Modeling Situations in Neural Chat Bots

Shoetsu Sato
The University of Tokyo
shoetsu@tkl.iis.u-tokyo.ac.jp

Naoki Yoshinaga
Institute of Industrial Science,
the University of Tokyo
ynaga@tkl.iis.u-tokyo.ac.jp

Masashi Toyoda
Institute of Industrial Science,
the University of Tokyo
toyoda@tkl.iis.u-tokyo.ac.jp

Masaru Kitsuregawa
Institute of Industrial Science,
the University of Tokyo
National Institute of Informatics
kitsure@tkl.iis.u-tokyo.ac.jp

Abstract

Social media accumulates vast amounts of online conversations that enable data-driven modeling of chat dialogues. It is, however, still hard to utilize the neural network-based SEQ2SEQ model for dialogue modeling in spite of its acknowledged success in machine translation. The main challenge comes from the high degrees of freedom of outputs (responses). This paper presents neural conversational models that have general mechanisms for handling a variety of situations that affect our responses. Response selection tests on massive dialogue data we have collected from Twitter confirmed the effectiveness of the proposed models with situations derived from utterances, users or time.

1 Introduction

The increasing amount of dialogue data in social media has opened the door to data-driven modeling of non-task-oriented, or chat, dialogues (Ritter et al., 2011). The data-driven models assume a response generation as a sequence to sequence mapping task, and recent ones are based on neural SEQ2SEQ models (Vinyals and Le, 2015; Shang et al., 2015; Li et al., 2016a,b; Xing et al., 2017). However, the adequacy of responses generated by these neural models is somewhat insufficient, in contrast to the acknowledged success of the neural SEQ2SEQ models in machine translation (Johnson et al., 2016).

The contrasting outcomes in machine translation and chat dialogue modeling can be explained by the difference in the degrees of freedom on output for a given input. An appropriate response to a given utterance is not monolithic in chat dialogue. Nevertheless, since only one ground truth response is provided in the actual dialogue data, the supervised systems will hesitate when choosing from the vast range of possible responses.

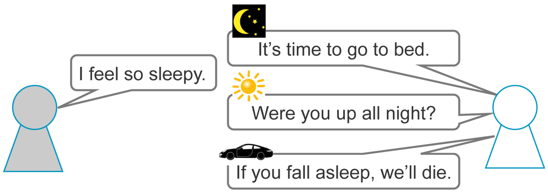

Figure 1: Conversational situations and responses.

So, how do humans decide how to respond? We converse with others while (implicitly) considering not only the utterance but also other various conversational situations (§ 2) such as time, place, and the current context of conversation and even our relationship with the addressee. For example, when a friend says "I feel so sleepy." in the morning, a probable response could be "Were you up all night?" (Figure 1). If the friend says the same thing at midnight, you might say "It's time to go to bed." Or if the friend is driving a car with you, you might answer "If you fall asleep, we'll die."

Modeling situations behind conversations has been an open problem in chat dialogue modeling, and this difficulty has partly forced us to focus on task-oriented dialogue systems (Williams and Young, 2007), the response of which has a low degree of freedom thanks to domain and goal specificity. Although a few studies have tried to exploit conversational situations such as speakers' emo-

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics- Student Research Workshop, pages 120–127
Vancouver, Canada, July 30 - August 4, 2017.
https://doi.org/10.18653/v1/P17-3020

tions (Hasegawa et al., 2013) or personal characteristics (Li et al., 2016b) and topics (Xing et al., 2017), the methods are specially designed for and evaluated using specific types of situations.

In this study, we explore neural conversational models that have general mechanisms to incorporate various types of situations behind chat conversations (§ 3.2). These models take into account situations on the speaker's side and the addressee's side (or those who respond) when encoding utterances and decoding its responses, respectively. To capture the conversational situations, we design two mechanisms that differ in how strong of an effect a given situation has on generating responses.

In experiments, we examined the proposed conversational models by incorporating three types of concrete conversational situations (§ 2): utterance, speaker/addressee (profiles), and time (season), respectively. Although the models are capable of generating responses, we evaluate the models with a response selection test to avoid known issues in automatic evaluation metrics of generated responses (Liu et al., 2016a). Experimental results obtained using massive dialogue data from Twitter showed that modeling conversational situations improved the relevance of responses (§ 4).

