A Key-Phrase Aware End2end Neural Response Generation Model

Jun XU[1], Haifeng WANG[2], Zhengyu NIU[2], Hua WU[2], and Wanxiang CHE[1]

[1] Research Center for Social Computing and Information Retrieval Harbin Institute of Technology, China {jxu,car}@ir.hit.edu.cn
[2] Baidu Inc., Beijing, China
{wanghaifeng,niuzhengyu,wu_hua}@baidu.com

Abstract. Previous Seq2Seq models for chitchat assume that each word in the target sequence has direct corresponding relationship with words in the source sequence, and all the target words are equally important. However, it is invalid since sometimes only parts of the response are relevant to the message. For models with the above mentioned assumption, irrelevant response words might have a negative impact on the performance in semantic association modeling that is a core task for open-domain dialogue modeling. In this work, to address the challenge of semantic association modeling, we automatically recognize key-phrases from responses in training data, and then feed this supervision information into an enhanced key-phrase aware seq2seq model for better capability in semantic association modeling. This model consists of an encoder and a two-layer decoder, where the encoder and the first layer sub-decoder is mainly for learning semantic association and the second layer sub-decoder is for responses generation. Experimental results show that this model can effectively utilize the key phrase information for semantic association modeling, and it can significantly outperform baseline models in terms of response appropriateness and informativeness.

Keywords: key-phrase · end2end · neural dialog model.

1 Introduction

Previous Seq2Seq model for chitchat [10, 14] based on the following assumption: each word in the target sequence have direct corresponding relationship with the source sequence, and all the target words are equally important. However, this assumption of seq2seq becomes invalid in the context of chitchat, sometimes only a part of the response has semantic association relationship with the message. Given an utterance, humans may initiate a new topic in response so that the dialogue can continue. We use an example to illustrate this kind of phenomenon in dialogue data, shown as follows:

message: Playing football exhausts me.
response: You can take a rest. By the way, how about going shopping tomorrow?

The phrase "take a rest" responds upon the phrase "exhausts me" in the message. But other content-word parts in the response, such as "going shopping tomorrow", are irrelevant to the message. Therefore, there is semantic association relationship between "exhausts me" and "take a rest", but not between "exhausts me" and "going shopping tomorrow". In this paper, phrases (from the response) being relevant to the message are called **key-phrases**, e.g., "take a rest" . This phenomenon is quite common[3]. Therefore, the underlying assumption of previous seq2seq based chitchat models is not valid anymore. For models with the above mentioned assumption, irrelevant response words, or non-key-phrase response words, might deteriorate their performance in semantic association modeling.

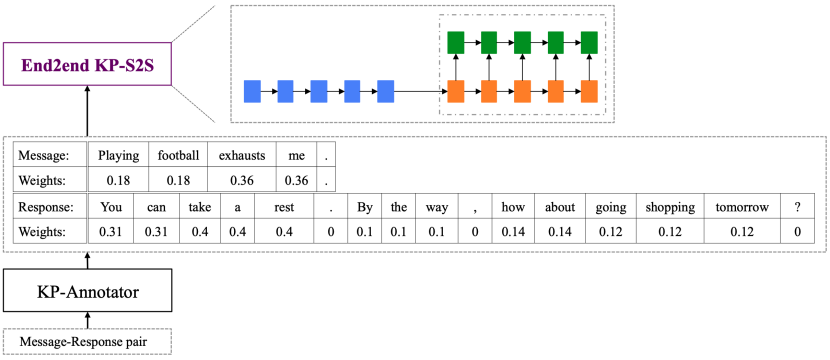

Fig. 1. The architecture of our system which consists of two modules.

