A Conditional Generative Matching Model for Multi-lingual Reply Suggestion

Budhaditya Deb[†] **Guoqing Zheng**[‡] **Milad Shokouhi**[†] **Ahmed Hassan Awadallah**[‡]
[†]Microsoft AI [‡]Microsoft Research
{budeb, zheng, milads, hassanam}@microsoft.com

Abstract

We study the problem of multilingual automated reply suggestions (RS) model serving many languages simultaneously. Multilingual models are often challenged by model capacity and severe data distribution skew across languages. While prior works largely focus on monolingual models, we propose Conditional Generative Matching models (CGM), optimized within a Variational Autoencoder framework to address challenges arising from multilingual RS. CGM does so with expressive message conditional priors, mixture densities to enhance multi-lingual data representation, latent alignment for language discrimination, and effective variational optimization techniques for training multi-lingual RS. The enhancements result in performance that exceed competitive baselines in relevance (ROUGE score) by more than 10% on average, and 16% for low resource languages. CGM also shows remarkable improvements in diversity (80%) illustrating its expressiveness in representation of multi-lingual data.

1 Introduction

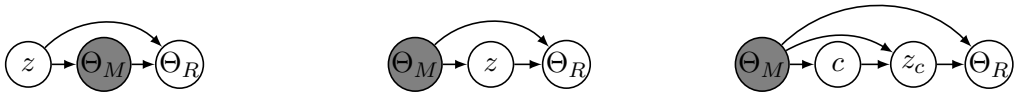
Automated reply suggestion (RS) helps users quickly process Email and chats, in popular applications like Gmail, Outlook, Microsoft Teams, and Facebook Messenger, by selecting a relevant reply generated by the system, without having to type in the response. Most existing RS systems are English mono-lingual models (Kannan et al., 2016; Henderson et al., 2017; Deb et al., 2019; Shang et al., 2015). We study the problem of creating multilingual RS models serving many languages simultaneously. Compared to mono-lingual models, a universal multilingual model offers several interesting research questions and practical advantages.

Universal models can save compute resources and maintenance overhead for commercial systems supporting many regions. In addition it can benefit languages with insufficient data by information sharing from high resource languages and thus enhance experiences for users especially in low-language resource regions. We investigate if a single multilingual RS model can replace multiple mono-lingual models with better performance, while overcoming the challenges in model capacity, data skew, and training complexities.

Trivially extending existing mono-lingual RS models to the multilingual setting (e.g. by jointly training with pre-trained multi-lingual encoders) tends to be sub-optimal, as multilingual models suffer from capacity dilution issue (Lample and Conneau, 2019), where it improves performance on low resource languages while hurting the high resource ones. This arises, not only due to the severe data imbalance and distribution skew across languages, but also due to insufficient capacity and lack of inductive biases in models to represent the multi-modal distribution of languages. We postulate that deep generative latent variable models with variational auto-encoders (VAE) (Kingma and Welling, 2014) are better suited to model the complex distribution of multi-lingual data, and be more data efficient for low resource languages.

To this end, we propose the Conditional Generative Matching Model (CGM), a VAE based retrieval architecture for RS to solve the above challenges. CGM enhances multilingual representation through: 1) expressive message conditional priors, 2) multi-component mixture density to represent different modalities of languages, and 3) alignment of latent components for language discrimination. In addition CGM incorporates training optimizations in the form of 1) loss regularizer, 2) learnable weights for loss components, 3) multi-sample loss estimation with variance scaling, and 4) focal loss, all of which lead to balanced representation and smooth convergence, a key challenge for variational training in multilingual settings.

We conducted extensive ablation studies and comparisons with two competitive baselines to



(a) Generative model for MCVAE (b) Generative model for CGM (c) Generative model for CGM-M

Figure 1: RS generative models in the continuous space. Text M-R pairs (in discrete space) are encoded into a common continuous space ($\Theta_M \sim \Theta_R$), where the encoders outputting Θ_M, Θ_R are considered extraneous to the generative model. The generative process is in the continuous space, with Θ_R generated conditioned on the input Θ_M and a Gaussian prior z. The figures show three variations of this generative process. In prior work MCVAE, z is sampled independently, while in CGM, it is conditional on Θ_M. CGM-M extends the message conditional prior with a Gaussian Mixture prior z_c and a categorical prior c.

show the impact of the above optimizations. Universal CGM models improve the relevance of RS (up to 13% excluding English) with even higher gains coming for low resource languages (16%), and when using CGM in a monolingual setting (19%). CGM also dramatically increases the diversity of suggested replies by 80% which is more illustrative of the improved representational capability of CGM in the multi-lingual landscape. CGM achieves this with relatively small increase in model sizes compared to the large pre-trained transformer stacks on which it is built, showing the modeling efficiencies that can be achieved through efficient training of latent variable models in a multi-lingual setting.

2 Background and Preliminaries

While RS has been modeled as a sequence to sequence model (Kannan et al., 2016), it more commonly appears as an information retrieval (IR) system by ranking responses from a fixed set (Henderson et al., 2017, 2019; Ying et al., 2021; Swanson et al., 2019; Zhou et al., 2016, 2018) due to better control over quality and relevance for practical systems. We briefly describe two retrieval architectures from prior literature which serves as the baselines for our multilingual RS model.

Matching model (Henderson et al., 2017; Ying et al., 2021) consists of two parallel encoders $[f_{\varphi_M}, f_{\varphi_R}]$ to encode message and reply (M-R) pairs into a common encoding space, $[\Theta_M, \Theta_R]$ and trained to maximize a normalized dot product $D = \Theta_M^\top \Theta_R$ between the M-R encodings. During prediction, the model finds the nearest neighbors of Θ_M with precomputed encodings from a fixed response set $R_{[s]}$. A language model bias is typically added to promote more common responses. The matching architecture is summarized as:

$$\mathcal{L}(\Theta_R|\Theta_M) = \log \frac{e^{D(\Theta_M,\Theta_R)}}{\sum_{r\in R_{[s]}} e^{D(\Theta_M,\Theta_r)}} \quad (1)$$

$$\text{Prediction} : Top_k\{\Theta_M^\top \Theta_r + \alpha LM(r) | r \in R_{[s]}\} \quad (2)$$

Matching Conditional VAE (MCVAE) (Deb et al., 2019) induces a deep generative latent variable model on the matching architecture, where a candidate response encoding is generated with $\Theta_{R'} = g_w(\Theta_M, z)$ conditioned on a latent prior $z \sim \mathcal{N}(0, I)$. The generated $\Theta_{R'}$ is used to match an actual response vector Θ_R from the fixed response set. The generative model of MCVAE is shown in figure 1a. In MCVAE, the encoders $[f_{\varphi_M}, f_{\varphi_R}]$ are pretrained using the matching formulation and kept frozen during the training. For prediction, MCVAE samples response vectors from g_w followed by scoring (eq 2) and a voting technique to rank replies over a fixed response set. MCVAE is trained in the variational framework by minimizing the negative evidence lower-bound (ELBO) in equation 3 with a Gaussian posterior q_ϕ (mean and co-variance parameterized from (Θ_M, Θ_R)) and the reconstruction loss $\mathcal{L}_M$ defined by Eq. (1).

$$\ell_{ELBO} = KL(q_\phi||p(z)) - \mathcal{L}_M(\Theta_R|\Theta_{R'}) \quad (3)$$

We extend the Matching and MCVAE models to a multi-lingual setting by using pretrained multilingual BERT (MBERT) (Devlin et al., 2019) for $[f_{\varphi_M}, f_{\varphi_R}]$ similar to (Ying et al., 2021) and jointly training the models for all languages.

3 CGM: A Conditional Generative Matching Model for Reply Suggestion

Our initial analysis with universal models (jointly training models with all languages), reveals that the universal MCVAE performs better than Matching. However, simply training models jointly is

not sufficient to achieve a models with high performance. First, the highly imbalanced nature of multi-lingual data leads to over- or under-fitting across languages resulting in performance worse than separately trained mono-lingual models. Second, training multi-lingual MCVAE proved is due to the reliance on a pretrained Matching model: it is not clear how to find a suitable Matching model checkpoint for initializing the MCVAE. Finally, since the text encoders for MCVAE are frozen during training, there is limited cross lingual transfer and improvement for low resource languages. Unfreezing the layers led to divergence of the model.

To address the limitations of MCVAE, we propose an enhanced Conditional Generative Matching (CGM) model, for the retrieval based RS with inductive biases for the multi-lingual data and effective training techniques for creating high quality universal models.