2 Conversational situations

Various types of conversational situations could affect our response (or initial utterance) to the addressee. Since neural conversational models need massive data to train a reliable model, our study investigates conversational situations that are naturally given or can be identified in an unsupervised manner to make the experimental settings feasible.

In this study, we represent conversational situations as discrete variables. That allows models to handle unseen situations in testing by classifying them into appropriate situation types via distributed representations or the like as described below, and helps to analyze the outputs. We consider the following conversational situations to each utterance and response in our dialogue dataset (§ 4), and cluster the situations to assign specific situation types to the utterances and responses in the training data of our conversational models.

Utterance The input utterance (to be responded to by the system) is a primary conversational situation and is already modeled by the encoder in the neural SEQ2SEQ model. However, we may be able to induce a different aspect of situations that are represented in the utterance but are not captured by the SEQ2SEQ sequential encoder (Sato et al., 2016). We first represent each utterance of utterance-response pairs in our dialogue dataset by a distributed representation obtained by averaging word2vec[1] vectors (pre-trained from our dialogue datasets (§ 4.1)) for words in the utterances. The utterances are then classified by k-means clustering to identify utterance types.[2]

User (profiles) User characteristics should affect his/her responses as Li et al. (2016b) have already discussed. We classify profiles provided by each user in our dialogue dataset (§ 4.1) to acquire conversational situations specific to the speakers and addressees. The same as with the input utterance, we first construct a distributed representation of each user's profile by averaging the pre-trained word2vec vectors for verbs, nouns and adjectives in the user profiles. The users are then classified by k-means clustering to identify user types.[3]

Time (season) Our utterances can be affected by when we speak as illustrated in § 1, so we adopted time as one conversational situation. On the basis of timestamp of the utterance and the response in our dataset, we split the conversation data into four season types: namely, spring (Mar. – May.), summer (Jun. – Aug), autumn (Sep. – Nov.), and winter (Dec. – Feb.). This splitting reflects the climate in Japan since our data are in Japanese whose speakers mostly live in Japan.

In training our neural conversational models, we use each of the above conversational situation types for the speaker side and addressee (who respond) side, respectively. Note that the utterance situation is only considered for the speaker side since its response is unseen in response generation. In testing, the conversational situation types for input utterances (or speaker and addressee's profiles) are identified by finding the closest centroid obtained by the k-means clustering of the utterances (profiles) in the training data.

3 Method

Our neural conversational models are based on the SEQ2SEQ model (Sutskever et al., 2014) and integrate mechanisms to incorporate various conversa-

[1] https://code.google.com/p/word2vec/

[2] We set k to 10. Although the space limitations preclude results when varying k, this does not affect our conclusions.

[3] We set k to 10, and add another cluster for users whose profiles were not available (6.3% of the users in our datasets).

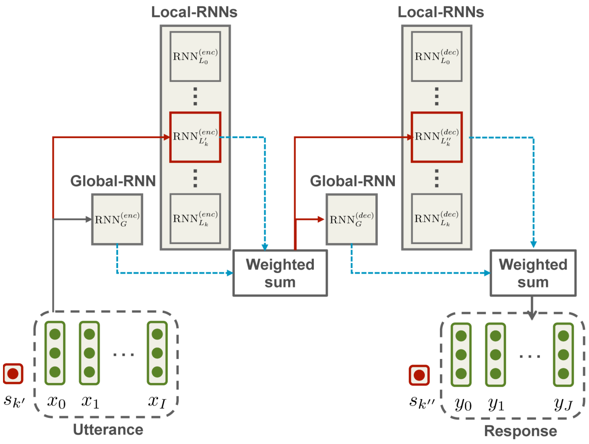
Local-RNNs
$\mathrm{RNN}^{(enc)}_{L_0}$
$\mathrm{RNN}^{(enc)}_{L_{k'}}$
$\mathrm{RNN}^{(enc)}_{L_k}$
Global-RNN
$\mathrm{RNN}^{(enc)}_{G}$
Weighted sum
Local-RNNs
$\mathrm{RNN}^{(dec)}_{L_0}$
$\mathrm{RNN}^{(dec)}_{L''_{k}}$
$\mathrm{RNN}^{(dec)}_{L_k}$
Global-RNN
$\mathrm{RNN}^{(dec)}_{G}$
Weighted sum
$s_{k'}$ x_0 x_1 x_I
Utterance
$s_{k''}$ y_0 y_1 y_J
Response