In this work, to address the challenge of semantic association modeling, we automatically recognize key-phrases from responses in training data, and then feed these supervision information into an enhanced seq2seq model for better capability in semantic association modeling. For key-phrase recognition in training data, we employ a key-phrase annotator (**KP-Annotator** for short) to calculate weights for each word in responses, where KP-Annotator is built on a manually annotated corpus in a supervised way. Each word weight indicates how much a word in the response is semantically associated with the message. It is expected that response words closely associated with messages have higher weights. Phrases formed by these high-weight words are considered as key-phrases. Then with message-response pairs and word weights as inputs, we employ a key-phrase aware two-layer-decoder based seq2seq model (**KP-S2S** for short) for semantic association modeling and response generation in a joint way, instead of the pipelined approach in previous works [15] [8] [18]. During training procedure,

[3] We annotated key-phrases for responses from randomly sampled 200 message-response pairs, extracted from Baidu Tieba. We found that there are 78% pairs in which non-key-phrases exist, or only a part of the response is relevant to the message.

we expect that KP-S2S will pay more attention on key-phrases, or important response words, with the use of word-weight information. Then it will significantly reduce the negative impact of non-key-phrase words on semantic association modeling.

KP-S2S consists of an encoder and a two-layer decoder. The first layer sub-decoder (a vanilla LSTM decoding unit) is for semantic association modeling, and the second layer sub-decoder (an enhanced LSTM decoding unit) is for response generation. For semantic association modeling, we employ weighted loss function mechanism and weighted learning rate mechanism in the encoder and the first layer sub-decoder during training procedure. With the two mechanisms, we "mask" non-key-phrase words during gradient calculation and parameter updating in a soft way, and thus it can help semantic association modeling between messages and relevant parts in responses. The first layer sub-decoder is expected to enable KP-S2S to promote informative and appropriate responses, and downgrade generic or inappropriate responses. The second layer sub-decoder is responsible for generation of the whole response. We expect that the use of the second layer sub-decoder can help generate fluent and appropriate responses. In test procedure of KP-S2S, no extra information (e.g., weights of words in a message) is required.

We conduct an empirical study of our model and a set of carefully selected state-of-the-art baseline models on a large Chinese dialogue corpus from Baidu Tieba. Experiment results confirm that in comparison with baselines, our model can produce a much higher ratio of appropriate and informative responses.

In summary, this paper makes the following contributions:

- To the best of our knowledge, this is the first work to explicitly reveal semantic association information in a message-response pair. These annotation results could be applicable for other data-driven conversation models. We demonstrate their effectiveness on HGFU model in our experiment.
- We propose an end2end key-phrase aware hierarchical response generation model with a two-layer decoder. It can effectively learn semantic association between words from message-response pairs, and generate a much higher ratio of informative and appropriate responses in comparison with baselines.

2 The Proposed Approach

2.1 The Architecture

Figure 1 provides the architecture of our system, consisting of (1) Key-phrase annotator (KP-Annotator for short), to annotate key-phrase information in training data and (2) Key-phrase aware seq2seq model (KP-S2S for short), to learn a response generation model.

2.2 Key-phrase Annotator (KP-Annotator)

Phrase Extraction We use a Chinese dependency parser [16] to obtain the tree-structure of an utterance, and then extract all grammatical phrases that meet following requirements:

- There is one and only one dependency edge in the phrase, where the edge comes from a word outside of the phrase, to ensure the phrase is "grammatical";
- Words within the phrase are consecutive in the utterance;
- The phrase contains at least two Chinese characters and at most four words;

Scoring Network We use two one-layer RNNs to encode a phrase Q and an utterance U into s_Q and s_U respectively, and calculate their relevance score as: $s(Q,U) = sigmoid(MLP([s_Q; s_U]))$, where $MLP(.)$ is a multi-layer perceptron. We call this relevance score the weight of the phrase Q.

