Zero-shot Cross-lingual Dialogue Systems with Transferable Latent Variables

Zihan Liu, Jamin Shin, Yan Xu, Genta Indra Winata,
Peng Xu, Andrea Madotto, Pascale Fung
Center for Artificial Intelligence Research (CAiRE)
Department of Electronic and Computer Engineering
The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong
{zliucr,jay.shin,yxucb,giwinata,pxuab}@connect.ust.hk,
pascale@ece.ust.hk

Abstract

Despite the surging demands for multilingual task-oriented dialog systems (e.g., Alexa, Google Home), there has been less research done in multilingual or cross-lingual scenarios. Hence, we propose a zero-shot adaptation of task-oriented dialogue system to low-resource languages. To tackle this challenge, we first use a set of very few parallel word pairs to refine the aligned cross-lingual word-level representations. We then employ a latent variable model to cope with the variance of similar sentences across different languages, which is induced by imperfect cross-lingual alignments and inherent differences in languages. Finally, the experimental results show that even though we utilize much less external resources, our model achieves better adaptation performance for natural language understanding task (i.e., the intent detection and slot filling) compared to the current state-of-the-art model in the zero-shot scenario.

1 Introduction

Task-oriented dialogue systems have been widely adopted in the industry (e.g., Amazon Alexa, Google Home, Apple Siri, Microsoft Cortana) as a virtual agent to tend to the needs of the users. However, these agents have mostly been trained with the monolingual dataset that is often expensive to build or acquire. In order to cope with the scarcity of low-resource language dialogue data, we are motivated to look into cross-lingual dialogue systems which can adapt with very little or no training data in the target language.

This task of zero-shot adaptation of dialogue systems to different languages is relatively new and has not been explored thoroughly enough yet. The main approach of previous work (Upadhyay et al., 2018; Chen et al., 2018; Schuster et al., 2019) in this task is using aligned cross-lingual
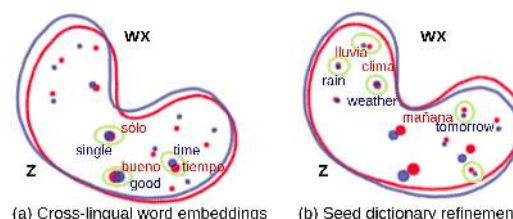

Figure 1: (a) The aligned word embeddings. (b) Refined mapping **W** with seed dictionary; green circle represents close words in different languages.

word embeddings between source and target languages. However, this method suffers from imperfect alignments between the source and target language embeddings. This can be attributed not only to the noise in aligning two different embeddings, but also to the inherent discrepancies in different languages such as Thai and English which come from entirely different roots. To address such variance in the alignment, we turn to probabilistic modeling with latent variables as it has been successfully used in several recent task-oriented dialogue systems (Wen et al., 2017; Zhao et al., 2017, 2018; Zhao and Eskenazi, 2018; Le et al., 2018).

However, we notice that naively using latent variables does not help the model improve much in slot filling and intent prediction. We hypothesize that the variance of the cross-lingual word embeddings is too large for the model to learn any meaningful latent variables. Hence, we propose to first refine the cross-lingual embeddings with ~10 seed word-pairs related to the dialogue domains. We then add Gaussian noise (Zheng et al., 2016) to further compensate the imperfect alignment of cross-lingual embeddings.

