentences are not meaningful, shown by extremely low BLEU scores.

As shown in Figure 2c, our EqHard-EM guarantees that all decoders are trained with the same number of samples, preventing non-training collapse; since our algorithm adopts hard assignments (either 0 or 1), it also has a high variance among different samples, preventing from synchronous-training collapse. Overall, our EqHard-EM achieves the highest BLEU-F score, as well as the best diversity scores (excluding the improperly trained Hard-EM). Appendix D presents sample responses generated by Soft-EM, Hard-EM, and EqHard-EM.

[6]This shows diversity scores (Dist and Pairwise-BLEU) are auxiliary metrics. They cannot serve as a standalone evaluation, as the metrics can be easily fooled by uniformly random sampling.

Dataset	Model	BLEU-1↑			BLEU-2↑			Distinct n-gram↑		Pairwise
		F	P	R	F	P	R	Dist-1	Dist-2	BLEU↓
Weibo	Beam search	18.59	25.59	14.60	8.51	**17.57**	5.62	10.41	14.17	65.46
	Soft-EM (Shen et al., 2019)	17.02	**27.55**	12.32	6.27	16.03	3.89	10.69	12.45	86.25
	Hard-EM (Shen et al., 2019)	4.81	3.09	10.84	2.15	1.61	3.22	4.01	4.11	0.00
	EqRandom-Dynamic	18.72	27.34	14.23	7.21	15.87	4.66	12.79	16.30	71.19
	EqRandom-Fixed	20.01	25.51	16.46	8.12	14.41	5.65	19.40	25.77	46.60
	EqHard-EM	**22.65**	26.17	**19.96**	**10.13**	14.55	**7.77**	27.03	42.02	22.32
OpenSubtitles	Beam search	14.39	11.23	20.01	3.46	2.40	6.22	13.38	20.71	42.76
	Soft-EM (Shen et al., 2019)	15.48	14.55	16.54	3.03	**2.76**	3.35	12.89	15.35	72.95
	Hard-EM (Shen et al., 2019)	6.84	4.13	19.81	1.25	0.73	4.30	4.83	4.81	1.50
	EqRandom-Dynamic	15.80	14.56	17.27	2.92	2.56	3.40	14.52	18.57	70.12
	EqRandom-Fixed	15.89	**14.59**	17.46	2.93	2.56	3.43	14.73	18.96	68.89
	EqHard-EM	**17.86**	12.98	**28.60**	**3.76**	2.36	**9.27**	34.09	50.41	13.15

Table 2: Comparing EM-like algorithms based on multi-decoder T5-small. The Soft-EM and Hard-EM variants follow Shen et al. (2019), who address diverse machine translation.

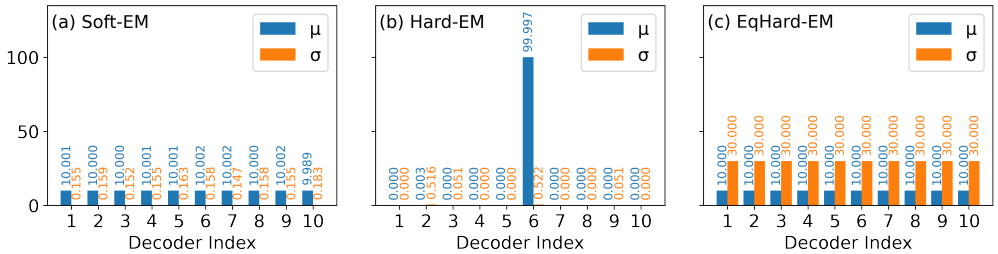

Figure 2: Decoder assignment statistics for (a) Soft-EM, (b) Hard-EM, and (c) EqHard-EM. Means and standard deviations are computed on the Weibo training set.

We further investigate whether EM-style, posterior-based assignment actually helps. We compare our EqHard-EM with two other equal-size hard assignment methods: `EqRandom-Fixed` and `EqRandom-Dynamic`. The `EqRandom-Fixed` variant pre-assigns samples to a decoder before training; this is essentially training each decoder on a partition of the training set. `EqRandom-Dynamic` assigns samples to decoders randomly and dynamically during training; this is essentially training a model with multiple initializations, if the number of epochs is large. In both cases, we ensure each decoder gets the same number of samples, so that they are equal-size hard assignments. As seen, these two methods perform similarly to each other. They are slightly worse than beam search, but much worse than our EqHard-EM in terms of both quality and diversity.

These controlled experiments convincingly show that our EqHard-EM is able to take advantage of the principled EM algorithm to learn a mixture of decoders, and that our model does not suffer from the collapse issue.

