Jointly Optimizing Diversity and Relevance in Neural Response Generation

Xiang Gao **Sungjin Lee** **Yizhe Zhang**
Chris Brockett **Michel Galley** **Jianfeng Gao** **Bill Dolan**
Microsoft Research, Redmond, WA, USA
{xiag,sule,yizzhang,chrisbkt,mgalley,jfgao,billdol}@microsoft.com

Abstract

Although recent neural conversation models have shown great potential, they often generate bland and generic responses. While various approaches have been explored to diversify the output of the conversation model, the improvement often comes at the cost of decreased relevance (Zhang et al., 2018). In this paper, we propose a SPACEFUSION model to jointly optimize diversity and relevance that essentially fuses the latent space of a sequence-to-sequence model and that of an autoencoder model by leveraging novel regularization terms. As a result, our approach induces a latent space in which the distance and direction from the predicted response vector roughly match the relevance and diversity, respectively. This property also lends itself well to an intuitive visualization of the latent space. Both automatic and human evaluation results demonstrate that the proposed approach brings significant improvement compared to strong baselines in both diversity and relevance. [1]

1 Introduction

The field of neural response generation is advancing rapidly both in terms of research and commercial applications (Gao et al., 2019; Zhou et al., 2018; Yoshino et al., 2019; Zhang et al., 2019). Nevertheless, vanilla sequence-to-sequence (S2S) models often generate bland and generic responses (Li et al., 2016a). Li et al. (2016a) encourage diversity by re-ranking the beam search results according to their mutual information with the conversation context. However, as beam search itself often produces lists of nearly identical sequences, this method can require a large beam width (e.g. 200). As a result, re-ranking can be extremely time-consuming, raising difficulties for real-time applications. This highlights the need to improve the diversity of candidates before re-ranking, and the need to optimize for diversity during training rather than just at the decoding stage.

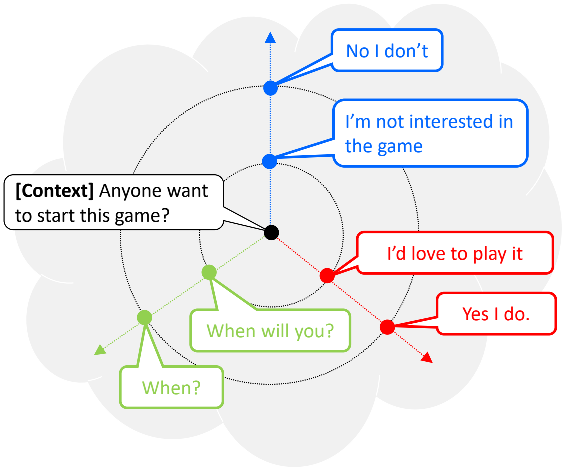
No I don't
I'm not interested in the game
[Context] Anyone want to start this game?
I'd love to play it
Yes I do.
When will you?
When?

Figure 1: Illustration of one **context** and its multiple **responses** in the latent space induced by our model. Distance and direction from the predicted response vector given the context roughly match the relevance and diversity, respectively. Based on the example in Table 2.[2]

While various approaches have been explored to diversify the output of conversation models, the improvement often comes at the cost of decreased response relevance along other dimensions. For instance, Zhao et al. (2017) present an approach to enhancing diversity by mapping diverse responses to a probability distribution using a conditional variational autoencoder (CVAE). Despite the improved response diversity, this approach reduces response relevance as measured against the baseline. One possible reason for this diversity-relevance trade-off is that such probabilistic approaches are not explicitly encouraged to induce a disentangled representation in latent space for

[1] An implementation of our model is available at https://github.com/golsun/SpaceFusion

[2] For simplicity, we omitted the response at the center: "I would love to play this game". See Table 2 for more details.

Proceedings of NAACL-HLT 2019, pages 1229–1238
Minneapolis, Minnesota, June 2 - June 7, 2019. ©2019 Association for Computational Linguistics

controlling diversity and relevance independently. Consider a Gaussian distribution, which is widely used for CVAE. A Gaussian distribution naturally brings frequent responses near its mean, and the resulting responses are often generic and boring. To generate diverse and interesting responses, one needs to sample a little distance from the mean. But doing so naturally leads to infrequent and thus even irrelevant responses.