As a result, a combination of these methods allows us to build a transferable latent variable model that learns the distribution of training language inputs that is invariant to noise in the cross-
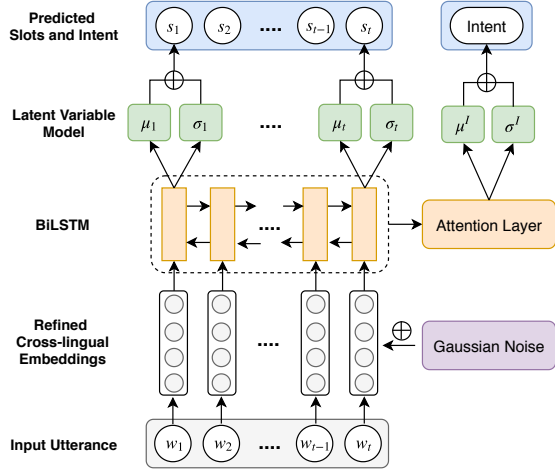
1297

Figure 2: The structure of our proposed model.

lingual word embeddings. This enables our model to capture the variance of semantically similar sentences across different languages, and achieve state-of-the-art results in zero-shot adaptation of English to Spanish and Thai for the natural language understanding task (i.e., the intent prediction and slot filling) on the dataset proposed by Schuster et al. (2019), even though we use much less external resources (i.e., ~10 seed word-pairs) while others utilize a large amount of bilingual corpus. We further visualize the learned latent variables to confirm that same-meaning words and sentences have similar distributions.

2 Related Work

Cross-lingual transfer learning which acts as one of the low-resource topics (Gu et al., 2018; Lee et al., 2019; Liu et al., 2019; Xu et al., 2018) has attracted more and more people recently, followed by the rapid development of cross-lingual word embeddings. Artetxe et al. (2017) proposed a self-learning framework and utilized a small size of word dictionary to learn the mapping between source and target word embeddings. Conneau et al. (2018) leveraged adversarial training to learn a linear mapping from a source to a target space without using parallel data. Joulin et al. (2018) utilized Relaxed CSLS loss to optimize this mapping problem. Winata et al. (2019) introduced a method to leverage cross-lingual meta-representations for code-switching name entity recognition by combining multiple monolingual word embeddings.

Chen et al. (2018) proposed a teacher-student framework leveraging bilingual data for cross-lingual transfer learning in dialogue state track-

ing. Upadhyay et al. (2018) leveraged joint training and cross-lingual embeddings to do zero-shot and almost zero-shot transfer learning in intent prediction and slot filling. Finally, Schuster et al. (2019) utilizes Multilingual CoVe embeddings obtained from training Machine Translation systems as in (McCann et al., 2017). The main difference of our work with previous work is that our model does not leverage any external bilingual data other than 11 word pairs for embeddings refinement.

3 Methodology

Our model consists of a refined cross-lingual embedding layer followed by a BiLSTM (Hochreiter and Schmidhuber, 1997) which parameterizes the Latent Variable Model, as illustrated in Figure 2. We jointly train our model to predict both slots and user intents. We denote $\mathbf{w} = [w_1, \ldots, w_T]$ as the input words and $\mathbf{e} = [e_1, \ldots, e_T]$ as the word embeddings of $\mathbf{w}$. The slot at time-step t is s_t, while the intent for each sentence $\mathbf{w}$ is denoted as I. Note that only matrices are bold-faced.

3.1 Cross-lingual Embeddings Refinement

To further refine the cross-lingual alignments to our task, we draw from the hypothesis that domain-related words are more important than others. Hence, as shown in Figure 1, we propose to refine the cross-lingual word embeddings (Joulin et al., 2018)[1] using very few parallel word pairs, which is obtained by selecting 11 English words related to dialogue domains (weather, alarm, and reminder) and translate them using bilingual lexicons. We refine the embeddings by leveraging the framework proposed in Artetxe et al. (2017).