Figure 2: Local-global SEQ2SEQ.

tional situations (§ 2) at speaker side and addressee side. In the following, we briefly introduce the SEQ2SEQ conversational model (Vinyals and Le, 2015) and then describe two mechanisms for incorporating conversational situations.

3.1 SEQ2SEQ conversational model

The SEQ2SEQ conversational model (Vinyals and Le, 2015) consists of two recurrent neural networks (RNNs) called an encoder and a decoder. The encoder takes each word of an utterance as input and encodes the input sequence to a real-valued vector representing the utterance. The decoder then takes the encoded vector as its initial state and continues to generate the most probable next word and to input the word to itself until it finally outputs EOS.

3.2 Situation-aware conversational models

The challenge in designing situation-aware neural conversational models is how to inject given conversational situations into RNN encoders or decoders. In this paper, we present two situation-aware neural conversational models that differ in how strong of an effect a given situation has.

3.2.1 Local-global SEQ2SEQ

Motivated by a recent success in multi-task learning for a deep neural network (Liu et al., 2016c,b; Gupta et al., 2016; Luong et al., 2016), our local-global SEQ2SEQ trains two types of RNN encoder and decoder for modeling situation-specific dialogues and universal dialogues jointly (Figure 2).

Local-RNNs are meant to model dialogues in individual conversational situations at both the speaker and addressee sides. Each local-RNN is trained (i.e., its parameters are updated) only on dialogues under the corresponding situation. A salient disadvantage of this modeling is that the size of training data given to each local-RNN decreases as the number of situation types increases.

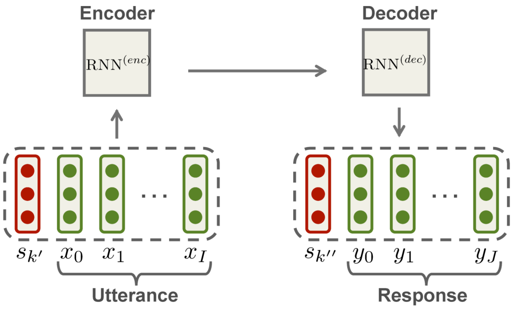

Figure 3: SEQ2SEQ with situation embeddings.

To address this problem, we combine another global-RNN encoder and decoder trained on all the dialogue data and take the weighted sum of the hidden states hs of the two RNNs for both the encoder and decoder to obtain the output as:

$$h_i^{(enc)} = \boldsymbol{W}_G^{(enc)} \mathrm{RNN}_G^{(enc)}(h_{i-1}^{(enc)}, x_i) + \boldsymbol{W}_L^{(enc)} \mathrm{RNN}_{L_{k'}}^{(enc)}(h_{i-1}^{(enc)}, x_i), \quad (1)$$

$$h_j^{(dec)} = \boldsymbol{W}_G^{(dec)} \mathrm{RNN}_G^{(dec)}(h_{j-1}^{(dec)}, y_{j-1}) + \boldsymbol{W}_L^{(dec)} \mathrm{RNN}_{L_{k''}}^{(dec)}(h_{j-1}^{(dec)}, y_{j-1}), \quad (2)$$

where $\mathrm{RNN}_G^{(\cdot)}(\cdot)$ and $\mathrm{RNN}_L^{(\cdot)}(\cdot)$ denote global-RNN and local-RNN, respectively, and the $\boldsymbol{W}$s are trainable matrices for the weighted sum. The embedding and softmax layers of the RNNs are shared.

3.2.2 SEQ2SEQ with situation embeddings

The local-global SEQ2SEQ (§ 3.2.1) assumes that dialogues with different situations involve different domains (or tasks) that are independent of each other. However, this assumption could be too strong in some cases and thus we devise another weakly situation-aware conversational model.