3.1 Message Conditional Prior

The implied generative process in MCVAE (Fig. 1a), is $p(z) \rightarrow p(\Theta_M|z) \rightarrow p(\Theta_R|\Theta_M, z)$, where the latent prior z is sampled independent of the message encoding Θ_M. However, in RS since Θ_M is always observed, ideally one would like to sample from $p(z|\Theta_M)$ to capture message-dependent information as well as rich multi-modality of the input space, particularly for multi-lingual data. In addition, although MCVAE works well empirically in the mono-lingual setting (Deb et al., 2019), the samples from $p(z)$ in general are not the same as $p(z|\Theta_M) \propto p(z)p(\Theta_M|z)$, unless $p(\Theta_M|z)$ is uniform across the space of Θ_M. This is a restrictive assumption, which motivates us to consider a prior conditioned on the input Θ_M for the generative model, by decomposing

$$p(\Theta_R, z|\Theta_M) = p(z|\Theta_M)p(\Theta_R|\Theta_M, z) \tag{4}$$

as shown in Figure 1b. The conditional prior $p(z|\Theta_M)$ is posed to encode message dependent information which can facilitate matching more relevant and diverse set of responses. We define the message-conditional prior $p(z|\Theta_M) = \mathcal{N}(\mu(\Theta_M), \Sigma(\Theta_M))$, where the prior parameters are learnt from data during training and used for prediction, to maximally capture the multiple modalities of intents and intrinsically complex distribution of multi-lingual data.

3.2 Prior with Mixture Density (CGM-M)

We postulate that a more expressive conditional prior, such as a mixture density, can better capture the multi-lingual data in contrast to the single prior density as used above. I.e., the different components of a mixture density can represent different languages and allow independent representation across languages. To this end we extend the message conditional prior with a Gaussian Mixture model (GMM) as,

$$p(z|\Theta_M) = \sum_{k=1}^{K} \pi_k(\Theta_M)\mathcal{N}\left(\mu_k(\Theta_M), \Sigma_k(\Theta_M)\right) \tag{5}$$

where $\mu_k(\Theta_M)$, $\Sigma_k(\Theta_M)$ are the message dependent means and diagonal covariances for the kth component of the GMM, and $\pi_k(\Theta_M)$ are the message dependent prior mixing coefficients. We hypothesize that components would correspond to different intents and languages, thus providing additional inductive bias for multi-lingual data. We refer to the mixture variant as CGM-M (Figure 1c).

3.3 Aligning Latent Space to Language

To further reinforce the notion that the CGM-M latent components encode language specific information from M-R pairs, we pose an additional constraint that the language of the message be inferred from the prior mixture coefficient. This is instantiated by building a simple classifier network with loss $\ell_{LC}(l|\Theta_M, \pi)$ to map the prior mixture coefficient $\pi(\Theta_M)$ onto the language l of the message. We also tested with mapping the 1) means and variances $[\mu_k(\Theta_M), \Sigma_k(\Theta_M)]$, and 2) samples z_k of the GMM, and found that mapping the $\pi(\Theta_M)$ leads to the best results. The classifier is learned jointly with the rest of the components.

3.4 Variational Training Architecture

The CGM models are formulated as a VAE in the continuous space of Θ_M, Θ_R. CGM includes two multi-lingual text encoders $[f_{\varphi_M}, f_{\varphi_R}]$, to convert the raw text of M-R into the common encoding space (encoders may be considered extraneous to the VAE but are learnt jointly with VAE layers), and a VAE with prior, posterior, and generation networks $[p_\psi(\mu, \Sigma), q_\phi(\mu, \Sigma), g_\theta]$.

The CGM-M extends the CGM version with category specific Gaussian components $[p_{\psi_c}, q_{\phi_c}]$ In addition it also includes a categorical prior and posterior $[\pi_c, \rho_c]$, and a language classifier l_c to

discriminate between languages. We use the standard reparameterization trick for the Gaussian variables and the Gumbel-Softmax trick (Jang et al., 2017) with hard sampling for the categorical variable. CGM-M (CGM is a special case with $K = 1$) is summarized as follows.

Generative Model : $p_\psi(\mu, \Sigma), g_\theta$

$$\pi = \text{Softmax}(\texttt{FFN}_1(\Theta_M)) \quad (6)$$

$$c = \text{GumbelSoftmax}(\texttt{FFN}_1(\Theta_M)) \quad (7)$$

$$\mu_\phi = \texttt{FFN}_2(\Theta_M),\ \Sigma_\phi = \text{Softplus}(\texttt{FFN}_3(\Theta_M)) \quad (8)$$

$$z_c = \mu_{\phi_c} + \varepsilon\Sigma_{\phi_c},\ \text{where } \varepsilon \sim \mathcal{N}(0, I) \quad (9)$$

$$\Theta_{R'} = \texttt{FFN}_4(\overleftrightarrow{z_c\Theta_M}) \quad (10)$$

Variational Posterior : $q_\phi(\mu, \Sigma)$

$$\rho = \text{Softmax}(\texttt{FFN}_5(\overleftrightarrow{\Theta_M\Theta_R})) \quad (11)$$

$$v = \text{GumbelSoftmax}(\texttt{FFN}_5(\overleftrightarrow{\Theta_M\Theta_R})) \quad (12)$$

$$\mu_\psi = \texttt{FFN}_6(\overleftrightarrow{\Theta_M\Theta_R}) \quad (13)$$

$$\Sigma_\psi = \text{Softplus}(\texttt{FFN}_7(\overleftrightarrow{\Theta_M\Theta_R})) \quad (14)$$

$$z_v = \mu_{\psi_v} + \xi\Sigma_{\psi_v},\ \text{where } \xi \sim \mathcal{N}(0, I) \quad (15)$$

Above, we expand the dimensions of projection vectors to $\mu : [h \times K], \Sigma : [h \times K]$ where h is the dimension of the forward projections and K is the number of categories in the mixture. After the category is selected (using Gumbel Softmax), we use the category index to select part of the expanded projections, as the k^{th} component of the means and variances (μ_k, Σ_k). Each $\texttt{FFN}_i$ denotes a two-layer feed-forward network (except $\texttt{FFN}_4$ which has 3 layers) with `tanh` activation and $\leftrightarrow$ denotes vector concatenation.

Note that the posteriors are conditioned on both Θ_M and Θ_R. This theoretically provides a richer representation of the M-R pairs and during inference allows us to score the combination of message and the selected response vectors. However, during training, it can lead to leakage through the network where the model simply ignores the message and uses the response vector for generation. We mitigate the leakage by applying a low-dimensional projection of response vector Θ_R before feeding into the variational network.

Following standard stochastic gradient variational bayes (SGVB) training, we minimize the negative ELBO to train the network. CGM-M adds the classifier loss to enforce alignment between latent vectors and language types. The training objectives for each are given as follows,

$$\ell_{\text{CGM}} = KL(q_\phi||p_\psi) - \mathcal{L}(\Theta_R|\Theta_{R'}) \quad (16)$$

$$\ell_{\text{CGM-M}} = KL_M(q_\phi||p_\psi) - \mathcal{L}(\Theta_R|\Theta_{R'}) + \ell_{LC} \quad (17)$$

where the reconstruction log-loss, $\mathcal{L}(\Theta_R|\Theta_{R'})$ is given by Eq. (1). For CGM, the KL divergence between the two multivariate Gaussian densities can be computed in closed form. However, for CGM-M, the KL divergence between two Gaussian mixtures does not admit a closed form. We estimate it with a variational approximation method described in (Hershey and Olsen, 2007)[1].

$$KL_M(q||p) \approx \sum_{i=1}^{K} \pi_i \log \frac{\sum_{j=1}^{K} \pi_j e^{KL(p_{\phi_i}||q_{\psi_j})}}{\sum_{k=1}^{K} \rho_k e^{KL(p_{\phi_i}||q_{\psi_k})}} \quad (18)$$

3.5 Training Optimizations

Training deep generative models with SGVB has been known to be notoriously tricky (Bowman et al., 2016; Fu et al., 2019). Our multilingual setting, and joint training of text encoders with VAE layers makes it even more challenging. We employed several optimizations to improve the convergence of the models.