We train this network with Max-Margin loss such that $s(Q^+,U)$ is larger than $s(Q^-,U)$ with at least Δ threshold, and the objective function is given by

$$L_s = max(\Delta - s(Q^+,U) + s(Q^-,U), 0), \tag{1}$$

where Q^- is a non-key-phrase co-occuring with a key-phrase Q^+ in the same utterance. Q^- and Q^+ are labeled manually. Finally, each word in the response is annotated with the maximum weight of phrases that contain the word (0 if no phrase includes this word). We calculate weights for words in messages similarly.

2.3 Model: KP-S2S

Figure 2 provides the architecture of KP-S2S, which consists of an encoder and a two-layer decoder. The encoder and the first layer sub-decoder are equipped with two weighted mechanism in the training procedure, to effectively capture semantic association between a message and a response. The second layer sub-decoder is designed to balance the generation of key-phrase information and other parts in the response.

The Model As shown in Figure 2, architecture of the encoder and the first layer sub-decoder is similar to seq2seq model with attention mechanism. The input to the second layer sub-decoder at time t consists of (1) the t-th output of the first layer sub-decoder m_t, which is calculated by combining the t-th hidden state of the first layer sub-decoder s_t with the t-th attention vector a_t, and (2) the word embedding $e'_{y_{t-1}}$. Note that $e'_{y_{t-1}}$ is independent of the embedding $e_{y_{t-1}}$ used in the first layer sub-decoder, as the first layer sub-decoder embedding e_{y-1} is designed to mainly focus on modeling semantic association between messages and responses, which is different from generating response utterances required by the second layer sub-decoder. The t-th output of response generator is calculated by the t-th hidden state r_t and m_t as follows:

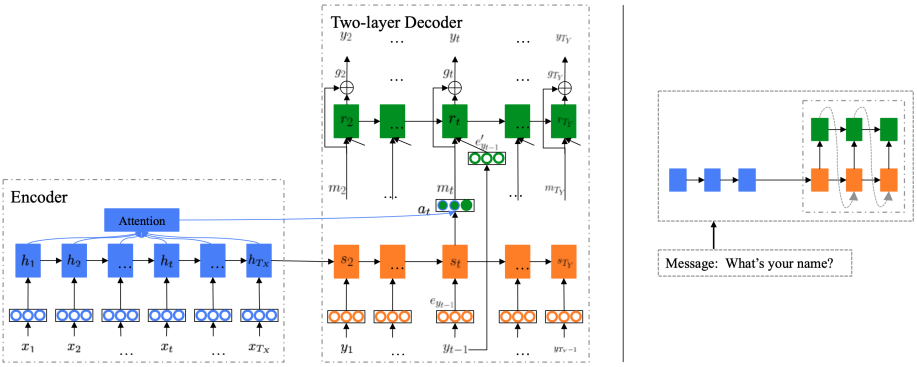

Fig. 2. An overview of the end2end KP-S2S model. The left is the proposed KP-S2S model, and the right is the testing procedure where only message is required for response generation.

$$g_t = W^o[r_t; m_t], \tag{2}$$

$$r_t = F(r_{t-1}, W^i[m_t; e'_{y_{t-1}}]), \tag{3}$$

where F is the LSTM cell, m_t is the output state, W^o and W^i are parameter matrices and $[;]$ denotes a concatenation operation.

With equation 3, the i-th output of the second layer sub-decoder is generated based on combination of the input message, predicted key-phrase information (i.e. the output of the first layer sub-decoder) and words ahead of y_i in the response. KP-S2S is expected to generate appropriate and informative responses since (1) the encoder and the first layer sub-decoder focus on modeling key-phrases association, which is helpful to generate informative responses, and (2) the second layer sub-decoder is for the generation procedure of the whole response, beneficial to generate fluent responses.

Two mechanism To effectively learn semantic association between a message and a response, two mechanisms, weighted loss function mechanism and weighted learning rate mechanism, are applied to the encoder and the first layer sub-decoder respectively .