Let $\mathbf{X}$ and $\mathbf{Z}$ be the aligned cross-lingual word embeddings between two languages. X_{i*} and Z_{j*} are the embeddings for the ith source word and j^{th} target word. We denote a binary dictionary matrix $\mathbf{D}$: $D_{ij} = 1$ if the ith source language word is aligned with the jth target language word and $D_{ij} = 0$ otherwise. The goal is to find the optimal mapping matrix $\mathbf{W}^*$ by minimizing:

$$\mathbf{W}^* = \arg \min_{\mathbf{W}} \sum_{i,j} D_{ij} ||X_{i*}\mathbf{W} - Z_{j*}||^2. \quad (1)$$

Following Artetxe et al. (2016), with orthogonal constraints, mean centering, and length normaliza-

---

[1]The embeddings are available in https://fasttext.cc

tion, we can maximize the following instead:

$$\mathbf{W}^* = \arg\max_{\mathbf{W}} \mathrm{Tr}(\mathbf{X}\mathbf{W}\mathbf{Z}^\mathrm{T}\mathbf{D}^\mathrm{T}). \qquad (2)$$

We iteratively optimize Equation 2 until distances between domain-related seed words are closer than a certain threshold after refinement. Figure 1 illustrates better alignment for domain-related words after refinement.

3.2 Gaussian Noise Injection

To cope with the noise in alignments, we inject Gaussian noise to English embeddings, so the trained model will be more robust to variance. This is a regularization method to improve the generalization ability to the unseen inputs in different languages, particularly languages from different roots such as Thai and Spanish. The final embeddings are $\mathbf{e}^* = [e_1 + N_1, \ldots, e_T + N_T]$, where $\mathbf{N} \sim \mathcal{N}(0, 0.1\mathbf{I})$.

3.3 Latent Variable Model (LVM)

Given a near-perfect cross-lingual embedding, there is still noise caused by the inherent discrepancies between source and target languages. This noise amplifies when combined with imperfect alignment, and makes point estimation vulnerable to the small, but not negligible differences across languages. Instead, using latent variables will allow us to model the distribution that captures the variance of semantically similar sentences across different languages. The whole training process is defined as follows:

$$[h_1...h_t...h_T] = \mathrm{BiLSTM}(\mathbf{e}^*), \qquad (3)$$

$$m_t = h_t w_a, \; a_t = \frac{exp(m_t)}{\sum_{j=1}^{T} exp(m_j)}, \; v = \sum_{t=1}^{T} a_t h_t, \qquad (4)$$

$$\begin{bmatrix} \mu_t^S \\ \log{(\sigma_t^S)^2} \end{bmatrix} = \mathbf{W}_r^S h_t, \begin{bmatrix} \mu^I \\ \log{(\sigma^I)^2} \end{bmatrix} = \mathbf{W}_r^I v, \qquad (5)$$

$$z_t^S \sim q_t^S(z|h_t), \; z^I \sim q^I(z|v), \qquad (6)$$

$$p_t^S(s_t|z_t^S) = \mathrm{Softmax}(\mathbf{W}_g^S z_t^S), \qquad (7)$$

$$p^I(I|z^I) = \mathrm{Softmax}(\mathbf{W}_g^I z^I), \qquad (8)$$

where attention vector (v) is obtained by following Felbo et al. (2017) and w_a is the weight matrix for the attention layer, $\mathbf{W}_{\{r,g\}}^{\{S,I\}}$ are trainable parameters, superscripts S and I refer to slot prediction and intent detection respectively, subscript "r" refers to "recognition" for obtaining the

mean and variance vectors while subscript "g" refers to "generation" for predicting the slots and intents, and $q_t^S \sim \mathcal{N}(\mu_t^S, (\sigma_t^S)^2\mathbf{I})$ and $q^I \sim \mathcal{N}(\mu^I, (\sigma^I)^2\mathbf{I})$ are the posterior approximations which we sample our latent vectors z_t^S and z^I from. Finally, p_t^S and p^I are the predictions for the slot of the *t-th* token and the intent of the utterance respectively. The objective functions for slot filling and intent prediction are:

$$\mathcal{L}^I = \mathbb{E}_{z^I}[\log p^I(I|z^I)], \qquad (9)$$

$$\mathcal{L}_t^S = \mathbb{E}_{z_t^S}[\log p_t^S(s_t|z_t^S)], \qquad (10)$$

$$\mathcal{L}^S = \sum_{t=1}^{T} L_t^S, \qquad (11)$$

hence, the final objective function to minimize is,

$$\mathcal{L} = \mathcal{L}^S + \mathcal{L}^I. \qquad (12)$$

The model prediction is not deterministic since the latent variables z_t^S and z^I are sampled from the Gaussian distributions. Therefore, in the inference time, we use the true mean μ_t^S and μ^I to replace z_t^S and z^I respectively to make the prediction deterministic.