Analyzing different decoders. We analyze how decoders are specialized in certain aspects of dialogue responses. Admittedly, the responses generated by different decoders would not exhibit clear clustering phenomena, because our inputs are open-domain dialogue utterances covering a wide range of topics. The generated responses have to fit the input in a certain way, so the between-sample variance would be larger than decoder variance. Despite this, we do find obvious patterns of certain decoders, including the length,[7] sentence structures (e.g., whether it is a question), and the preference of certain words (Figure 3). For example, Decoder 7 tends to generate short question responses, whereas Decoders 9 and 10 tend to generate positive-sentiment utterances with the word "like/love". For comparison, we also show beam search results, which are nearly uniform among different samples in the beam. Our decoder analysis, despite its simplicity, is conducted in a quantitative fashion, as opposed to previous case study-based claims (Gu et al., 2019; Shen et al., 2019). This provides rigorous evidence that our EqHard-EM is able to train multiple decoders to capture different "modes" of open-domain dialogue utterances.

Effect of the number of decoders. We investigate how well our approach scales with the number of decoders. Figure 4 shows a comparison between beam search and our multi-decoder model under

[7]We find the EqHard-EM generates slightly shorter responses than beam search (Figures 3a and 3d). By examining the utterances, we find beam search tends to generate repetitive words such as "ha ha ha. . ."

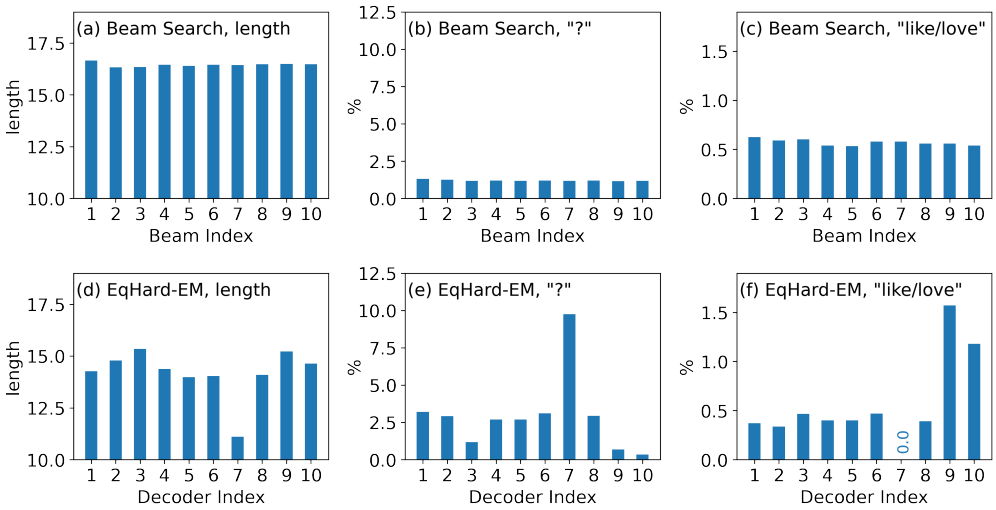

Figure 3: Analyzing different decoders trained by EqHard-EM in comparison to different samples obtained by beam search on the Weibo dataset. Subplots (a) and (d) compare the length of generations; Subplots (b) and (e) compare the percentages of the question mark; and Subplots (c) and (f) compare the percentages of the word "like/love" (喜欢) generated.

5-, 10-, and 15-generation settings. Due to the limit of space and resources, we conducted the analysis only on the Weibo dataset. As seen, our multi-decoder model performs similarly to beam search when there are only 5 decoders. However, the performance gap increases when we have more decoders, which verifies the need for our multi-decoder model.

Additional results. We conducted human evaluation (Appendix C), where results are highly consistent with automatic metrics, showing that our model generates high-quality and diverse responses. Appendix D shows example outputs of different EM algorithms, confirming that our EqHard-EM variant has largely alleviated the collapse issues. Appendix E shows an ablation study for the recurrent dropout trick (Shen et al., 2019).

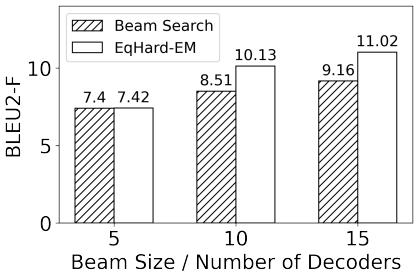

Figure 4: Comparison between beam search and our multi-decoder model in 5-, 10-, and 15-generation settings on the Weibo dataset.

5 Conclusion

In this work, we propose an Equal-size Hard Expectation–Maximization (EqHard-EM) algorithm for multi-decoder diverse dialogue generation. Our EqHard-EM adopts the hard assignment as in Hard-EM, but additionally imposes an equal-assignment constraint to ensure all decoders are well-trained. We provide detailed theoretical analysis to justify our algorithm. Extensive experiments on two large-scale datasets further confirm that our EqHard-EM algorithm enables decoders to specialize and generate high-quality and diverse responses.