We represent the given situations at speaker and addressee sides, $s_{k'}$ and $s_{k''}$, as situation embeddings and then feed them to the encoder and decoder prior to processing sequences (Figure 3) as:

$$h_0^{(enc)} = \mathrm{RNN}(h_{init}, s_{k'}), \quad (3)$$

$$h_i^{(enc)} = \mathrm{RNN}(h_{i-1}^{(enc)}, x_{i-1}), \quad (4)$$

$$h_0^{(dec)} = \mathrm{RNN}(h_{I+1}^{(enc)}, s_{k''}), \quad (5)$$

$$h_j^{(dec)} = \mathrm{RNN}(h_{j-1}^{(dec)}, y_{j-1}), \quad (6)$$

where h_{init} is a vector filled with zeros and $h_{I+1}^{(enc)}$ is the last hidden state of the encoder.

Average length in words (utterances)	15.7
Average length in words (responses)	10.1
Average length in words (user profiles)	37.4
Number of users	386,078

Table 1: Statistics of our dialogue datasets (training, validation, and test portions are merged here).

This encoding was inspired by a neural machine translation system (Johnson et al., 2016) that enables multilingual translation with a single model. Whereas it inputs the target language embedding only to the encoder to control the target language, we input the speaker-side situation to the encoder and the addressee-side one to the decoder.

4 Evaluation

In this section, we evaluate our situation-aware neural conversational models on massive dialogue data obtained from Twitter. We compare our models (§ 3.2) with SEQ2SEQ baseline (§ 3.1) using a response selection test instead of evaluating generated responses, since Liu et al. (2016a) recently pointed out several problems of existing metrics such as BLEU (Papineni et al., 2002) for evaluating generated responses.

4.1 Settings

Data We built massive dialogue datasets from our Twitter archive that have been compiled since March, 2011. In this archive, timelines of about 1.5 million users[4] have been continuously collected with the official API. It is therefore suitable for extracting users' conversations in timelines.

On Twitter, a post (tweet) and a mention to it can be considered as an utterance-response pair. We randomly extracted 23,563,865 and 1,200,000 pairs from dialogues in 2014 as training and validation datasets, and extracted 6000 pairs in 2015 as a test dataset in accordance with the following procedure. Because we want to exclude utterances that need contexts in past dialogue exchanges to respond from our evaluation dataset, we restrict ourselves to only tweets that are not mentions to other tweets (in other words, utterances without past dialogue exchanges are chosen for evaluation). For each utterance-response pair in the test dataset, we randomly chose four (in total, 24,000) responses in 2015 as false response candidates which together constitute five response candidates for the response selection test. Each utterance and response (candidate) is tokenized by MeCab[5] with NEologd[6] dictionary to feed the sequence to the word-based encoder decoder.[7] Table 1 shows statistics on our dialogue datasets.

Vocabulary size	100,000
Dropout rate	0.25
Mini-batch size	800
Dimension of embedding vectors	100
Dimension of hidden states	100
Learning rate	1e-4
Number of samples in sampled softmax	512

Table 2: Hyperparameters for training.

Models In our experiments, we compare our situation-aware neural conversational models (we refer to the model in § 3.2.1 as **L/G SEQ2SEQ** and the model in § 3.2.2 as **SEQ2SEQ emb**) with situation-unaware **baseline** (§ 3.1) for taking each type of conversational situations (§ 2) into consideration. We also evaluate the model in § 3.2.1 without global-RNNs (referred to as **L SEQ2SEQ**) to observe the impact of global-RNNs.

We used a long-short term memory (LSTM) (Zaremba et al., 2014) as the RNN encoder and decoder, sampled softmax (Jean et al., 2015) to accelerate the training, and TensorFlow[8] to implement the models. Our LSTMs have three layers and are optimized by Adam (Kingma and Ba, 2015). The hyperparameters are fixed as in Table 2.