4 Experiments

4.1 Dataset

We conduct our experiments under the zero-shot scenario of multilingual task-oriented dialogue dataset presented by Schuster et al. (2019). Our model is trained only with the English data and then do a zero-shot test on Spanish and Thai test set. We delexicalize words by replacing the tokens which represent numbers, time (such as am, pm), and duration (such as 30min) with special tokens <number>, <time>, and <last> respectively.

4.2 Training Details

In the training procedure, we freeze the word embeddings of the primary language, and then replace them with the corresponding aligned word embeddings of the unseen languages for a zero-shot test. We use bi-directional LSTM model with hidden dimension size of 250, and the latent variable model with both mean and variance in the size of 100. Gaussian noise with zero mean and variance of 0.1 is injected dynamically in different iterations. We use the accuracy to evaluate the performance of intent prediction and the standard

Model	Spanish				Thai			
	Intent acc.		Slot F1		Intent acc.		Slot F1	
	LVM	CRF	LVM	CRF	LVM	CRF	LVM	CRF
Vanilla BiLSTM	46.36	44.13	15.64	11.32	35.12	33.57	5.82	5.24
+ *noise (N)*	72.97	66.95	46.56	20.27	40.37	37.53	10.66	6.51
+ *refinement (R)*	87.69	88.23	61.63	42.62	59.40	59.28	21.84	16.53
+ *noise & refinement*	89.21	88.79	64.04	43.98	70.81	64.48	29.54	17.46
+ *N & R & delexicalization*	**90.20**	89.98	**65.79**	47.70	**73.43**	69.62	**32.24**	23.11
Zero-shot SLU [‡]	46.64		15.41		35.64		12.11	
Multi. CoVe w/ auto	53.89		19.25		70.70		35.62	
Translate Train [†]	*85.39*		*72.87*		*95.85*		*55.43*	

Table 1: Results on different models including baseline models, where N refers to the Gaussian noise injection and R refers to the cross-lingual embeddings refinement. [‡] We implemented Upadhyay et al. (2018) model and evaluated with our test set. [†] Schuster et al. (2019) translated English data to Spanish and Thai with a trained supervised machine translation system and it is considered as our *upper bound* result.

BIO structure to calculate the F1 score for evaluating the performance of slot filling. In the zero-shot cross-lingual adaptation, we simply replace the training language (i.e., English) word embeddings with the cross-lingual target language (i.e., Spanish or Thai) word embeddings. Note that we never use any target language evaluation data to select the model for zero-shot cross-lingual adaptation, instead, we utilize the English validation set and early stop strategy according to the slot F1 score.

4.3 Word Pairs

We choose the number of word pairs based on the vocabulary size of the corpus. Intuitively, the larger the vocabulary size is, the more words we need to align across languages, and the more word pairs we need to achieve good performance. We select 11 domain-related words which frequently exist in the English training set. The number of words we select is around 0.25% of the vocabulary size for the English training set. The concrete information of the 11 word pairs is as follows:

The English seed words we selected are *weather, forecast, temperature, rain, hot, cold, remind, forget, alarm, cancel, tomorrow*, which are related to the three dialogue domains (weather, alarm, and reminder). We translate them by leveraging bilingual dictionaries[2]. The corresponding translations in Spanish and Thai are *clima, pronóstico, temperatura, lluvia, caliente, frío, recordar, olvidar, alarma, cancelar, mañana* and อากาศ, พยากรณ์, อุณหภูมิ, ฝน, ร้อน, หนาว,