Limitation and future work. Our paper addresses the generic responses for dialogue systems by training a mixture of decoders. This inevitably slows down the training process similar to standard EM algorithms (Shen et al., 2019), because we need to perform inference for all decoders even though our EqHard-EM propagates gradient to one. However, our Hungarian algorithm has little overhead (less than 1%) for the training process, which is negligible. It should also be mentioned that, when the model is deployed, the inference speed would not be affected much, because different decoders can be parallelized if samples come one by one. Moreover, our multi-adapter architecture saves 76% memory without hurting much performance compared with full-Transformer decoders.

In the future, we plan to explore more effective initialization by meta-learning to better specialize decoders during our EqHard-EM training.

6 Reproducibility Statement

We provided a detailed proof of Theorem 1 in Appendix A. For reproducing our results, we published the source code of our work as a Git repository (see the footnote from the abstract). The implementation details are also thoroughly discussed in Appendix B. We further conducted training and inference efficiency experiments on an i9-9900X CPU and a TITAN RTX GPU. Notably, the computational overhead of the Hungarian algorithm is negligible, as the E-step, Hungarian algorithm, and M-step take 75.69%, 0.97%, and 24.31% of the total training time, respectively. At inference, EqHard-EM is actually more efficient than beam search of a single-decoder model. Our multi-adapter model processes 79.1 samples per second, whereas beam search processes 16.9 samples per second. Our approach achieves a 4.7x speedup because inference for each decoder can be computed in parallel across multiple GPUs, whereas beam search cannot leverage the same level of parallelism.

References

Abien Fred Agarap. Deep learning using rectified linear units (ReLU). *arXiv preprint arXiv:1803.08375*, 2018. URL https://arxiv.org/abs/1803.08375.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. In *ACL*, pp. 4623–4637, 2020. URL https://aclanthology.org/2020.acl-main.421.

Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. PLATO: Pre-trained dialogue generation model with discrete latent variable. In *ACL*, pp. 85–96, 2020. URL https://aclanthology.org/2020.acl-main.9.

Leonard E. Baum, Ted Petrie, George Soules, and Norman Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171, 1970. URL http://www.jstor.org/stable/2239727.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. In *CoNLL*, pp. 10–21, 2016. URL https://aclanthology.org/K16-1002.

Peter Dayan and Geoffrey E. Hinton. Using Expectation-Maximization for Reinforcement Learning. *Neural Computation*, 9(2):271–278, 1997. URL https://doi.org/10.1162/neco.1997.9.2.271.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–22, 1977. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1977.tb01600.x.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pp. 4171–4186, 2019. URL https://aclanthology.org/N19-1423.

Shaoxiong Feng, Hongshen Chen, Kan Li, and Dawei Yin. Posterior-GAN: Towards informative and coherent response generation with posterior generative adversarial network. In *AAAI*, pp. 7708–7715, 2020. URL https://ojs.aaai.org/index.php/AAAI/article/view/6273.

Jun Gao, Wei Bi, Xiaojiang Liu, Junhui Li, and Shuming Shi. Generating multiple diverse responses for short-text conversation. In *AAAI*, pp. 6383–6390, 2019. URL https://ojs.aaai.org/index.php/AAAI/article/view/4601.

Xiaodong Gu, Kyunghyun Cho, Jung-Woo Ha, and Sunghun Kim. DialogWAE: Multimodal response generation with conditional Wasserstein auto-encoder. In *ICLR*, 2019. URL https://openreview.net/forum?id=BkgBvsC9FQ.

Xiaodong Gu, Kang Min Yoo, and Jung-Woo Ha. DialogBERT: Discourse-aware response generation via learning to recover and rank utterances. In *AAAI*, pp. 12911–12919, 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/17527.

John B. Hampshire and Alex Waibel. The meta-pi network: Building distributed knowledge representations for robust multisource pattern recognition. *TPAMI*, 14(7):751–769, 1992. URL https://ieeexplore.ieee.org/document/142911.

Tianxing He, Jingzhao Zhang, Zhiming Zhou, and James Glass. Exposure bias versus self-recovery: Are distortions really incremental for autoregressive text generation? In *EMNLP*, pp. 5087–5102, 2021. URL https://aclanthology.org/2021.emnlp-main.415.