Evaluation procedure We use the above models to rank response candidates for a given utterance in the test set. We compute the averaged cross-entropy loss for words in each response candidate (namely, its perplexity) by giving the candidate following the input utterance to each conversational model, and used the resulting values for ranking candidates to choose top-k plausible ones. We adopt **1 in t P@k** (Wu et al., 2016) as the evaluation metric, which indicates the ratio of utterances that are provided the single ground truth in top k responses chosen from t candidates. Here we use **1 in 2 P@1**,[9] **1 in 5 P@1**, and **1 in 5 P@2**.

[4] Our collection started from 26 popular Japanese users in March 2011, and the user set has iteratively expanded to those who are mentioned or retweeted by already targeted users.

[5] http://taku910.github.io/mecab/

[6] https://github.com/neologd/mecab-ipadic-neologd

[7] The number of words in the utterances and the response candidates in the test set is limited to equal or less than 20, since very long posts do not constitute usual conversation.

[8] https://www.tensorflow.org/

[9] We randomly selected one false response candidate from the four pre-selected ones when $t = 2$.

Model	1 in 2 P@1	1 in 5 P@1	1 in 5 P@2
Baseline	64.5%	33.9%	56.6%
Situation: utterance			
L Seq2Seq	67.2%	37.2%	60.6%
L/G Seq2Seq	**68.5%**	**38.2%**	**62.1%**
Seq2Seq emb	65.6%	35.4%	58.2%
Situation: speaker/addressee (profiles)			
L Seq2Seq	67.3%	**38.0%**	60.9%
L/G Seq2Seq	66.4%	36.4%	59.2%
Seq2Seq emb	**67.8%**	37.5%	**61.1%**
Situation: time (season)			
L Seq2Seq	*62.0%*	*30.8%*	*54.8%*
L/G Seq2Seq	65.9%	35.8%	58.1%
Seq2Seq emb	**67.3%**	**37.6%**	**60.7%**

Table 3: Results of the response selection test.

4.2 Results

Table 3 lists the results of the response selection test. The proposed conversational models successfully improved the relevance of selected responses by incorporating conversational situations.

The proposed model that performed best is different depending on the situation type. We found from the dataset that many of the conversations did not seem to be affected by the seasons, that is, time (season) situation is less influential than other situations. This explains the poor performance of **L Seq2Seq** with time (season) situations due to the data sparseness in training local-RNNs, although the sparseness is mostly addressed by global RNNs in **L/G Seq2Seq**.

As stated in § 3.2.2, **L/G Seq2Seq** is expected to capture situations more strongly than **Seq2Seq emb**. To confirm this, we plotted scattergrams of the utterance vectors (Figure 4) and the user profile vectors (Figure 5) in the training data by using t-SNE (Maaten and Hinton, 2008). We provide cluster descriptions by manually looking into the content of the utterances and user profiles in each cluster. The descriptions are followed by ↗ if **L/G Seq2Seq** performed better than **Seq2Seq emb** in terms of 1 in 5 P@1 for test utterances with the corresponding situation type, by ↘ if the opposite and by → if comparable (differences are within ± 1.0%). Elements of clusters were randomly sampled.

L/G Seq2Seq tends to perform better for utterances with densely concentrated (or coherent) speaker profile clusters (Figure 5). This is because utterances given by the speakers in these coherent clusters (and the associate responses) have similar conversations, situations of which are captured by local-RNNs in the local-global Seq2Seq.

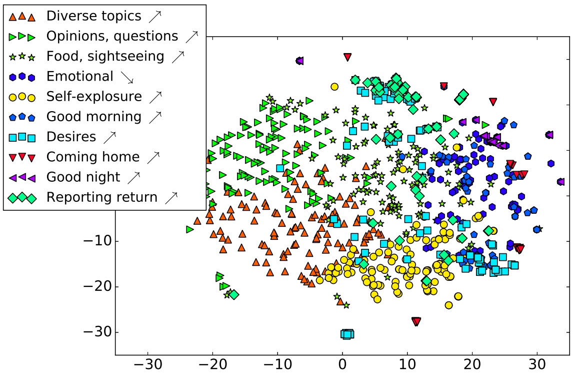

Figure 4: The scattergram of sampled utterance vectors visualized using t-SNE.