In this paper, we propose a novel geometrical approach that explicitly encourages a structured latent space in which the distance and direction from a predicted response vector roughly match the relevance and diversity, respectively, as illustrated in Figure 1. To induce such a latent space, we leverage two different models: 1) a S2S model, producing the predicted response vector (the black dot at the center in Figure 1), and 2) an autoencoder (AE) model, yielding the vectors for potential responses (the colored dots). In order to make the S2S and AE share the same latent space (the cloud), we use the same decoder for both and train them jointly end-to-end with novel regularization terms. As this fuses the two latent spaces, we refer to our model as SPACEFUSION.

Regularization is necessary because only sharing the decoder, as in (Luan et al., 2017), does not necessarily align the latent spaces obtained by S2S and AE respectively or impose a disentangled structure onto the space. We introduce two regularization terms to tackle this issue. 1) interpolation term: we encourage a smooth semantic transition along the path between the predicted response vector and each target response vector (arrowed lines in Figure 1). This term effectively prevents semantically different responses from aligning in the same direction, essentially scattering them over different directions. 2) fusion term: we want the vectors from the two models to be distributed in a homogeneous manner, rather than forming two separate clusters (Figure 5) that can potentially make sampling non-trivial. With the resulting latent space, we can control relevance and diversity by respectively adjusting distance and direction from a predicted response vector, without sacrificing each other greatly.

Our approach also lends itself well to the intuitive visualization of latent space. Since our model allows us to geometrically find not only the predicted response vector but also the target response vector as in Figure 5, we can visually interpret the structure of latent space and identify major issues thereof. We devote Section 5.1 to show comprehensive examples for visualization-based analysis.

Automatic and human evaluations demonstrate that the proposed approach improves both the diversity and relevance of the responses, compared to strong baselines on two datasets with one-to-many context-response mapping.

2 Related Work

Grounded conversation models utilize extra context inputs besides conversation history, such as persona (Li et al., 2016b), textual knowledge (Ghazvininejad et al., 2017; Galley et al., 2019), dialog act (Zhao et al., 2017) and emotion (Huber et al., 2018). Our approach does not depend on such extra input and thus is complementary to this line of studies.

Variational autoencoder (VAE) models explicitly model the uncertainty of responses in latent space. Bowman et al. (2016) used VAE with Long-Short Term Memory (LSTM) cells to generate sentences. The basic idea of VAE is to encode the input x into a probability distribution (e.g. Gaussian) z instead of a point encoding. However, it suffers from the *vanishing latent variable problem* (Bowman et al., 2016; Zhao et al., 2017) when applied to text generation tasks. Bowman et al. (2016); Fu et al. (2019) proposed to tackle this problem with word dropping and specific KL annealing methods. Zhao et al. (2017) proposed to add a *bag-of-word* loss, complementary to KL annealing. Applying this to a CVAE conversation model, they showed that even greedy decoding can generate diverse responses. However, as VAE/CVAE conversation models can be limited to a simple latent representations such as standard Gaussian distribution, Gu et al. (2018) proposed to enrich the latent space by leveraging a Gaussian mixture prior. Our work takes a geometrical approach that is fundamentally different from probabilistic approaches to tackle the limitations of parameteric distributions in representation and difficulties in training.

Decoding and ranking encourage diversity during the decoding stage. As “vanilla” beam search often produces lists of nearly identical sequences, Vijayakumar et al. (2016) propose to include a dissimilarity term in the objective of beam search decoding. Li et al. (2016a) re-ranked the results obtained by beam search based on mutual information with the context using a separately trained response-to-context S2S model.

Multi-task learning is another line of studies related to the present work (see Section 3.2). Sennrich et al. (2016) use multi-task learning to improve neural machine translation by utilizing monolingual data, which usually far exceeds the amount of parallel data. A similar idea is applied by Luan et al. (2017) to conversational modeling, involving two tasks: 1) a S2S model that learns a context-to-response mapping using conversation data, and 2) an AE model that utilizes speaker-specific non-conversational data. The decoders of S2S and AE were shared, and the two tasks were trained alternately.

3 The SpaceFusion Model

3.1 Problem statement

Let $\mathcal{D} = [(x_0, y_0), (x_1, y_1), \cdots, (x_n, y_n)]$ denote a conversational dataset, where x_i and y_i are a context and its response, respectively. x_i consists of one or more utterances. Our aim is to train a model on $\mathcal{D}$ to generate relevant and diverse responses given a context.