Thomas Hofmann. Learning the similarity of documents: An information-geometric approach to document retrieval and categorization. In *NeurIPS*, pp. 915–920, 1999. URL https://proceedings.neurips.cc/paper/1999/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *ICLR*, 2020. URL https://openreview.net/pdf?id=rygGQyrFvH.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *ICML*, pp. 2790–2799, 2019. URL http://proceedings.mlr.press/v97/houlsby19a.html.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991. URL https://doi.org/10.1162/neco.1991.3.1.79.

R. Jonker and A. Volgenant. A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing*, pp. 325–340, 1987. URL https://doi.org/10.1007/BF02278710.

Kashif Khan, Gaurav Sahu, Vikash Balasubramanian, Lili Mou, and Olga Vechtomova. Adversarial learning on the latent space for diverse dialog generation. In *COLING*, pp. 5026–5034, 2020. URL https://aclanthology.org/2020.coling-main.441.

Harold W Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955. URL https://onlinelibrary.wiley.com/doi/10.1002/nav.3800020109.

Chunyuan Li, Xiang Gao, Yuan Li, Baolin Peng, Xiujun Li, Yizhe Zhang, and Jianfeng Gao. Optimus: Organizing sentences via pre-trained modeling of a latent space. In *EMNLP*, pp. 4678–4699, 2020. URL https://aclanthology.org/2020.emnlp-main.378.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *NAACL-HLT*, pp. 110–119, 2016a. URL https://aclanthology.org/N16-1014/.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. Deep reinforcement learning for dialogue generation. In *EMNLP*, pp. 1192–1202, 2016b. URL https://aclanthology.org/D16-1127.

Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. Adversarial learning for neural dialogue generation. In *EMNLP*, pp. 2157–2169, 2017a. URL https://aclanthology.org/D17-1230.

Xia Li, Zhisheng Zhong, Jianlong Wu, Yibo Yang, Zhouchen Lin, and Hong Liu. Expectation-Maximization attention networks for semantic segmentation. In *ICCV*, pp. 9167–9176, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Li_Expectation-Maximization_Attention_Networks_for_Semantic_Segmentation_ICCV_2019_paper.html.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. DailyDialog: A manually labelled multi-turn dialogue dataset. In *IJCNLP*, pp. 986–995, 2017b. URL https://aclanthology.org/I17-1099/.

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *LREC*, pp. 1742–1748, 2018. URL https://aclanthology.org/L18-1275.pdf.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *EMNLP*, pp. 2122–2132, 2016. URL https://aclanthology.org/D16-1230.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008. URL https://dl.acm.org/doi/10.5555/1394399.

Toshiaki Nakazawa, Hideki Nakayama, Chenchen Ding, Raj Dabre, Shohei Higashiyama, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, Chenhui Chu, Akiko Eriguchi, Kaori Abe, Yusuke Oda, and Sadao Kurohashi. Overview of the 8th workshop on Asian translation. In *Proceedings of the Workshop on Asian Translation*, pp. 1–45, 2021. URL https://aclanthology.org/2021.wat-1.1.

Nghia Ngo Trung, Duy Phung, and Thien Huu Nguyen. Unsupervised domain adaptation for event detection using domain-specific adapters. In *ACL-IJCNLP Findings*, pp. 4015–4025, 2021. URL https://aclanthology.org/2021.findings-acl.351.

Lawrence Rabiner and Biing-Hwang Juang. An introduction to hidden Markov models. *IEEE ASSP Magazine*, 3(1):4–16, 1986. URL https://ieeexplore.ieee.org/document/1165342.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text Transformer. *JMLR*, 21(140):1–67, 2020. URL https://jmlr.org/papers/v21/20-074.html.

Iulian Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, pp. 3776–3783, 2016. URL https://ojs.aaai.org/index.php/AAAI/article/view/9883.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *ICLR*, 2017. URL https://openreview.net/forum?id=B1ckMDqlg.

Tianxiao Shen, Myle Ott, Michael Auli, and Marc'Aurelio Ranzato. Mixture models for diverse machine translation: Tricks of the trade. In *ICML*, pp. 5719–5728, 2019. URL https://proceedings.mlr.press/v97/shen19c.html.

Christian D. Sigg and Joachim M. Buhmann. Expectation-maximization for sparse and non-negative PCA. In *ICML*, pp. 960–967, 2008. URL https://doi.org/10.1145/1390156.1390277.

Koustuv Sinha, Prasanna Parthasarathi, Jasmine Wang, Ryan Lowe, William L. Hamilton, and Joelle Pineau. Learning an unreferenced metric for online dialogue evaluation. In *ACL*, pp. 2430–2441, 2020. URL https://aclanthology.org/2020.acl-main.220.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. RUBER: An unsupervised method for automatic evaluation of open-domain dialog systems. In *AAAI*, pp. 722–729, 2018. URL https://ojs.aaai.org/index.php/AAAI/article/view/11321.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*, 2016. URL https://arxiv.org/abs/1610.02424.