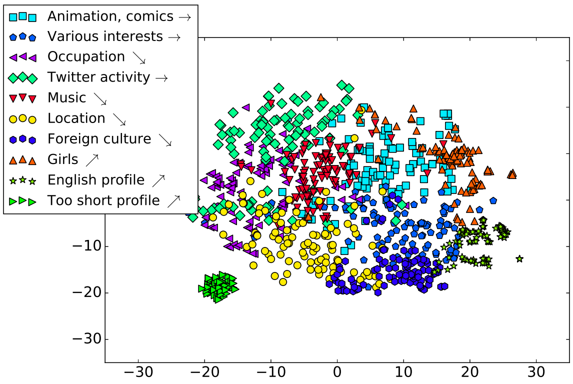

Figure 5: The scattergram of sampled user profile vectors visualized using t-SNE.

This explains the reason why **L/G Seq2Seq** outperformed the other situation-aware conversational models when utterance situations are considered (Figure 4). Conversations in the same clusters are naturally consistent, and conversations assigned to the same clusters form typical activities or specific tasks (*e.g.*, greetings, following other users, and questions (and answering)) in Twitter conversation. **L/G Seq2Seq**, designed as a kind of multi-task Seq2Seq, literally captures these task-specific behaviors in the conversations.

Although some utterance clusters have general conversations (*e.g.*, diverse topics), the response performances in those clusters have still improved. This is because these general clusters are free from harmful common responses that are quarantined into situation-specific clusters (*e.g.*, greetings etc.) and the corresponding local-RNNs should avoid generating those common responses. Note that this problem has been pointed out and addressed by Li et al. (2016a) in a totally different way.

Situation: utterance (opinions, questions)	
Input:	ちょっと最近BOTのフォロー多いんですけど (I've recently been followed by many bot accounts.)
Baseline	お疲れ様やで (You've gotta be tired.)
L/G SEQ2SEQ	ブロックしちゃいましょう (Let's block them.)
Situation: addressee profiles (girls)	
Input:	なにグラブル始めてるんだ原稿しろ (Why am I starting Granblue Fantasy? I have to write the paper...)
Baseline	おい、大丈夫か？ (Hey, are you okay?)
SEQ2SEQ emb	フレンドなろ♡ (Let's be friends♡)
Situation: time (season) (summer)	
Input:	7月になって、流石にパーカーは暑くなってきた (July is too warm to wear a hoodie.)
Baseline	そうなんです! (Yes!)
SEQ2SEQ emb	まだ着てたの!? (Do you still wear one?)

Table 4: Responses selected by the systems.

Examples Table 4 lists the response candidate selected by the baseline and our models. As we had expected, the situation-aware conversational models are better at selecting ground-truth responses for situation-specific conversations.

5 Related Work

Conversational situations have been implicitly addressed by preparing datasets specific to the target situations and by solving the problem as a task-oriented conversation task (Williams and Young, 2007); examples include troubleshooting (Vinyals and Le, 2015), navigation (Wen et al., 2015), interviewing (Kobori et al., 2016), and restaurant search (Wen et al., 2017). In what follows, we introduce non-task-oriented conversational models that explicitly consider conversational situations.

Hasegawa et al. (2013) presented a conversational model that generates a response so that it elicits a certain emotion (*e.g.*, joy) in the addressee mind. Their model is based on statistical machine translation and linearly interpolates two conversational models that are trained from a small emotion-labeled dialogue corpus and a large non-labeled dialogue corpus, respectively. This model is similar to our local-global SEQ2SEQ but differs in that it has hyperparameters for the interpolation, whereas our local-global SEQ2SEQ automatically learns $\boldsymbol{W}_G$ and $\boldsymbol{W}_L$ from the training data.

Li et al. (2016b) proposed a neural conversational model that generates responses taking into consideration speakers' personalities such as gender or living place. Because they fed a specific speaker ID to their model and represent individual (known) speakers with embeddings, Their model cannot handle unknown speakers. In contrast, our model can consider any speakers with profiles because we represent each cluster of profiles with an embedding and find an appropriate profile type for the given profile by nearest-neighbor search.