1) Matching loss regularization: In CGM, the encoders for Θ_M, Θ_R are learnt jointly with the VAE layers in order to maximize richness of shared latent representation across languages. Thus Θ_R is a moving target for the VAE generator outputting $\Theta_{R'}$ and causes the training to diverge without additional constraints. In MCVAE, this was mitigated by initializing and freezing the text encoders from a trained Matching model, but can be counterproductive in the multilingual scenario. To enable joint training of text encoders and the VAE layers, and mitigate the issue of a moving target for reconstruction, we introduce a regularization in the form of a matching score between Θ_M and Θ_R,

$$\ell_{\text{CGM-M}} = KL_M(q_\phi||p_\psi) - \mathcal{L}(\Theta_R|\Theta_{R'}) + \ell_{LC} - \mathcal{L}(\Theta_R|\Theta_M) \quad (19)$$

which constrains the response vector to have a representation close to the message vector. This provides an independent anchor for the reconstruction and allows the end-to-end training of the model utilizing the full parameter space of the encoders for enhanced representation.

2) Multi-sample variance scaling: In SGVB, using a single sample of z usually results in high variance in the ELBO estimate. One remedy is to estimate the ELBO with multiple samples, either in the non-weighted and or importance

[1] Another approach with Monte-Carlo sampling requires a large number of samples and was not as effective.

weighted (Burda et al., 2016) versions. However, these led to only minor improvements.

In multi-sample training we take the expectation of the ELBO over the samples. We found that if instead we first take the expectation of the samples $z' = \sum_{i=1}^{k} z_i/k$ before computing the ELBO loss, we can reduce the variance and stabilize the training. Since z' follows an equivalent distribution $z' \sim \mathcal{N}(\mu, \frac{\Sigma}{k})$, we can estimate ELBO with multiple samples drawn from the scaled distribution and compute the expectation as follows. The adjustment provides significant improvements in training convergence and metrics.

$$\ell_{CGM} = \mathbb{E}_{z'}[-KL_{z'}(q_\phi||p_\psi) + \mathcal{L}(\Theta_R|\Theta_{R'})] \quad (20)$$

3) Weighting loss components with Homoscedastic Uncertainty (HSU): The final loss formulations for both CGM and CGM-M have several components. For finer control of training, we introduce learnable weights w_i for each of the components. Weighting different components of the ELBO loss has shown to improve performance (Higgins et al., 2017) in SGVB and thus even without additional components, such a weighting process is recommended.

Following (Cipolla et al., 2018), we view the loss formulation as a multi-task learning objective with different *homo-scedastic* uncertainties (HSU) for each task. Assuming the components factorize to Gaussian (continuous) and discrete (cross-entropy) likelihoods, the loss with HSU can be viewed as:

$$\begin{aligned}\ell_{HSU} =& \frac{1}{2\sigma_1^2} KL(q_\phi||p_\psi) - \frac{1}{2\sigma_2^2}\mathcal{L}(\Theta_R|\Theta_{R'}) \\ &- \frac{1}{2\sigma_3^2}\mathcal{L}(\Theta_R|\Theta_M) + \frac{1}{2\sigma_4^2}\ell_{LC} \\ &+ \log(\sigma_1) + \log(\sigma_2) + \log(\sigma_3) + \log(\sigma_4)\end{aligned} \quad (21)$$

Equating the uncertainties with the weights in our loss equation, this can be seen as learning the relative weights for each component where $w_i \sim 1/\sigma_i^2$ and provides a smooth, regularized and differentiable interpretation of weights. We introduce the weights as parameters in the model and learn them jointly with rest of the network.

4) Handling data skew with Focal Loss (FL): Multilingual training can have different convergence rates across languages and akin to behaviors observed in multi-modal training (Wang et al., 2020b). Carefully configured sampling ratios for different languages can alleviate this problem but requires costly hyper-parameter search. Instead we employ a popular technique for handling skewed data distribution: the focal loss (FL) (Lin et al., 2020).

$$\mathcal{L}_{FL}(\Theta_R|\Theta_{R'}) = (1 - e^{\mathcal{L}(\Theta_R|\Theta_{R'})})^\alpha \mathcal{L}(\Theta_R|\Theta_{R'}) \quad (22)$$

The FL (with $\alpha = 1$) is applied on the reconstruction log-probability component of ELBO, such that strongly reconstructed vectors are given lower weights than the weakly reconstructed ones which balances the convergence across languages.

3.6 Prediction and Ranking Responses

During prediction, we rank and select responses from a fixed response set $R_{[s]}$. Since the models generate response vectors in the continuous space, the prediction process needs to convert the samples into ranking in the discrete space of responses. The process is described as follows.

$$\log p_i(\Theta_{R_{[s]}}|\Theta_M) = \mathcal{L}(\Theta_{R'_i}|\Theta_{R_{[s]}}) - KL_z(q||p) \quad (23)$$

$$MRR(R_{[s]}) = \frac{1}{N}\sum_{i}^{N}[Rank_{R_{[s]}} \log p_i(\Theta_{R_{[s]}}|\Theta_M)]^{-1} \quad (24)$$

For each message we generate 1000 samples of latent conditional priors from $z \sim \mathcal{N}(\mu_\phi, \Sigma_\phi)$ and from categorical prior for CGM-M. Next, we generate samples of the response vectors using the generator network, $\Theta_{R'_i} \sim g_\theta(\Theta_{R'_i}|\Theta_M, z_i)$. We compute the scores for the i^{th} generated sample w.r.t to the fixed response set $\log p_i(\Theta_{R_{[s]}}|\Theta_M)$ in eq. 23, where the KL divergence is directly computed on the samples z under a Normal or GMM distribution for the prior and posterior. To reduce the scoring overhead over 40k responses with 1000 samples, we pre-select top k ($k = 100$ provides sufficiently diverse candidates) using the matching score (eq. 2). Finally, the mean reciprocal ranks (MRR) over all the samples (eq. 24) are used to select the top 3 as our predicted responses.

4 Experiments

Multi-lingual data: We use the MRS (Mulitlingual Reply Suggestions) data set (Zhang et al., 2021) for our experiments. MRS consists of message-reply (M-R) pairs separated into different languages from Reddit conversations (Baumgartner et al., 2020) using the FastText detector (Joulin et al., 2016). We select the top 15 languages for experimentation (data volume was insufficient for

	Latent Factors	Cond. Prior	Mix. Density	Language alignment	Multilingual training opts
Matching	-	-	-	-	-
MCVAE	✓	-	-	-	-
CGM	✓	✓	-	-	✓
CGM-M	✓	✓	✓	✓	✓

Table 1: Comparison of components of Matching, MCVAE (Sec 2), CGM, and CGM-M (Sec 3)

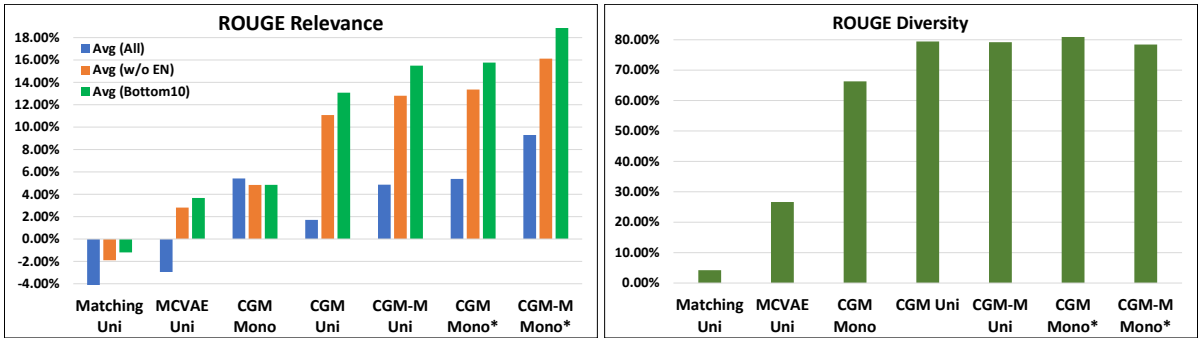


Figure 2: Main results. With the Matching monolingual models as baseline, the figures show the % changes in metrics for model variants (see Sec 4 for model description and Sec 4.1 for discussion). For each model variant, we show the metrics across three languages groups (All, w/o-EN and bottom 10 low resource languages. (Left) Relevance (Right) Diversity.

others) with 80% split for training (2nd column in Table 4) and the rest for validation and test. We create response sets with most frequent responses (>20 frequency) in the m-r pairs. For low resource languages, we augment this natural set with machine translated responses from EN, resulting in $\sim$ 40k responses for each language.

Metrics: We use ROUGE (Lin, 2004) for scoring the *relevance* of the 3 predicted responses against the reference response. We also compute the self-ROUGE (Celikyilmaz et al., 2020) within the 3 responses as a measure of *diversity*. For both, we report the average of the ROUGE-F1 for 1/2/3-grams across the three responses.