Oriol Vinyals and Quoc Le. A neural conversational model. *arXiv preprint arXiv:1506.05869*, 2015. URL https://arxiv.org/abs/1506.05869.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. K-adapter: Infusing knowledge into pre-trained models with adapters. In *ACL-IJCNLP Findings*, pp. 1405–1418, 2021a. URL https://aclanthology.org/2021.findings-acl.121.

Yida Wang, Yinhe Zheng, Yong Jiang, and Minlie Huang. Diversifying dialog generation via adaptive label smoothing. In *ACL-IJCNLP*, pp. 3507–3520, 2021b. URL https://aclanthology.org/2021.acl-long.272.

Bolin Wei, Shuai Lu, Lili Mou, Hao Zhou, Pascal Poupart, Ge Li, and Zhi Jin. Why do neural dialog systems generate short and meaningless replies? A comparison between dialog and translation. In *ICASSP*, pp. 7290–7294, 2019. URL https://ieeexplore.ieee.org/document/8682634.

Yuqiao Wen, Guoqing Luo, and Lili Mou. An empirical study on the overlapping problem of open-domain dialogue datasets. In *LREC*, pp. 146–153, 2022. URL http://www.lrec-conf.org/proceedings/lrec2022/pdf/2022.lrec-1.16.pdf.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019. URL https://arxiv.org/abs/1910.03771.

Sixing Wu, Ying Li, Dawei Zhang, Yang Zhou, and Zhonghai Wu. Diverse and informative dialogue generation with context-specific commonsense knowledge awareness. In *ACL*, pp. 5811–5820, 2020. URL https://aclanthology.org/2020.acl-main.515.

Sixing Wu, Ying Li, Minghui Wang, Dawei Zhang, Yang Zhou, and Zhonghai Wu. More is better: Enhancing open-domain dialogue generation via multi-source heterogeneous knowledge. In *EMNLP*, pp. 2286–2300, 2021. URL https://aclanthology.org/2021.emnlp-main.175.

Lili Yao, Yaoyuan Zhang, Yansong Feng, Dongyan Zhao, and Rui Yan. Towards implicit content-introducing for generative short-text conversation systems. In *EMNLP*, pp. 2190–2199, 2017. URL https://aclanthology.org/D17-1233.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *ACL*, pp. 654–664, 2017. URL https://aclanthology.org/P17-1061.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. Commonsense knowledge aware conversation generation with graph attention. In *IJCAI*, pp. 4623–4629, 2018. URL https://www.ijcai.org/proceedings/2018/643.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. The design and implementation of XiaoIce, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93, 2020. URL https://aclanthology.org/2020.cl-1.2.

Yicheng Zou, Zhihua Liu, Xingwu Hu, and Qi Zhang. Thinking clearly, talking fast: Concept-guided non-autoregressive generation for open-domain dialogue systems. In *EMNLP*, pp. 2215–2226, 2021. URL https://aclanthology.org/2021.emnlp-main.169.

A PROOF OF THEOREM 1

We introduce two lemmas for proving Theorem 1. The first lemma controls the error between the empirical average and the true expectation. It is a direct application of Hoeffding's inequality. The second lemma bounds the difference between the learned probability and the true underlying probability. Following the notation in Eqn. (2), we use $Q(z|\mathbf{c},\mathbf{r})$ as the estimated posterior.

Lemma 1. *Given a dataset $\mathcal{D}$ consisting of context-response pairs $(\mathbf{c},\mathbf{r})$, for any z, and $\epsilon > 0$, we have*

$$\Pr\left(\left|\frac{1}{|\mathcal{D}|}\sum_{(\mathbf{c},\mathbf{r})\in\mathcal{D}} Q(z|\mathbf{c},\mathbf{r}) - \mathbb{E}_{\mathbf{c},\mathbf{r}}[Q(z|\mathbf{c},\mathbf{r})]\right| \geq \epsilon\right) \leq 2\exp\left(-2\epsilon^2|\mathcal{D}|\right)$$

Proof. Observe that $Q(z|\mathbf{c},\mathbf{r})$ is real-value random variable. The proof is then clear by using Hoeffding's inequality. □

Lemma 1 is a quantitative description of the law of large numbers.