---
[2]https://github.com/facebookresearch/MUSE

Model	Spanish		Thai	
	Intent	Slot	Intent	Slot
Our model	**90.20**	**65.79**	**73.43**	**32.24**
- *LVM*	85.85	61.86	66.01	25.22
- *LVM + MLP*	86.02	62.34	66.56	28.35

Table 2: Ablation Study on LVM models. {- *LVM*} means removing LVM and {- *LVM + MLP*} means replacing LVM with a Multi-Layer Perceptron which has the same size as the LVM.

เตือน, ลืม, เตือน, ยกเลิก, พรุ่ง respectively.

4.4 Evaluation

We implement and evaluate the following models:

Zero-shot SLU Upadhyay et al. (2018) used cross-lingual embeddings (Bojanowski et al., 2017) to do zero-shot transfer learning.

Conditional Random Fields (CRF) We reproduce the baseline model in Schuster et al. (2019), and also add embedding noise, cross-lingual refinement, and delexicalization.

Latent Variable Model (LVM) - Ours We replace the CRF module with latent variables and also apply it to intent prediction.

Besides, we directly compare with the baseline models illustrated in Schuster et al. (2019):

Multi. CoVe w/ auto They combined Multilingual CoVe (Yu et al., 2018) with an auto-encoder objective and then used the trained encoder with the CRF model.
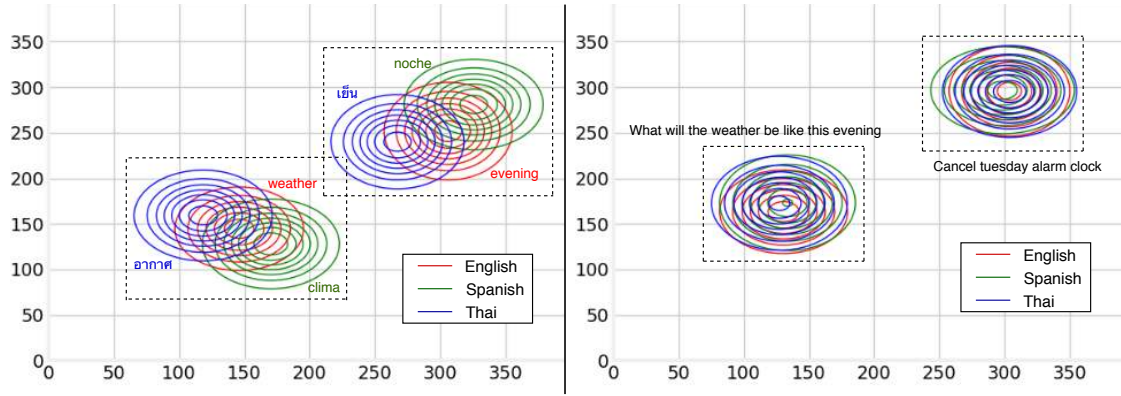
Figure 3: Visualization of latent variables on words **(left)** and sentences **(right)**. **Left:** We choose "weather-clima-อากาศ" and "evening-noche- เย็น" from parallel sentences. *English: "What will the weather be like this evening"*, *Spanish: "Cómo será el clima esta noche"*, *Thai:* "ตอน เย็น นี้ อากาศ จะ เป็น อย่างไร". **Right:** We choose two English sentences and show their distributions and those of the corresponding Spanish and Thai translations.

Translate Train They trained a supervised machine translation system to translate English data into the target language and then trained the CRF model on this translated dataset.

5   Results & Discussion

From Table 1, in general, LVM outperforms CRF models. This is because for semantically same words (e.g., weather and clima) LVM considers such close enough points as the same distribution, but CRF is more likely to classify them differently. This can be shown very clearly from Figure 3, in which the latent variables demonstrate similar distributions for semantically similar sentences and words. In addition, we can see that adding only Gaussian noise to the Vanilla BiLSTM improves our prediction performance significantly, which implies that the robustness of our model towards the noisy signals which come from the target embedding inputs.