– Weighted Loss Function Mechanism We introduce a weighted loss function in which the contribution of each response word to the loss is directly weighted by its weight calculated by the KP-Annotator. We call this weighted loss function as L_{focus} as it focus on words in key-phrases (typically have high weight), it is defined as:

$$L_{focus}(X, Y) = \sum_{t=2}^{T_Y} w_{y_t} L(y_t, m_t; M), \tag{4}$$

where L is the standard loss function in Seq2Seq, $M \in \mathbb{R}^{|V| \times d_h}$ is the output word decoding matrix, m_t is the output of the decoder at time t, y_t is the (t)-th word in response utterance Y. Equation 4 means that response words from key-phrases are more important for the model loss, in comparison with words from non-key-phrases.
$L_{focus}(X, Y)$ is utilized as an auxiliary loss to help capture semantic association with the help of key phrase information (i.e. response words' weight assigned by KP-Annotator).

– Weighted Learning Rate Mechanism
Semantic association relationships captured by the encoder and the first layer decoder are partly stored in the embeddings of words in responses, it is unreasonable to update embeddings of words that not relevant to message. However, Stochastic Gradient Descent (**SGD**) usually keeps a fixed learning rate for all parameters at each optimizing step. In this work, we propose a weighted learning rate mechanism, in which learning rate are adaptive for each word based on its weight. Specifically, embeddings of words in responses are updated at each optimization step as follows:

$$e_{y_t}^{new} = e_{y_t} - (w_{y_t}\lambda)\nabla_{y_t}, \tag{5}$$

where e_{y_t} is the decoder embedding of the t-th word in a response, w_{y_t} is the weight of word y_t and λ is the learning rate kept by SGD.
With equations 6, representation learning of words from key-phrases will be guided by the loss more than those from non-key-phrases.
Meantime, as illustrated in [15], humans tend to focus on certain aspects in message to respond, rather than responds on all content in message. It indicates that sometimes not all content in message are semantically associated with response. Taken this into consideration, we utilize weighted learning rate mechanism in the updating of encoder embeddings as follows:

$$e_{x_t}^{new} = e_{x_t} - (w_{x_t}\lambda)\nabla_{x_t}, \tag{6}$$

where e_{x_t} is the encoder embedding of the t-th word in a message, w_{x_t} is the weight of x_t.

With the use of above-mentioned mechanisms, we make our model pay more attention to words from key-phrases. It helps our model to learn semantic association between messages and relevant parts in responses.

2.4 Optimization and Prediction

The loss function is defined as the sum of the standard negative log-likelihood loss utilized in optimizing Se2Seq $L_{generator}(X, Y)$ and $L_{focus}(X, Y)$:

$$L_{KP\_S2S}(X, Y) = L_{generator}(X, Y) + L_{focus}(X, Y).$$

We adopt SGD for model optimization, except that word embeddings the first layer sub-decoder and the encoder are updated with Equations 5 and 6.

In the testing procedure, **No extra information** (e.g. weights) besides message itself is needed for response generation as shown in Figure 2, and there is not internal step of predicting key-phrases.

3 Experiments and Results

3.1 Datasets

For an empirical study of our system, we first collect 20,000,000 message-response (m-r) pairs from Baidu Tieba[4]. Then, we perform Chinese word segmentation for each pair with an open-source lexical analysis tool[5] [1]. We split D into three subsets, D_{train} (3.2m pairs), $D_{validation}$ (9k pairs) and D_{test} (9k pairs). We take the most frequent 50,000 words in D_{train} as the vocabulary and other out-of-vocabulary words as UNKs. We perform word weight calculation on message-response pairs from D_{train} with the use of KP-Annotator.

The training/validation/testing set of KP-Annotator consists of 78k/19k/19k response-phrase (from corresponding message) pairs labeled by human, with equal share of positive pairs and negative pairs.

3.2 Evaluation metrics

Automatic Metrics Following previous works, we apply three kinds of embedding-based metrics introduced in [6] and Distinct-i metric was proposed in [3].