Lemma 2. *Given any z and estimation $Q(z|\mathbf{c},\mathbf{r})$ of the true $P(z|\mathbf{c},\mathbf{r})$, we have*

$$\left|\mathbb{E}_{\mathbf{c},\mathbf{r}}[Q(z|\mathbf{c},\mathbf{r})] - \frac{1}{K}\right| \leq \mathbb{E}_{\mathbf{c},\mathbf{r}}\left[\left|Q(z|\mathbf{c},\mathbf{r}) - P(z|\mathbf{c},\mathbf{r})\right|\right]$$

Proof.

$$\begin{aligned}
\left|\mathbb{E}_{\mathbf{c},\mathbf{r}}[Q(z|\mathbf{c},\mathbf{r})] - \frac{1}{K}\right| &= \left|\sum_{\mathbf{c}}\sum_{\mathbf{r}} Q(z|\mathbf{c},\mathbf{r})P(\mathbf{c},\mathbf{r}) - \frac{1}{K}\right| \\
&= \left|\sum_{\mathbf{c}}\sum_{\mathbf{r}} Q(z|\mathbf{c},\mathbf{r})P(\mathbf{c},\mathbf{r}) - \sum_{\mathbf{c}}\sum_{\mathbf{r}} P(z|\mathbf{c},\mathbf{r})P(\mathbf{c},\mathbf{r})\right| \qquad (11)\\
&= \left|\sum_{\mathbf{c}}\sum_{\mathbf{r}} P(\mathbf{c},\mathbf{r})\left(Q(z|\mathbf{c},\mathbf{r}) - P(z|\mathbf{c},\mathbf{r})\right)\right| \\
&\leq \sum_{\mathbf{c}}\sum_{\mathbf{r}} P(\mathbf{c},\mathbf{r})\left|Q(z|\mathbf{c},\mathbf{r}) - P(z|\mathbf{c},\mathbf{r})\right| \qquad (12)\\
&= \mathbb{E}_{\mathbf{c},\mathbf{r}}\left[\left|Q(z|\mathbf{c},\mathbf{r}) - P(z|\mathbf{c},\mathbf{r})\right|\right]
\end{aligned}$$

Here, Eqn. (11) replaces $1/K$ with the marginalization of $P(z,\mathbf{c},\mathbf{r})$ over $\mathbf{c}$ and $\mathbf{r}$. □

Lemma 2 bounds the absolute difference between the estimation $Q(z|\mathbf{c},\mathbf{r})$ and the true $P(z|\mathbf{c},\mathbf{r})$.

Theorem 1. *With probability at least $1-\delta$, we have*

$$\left|\frac{1}{|\mathcal{D}|}\sum_{(\mathbf{r},\mathbf{c})\in\mathcal{D}} Q(z|\mathbf{r},\mathbf{c}) - \frac{1}{K}\right| \leq \sqrt{\frac{1}{2|\mathcal{D}|}\log\left(\frac{2K}{\delta}\right)} + \mathbb{E}_{\mathbf{c},\mathbf{r}}\left[\left|Q(z|\mathbf{c},\mathbf{r}) - p(z|\mathbf{c},\mathbf{r})\right|\right]$$

for all $z \in \{1,\cdots,K\}$. Here, $Q(z|\mathbf{c},\mathbf{r})$ is an estimated posterior and $\mathcal{D}$ is a dataset with finite samples.

Proof. Lemma 1 shows a high probability inequality for a single z. The probability for the original inequality to hold for all z is then:

$$\Pr\left(\forall z\in[K] : \left|\frac{1}{|\mathcal{D}|}\sum_{(\mathbf{c},\mathbf{r})\in\mathcal{D}} Q(z|\mathbf{c},\mathbf{r}) - \mathbb{E}_{\mathbf{c},\mathbf{r}}[Q(z|\mathbf{c},\mathbf{r})]\right| \geq \epsilon\right) \leq 2K\exp\left(-2\epsilon^2|\mathcal{D}|\right)$$

Choose an ϵ such that $\delta = 2K \exp\left(-2\epsilon^2|\mathcal{D}|\right)$. We have

$$\Pr\left(\forall z \in [K] : \left|\frac{1}{|\mathcal{D}|}\sum_{(\mathbf{c},\mathbf{r})\in\mathcal{D}} Q(z|\mathbf{c},\mathbf{r}) - \mathbb{E}_{\mathbf{c},\mathbf{r}}[Q(z|\mathbf{c},\mathbf{r})]\right| \geq \sqrt{\frac{1}{2|\mathcal{D}|}\log\left(\frac{2K}{\delta}\right)}\right) \leq \delta$$