Sordoni et al. (2015) encoded a given utterance and the past dialogue exchanges, and combined the resulting representations for RNN to decode a response. Zhao et al. (2017) used a conditional variational autoencoder and automatically-induced dialogue acts to handle discourse-level diversity in the encoder. While these sophisticated architectures are designed to take dialogue histories into consideration, our simple models can easily exploit various situations.

Recently, Xing et al. (2017) proposed to explicitly consider topics of utterances to generate topic-coherent responses. Although they used latent Dirichlet allocation while we use k-means clustering, both methods confirmed the importance of utterance situations. The way to obtain specific situations is still an open research problem. As demonstrated in this study, our primary contribution is the invention of neural mechanisms that can consider various conversational situations.

Our local-global SEQ2SEQ model is closely related to a many-to-many multi-task SEQ2SEQ proposed by Luong et al. (2016). The critical difference is in that their model assumes only local tasks, while our model assumes many local tasks (situation-specific dialogue modeling) and one global task (general dialogue modeling).

6 Conclusion

We proposed two situation-aware neural conversational models that have general mechanisms for handling various conversational situations represented by discrete variables: (1) local-global SEQ2SEQ that combines two SEQ2SEQ models (§ 3.2.1) to handle situation-specific dialogues and universal dialogues jointly, and (2) SEQ2SEQ with situation embeddings (§ 3.2.2) that feeds the situations directly to a SEQ2SEQ model. The response selection tests on massive Twitter datasets confirmed the effectiveness of using situations such as utterances, user (profiles), or time.

Acknowledgements This work was partially supported by JSPS KAKENHI Grant Number 16K16109 and 16H02905.

References

Pankaj Gupta, Hinrich Schütze, and Bernt Andrassy. 2016. Table filling multi-task recurrent neural network for joint entity and relation extraction. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING-16)*. pages 2537–2547.

Takayuki Hasegawa, Nobuhiro Kaji, Naoki Yoshinaga, and Masashi Toyoda. 2013. Predicting and eliciting addressee's emotion in online dialogue. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*. pages 964–972.

Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. On using very large target vocabulary for neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP-16)*. pages 1–10.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2016. Google's multilingual neural machine translation system: Enabling zero-shot translation. *arXiv preprint arXiv:1611.04558* .

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the third International Conference on Learning Representations (ICLR-15)*.

Takahiro Kobori, Mikio Nakano, and Tomoaki Nakamura. 2016. Small talk improves user impressions of interview dialogue systems. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL-16)*. pages 370–380.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-16)*. pages 110–119.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. 2016b. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL-16)*. pages 994–1003.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016a. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP-16)*. pages 2122–2132.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016b. Deep multi-task learning with shared memory for text classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP-16)*. pages 118–127.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016c. Recurrent neural network for text classification with multi-task learning. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)*. pages 2873–2879.

Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence to sequence learning. In *Proceedings of the fifth International Conference on Learning Representations (ICLR-16)*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research* 9(Nov):2579–2605.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*. pages 311–318.

Alan Ritter, Colin Cherry, and William B Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP-11)*. pages 583–593.

Shoetsu Sato, Naoki Yoshinaga Shonosuke Ishiwatari, Masashi Toyoda, and Masaru Kitsuregawa. 2016. UT dialogue system at NTCIR-12 STC. In *Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies*. pages 518–522.

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 (ACL-IJCNLP-15)*. pages 1577–1586.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-15)*. pages 196–205.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS-14)*. pages 3104–3112.

Oriol Vinyals and Quoc V. Le. 2015. A neural conversational model. In *Proceedings of Deep Learning Workshop held at the 31st International Conference on Machine Learning (ICML-15)*.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP-15)*. pages 1711–1721.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gasic, Lina M. Rojas Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL-17)*. pages 438–449.

Jason D. Williams and Steve Young. 2007. Partially observable markov decision processes for spoken dialog systems. *Computer Speech and Language* 21(2):393–422.

Bowen Wu, Baoxun Wang, and Hui Xue. 2016. Ranking responses oriented to conversational relevance in chat-bots. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING-16)*. pages 652–662.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generation. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI-17)*. pages 3351–3357.

Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. 2014. Recurrent neural network regularization. In *Proceedings of the third International Conference on Learning Representations (ICLR-14)*.

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 (ACL-17)*.