Furthermore, it is clearly visible that cross-lingual embeddings refinement is more effective in Spanish than Thai. This is attributed to the quality of alignments in the two languages. Spanish is much more lexically and grammatically similar to English than Thai, so word-level embedding refinement is reasonably good. Jointly incorporating all three methods (Gaussian noise injection, cross-lingual embeddings refinement, and delexicalization) further reduces the noise in the inputs as well as makes the model more robust to noise, which help LVM to more easily approximate the distribution.

Finally, in Table 2, we ablate the usage of LVM to see whether the boost of performance comes simply from the increase of parameter size. By removing or replacing LVM with MLP, we can see the clear performance gains by using LVM.

6   Conclusion

In this paper, we propose a transferable latent variable that focuses on improving the zero-shot cross-lingual adaptation of natural language understanding task to low-resource languages. We show that a combination of 1) cross-lingual embeddings refinement, 2) Gaussian noise injection, and 3) latent variables are effective in coping with the variance of semantically similar sentences across different languages, and the visualizations of the latent variables confirm such. We leverage very few resources (i.e., 11 seed word pairs) and achieve state-of-the-art performance for English-to-Spanish and English-to-Thai in the zero-shot cross-lingual scenario.

Acknowledgments

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Wenhu Chen, Jianshu Chen, Yu Su, Xin Wang, Dong Yu, Xifeng Yan, and William Yang Wang. 2018. Xl-nbt: A cross-lingual neural belief tracking framework. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 414–424.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations (ICLR)*.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1615–1625.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor OK Li. 2018. Universal neural machine translation for extremely low resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 344–354.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2984.

Hung Le, Truyen Tran, Thin Nguyen, and Svetha Venkatesh. 2018. Variational memory encoder-decoder. In *Advances in Neural Information Processing Systems*, pages 1508–1518.

Nayeon Lee, Zihan Liu, and Pascale Fung. 2019. Team yeon-zi at semeval-2019 task 4: Hyperpartisan news detection by de-noising weakly-labeled data. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1052–1056.

Zihan Liu, Yan Xu, Genta Indra Winata, and Pascale Fung. 2019. Incorporating word and subword units in unsupervised machine translation using language model rescoring. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 275–282.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems*, pages 6294–6305.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. 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 3795–3805.

Shyam Upadhyay, Manaal Faruqui, Gokhan Tür, Hakkani-Tür Dilek, and Larry Heck. 2018. (almost) zero-shot cross-lingual spoken language understanding. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6034–6038. IEEE.

Tsung-Hsien Wen, Yishu Miao, Phil Blunsom, and Steve Young. 2017. Latent intention dialogue models. In *International Conference on Machine Learning*, pages 3732–3741.

Genta Indra Winata, Zhaojiang Lin, and Pascale Fung. 2019. Learning multilingual meta-embeddings for code-switching named entity recognition. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 181–186.

Peng Xu, Andrea Madotto, Chien-Sheng Wu, Ji Ho Park, and Pascale Fung. 2018. Emo2vec: Learning generalized emotion representation by multi-task training. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 292–298.

Katherine Yu, Haoran Li, and Barlas Oguz. 2018. Multilingual seq2seq training with similarity loss for cross-lingual document classification. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 175–179.

Tiancheng Zhao and Maxine Eskenazi. 2018. Zero-shot dialog generation with cross-domain latent actions. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 1–10.

Tiancheng Zhao, Kyusong Lee, and Maxine Eskenazi. 2018. Unsupervised discrete sentence representation learning for interpretable neural dialog generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1098–1107.

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)*, pages 654–664.

Stephan Zheng, Yang Song, Thomas Leung, and Ian Goodfellow. 2016. Improving the robustness of deep neural networks via stability training. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.