Then with probability at least $1-\delta$, we have:

$$\begin{aligned}
&\left|\frac{1}{|\mathcal{D}|}\sum_{(\mathbf{c},\mathbf{r})\in\mathcal{D}} Q(z|\mathbf{c},\mathbf{r}) - \frac{1}{K}\right| \\
\leq &\left|\frac{1}{|\mathcal{D}|}\sum_{(\mathbf{r},\mathbf{c})\in\mathcal{D}} Q(z|\mathbf{c},\mathbf{r}) - \mathbb{E}_{\mathbf{c},\mathbf{r}}[Q(z|\mathbf{c},\mathbf{r})]\right| + \left|\mathbb{E}_{\mathbf{c},\mathbf{r}}[Q(z|\mathbf{c},\mathbf{r})] - \frac{1}{K}\right| \\
\leq &\sqrt{\frac{1}{2|\mathcal{D}|}\log\left(\frac{2K}{\delta}\right)} + \left|\mathbb{E}_{\mathbf{c},\mathbf{r}}[Q(z|\mathbf{c},\mathbf{r})] - \frac{1}{K}\right| && \text{(w.p. } 1-\delta\text{, by Lemma 1)} \\
\leq &\sqrt{\frac{1}{2|\mathcal{D}|}\log\left(\frac{2K}{\delta}\right)} + \mathbb{E}_{\mathbf{c},\mathbf{r}}\left[\left|Q(z|\mathbf{c},\mathbf{r}) - P(z|\mathbf{c},\mathbf{r})\right|\right] && \text{(by Lemma 2)}
\end{aligned}$$

concluding the proof. □

B Implementation Details

Our multi-adapter model used T5-small (Raffel et al., 2020) from HuggingFace (Wolf et al., 2019) as the backbone: 6 layers, 8 attention heads, an input/output dimension of 512, and an inner dimension of 2048 for both the encoder and decoder. We implemented our multi-decoder model by inserting an adapter layer with an inner dimension of 256 after each Transformer decoder layer. We adopted 10 decoders based on the common evaluation metric of diverse text generation (Zhao et al., 2017; Gu et al., 2019). Table 3 shows a comparison between our multi-adapter architecture and the full-Transformer architecture. As seen, our multi-adapter model retains a competitive BLEU performance while only using 24% of the parameters. Since the goal of our research is not to push the scores, we adopted the multi-adaptor architecture in the main experiments, because it helped our development process much.

For training, we had two stages: 1) Pretraining of the shared Transformer model, and 2) EM training of the adapter layers. In both stages, we used a learning rate of 0.001 with the Adam optimizer with $\boldsymbol{\beta} = (0.9, 0.999)$. We used a batch size of 64 for pretraining. For our 10-decoder EqHard-EM training, we considered 640 samples in an E-step, assigning 64 samples for each decoder.

C Human Evaluation

We conducted human evaluation to compare both the quality and diversity of the responses from T5-small (beam search) and our multi-decoder model. We again chose the Weibo dataset, because it is less noisy. Five native Chinese speakers were provided with a dialogue context and generated responses of two systems, and were asked to select the better system or indicate a tie. For quality, annotators were presented with only one response from each model, preventing them to identify the systems based on diversity. For diversity, all 10 responses were presented. In both cases, the presentation order of the two systems was randomly shuffled, so the annotators could not tell which system generated which response.

Decoder Architecture	#Parameters	BLEU2-F
Adapter	108M	10.13
Transformer	451M	11.08

Table 3: Comparison between adapter and Transformer architectures for our 10-decoder model.

	Model	Wins	Ties	Losses	p-value
Overall Quality	T5-small beam search	32.8%	19.0%	48.2%	<0.0001
	Multi-adapter + EqHard-EM	**48.2%**	19.0%	**32.8%**	
Diversity	T5-small beam search	4.8%	8.0%	87.2%	<0.0001
	Multi-adapter + EqHard-EM	**87.2%**	8.0%	**4.8%**	

Table 4: Human evaluation for the overall quality and diversity of the generated responses. Five annotators were invited to each evaluate 100 samples on the Weibo dataset. The p-value is given by the one-sided binomial test (ignoring ties).

Table 4 shows the results of human evaluation. Our multi-decoder model shows statistically significant improvement in terms of overall quality and diversity. Especially, annotators believe our approach generates more diverse utterances on 18x more samples. The results are highly consistent with the automatic metrics in Table 4, confirming the effectiveness of our approach.

D Case Study

Table 5 shows example responses from Weibo and OpenSubtitles. As seen, Soft-EM exhibits the synchronous-training collapse, where all decoders generate highly similar responses. Hard-EM, on the other hand, suffers from the non-training collapse, where a large number of decoders are not properly trained and generate low-quality text (e.g., certain words being repeated). By contrast, our EqHard-EM algorithm enables us to properly train all decoders to generate high-quality, diverse responses.