Human evaluation We randomly sample 300 cases and invite three annotators to evaluate the quality of generated responses from 4 models. For each message-response pair, annotators are asked to rate with a score from {"0", "+1", "+2"}. A response will be rated with "0" if it is inappropriate as an reply to a message. We define inappropriateness from following aspects: (1) disfluenency: a response is not fluent, (2) irrelevance: a response is not semantically relevant to a message, (3) self-contradiction: there is internal semantic conflict within a response. If a response is appropriate but uninformative, it will be rated with "+1". If it is both appropriate and informative, then it will be rated with "+2".

Moreover, we report the appropriate rate (p_{cue_words}) of predicted cue word or coarse words in pipelined models, annotators are invited to label whether predicted words are appropriate to respond given message or not.

3.3 Systems

We conduct empirical comparison of our model with four state-of-the-art models, including (1) **MMI-bidi** [3] which is a seq2seq model using Maximum Mutual Information (MMI) as the objective; (2) **CMHAM** [15] which is Seq2Seq model

[4] https://tieba.baidu.com
[5] https://github.com/baidu/lac

enhanced with constrained multi-head attention mechanism; (3) **MrRNN** [8] which is a pipelined content-introducing model; and (4) **HGFU** [18] which incorporates auxiliary cue word information into seq2seq.

For fair comparison, decoder in all baseline models are set as a two-layer RNN. The vocab size is 50k, hidden size is 512, embedding size is 512, and model are optimized with adam (lr=0.001). Embeddings in encoder and decoder are separated.

HGFU$^+$: We utilize a weighted PMI statistic on the same 0.4 billion Tieba message-response pairs, which only influent the prediction of cue word in the testing procedure. Specifically, the concurrency times of word x_i and word y_j is counted as w_{y_j}, the occurrence times of word x_i is still counted as 1 and the occurrence times of word y_j is counted as w_{y_j}. We use the model trained in HGFU for the testing procedure.

KP-S2S: We implement KP-S2S shown in Figure 2. For training of KP-Annotator, we manually label key-phrases in a message for given response on 100k message-response pairs. Δ is set to 0.1

3.4 Evaluation Results for KP-Annotator

The KP-Annotator can be regarded as a binary classification task. The AUC score of KP-Annotator is 0.864. It indicates that given s message, the weight of key-phrases (i.e. positive) in response will be higher than non-key-phrases (i.e. negative) in 86.4% of cases.

3.5 Evaluation Results for KP-S2S

Table 1 presents human evaluation results of KP-S2S and baseline models. KP-S2S is significantly better (sign test, p-value less than 0,0001) than all the baselines on test set.

s	+2	+1	0	Kappa	Avg-score	p_{cue_words}
MMI-bidi	0.16	0.23	0.61	0.79	0.55	-
CMHAM	0.27	0.15	0.58	0.78	0.68	-
MrRNN	0.19	0.08	0.73	0.56	0.46	0.45
HGFU	0.27	0.10	0.63	0.69	0.64	0.53
HGFU$^+$	0.32	0.09	0.57	0.62	0.73	0.69
KP-S2S	**0.47**	**0.07**	**0.46**	0.65	**1.01**	-

Table 1. Results of human evaluation. p_{cue_words} stands for the appropriate rate of predicted words for given message. The kappa values of models are all higher than 0.5

We see that in terms of both appropriateness and informativeness, KP-S2S significantly outperforms baseline models. Moreover, KP-S2S tends to generate less inappropriate responses than the baselines, its ratio of responses being rated with "0" is significantly lower than the baselines. It is noticed that with the help

of MMI-bidi, S2SAtt still tends to generate inappropriate or generic responses, as shown in Table 1. Moreover, for CMHAM, its multi-head attention mechanism leads to better performance in comparison with MMI-bidi/MrRNN/HGFU, a higher ratio of "+2", and a lower ratio of "0". It indicates CMHAM has a better capability in semantic association modeling. This result is consistent with the conclusion in [15]. But CMHAM still generates more than half of inappropriate responses in test set, and its ratio of responses being rated with "1" (typically safe responses) is even higher than other baselines except MMI-bidi . It indicates that CMHAM cannot effectively deal with irrelevant response words in training data, which interfere with the alignment between message words and relevant response words.