Train parameters: We use the multi-lingual version of the pretrained BERT model (MBERT) (Devlin et al., 2019) as out text encoders for which we use the Huggingface's transformers library (Wolf et al., 2020). We freeze the embedding layer of MBERT encoders, which reduces training overhead, and preserves cross-lingual representation without impacting performance (Lee et al., 2019; Peters et al., 2019). We use dimension size of 512 for the VAE layers. For CGM-M we set the number of categories to $K = 20$.

We train with the Adam optimizer (peak rate: $1e-5$, exp. decay: 0.999 after warm up of 1000 steps), batch size of 256, and m-r pairs truncated to length 64 and 32 respectively. We add language tokens (e.g. EN, PT) before m-r pairs as additional language identifier. All the model sizes are relatively similar (1.3GB to 1.5GB) since most parameters are in the two MBERT encoders with 12 transformer layers (each around 700MB).

Multilingual training: We uniformly sample languages such that models have equal exposure to each language during training. This leads to good performance across all languages except EN. Alternatively, sampling proportionate to data volumes, had good performance for EN but led to severe under-fitting for most languages other than EN as EN dominates the training with orders of magnitude more data. The ideal sampling is somewhere in between, but requires extensive search to optimize. On single NVidia V100 GPUs, models converge within 1-2 epochs $\sim 48hrs$ over the entire data (i.e., 1-2 epochs for EN and multiple epochs for others). Joint training amortizes the training costs, and can be used even when targeting monolingual models, by saving per-language checkpoints.

Model variants: We analyze 4 models: Matching, MCVAE, CGM and CGM-M (Table 1). For each we consider 3 multilingual model variants. **[Mono]**: individually trained monolingual models on each language. **[Uni]**: jointly trained universal model with a single checkpoint for evaluation. **[Mono*]**: jointly trained model with per language

checkpoints (saved when the validation metrics peak for each language) for evaluation. Since models peak at different point for each language, Mono* is expected to have a better performance than the Universal counterpart with a single checkpoint.

4.1 Main Results

Figure 2 shows the relevance and diversity metrics for different model variants. With Matching-Mono models (trained individually per language) as the baseline, we plot the % changes in metrics for the other model variants. Models are trained on all languages, with relevance metrics shown in 3 language groups: 1) All, 2) All w/o EN, and 3) Bottom 10 low resource languages, to highlight the differences from data volumes in languages.[2]

Relevance (Figure 2-Left): Compared to individually trained monolingual Matching model, the universally trained Matching-Uni regresses on all the three language group while MCVAE-Uni improves for latter two groups (w/o EN and bottom 10 languages). The CGM-Mono improves the metrics across all three languages. Thus even without joint training, CGM by itself is better than the baselines and thus raises the bar which the universal models needs to match or overcome.

The CGM and CGM-M universal models improve on all the language groups although for the CGM-uni, there is regression in the *All*-languages group compared to the CGM-mono (more discussion later). However, CGM-M-Uni with around 5% increase is actually slightly better than CGM-mono, showing that we can replace the monolingual models with a single universal model. Next, the Mono* models (universally trained but with best per-language checkpoints saved) can achieve even bigger gains and CGM-M-Mono* surpasses other models in every language group.

Within language groups, we observe increase upto 16% without EN and upto 19% for bottom 10 languages. EN with two orders of magnitude more data, remains severely under-fitted in all the jointly trained model, due to which the metrics improvements in *All* languages group remains low.

Diversity (Figure 2-Right): The CGM performance is most striking for diversity metrics where we see 80% improvements. Diversity improvements more than the relevance gains, illustrate that deep generative modeling enhancements in CGM leads to richer representation of multilingual data with improved discrimination and disentanglement of language and latent intents in M-R pairs. CGM-M achieves high diversity on top of the best relevance metrics, showing the enhanced representation through mixture models.

[2]Here we present quantitative results. For qualitative analysis, multi-lingual text predictions are provided in the appendix.

Line #	Baselines (Uni w/o EN)	ROUGE (Rel)	ROUGE (Div)
1	Matching	0.0353 (0%)	0.3940 (0%)
2	MCVAE	0.0369 (+4.80%)	0.289 (-26.65%)
	CGM (Uni w/o EN)		
3	Basic CGM	0.0378 (+7.25%)	0.354 (-10.16%)
4	+Variance Scaling (100 Samples)	0.0393 (+11.50%)	0.171 (-56.44%)
5	+Focal Loss, HSU	0.0398 (+12.78%)	0.161 (-59.08%)
6	+Rsp Vector in Posterior	0.0399 (+13.23%)	0.081 (-79.42%)
	CGM-M (Uni w/o EN)		
7	Basic CGM-M	0.0386 (+9.52%)	0.299 (-24.10%)
8	+Variance Scaling (100 Samples)	0.0396 (+12.23%)	0.189 (-51.96%)
9	+Focal Loss, HSU	0.04017(+13.87%)	0.172 (-56.30%)
10	+Lang Classifier	0.04043 (+14.60%)	0.164 (-58.33%)
11	+Rsp Vector in Posterior	0.0406 (+14.98%)	0.082 (-79.18%)

Figure 3: Ablation studies for different training optimizations (Sec 3.5) with results discussed in Sec 4.2.

4.2 Ablation Studies

We conducted extensive ablation studies with the different model variants, and training optimizations and summarize the results in Figure 3. For ablations we report the metrics for language group without EN, as the significantly higher data volume in EN can conflate the results.

Baselines: We use the Matching-uni model (line 1) as the baseline. MCVAE (line 2) improves both relevance (4.8%) and diversity (27%) which shows the potential of deep generative models.

Training optimizations with CGM: The basic CGM-Uni model (line 3) and CGM-M (Line 7) shows modest relevance gains compared to MCVAE. We attribute the modest gains due to complexities with end-to-end training of the CGM. Through training optimizations of variance scaling, and FL and HSU (lines 4, 5), CGM can comfortably surpass MCVAE in relevance (12.8%) and double the diversity (59%). CGM-M, shows similar increase (13.87%) with variance scaling (line 8), and FL and HSU (line 9) outperforming the best achieved with CGM. The biggest improvements come from multi-sample variance scaling (lines 4, 8) with additional improvements from FL and HSU (lines 5, 9). Overall, the optimizations lead to more stable training, and faster convergence across languages. They also alleviate the need for manual tuning for skewed data and loss component weights, making the training process virtually hyper-parameter free.

Language Mapping in CGM-M: One key reason for improved performance with CGM-M is the potential inductive bias for languages through

Language & Size	Matching Mono	CGM Mono	Matching Uni	MCVAE Uni	CGM Uni	CGM-M Uni	CGM Mono*	CGM-M Mono*
EN (49M)	0.117	7.89%	-28.37%	-27.64%	-38.53%	-29.26%	-28.94%	-19.99%
ES (1.86M)	0.035	4.55%	-3.45%	1.29%	5.59%	7.92%	6.57%	9.24%
DE (1.49M)	0.034	8.30%	-7.86%	-1.83%	-8.26%	-1.71%	2.57%	8.97%
PT (1.45M)	0.071	0.96%	-6.60%	-4.22%	1.85%	1.21%	3.78%	3.22%
FR (1.12M)	0.036	6.86%	-3.69%	3.03%	6.49%	6.37%	9.02%	12.43%
SV (590K)	0.032	8.32%	0.51%	5.15%	13.05%	16.51%	13.05%	20.88%
IT (589K)	0.036	3.62%	-5.04%	-2.34%	16.30%	18.57%	17.24%	18.57%
JA (582K)	0.031	-7.35%	-5.89%	-0.44%	-8.20%	-5.66%	-6.38%	-3.90%
NL (510K)	0.032	6.70%	-0.42%	3.59%	8.80%	8.42%	8.80%	11.14%
RU (413K)	0.025	12.32%	4.10%	11.63%	18.45%	18.14%	18.72%	21.95%
FI (308K)	0.018	9.76%	-0.18%	6.56%	16.82%	17.35%	18.49%	19.59%
DA (301K)	0.032	11.47%	5.11%	11.10%	22.31%	23.54%	22.31%	28.41%
RO (250K)	0.030	9.19%	7.12%	2.51%	12.83%	16.57%	17.83%	21.35%
TR (173K)	0.063	0.63%	1.03%	8.95%	35.51%	40.30%	39.31%	40.30%
PL (136K)	0.028	4.50%	-5.05%	1.56%	6.22%	2.69%	6.22%	9.20%
Avg (All)	0.041	5.41%	-6.90%	-2.94%	1.71%	4.86%	5.37%	9.30%
Avg (w/o EN)	1.041	4.83%	-1.90%	2.81%	11.08%	12.80%	13.36%	16.12%
Avg (Bottom10)	2.041	4.84%	-1.19%	3.67%	13.08%	15.49%	15.76%	18.86%

Figure 4: Relevance metrics across 15 languages. (Model description in Sec 4 and discussion in Sec 4.3)

the mixture components, which can be further boosted by explicit mapping of latent vectors to languages. Language mapping improves the relevance to 14.6% (line 10) over the baseline. We also see a slight boost in diversity showing the improved modeling of the multi-lingual distribution using this approach.