3.2 Fusing latent spaces

We design our model to induce a latent space where different responses for a given context are in different directions around the predicted response vector, as illustrated in Figure 1. Then we can obtain diverse responses by varying the direction and keep their relevance by sampling near the predicted response vector.

To fulfill this goal, we first produce the predicted response representation z_{S2S} and target response representations z_{AE} using an S2S model and an AE model, respectively, as illustrated in Figure 2. Both encoders are implemented using stacked Gated Recurrent Unit (GRU) (Cho et al., 2014) cells followed by a noise layer that adds multivariate Gaussian noise $\epsilon \sim N(0, \sigma^2\mathbf{I})$. We then explicitly encourage smooth semantic transition along the path from z_{S2S} to z_{AE} by imposing any interpolation between them to generate the same response via the following loss term:

$$\mathcal{L}_{\text{interp}} = -\frac{1}{|y|} \log p(y|z_{\text{interp}}) \qquad (1)$$

where $z_{\text{interp}} = u z_{\text{S2S}} + (1-u) z_{\text{AE}}$ and $u \sim U(0,1)$ is a uniformly distributed random variable. $|y|$ is the number of words in y. Note that it is this regularization term that effectively prevents significantly different responses from aligning in the same direction, essentially scattering them over different directions. In order for this interpolation loss to work, we share the same decoder for both AE and S2S models as in (Luan et al., 2017). The decoder consists of stacked GRU cells followed by a softmax layer. It is worth mentioning that z_{interp} is not just randomly drawn from a single line but from a richer probabilistic region as both z_{interp} and z_{S2S} are stochastic due to the random component ϵ.

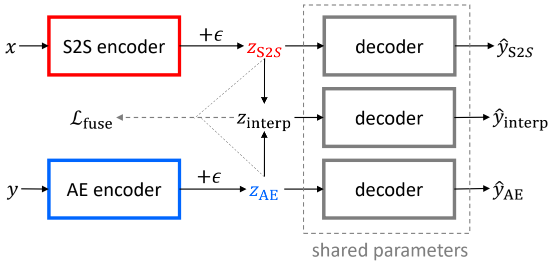

Figure 2: SpaceFusion model architecture.

Now, we want vectors from both the AE and S2S models to be distributed in a homogeneous manner scattered over the entire space while keeping the distance between z_{S2S} and z_{AE} as small as possible for any (context-response) pair in the training data. This objective is represented in the following regularization term:

$$\begin{aligned}\mathcal{L}_{\text{fuse}} = &\sum_{i \in \text{batch}} \frac{d(z_{\text{S2S}}(x_i), z_{\text{AE}}(y_i))}{n} \\ &- \sum_{i,j \in \text{batch}, i \neq j} \frac{d(z_{\text{S2S}}(x_i), z_{\text{S2S}}(x_j))}{n^2 - n} \\ &- \sum_{i,j \in \text{batch}, i \neq j} \frac{d(z_{\text{AE}}(y_i), z_{\text{AE}}(y_j))}{n^2 - n} \qquad (2)\end{aligned}$$

where n is the batch size and $d(a, b)$ is the root mean square of the difference between a and b. For each batch, we basically disperse vectors obtained by the same model and pull the predicted response vectors to the corresponding target response vectors. In practice, we found that the performance is better if the Euclidean distance is clipped to a prescribed maximum value.[3]

Finally, with weight parameters α and β, the

[3] This value is set as 0.3 for the present experiments

loss function is defined as:

$$\begin{aligned}\mathcal{L} = &- \frac{1}{|y|} \log p(y|z_{\text{S2S}}) \\ &- \frac{1}{|y|} \log p(y|z_{\text{AE}}) \\ &+ \alpha \mathcal{L}_{\text{interp}} + \beta \mathcal{L}_{\text{fuse}} \end{aligned} \tag{3}$$

As $\mathcal{L}_{\text{interp}}$ and $\mathcal{L}_{\text{fuse}}$ encourage the path between z_{S2S} and z_{AE} to be smooth and short while scattering vectors over the entire space, they effectively fuse the z_{S2S} latent space and the z_{AE} latent space. Accordingly we refer this approach as SPACEFUSION with path regularization.