E Evaluating the Recurrent Dropout Trick

Shen et al. (2019) address the collapse issues in the EM algorithm by implementing an ad hoc recurrent dropout trick: disabling dropout layers for the E-step while retaining dropout layers for the M-step. They claim that the dropout layers in the E-step interfere with the specialization of different decoders, causing the mixture model to collapse. We conduct an ablation study and show that their trick is ineffective in our dialogue setting. As shown by Table 6, the recurrent dropout trick has minimum effect on the EM algorithms, and it does not prevent the non-training and synchronous-training collapses for Hard-EM and Soft-EM. This is not surprising because even Shen et al. (2019) acknowledge that their recurrent trick does not work consistently, as shown in Table 7 of their Appendix (red and blue numbers). The one-to-many phenomenon in our dialogue setting is more severe than their machine translation setting, and it is understandable that such an ad hoc trick, without any theoretical justifications, may fail in the more difficult task.

Weibo Input: 终于买了个爱疯，发觉也没啥用，还不如ipad *Finally got an iPhone, found it useless, even worse than an ipad*		
Soft-EM	Hard-EM	EqHard-EM
我也有一个，不过没有那么大 *I have one too, but not so big*	江江江[repeat] *river river river [repeat]*	这个是什么牌子的? *What brand is this?*
我也有一个，不过没有那么大 *I have one too, but not so big*	[eos] *[eos]*	我也想买一个，可惜没钱买 *I want to buy one too, too bad I don't have money*
我也有一个，不过没有那么大 *I have one too, but not so big*	珍珍珍[repeat] *treasure treasure treasure [repeat]*	我也是，还不如ipad呢 *Me too, even worse than an iPad indeed*
我也有一个，不过没有那么大 *I have one too, but not so big*	[eos] *[eos]*	我也想买一个，可惜没钱买 *I want to buy one too, too bad I don't have money*
我也有一个，不过没有那么大 *I have one too, but not so big*	[eos] *[eos]*	这个东意不错，我也想买一个 *This stuff is good, I want to buy one too*
我也有一个，不过没有那么大 *I have one too, but not so big*	我也有一个，不过没有那么大 *I have one too, but not so big*	我也有一个，不过没有ipad *I have one too, but not the ipad*
我也有一个，不过没有那么大 *I have one too, but not so big*	想想想[repeat] *think think think [repeat]*	你这是在哪里啊？我也想买一个 *Where are you at? I want to buy one too*
这个是什么？求解 *What is this? Please exp[lain]*	苇苇苇[repeat] *reed reed reed [repeat]*	我也想买一个，但是没钱买 *I want to buy one too, but I don't have money*
我也有一个，不过没有那么大 *I have one too, but not so big*	雪雪雪[repeat] *snow snow snow [repeat]*	我也买了一个，不过还是很不错的 *I bought one too, but it's actually pretty good*
这个是什么？求解释 *What is this? Please explain*	。。。[repeat] *... [repeat]*	这个不错，不过我喜欢。 *This is not bad, but I like it.*
OpenSubtitles Input: and you know what happens if we object?		
Soft-EM	Hard-EM	EqHard-EM
i'm not gonna object.	what what what [repeat]	i don't know.
i do.	i'm not a victim.	what do you mean?
i don't know.	and and and [repeat]	if we object, we 'll be able to.
i don't know.	if really really really [repeat]	i don't know, i don't know.
i do.	no pe pe pe [repeat]	i don't know.
i don't know.	we're gonna have to do this.	i'm not gonna be able to do anything.
i don't know.	okay okay okay [repeat]	we're gonna have to.
i don't know.	hurry hurry hurry [repeat]	what?
i do.	what?	i don't know.
i don't know.	then then then [repeat]	i'm not gonna be able to do this, mr. santos.

Table 5: Case studies. Translations are in *italic*. [repeat] indicates that further repetitions are omitted, and [eos] indicates an end-of-sentence token is generated as the first token.

Model	BLEU-1↑			BLEU-2↑			Distinct n-gram↑		Pairwise
	F	P	R	F	P	R	Dist-1	Dist-2	BLEU↓
Beam search	18.59	25.59	14.60	8.51	**17.57**	5.62	10.41	14.17	65.46
Soft-EM (Shen et al., 2019)	17.02	**27.55**	12.32	6.27	16.03	3.89	10.69	12.45	86.25
Soft-EM w/ dropout trick (Shen et al., 2019)	17.13	27.53	12.43	6.30	16.02	3.92	10.88	12.77	84.77
Hard-EM (Shen et al., 2019)	4.81	3.09	10.84	2.15	1.61	3.22	4.01	4.11	0.00
Hard-EM w/ dropout trick (Shen et al., 2019)	5.17	3.39	10.86	2.25	1.71	3.27	3.51	3.60	2.22
EqHard-EM	**22.65**	26.17	**19.96**	**10.13**	14.55	**7.77**	27.03	42.02	22.32

Table 6: Ablation study for the recurrent dropout trick Shen et al. (2019) on the Weibo dataset.