Posterior conditioned on both message and response: The joint conditioning of the posterior with both the Θ_M, Θ_R vectors[3] gives the best relevance for both CGM and CGM-M (lines 6, 11) with CGM-M exceeding all other variants. More interesting is the substantial improvement in diversity (80%), which illustrates that it encourages a richer representation in the prior by perhaps disentangling latent intents and language characteristics better. We note here that, in CGM-M, using the full Θ_R dimension (768) led to high level of leakage through the posterior (multiple components of the mixture further aids the leakage). We use a low dimensional projection of size 16 in CGM-M to mitigate the issue.

4.3 Analysis across Languages Groups

Next, we discuss the performance breakdown of models across individual languages. Figure 4 expands the Relevance metrics from Figure 2 for all languages. As before, we use the the Matching-Mono as the baseline, and list the % changes over this baseline for each model and language.

We see that, all jointly trained variants (Uni and Mono*) have severe under fitting for EN. In fact if we simply remove EN from the metrics the CGM variants vastly improve upon the monolingual versions. With almost two orders of magnitude more data in EN (49M), it remains challenging to have good performance simultaneously for EN and other languages without additional tricks. In general the improvements are less for the top 5 high-resource languages which can be attributed to lesser impact from information sharing and lower exposure of these languages due to uniform sampling. Such issues have been reported in prior literature as capacity dilution (Johnson et al., 2017; Conneau et al., 2020; Wang et al., 2020a) where there is always a trade-off between low and high resource languages. CGM while not completely eliminating it, largely mitigates the issue.

The impact of CGM with joint training is more pronounced for the bottom 10 language group. For example we see 15.49% improvement for CGM-M compared to only 3.67% for MCVAE-Uni. Finally, we see improvements of 15.76% for CGM-Mono* and 18.86% for CGM-M Mono* models, illustrating that even if we target mono-lingual models, CGM can take advantage of shared learning through joint training while saving compute.

The improvements for low resource languages, show that CGM is more data efficient due to model enhancements, while the prevention of regressions for high resource languages show a more balanced learning through training optimizations. The fact that these relevance improvements come in addition to 80% improvements in diversity, shows the remarkable effectiveness of CGM to represent the multi-modal landscape of multi-lingual RS.

5 Related Work

VAEs have been used in retrieval based Q&A (Yu et al., 2020), document matching (Chaidaroon and Fang, 2017), and recommendations (Chen and

[3]We had excluded Θ_R in the posterior of other configurations to show this effect.

de Rijke, 2018). CGM for RS is most closely related to MCVAE (Deb et al., 2019) but differs in the expressive conditional priors, multi-component mixture density priors, language alignment, and training optimizations which makes it effective in a multi-lingual setting.

For multi-task scenarios, VAEs can offer significant modeling efficiencies (Cao and Yogatama, 2020; Rao et al., 2019) with additional improvements through mixture model priors, e.g. in (Dilokthanakul et al., 2017; Yang et al., 2019) for unsupervised clustering, in (Lee et al., 2021) for unsupervised meta-learning, and in (Shi et al., 2019) as a multi-modal variational mixture-of-experts.

VAEs can also improve multilingual representation for low resource languages, e.g. in models like BERT (Li et al., 2020), in (Wei and Deng, 2017) for document classification, in (Chorowski et al., 2019) for disentangling phonemes for speech synthesis, and in (Zhang et al., 2016; Eikema and Aziz, 2019) for neural machine translation. VAEs can improve diversity in language generation and retrieval tasks (Zhao et al., 2017; Tran et al., 2017; Shen et al., 2017; Deb et al., 2019) through better modeling efficiencies. Such results motivated us to apply VAEs for multilingual RS.

We may also consider alternative to VAEs such as training auxiliary tasks with adapters (Houlsby et al., 2019), adversarial learning (Chen et al., 2018, 2019; Huang et al., 2019), and mixing pre-training and fine-tuning (Phang et al., 2020) to improve modeling in multilingual setting. This is subject of future work. We also plan to experiment with higher capacity multilingual encoders such XLM-R (Lample and Conneau, 2019) and InfoXLM (Chi et al., 2021) to further improve the performance. However, the choice of the base encoder is orthogonal to the improvements (especially on diversification) shown in this paper.

As noted in prior work, multilingual training can have capacity dilution issues (Johnson et al., 2017; Conneau et al., 2020; Wang et al., 2020a). Overall, multilingual models are closing the gap with monolingual counterparts for wide range of tasks (Ying et al., 2021; Ranasinghe and Zampieri, 2020; Yang et al., 2020), and as shown in this paper, even surpass them. Careful sampling strategies, and techniques such as Translation Language Model (TLM) can alleviate the "curse of multilinguality" (Lample and Conneau, 2019) but we show improvements without additional data augmentation (translation pairs), and with simple uniform sampling.

6 Conclusions

In this paper we present a conditional generative Matching model (CGM) for retrieval based suggested replies. CGM not only provides relevance gains (15%), but also substantial improvements in diversity (80%). While CGM clearly advances the state of art for modeling multi-lingual RS systems, it also illustrates that through proper model choices and training optimizations, we can surpass and replace monolingual models. This is important for both industry and academia and suggests similar strategies to be applied across diverse tasks. This is subject of future work.

References

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Józefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *CoNLL*.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. 2016. Importance weighted autoencoders. In *ICLR*.

Kris Cao and Dani Yogatama. 2020. Modelling latent skills for multitask language generation. *arXiv preprint arXiv:2002.09543*.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Suthee Chaidaroon and Yi Fang. 2017. Variational deep semantic hashing for text documents. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 75–84.

Xilun Chen, Ahmed Hassan, Hany Hassan, Wei Wang, and Claire Cardie. 2019. Multi-source cross-lingual model transfer: Learning what to share. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3098–3112.

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. 2018. Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics*, 6:557–570.

Yifan Chen and Maarten de Rijke. 2018. A collective variational autoencoder for top-n recommendation with side information. In *Proceedings of the*

3rd Workshop on Deep Learning for Recommender Systems, pages 3–9.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, He-Yan Huang, and Ming Zhou. 2021. Infoxlm: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588.

Jan Chorowski, Ron J. Weiss, Samy Bengio, and Aaron van den Oord. 2019. Unsupervised speech representation learning using wavenet autoencoders. In *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

R. Cipolla, Y. Gal, and A. Kendall. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7482–7491.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

Budhaditya Deb, P. Bailey, and M. Shokouhi. 2019. Diversifying reply suggestions using a matching-conditional variational autoencoder. In *NAACL-HLT*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Nat Dilokthanakul, Pedro A. M. Mediano, Marta Garnelo, Matthew C. H. Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. 2017. Deep unsupervised clustering with gaussian mixture variational autoencoders.

Bryan Eikema and Wilker Aziz. 2019. Auto-encoding variational neural machine translation. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 124–141, Florence, Italy. Association for Computational Linguistics.

Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, and Lawrence Carin. 2019. Cyclical annealing schedule: A simple approach to mitigating KL vanishing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 240–250, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv:1705.00652*.

Matthew Henderson, Ivan Vulić, Daniela Gerz, Iñigo Casanueva, Paweł Budzianowski, Sam Coope, Georgios Spithourakis, Tsung-Hsien Wen, Nikola Mrkšić, and Pei-Hao Su. 2019. Training neural response selection for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5392–5404.

J. R. Hershey and P. A. Olsen. 2007. Approximating the kullback leibler divergence between gaussian mixture models. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, volume 4, pages IV–317–IV–320.

Irina Higgins, Loic Matthey adnd Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Lifu Huang, Heng Ji, and Jonathan May. 2019. Cross-lingual multi-level adversarial transfer to enhance low-resource name tagging. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3823–3833.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In *ICLR*.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. FastText.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Anjuli Kannan, Karol Kurach, Sujith Ravi, Tobias Kaufmann, Andrew Tomkins, Balint Miklos, Gregory S. Corrado, László Lukács, Marina Ganea, Peter Young, and Vivek Ramavajjala. 2016. Smart Reply: Automated Response Suggestion for Email. In *KDD*.

Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. *ICLR*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Dong Bok Lee, Dongchan Min, Seanie Lee, and Sung Ju Hwang. 2021. Meta-gmvae: Mixture of gaussian vae for unsupervised meta-learning. In *ICLR*.

Jaejun Lee, Raphael Tang, and Jimmy Lin. 2019. What would elsa do? freezing layers during transformer fine-tuning. *arXiv preprint arXiv:1911.03090*.

Chunyuan Li, Xiang Gao, Yuan Li, Xiujun Li, Baolin Peng, Yizhe Zhang, and Jianfeng Gao. 2020. Optimus: Organizing sentences via pre-trained modeling of a latent space.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. 2020. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327.

Matthew E Peters, Sebastian Ruder, and Noah A Smith. 2019. To tune or not to tune? adapting pretrained representations to diverse tasks. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 7–14.

Jason Phang, Iacer Calixto, Phu Mon Htut, Yada Pruksachatkun, Haokun Liu, Clara Vania, Katharina Kann, and Samuel Bowman. 2020. English intermediate-task training improves zero-shot cross-lingual transfer too. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 557–575.

Tharindu Ranasinghe and Marcos Zampieri. 2020. Multilingual offensive language identification with cross-lingual embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5838–5844.

Dushyant Rao, Francesco Visin, Andrei A. Rush, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. 2019. Continual unsupervised representation learning. In *NeurIPS*.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1577–1586.

Xiaoyu Shen, Hui Su, Yanran Li, Wenjie Li, Shuzi Niu, Yang Zhao, Akiko Aizawa, and Guoping Long. 2017. A Conditional Variational Framework for Dialog Generation. In *ACL*.

Yuge Shi, Siddharth N ad Brooks Paige, and Philip Torr. 2019. Variational mixture-of-experts autoencoders for multi-modal deep generative models. In *NeurIPS*.

Kyle Swanson, Lili Yu, Christopher Fox, Jeremy Wohlwend, and Tao Lei. 2019. Building a production model for retrieval-based chatbots. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 32–41, Florence, Italy. Association for Computational Linguistics.

Quan Hung Tran, Gholamreza Haffari, and Ingrid Zukerman. 2017. A Generative Attentional Neural Network Model for Dialogue Act Classification. In *ACL*.

Zirui Wang, Zachary C Lipton, and Yulia Tsvetkov. 2020a. On negative interference in multilingual language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4438–4450.

Zirui Wang, Yulia Tsvetkov, Orhan Firat, and Yuan Cao. 2020b. Gradient vaccine: Investigating and improving multi-task optimization in massively multilingual models. In *ICLR*.

Liangchen Wei and Zhi-Hong Deng. 2017. A variational autoencoding approach for inducing cross-lingual word embeddings. In *IJCAI*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Linxiao Yang, Ngai-Man Cheung, Jiaying Li, and Jun Fang. 2019. Deep clustering by gaussian mixture variational autoencoders with graph embedding. In *ICCV*.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual universal sentence encoder for semantic retrieval. In *ACL*.

Qianlan Ying, Payal Bajaj, Budhaditya Deb, Yu Yang, Wei Wang, Bojia Lin, Milad Shokouhi, Xia Song, Yang Yang, and Daxin Jiang. 2021. Language scaling for universal suggested replies model. In *NAACL-HLT, Industrial Track*.

Wenhao Yu, Lingfei Wu, Qingkai Zeng, Shu Tao, Yu Deng, and Meng Jiang. 2020. Crossing variational autoencoders for answer retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5641, Online. Association for Computational Linguistics.

Biao Zhang, Deyi Xiong, Jinsong Su, Hong Duan, and Min Zhang. 2016. Variational neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 521–530.

Mozhi Zhang, Wei Wang, Budhaditya Deb, Guoqing Zheng, Milad Shokouhi, and Ahmed Hassan Awadallah. 2021. A dataset and baselines for multilingual reply suggestion. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1207–1220, Online. Association for Computational Linguistics.

Tiancheng Zhao, Ran Zhao, and Maxine Eskénazi. 2017. Learning Discourse-level Diversity for Neural Dialog Models using Conditional Variational Autoencoders. In *ACL*.

Xiangyang Zhou, Daxiang Dong, Hua Wu, Shiqi Zhao, Dianhai Yu, Hao Tian, Xuan Liu, and Rui Yan. 2016. Multi-view response selection for human-computer conversation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 372–381, Austin, Texas. Association for Computational Linguistics.

Xiangyang Zhou, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, and Hua Wu. 2018. Multi-turn response selection for chatbots with deep attention matching network. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1118–1127, Melbourne, Australia. Association for Computational Linguistics.

A Text Samples from Model Predictions

A.1 Relevance and Diversity

We created sample messages in EN manually, and predict the responses from different models: Matching in Figure 5, CGM in Figure 6 and CGM-M in Figure 7.

We see that in terms of relevance while it is hard to notice the differences on such a small sample, overall the predictions from the Matching model are less relevant than CGM. However, we can clearly distinguish the diversity of responses: predictions from Matching have a high level of duplicates where some of the responses differ by just a punctuation. While this can be easily de-duplicated using simple heuristics, the results show that inherently the Matching model ranks very similar responses at the top. The CGM models in contrast, show a lot of diversity in responses without reducing the relevance of the responses.

We also see that some of the responses are quite specific and not relevant, with some responses being rude or mildly inappropriate. It shows the issues with using responses from the Reddit dataset without careful curation (the MRS dataset does clean up for inappropriate responses but cannot completely eliminate them without human curation). Production systems usually have human curated response sets which can tackle these issues.

A.2 Multi-lingual Behavior

Next we look at the multilingual ability of CGM. We translate the same set of messages used for EN for predicting responses, so as to have better comparative understanding of the quality different languages.

We present the predictions for **ES** in Fig. 8 and **JA** in Fig. 9. We see that the responses are relevant and diverse in these languages and thus CGM performs adequately in languages other than EN.

A.3 Cross-lingual Behavior

Finally we investigate the cross lingual nature of the CGM model, in order to understand if the multilingual models share representations and learnings across languages.

In Figure 10 we use EN messages and force the model to predict responses from the ES set. Surprisingly such a system is able to select relevant results in the target language. While the quality here is not as good, but it is interesting to see that such cross lingual prediction works quite well.

In Figure 11 we use messages in German and predict with English responses. Again the results are quite acceptable. This may be expected as English and German are closely related languages. To see slightly different pairs of languages, we look at Japanese messages with predictions in English in Figure 12. Here the quality is actually much worse, but we still see some match with the overall intent of the messages showing good cross lingual representation even for weakly related languages.

Messages	Matching-1	Matching-2	Matching-3
Good morning! How was last night's party? Had fun?	It was great!	It was fun	It was awesome!
Did you see the new movie? It was really funny! Wish we saw it together	It was awesome!	What movie is it?	Which movie?
Please have a look and review. Its the latest update I in put last night.	Thanks for the update!	Thanks for the update.	Thank you for the update!
I am not feeling very well. Will try to get some sleep.	Sleep is for the weak	Hope you feel better soon.	Hope you feel better soon!
I have solved the issue with out of memory. That was some bug!	What bug?	It's not a bug, it's a feature!	How did you solve it?
It is so nice outside! I think will go fishing in the lake today.	I'll be there too!	Awesome! Good luck!	Good luck tomorrow!!
Did you hear the latest album. Its simply awesome, blown away!!!	What album?	Great album	The project was famous. good sales program, congratulations!
Stunning win today, what a goal!! He is a genius :D	He really is!	He really is	He really is.
The new netflix show, just watch the 1st season. Rest is not too good	What series?	What show is it?	What show?
Did you know that tomato is fruit, not a vegetable? I dont really care	What is a potato?	I can't believe it's not butter	What's a potato?
I am soo looking forward to the holiday! I am planning to take a few days off in July.	Good luck tomorrow!!	Good luck tomorrow!	Good luck tomorrow.
Augmented Reality is so awesome! I filled my room with virtual confetti!	Project updates continuously, team work actively. congratulations!	Project updates continuously, team work actively. congratulations!	Congrats to you as well!
A new bookshop opened down the road. I plan to go there for a reading.	What book?	What book	Which book?
It was pretty shocking what happened. It was all over the news. Had nightmares	I'm glad I wasn't the only one.	Glad I wasn't the only one.	Glad I wasn't the only one
Forrest Gump has an amazing soundtrack. Been listening to since childhood, and see it in new light as i grow up.	I love it too!	Love it too.	What song is it?
I am pretty bored these days. Need a new vocation.	What career?	What do you want to do?	What field are you in?
The new wired article is pretty revealing about corporate politics	What article?	Could you keep it down please? This is a public forum.	What section?
Christmas has come early. Enjoy while it lasts!	Thanks! Good luck to you too!	Congrats to you too!	Congrats to you!
Did some slow roasting in the oven yesterday. The stuff came out pretty tender and juicy.	How did it taste?	What did it taste like?	What size did you get?