3.3 Training

In contrast to previous multi-task conversation model (Luan et al., 2017), where S2S and AE are trained alternately, our approach trains S2S and AE at the same time by minimizing the loss function of Equation 3.

3.4 Inference

Like Zhao et al. (2017); Bowman et al. (2016), for a given context, we sample different latent vectors to obtain multiple hypotheses. This is done by adding a random vector r that is uniformly sampled from a hypersphere of radius $|r|$ to the prediction $z_{\text{S2S}}(x)$.

$$z(x, r) = z_{\text{S2S}}(x) + r \tag{4}$$

where $|r|$ is tuned on the validation set to optimize the trade-off between relevance and diversity. $z(x, r)$ is then fed to the decoder as the initial state of GRU cells. We then generate responses using greedy decoding.[4]

4 Experiment Setup

4.1 Datasets

We used the following datasets. Some of their key features are presented in Table 1.

Switchboard: We use the version offered by Zhao et al. (2017), which is an extension of the original version by Godfrey and Holliman (1997). Zhao et al. (2017) collected multiple references for the test set using information retrieval (IR) techniques followed by human filtering, and randomly split the data into 2316/60/62 conversations for train/validate/test, respectively. Each conversation has multiple turns and thus multiple (x, y) pairs, as listed in Table 1. As our approach does not utilize extra information except conversation history, we removed the meta data (e.g. gender, age, prompt) from this dataset.

	Switchboard	Reddit
train (x, y) samples	0.2M	7.3M
test (x, y) samples	5418	5000
ref. source	IR+filtering	natural
ref. availability	test only	train/vali/test
ref. per context	7.7	24.1

Table 1: Key features of the datasets.

Reddit: As the Switchboard dataset is relatively small and multiple references are synthetically constructed, we have developed another multi-reference dataset by extracting posts and comments on Reddit.com during 2011 collected by a third party.[5] As each Reddit post and comment may have multiple comments, it is a natural source of multi-reference responses. We further filtered the data based on the number of replies to obtain the final conversation dataset in which each context has at least 10 different responses, and on average the number of responses is 24.1 for a given context. The size is significantly larger than Switchboard, as listed in Table 1. The conversations are randomly shuffled before being split into train/valid/test subsets.

4.2 Model setup

Both encoders and the shared decoder consist of two GRU cells, each with 128 hidden units. The variance of the noise layer in each decoder is $\sigma^2 = 0.1^2$. The word embedding dimension is 128. The weight parameters (see Equation 3) are set as $\alpha = 1$ and $\beta = 30$. For both datasets, the inference radius $|r|$ (see Equation 4) is set to 1.5 which optimizes F1 score on the validation set. All models are trained using the Adam method (Kingma and Ba, 2014) with a learning rate of 0.001 on both datasets until convergence (around 4 epochs for Reddit and 10 epochs for Switchboard).

4.3 Automatic evaluation

For a given context x, we have N_r reference responses and generate the same number of hypothe-

[4] Although we use greedy decoding in this work, other decoding techniques, such as beam search, can be applied.

[5] http://files.pushshift.io/reddit/comments/

ses.[6] We define the following metrics based on 4-gram BLEU (Papineni et al., 2002), as suggested by Zhao et al. (2017).

$$\text{Precision} = \frac{1}{N_r}\sum_{i=1}^{N_r} \max_{j\in[1,N_r]} \text{BLEU}(r_j, h_i)$$

$$\text{Recall} = \frac{1}{N_r}\sum_{j=1}^{N_r} \max_{i\in[1,N_r]} \text{BLEU}(r_j, h_i)$$

$$\text{F1} = 2\,\frac{\text{precision}\cdot\text{recall}}{\text{precision}+\text{recall}}$$

We use Precision as an approximate surrogate metric for relevance and Recall for diversity. It should be noted that recall is not equivalent to other diversity metrics, e.g., distinct (Li et al., 2016a) and entropy (Zhang et al., 2018), which only depend on hypotheses. One potential issue of these metrics is that even randomly generated responses may yield a high diversity score. F1 is the harmonic average of these two and is used to measure the overall response quality.

4.4 Human evaluation

We conduct a human evaluation using crowdworkers. For each hypothesis, given its context, we ask three annotators to individually measure the quality, on a scale of 1 to 5, in terms of two aspects: relevance and interest. Interestingness is treated as an estimation of the diversity, as these two are often correlated. The hypotheses from all systems are shuffled before being provided to annotators. System names are invisible to the annotators.

4.5 Baselines

We compare the proposed model with the following baseline models:

S2S+Sampling: We consider a vanilla version of S2S model. The dimensions are similar to our model: both encoder and decoder consist of two stacked GRU cells with 128 hidden units, and the word embedding size is 128. As in the baseline in Zhao et al. (2017), we applied softmax sampling at inference time to generate multiple hypotheses.

CVAE+BOW: For the CVAE conversation model, we use the original implementation and hyperparameters of Zhao et al. (2017) with the bag-of-words (BOW) loss. The number of trainable model parameters is 15.4M, which is much larger than our model (3.2M).

MTask: Since our approach utilizes a multi-task learning scheme, we also compare it against a vanilla multi-task learning model, MTask, similar to (Luan et al., 2017), to illustrate the effect of space fusion. The model architecture and hyperparameters are identical to the proposed model except that the loss function is $\mathcal{L} = -\log p(y|z_{\text{S2S}}) - \log p(y|z_{\text{AE}})$.

5 Results and Analysis

5.1 In-depth analysis of latent space

In this section, we undertake an in-depth analysis to verify whether the latent space induced by our method manifests desirable properties, namely: 1) disentangled space structure between relevance and diversity, 2) homogeneous space distribution in which semantics changes smoothly without holes. We first provide a qualitative investigation based on real examples. Then, we present a set of corpus-level quantitative analyses focused on geometric properties.

5.1.1 Qualitative examples

In Table 2, we investigate three different directions from the context "`Anyone want to start this game?`" , which is a real example taken from Reddit. The three different directions correspond to clearly different semantics: "`No I don't`", "`when?`" and "`Yes I do.`" If we generate a response with the vector predicted by the S2S model (u = 0), our model outputs "`I would love to play this game`" which is highly relevant to the context. Now as we move along each direction, we can see our model gradually transforms the response toward the corresponding responses of each direction. For instance, towards "`No I don't`", our model gradually transforms the response to "`I am not interested in the game`" (u = 0.18) and then "`I am not interested.`" (u = 0.21). In contrary, towards "`Yes I do`", the response transforms to "`I would love to play it.`" (u = 0.15). Besides the positive or negative directions, the same transition applies to other directions such as "`When?`". This example clearly shows that there is a rough correspondence

[6] We set the number of hypotheses equal to the number of references to encourage precision and recall have comparable impact on F1

context x: Anyone want to start this game?
response at $u = 0$: I would love to play this game.

u	**towards "No I don't."**	u	**towards "when?"**	u	**towards "Yes I do."**
0.18	I am not interested in the game.	0.15	I'd be interested in the game	0.15	I'd love to play it.
0.21	I am not interested.	0.31	When is it?	0.27	Yes I do.
0.30	No I don't.	0.40	When will you?		
		1.00	When?		

Table 2: Semantic interpolation along different directions y. Results decoded from z_{interp} See Fig. 1 for a visualization.

.

context x: Anyone want to start this game?
towards one possible target y: Yes I do.

u	**with regularization**	u	**without regularization**
0.00	I would love to play this game.	0.00	I would have to play with the game.
0.15	I would love to play it.	0.29	Dude, I know, but, or etc.
0.30	Yes I do	0.61	Op I was after though today
		0.85	I'm single :(though
		0.90	Yes I do.

Table 3: Semantic interpolation with and without regularization. Results decoded from z_{interp} .

between geometric properties and semantic properties in the latent space induced by our method as shown in Figure 1– the relevance of the response decreases as we move away from the predicted response vector and different directions are associated with semantically different responses.

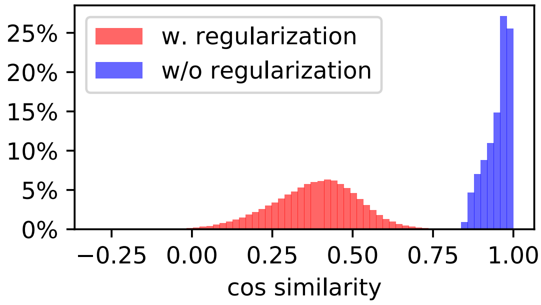

Figure 3: Distribution of the directions from a given context to its multiple responses, measured by the cosine similarity between $z_{\text{AE}}(y_{k,i}) - z_{\text{S2S}}(x_k)$ and $z_{\text{AE}}(y_{k,j}) - z_{\text{S2S}}(x_k)$. Histogram calculated based on 1000 x_k from Reddit test data and visualized with bin width of 0.02.