Figure 5: Some samples of English message predicted with English replies using the Matching Model. The replies marked in red shows the duplicate responses.

Messages	CGM-1	CGM-2	CGM-3
Good morning! How was last night's party? Had fun?	It was delicious	Today was a good day	Was great!
Did you see the new movie? It was really funny! Wish we saw it together	I'd love to see it!	No I didn't.	No, it was Ex Machina
Please have a look and review. Its the latest update I in put last night.	I think it looks great!	I don't see anything I need. Sorry.	What's the app?
I am not feeling very well. Will try to get some sleep.	Sleep	Sleep is for the weak	Hope you are too.
I have solved the issue with out of memory. That was some bug!	Happened to me too	Thanks! It worked!	Where did you find it?
It is so nice outside! I think will go fishing in the lake today.	Do you like fish sticks?	There's always a bigger fish.	I think it looks great!
Did you hear the latest album. Its simply awesome, blown away!!!	Glad you think so!	What are you listening to?	Great album
Stunning win today, what a goal!! He is a genius :D	Good for him!	He's so good x4	A surprise, to be sure, but a welcome one!
The new netflix show, just watch the 1st season. Rest is not too good	Breaking Bad	What series?	What episode was this?
Did you know that tomato is fruit, not a vegetable? I dont really care	No I didn't.	No, it is not.	I'm vegan
I am soo looking forward to the holiday! I am planning to take a few days off in July.	What's your budget?	Mind if I check with you at 10 weeks?	What year is this?
Augmented Reality is so awesome! I filled my room with virtual confetti!	It really ties the room together.	It was delicious!	This will make a fine addition to my collection! (/r/GrievousCollection)
A new bookshop opened down the road. I plan to go there for a reading.	This is library	Which store?	Still open?
It was pretty shocking what happened. It was all over the news. Had nightmares	What news?	I'm glad I wasn't the only one.	What was so bad about it?
Forrest Gump has an amazing soundtrack. Been listening to since childhood, and see it in new light as i grow up.	Lil Pump	Forrest Gump	Thanks for listening!
I am pretty bored these days. Need a new vocation.	What field are you in?	You need new friends	You can do it! I believe in you!
The new wired article is pretty revealing about corporate politics	What shower thought has a source?	Wallpaper?	What kind of business?
Christmas has come early. Enjoy while it lasts!	And to you!	Better late than never!	Thanks! Enjoy!
Did some slow roasting in the oven yesterday. The stuff came out pretty	How much were they?	How did it turn out?	I'll try spinning, that's a good

Figure 6: Some samples of English message predicted with English replies using the CGM Model.

Messages	CGM-M-1	CGM-M-2	CGM-M-3
Good morning! How was last night's party? Had fun?	It was ok	Today was a good day	Pretty good!
Did you see the new movie? It was really funny! Wish we saw it together	We did!	What movie is it?	I saw it!
Please have a look and review. Its the latest update I in put last night.	Done check your inbox :)	Added an update to the OP	Will do! Good luck!
I am not feeling very well. Will try to get some sleep.	Have a good time!	Sleep is for the weak	I hope you feel better soon.
I have solved the issue with out of memory. That was some bug!	I did the same thing!	Thanks! It worked!	Did you try it?
It is so nice outside! I think will go fishing in the lake today.	We all float down here.	Go fish	Sounds like a good time!
Did you hear the latest album. Its simply awesome, blown away!!!	Thank you for checking it out!	I heard it too	It did!
Stunning win today, what a goal!! He is a genius :D	Absolutely incredible!	Goals!	Thanks for playing!
The new netflix show, just watch the 1st season. Rest is not too good	What's the first?	You get the show.	I'd watch it
Did you know that tomato is fruit, not a vegetable? I dont really care	It's what plants crave.	What is a potato?	r/contagiouslaughter
I am soo looking forward to the holiday! I am planning to take a few days off in July.	Hope to see you there!	RemindMe! 3 weeks	Sounds like a good time!
Augmented Reality is so awesome! I filled my room with virtual confetti!	How did you like it?	So exciting!	I really like it!
A new bookshop opened down the road. I plan to go there for a reading.	What book is this?	What are you reading?	Way to go!
It was pretty shocking what happened. It was all over the news. Had nightmares	What news?	What story?	I'm sorry to hear that :(
Forrest Gump has an amazing soundtrack. Been listening to since childhood, and see it in new light as i grow up.	Great album	I like it too.	And I love it.
I am pretty bored these days. Need a new vocation.	What do you want to do?	We will watch your career with great interest!	Teacher?
The new wired article is pretty revealing about corporate politics	So business as usual?	Facts are facts.	The project has great potential success.
Christmas has come early. Enjoy while it lasts!	I hope you're right!	It never ends	Thanks, same to you!
Did some slow roasting in the oven yesterday. The stuff came out pretty tender and juicy.	That's awesome to hear!	It was delicious!	How did it turn out?

Figure 7: Some samples of English message predicted with English replies using the CGM-M Model.

Message	CGM-1	CGM-2	CGM-3
¡Buenos días! ¿Cómo estuvo la fiesta de anoche? ¿Te divertías?	Fui a ir de compras :)	Muy bien.	¡Buenos días!
¿Viste la nueva película? ¡Fue muy gracioso! Ojalá lo vimos juntos	Estuvo bien.	¡Me encanta esa película!	Vi
¡Bienvenido! Me alegra tenerte de vuelta en el trabajo.	¡Gracias por las amables palabras!	¡Gracias! Te lo :)	¡Gracias! ¿le hará :)
¡Me voy de vacaciones! Necesitaba un descanso. Nos vemos en un par de semanas :-)	Vacaciones	¡Impresionante! ¡Disfrutar!	¡Viajes seguros!
¿Puede enviarme el enlace al documento? Parece que no encuentro el enlace.	Imposible. Tal vez los archivos están incompletos.	Claro que puedes.	¡Si no lo he enviado, avísame!
No me siento muy bien. Trataré de dormir un poco.	Me alegro de no estar solo.	¿Depresión?	Yo también lo siento.
El tráfico es bastante malo. Debería ser otra hora, pero no estoy seguro.	Siempre es soleado en Filadelfia	¿Qué te hace estar tan seguro?	Ningún lugar es seguro.
He resuelto el problema con fuera de la memoria. ¡Eso fue un bicho!	¿Besaste a tu madre con esa boca?	Hecho.. Reciprocate biko	No es un error, es una característica.
¡Es tan agradable afuera! Creo que hoy pescará en el lago.	Siempre hay un pez más grande.	¡Especialmente más tarde en el verano!	Espero que también sea :)
¿Oíste el último álbum? Es simplemente impresionante, impresionado!!!	Por el momento no, lo siento.	Me alegra escucharlo!!	¡Ese es! ¡Muchas gracias!
Impresionante victoria hoy, ¡qué gol! Es un genio :D	Espero que también sea :)	Goles tbh	Absolutamente increíble
La nueva serie de Netflix, sólo mira la primera temporada. El descanso no es demasiado bueno	Esa es la mordaza de la temporada	¿De qué episodio es ese?	¿Está buena?
Una nueva librería abrió el camino. Planeo ir allí para una lectura.	Voy a echarle un vistazo	Hecho por favor, vuelve a volver a :)	He estado allí, hecho eso.
Acabo de terminar su presentación a la conferencia, a la espera de escuchar al revisor #2.	Se ha superado la llamada de prueba. Comenzando la primera etapa	¿Tiene un enlace?	¿Cuál fue su reacción?
Fue bastante impactante lo que pasó. Estaba en todas las noticias. Tuvo pesadillas	Diferentes golpes para diferentes personas.	Noticias - Fox	#NAME?
Estoy bastante aburrido estos días. Necesito una nueva vocación.	Negociado, disfrutar!	Necesitas nuevos amigos.	Aprendes algo nuevo todos los días

Figure 8: Some samples of Spanish messages and predicted with Spanish replies using the CGM-M Model.

Message	CGM-1	CGM-2	CGM3
?			
!			!
		!	*再び
	?		Imgur
		!	
1		TS !	
!		!	!
? !!!			!
Netflix 1	2018	!	?
		!	
		!	*彼の

Figure 9: Some samples of Japanese messages and predicted with Japanese replies using the CGM-M Model.