5.1.2 Direction vs. diversity

In order to quantitatively verify the correspondence between *direction* and *diversity*, we visualize the distribution of cosine similarities among multiple references for each context for a set of 1000 random samples drawn from the test dataset. Specifically, for a context x_k and its associated reference responses $[y_{k,0}, y_{k,1}, \cdots]$, we compute the cosine similarity between $z_{\text{AE}}(y_{k,i}) - z_{\text{S2S}}(x_k)$ and $z_{\text{AE}}(y_{k,j}) - z_{\text{S2S}}(x_k)$. In Figure 3, we compare the distribution of our model with that of MTask, which does not employ our regularization terms. While our method yields a bell shaped curve with average cosine similarity being close to zero (0.38), the distribution of MTask is extremely skewed with average cosine similarity being close to 1 (0.95). This indicates that the directions of the reference responses are more evenly distributed in our latent space whereas everything is packed in a narrow band in the MTask's space. This essentially makes the inference process simple and robust in that one can choose arbitrary directions to generate diverse responses.

5.1.3 Distance vs. relevance

In order to quantitatively verify the correspondence between *distance* and *relevance*, we visualize the perplexity of reference responses along the path from the associated z_{S2S} ($u = 0$) to the z_{AE} ($u = 1$) corresponding to the predicted response. In Figure 4, we compare our model with MTask, which as already noted, does not employ our regularization terms. While our model shows a gradual increase in perplexity, there is a huge bump for MTask's line. This clearly indicates that there is a rough correspondence between distance and relevance in our latent space whereas even a slight change can lead to an irrelevant response in the MTask's space.

We further illustrate the smooth change in relevance according to distance for a specific example in Table 3. Given the context "`Anyone want to start this game?`", our model

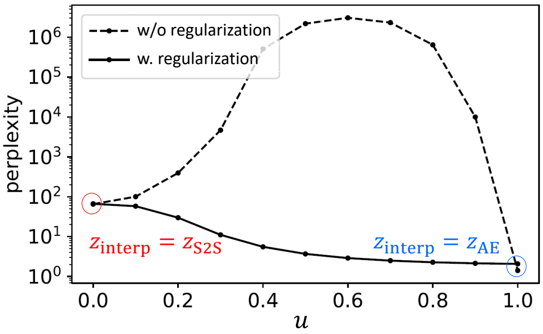

Figure 4: Perplexity of z_{interp} on the Reddit test dataset as a function of u for simple multi-task model (without regularization, dashed line) and SPACEFUSION (with regularization, solid line).

is able to transition from the predicted response "`I would love to play this game`" to a one of reference responses "`Yes I do`". The relevance smoothly descreases, generating intermediate responses such as "`I would love to play it.`" In contrary, the MTask model tends to produce irrelevant or ungrammatical responses as it moves away from the predicted response.

5.1.4 Homogeneity and Convexity

Other desirable properties, with which we want to equip our latent space are *homogeneity* and *convexity*. If the space is not homogeneous, we have to sample differently depending on the regional traits. If the space is not convex, we have to worry about running into the holes that are not properly associated with valid semantic meanings. In order to verify homogeneity and convexity, we visualize our latent space in a 2D space produced by the multidimensional scaling (MDS) algorithm (Borg and Groenen, 2003), which approximately preserves pairwise distance. For comparison, we also provide a visualization for MTask. As shown in Figure 5, our latent space offers great homogeneity and convexity regardless of which model is used to produce a dot (i.e. z_{S2S} or z_{AE}). In contrary, MTask's latent space forms two separate clusters for z_{S2S} and z_{AE} with a large gap in-between where no training samples were mapped to.

5.2 Automatic evaluation

We let each system generate 100 hypotheses $\{h_j\}$ for each context x_i in the test dataset. Assuming x_i has $N_{r,i}$ references, we pick the top $N_{r,i}$ distinct hypotheses ranked by $\log p(h_j|x_i) + \lambda|h_j|$. Similar to (Li et al., 2016a; Wu et al., 2016), we takes $|h_j|$ into consideration, as BLEU is sensitive to length.