Message	CGM-1	CGM-2	CGM-3
Good morning! How was last night's party? Had fun?	¿Cómo duermes por la noche?	Sí, fue	¡Buenos días!
Did you see the new movie? It was really funny! Wish we saw it together	¿Por qué no se me ocurrió?	¡La mejor de las suertes para ti!	¡Uno de nosotros!
Welcome back! Glad to have you back at work.	¡Gracias! ¡Gracias! ¡Gracias!	¡Gracias, buena suerte!	¡Gracias por tu tiempo!
Can you send me the link to the document? I cant seem to find the link.	¡Yo también necesito saberlo!	Bien, ¿quieres compartir un enlace por qué?	¿Por qué no puedes?
I am not feeling very well. Will try to get some sleep.	¡Qué salvación!	¿Me puedes ayudar?	¿Por qué estamos aquí? ¿Sólo para sufrir?
The traffic is pretty bad. Should be another an hour, but not sure.	¿Hay alguna posibilidad de que la pista se doble?	¿Por qué es un problema?	¿Por qué esto es una cosa
I have solved the issue with out of memory. That was some bug!	¡Eliminar! ¡Eliminar! ¡Eliminar!	¿Quizás los archivos están incompletos?	¿Has hecho comprobar tu bandeja de entrada :)
It is so nice outside! I think will go fishing in the lake today.	¡Mucho espacio para actividades!	¿Cómo duermes por la noche?	Hasta luego y gracias por todos los peces.
Did you hear the latest album. Its simply awesome, blown away!!!	¡Qué salvación!	¡Buenos días!	¡Me alegro de oírlo, gracias!
Stunning win today, what a goal!! He is a genius :D	¡Suficientemente bueno para mí!	¡Es un hombre increíble!	¡Los jugadores se levantan!
The new netflix show, just watch the 1st season. Rest is not too good	¡Qué salvación!	¿Por qué no los 3?	¡Me gusta mucho!
A new bookshop opened down the road. I plan to go there for a reading.	Tienes mucho que aprender sobre esta ciudad, cariño.	¿Qué libro es éste?	¿Cuál es tu dirección?
It was pretty shocking what happened. It was all over the news. Had nightmares	Nuestras vidas comienzan a terminar el día en que nos quedamos callados sobre las cosas que importan.	¿Qué noticias?	¡Uno de nosotros!
Forrest Gump has an amazing soundtrack. Been listening to since childhood, and see it in new light as i grow up.	¿Qué tipo de música te gusta?	¡Uno de nosotros!	¡La mejor de las suertes para ti!
I am pretty bored these days. Need a new vocation.	¿Qué es lo mejor que ser genial?	¿Qué es lo que quieres?	¿Cuál es tu especialidad?
The new wired article is pretty revealing about corporate politics	¿Así que los negocios como siempre?	¡Gracias por leerlo!	Sus ideas son intrigantes para mí y deseo suscribirme a su boletín de noticias.

Figure 10: Some samples of English messages and predicted with Spanish replies using the CGM-M Model. While the quality is not as good as when the input message is in Spanish, the general close match of intents of the message and responses illustrates the cross lingual ability of of the model.

Message	CG-1	CGM-2	CGM-3
Guten Morgen! Wie war die Party gestern Abend? Hatten Sie Spaß?	Really good.	Today was a good day	It was ok
Haben Sie den neuen Film gesehen? Es war wirklich lustig! Wunsch, dass wir es zusammen gesehen haben	This film is older.	Yes I did!	It was awesome!
Willkommen zurück! Froh, Sie wieder bei der Arbeit zu haben.	Thank you. (:	You're back!	Thanks, same to you!
Können Sie mir den Link zum Dokument zusenden? Ich kann den Link nicht finden.	Video is up on this sub!	you can edit since I gave new info	Infowars.com
Mir geht es nicht sehr gut. Wird versuchen, etwas Schlaf zu bekommen.	Freudian slip	Try it!	Courage
Der Verkehr ist ziemlich schlecht. Sollte eine weitere Stunde sein, aber nicht sicher.	Even a broken clock is right twice a day.	What time zone are you in?	Gotta go fast!
Ich habe das Problem mit unzusamten gelöst. Das war ein Fehler!	Not a problem!	You're not my supervisor!	Thank you for your service!
Es ist so schön draußen! Ich denke, ich werde heute im See angeln gehen.	To the moon!	You will!	Go fish.
Hast du das neueste Album gehört? Es ist einfach genial, weggeblasen!!!	r/fakealbumcovers	It really was!	What was the original?
Atemberaubende Sieg heute, was für ein Ziel!! Er ist ein Genie :D	He really does!	He deserves it.	Thanks for playing!
Die neue Netflix-Show, schauen Sie sich einfach die 1. Staffel an. Ruhe ist nicht zu gut	What series?	Season 2	I'd watch it.
Eine neue Buchhandlung wurde eröffnet. Ich habe vor, dort für eine Lesung zu gehen.	What book is this?	I want to go to there.	Where was it?
Es war ziemlich schockierend, was passiert ist. Es war alles über die Nachrichten. Hatte Alpträume	Our lives begin to end the day we become silent about things that matter.	What news?	Patrolling the Mojave almost makes you wish for a nuclear winter.
Forrest Gump hat einen erstaunlichen Soundtrack. Habe seit seiner Kindheit zugehört und sie in neuem Licht gesehen, wenn ich erwachsen bin.	This film is older.	I love it too.	Movie?
Ich bin ziemlich gelangweilt in diesen Tagen. Brauchen Sie eine neue Berufung.	r/stoppedworking	Be the change you want to see!	Becoming?
Der neue verkabelte Artikel ist ziemlich aufschlussreich über Unternehmenspolitik	So business as usual?	The project has great potential success.	Satire?

Figure 11: Some samples of German messages and predicted with English replies using the CGM-M Model. While the quality is not as good as when the input message is in German, the general close match of intents of the message and responses illustrates the cross lingual ability of of the model.

Message	CGM-1	CGM-2	CGM-3
おはようございます！昨夜のパーティーはどうでしたか。楽しかった?	Absolutely nothing!	What did you not like about it?	Today was a good day
新しい映画を見ましたか。それは本当に面白かったです!一緒に見て欲しい	Thank you! I'm glad you enjoyed it.	It was amazing!	It was awesome!
再びようこそ！仕事に戻ってきてうれしいです。	Have a great time!	Thank you! I definitely will!	Glad to hear it! :)
ドキュメントへのリンクを送って下さい。私はリンクを見つけることができないようです。	Please, read and follow the instructions at the top of the page. Thanks!	clicked	Done. Check your inbox!
トラフィックはかなり悪いです。もう1時間になるはずですが、わかりません。	Thank you for your positive feedback! :)	Thank you, I will.	I will :)
私はメモリ不足の問題を解決しました。それはいくつかのバグでした!	Appreciated!	Good project, congratulations!	Great work
外はとても素敵です!今日は湖で釣りに行くと思います。	Thank you! I definitely will!	Pics please!	Thanks! Me too!
最新アルバムを聞きましたか?その単に素晴らしい、吹き飛ばされた!!!	r/fakealbumcovers	Another!	Yes I did :)
今日の見事な勝利、何ゴール!!彼は天才:D	He sure is!	Love him!	So much winning!
新しいNetflixショーは、ちょうど第1シーズンを見ます。休息はあまり良くない	Wabbit season!	r/nhlstreams	Six seasons and a movie!
道の下に新しい書店が開いた。私は読書のためにそこに行く予定です。	You're going down a path I can't follow!	Thank you! !translated	Freedom!
何が起こったのかかなり衝撃的でした。それはニュースのいたるところにあった。悪夢を見た	r/notinteresting	What evidence?	What was his reaction?
フォレストガンプは素晴らしいサウンドトラックを持っています。子供の頃から耳を傾け、私が成長するにつれて新しい光の中でそれを見てください。	Recorded!	Love it! Thank you!	Thank you so very much.
私は最近かなり退屈です。新しい職業が必要です。	Yes you are!	You are!	That means a lot, thank you!
新しい有線記事は、企業政治についてかなり明らかです	Your ideas are intriguing to me and I wish to subscribe to your newsletter.	Please lower your voice. This is a public forum.	Please, read and follow the instructions at the top of the page. Thanks!

Figure 12: Some samples of Japanese messages and predicted with English replies using the CGM-M Model. The quality here is definitely poorer that German to English, perhaps since EN and JA are not as closely related. However we still get the general close match of intents of the message and responses.