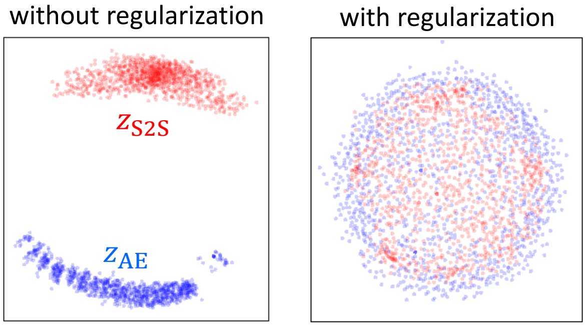

Figure 5: MDS visualization of the two latent spaces: z_{s2s} (red dots) and z_{AE} (blue dots) of 1000 randomly picked (x, y) pairs from the Reddit test dataset. Left: multi-task model (without regularization); right: SPACEFUSION (with regularization).

For fair comparison, λ is tuned such that the average hypothesis length becomes roughly the same for all systems and approaches the average length of the references.[7]

The automatic evaluation results are reported in Table 4. On both datasets, the proposed system consistently outperforms the baselines by a large margin in Precision, Recall, and F1.

Examples of system outputs and human references can be found in Table 5 and Table 6 for Reddit and Switchboard, respectively. As shown in the examples, CVAE+BOW and other baseline models may generate diverse but not-so-relevant responses.

dataset	model	Precision	Recall	F1
Switchboard	SPACEFUSION	**1.22**	**0.66**	**0.86**
	CVAE+BOW	0.76	0.57	0.65
	MTask	0.75	0.43	0.54
	S2S+Sampling	0.57	0.48	0.52
Reddit	SPACEFUSION	**0.40**	**0.26**	**0.31**
	CVAE+BOW	0.16	0.18	0.17
	MTask	0.31	0.18	0.23
	S2S+Sampling	0.10	0.11	0.11

Table 4: Performance of each model on automatic measures. The highest score in each row is in **bold** for each dataset. Note that our BLEU scores are normalized to $[0, 100]$.

5.3 Human evaluation

We randomly sampled 500 contexts from the Reddit test dataset and picked the top 1 hypothesis generated for each context ranked by $\log p(h_j|x_i) + \lambda|h_j|$. As in the automatic evaluation, we tuned λ such that all systems have roughly

[7] Approximately 10 words/tokens for Switchboard and 12 for Reddit

context	Everything about this movie is awesome!
SpaceFusion	○ I love this movie. ○ It's so awesome!!! I have no idea how to watch this movie. I can't wait for the trailer. ○ I don't think i'm a fan of the movie. ○ I would love to see this. ○ I want to watch this movie.
CVAE +BOW	○ Smartphones of the best games!. ○ I'm in the same boat! I feel the same way about this ○ I don't know why but can't tell if the New York will be the only one. ○ Caves would never say yes, but I'd love to know. ○ I know where this movie is.
MTask	○ This is so funny. ○ I think I saw this. ○ I don't know how many people do you. ○ I would. ○ I remember watching was was disappointed.
S2S+ sampling	○ Whoa man, this is amazing!!! ○ Man, I love Tom Cruise and I love the show. ○ Is this a bill on the right? ○ More like samurai jack. ○ I weep for the hivemind.
human	○ I love this movie. ○ Watched this yesterday! so amazing. ○ My favorite Muppet movie by far. ○ Seriously. Good times. Great oldies. ○ Now you have got the song into my head.

Table 5: Example system outputs and human reference on Reddit dataset

the same average hypothesis length. We also randomly select one reference for each context and compare them with the systems (labeled "human" in Table 7)

As illustrated in Table 7, the proposed model outperforms all systems except human, consistent with our automatic evaluation results.

	relevance	interest	average
SpaceFusion	**2.72**	**2.53**	**2.63**
CVAE+BOW	2.51	2.37	2.44
Multi-Task	2.34	2.14	2.24
S2S+Sampling	2.58	2.43	2.50
human	3.59	3.41	3.50

Table 7: Performance of each model on human evaluation. The highest score, except human, in each row is in **bold**.

6 Conclusion

We propose a SpaceFusion model to jointly optimize diversity and relevance that leverages novel regularization terms to essentially fuse the latent space of a S2S model with that of an autoencoder model. This fused latent space exhibits desirable properties such as smooth semantic interpolation between two points. The distance and direction from the predicted response vector roughly match relevance and diversity, respectively. These properties also enable intuitive visualization of the latent space. Both automatic and human evaluation results demonstrate that the proposed approach brings significant improvement compared to strong baselines in terms of both diversity and relevance. In future work, we will provide theoretical justification of the effectiveness of the proposed regularization terms. We expect that this technique will find application as an efficient "mixing board" for conversation that draws on multiple sources of information.

context	**A**: Are they doing a lot of recycling out in ***? **B**: Well at my workplace they are. We have places for aluminum cans and we have everybody been issued a separate trash can for recyclable paper. Let's see them get that word straight. **A**: Uh-huh **B**: As far as the community goes, it's pretty much voluntary. Nothing's been done. ***. We have just private garbage services.
SpaceFusion	○ They're not recycling as a matter of fact. ○ And it's a waste of time ○ You know the paper is a big recycling program. ○ But they just put it in the trash ○ Or if you have a recycling bag or anything like that.
CVAE +BOW	○ Um - hum. ○ Oh that would be nice. ○ Is that right? ○ It's a it's a luxury ○ Well that's interesting.
MTask	○ Um - hum. ○ That's right. ○ I don't know how much money you're going to do it. ○ That's right. ○ Yeah
S2S+ sampling	○ That's great. ○ What do you feel about this uh - huh. ○ It's just kind of a glass of town and it's even irritating to me. I don't know ○ Somebody and it's so much better. ○ It now and it's hard to go.
human	○ Yeah it is to some degree inconvenient i'd have to say that

Table 6: Example system outputs and human reference on Switchboard dataset. Controversial or offensive words are replaced by ***.

References

Ingwer Borg and P Groenen. 2003. Modern multidimensional scaling: theory and applications. *Journal of Educational Measurement*, 40(3):277–280.

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111.

Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, and Lawrence Carin. 2019. Cyclical annealing schedule: A simple approach to mitigating kl vanishing. *arXiv preprint arXiv:1903.10145*.

Michel Galley, Chris Brockett, Xiang Gao, Jianfeng Gao, and Bill Dolan. 2019. Grounded response generation task at dstc7.

Jianfeng Gao, Michel Galley, and Lihong Li. 2019. Neural approaches to conversational ai. *Foundations and Trends® in Information Retrieval*, 13(2-3):127–298.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2017. A knowledge-grounded neural conversation model. *arXiv preprint arXiv:1702.01932*.

John J Godfrey and Edward Holliman. 1997. Switchboard-1 release 2. *Linguistic Data Consortium, Philadelphia*, 926:927.

Xiaodong Gu, Kyunghyun Cho, Jungwoo Ha, and Sunghun Kim. 2018. DialogWAE: Multimodal response generation with conditional wasserstein auto-encoder. *arXiv preprint arXiv:1805.12352*.

Bernd Huber, Daniel McDuff, Chris Brockett, Michel Galley, and Bill Dolan. 2018. Emotional dialogue generation using image-grounded language models. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 277. ACM.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016b. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 994–1003.

Yi Luan, Chris Brockett, Bill Dolan, Jianfeng Gao, and Michel Galley. 2017. Multi-task learning for speaker-role adaptation in neural conversation models. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 605–614.

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 86–96.

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

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Koichiro Yoshino, Chiori Hori, Julien Perez, Luis Fernando D'Haro, Lazaros Polymenakos, Chulaka Gunasekara, Walter S Lasecki, Jonathan K Kummerfeld, Michel Galley, Chris Brockett, et al. 2019. Dialog system technology challenge 7. *arXiv preprint arXiv:1901.03461*.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018. Generating informative and diverse conversational responses via adversarial information maximization. In *Advances in Neural Information Processing Systems*, pages 1813–1823.

Yizhe Zhang, Xiang Gao, Sungjin Lee, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. 2019. Consistent dialogue generation with self-supervised feature learning. *arXiv preprint arXiv:1903.05759*.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 654–664.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2018. The design and implementation of xiaoice, an empathetic social chatbot. *arXiv preprint arXiv:1812.08989*.