Further, in HGFU$^+$, weighted PMI statistic is utilized to help capturing semantic association between messages and relevant parts in responses, irrelevant parts in responses are largely ignored due to their relatively low weights. In Table 1, we see the appropriate rates of predicted cue word by weighted PMI is increased to 69%, achieving a 16% absolute promotion compared to original PMI. Higher appropriate rates of predicted cue word leads to better generation performance in comparison to original HGFU, the average score increased by 14%. Meantime, we can see that the appropriate rates of predicted cue word or coarse words have a close positive correlation with the average scores of models.

Models	Emb. Average	Emb. Greedy	Emb. Extrema	Distinct-1	Distinct-2
MMI-bidi	0.74	0.58	0.51	0.07	0.22
CMHAM	0.80	0.61	0.52	0.06	0.22
MrRNN	0.79	0.59	0.52	0.04	0.22
HGFU	0.77	0.59	0.50	0.10	0.36
HGFU$^+$	0.79	0.60	0.52	0.10	0.37
KP-S2S	**0.82**	**0.62**	**0.54**	**0.13**	**0.44**

Table 2. Results of automatic evaluation. KP-S2S is significantly better (except p-value of CMHAM vs KP-S2S is 0.0252) than all the baselines.

Table 2 provides the results on automatic evaluation. We see that in terms of embedding based metrics, KP-S2S performs better than the baselines. This results are consistent with human evaluation result in Table 1.

3.6 Case Study

For case study, we compared KP-S2S with baselines using some examples, as shown in Figure 3. In Case 1, the response generated by KP-S2S, "Can't a single dog go out to make friends?", contains a phrase "single dog" which is semantically associated with word "loneliness" in the message. In other Cases, there is also such kind of semantic association between the response generated by KP-S2S and the message, e.g., "catch cold" and "wrapped up in my clothes",

Messages	Generated Responses
去与不去，在于你能不能耐得住寂寞了 (Go or not, depends on whether you can stand loneliness.)	[MMI-bidi] 好的，谢谢，我去看看。(OK, thank you. I'll check it out.) [CMHAM] 我是寂寞了。(I'm lonely.) [MrRNN] 好的，我也寂寞了。(Okay, I'm lonely.) [HGFU] *我耐不住寂寞的寂寞 (I can't bear the loneliness of loneliness.) [**HGFU$^+$**] 我不寂寞啊。(I'm not lonely.) [**KP-S2S**] 单身狗出去交个朋友不行吗？(Can't a single dog go out to make friends?)
多注意，我都是裹的严严实实 (Pay more attention to it. I am wrapped up in my clothes.)	[MMI-bidi] *好的，谢谢，我去试试。(OK, thank you. I'll try.) [CMHAM] *我的是的，你可以试试。(Mine is like this, you can try.) [MrRNN] 好的。(All right.) [HGFU] 嗯，我也裹着严严实实的羽绒服 (Well, I'm also wrapped in a tight down jacket.) [**HGFU$^+$**] 嗯，我裹着被子了。(Well, I'm wrapped in quilts.) [**KP-S2S**] 裹得严严实实，不要感冒 (Wrap up tightly and do not catch cold.)
银行开户慢死了，等了第六天了，还在等 (It is too slowly to open a bank account, I have waited for six days, and still waiting.)	[MMI-bidi] *是啊，我也是这么想的。(Yes, I think so too.) [CMHAM] 好的，我也在等了。(Okay, I'm waiting, too.) [MrRNN] *是什么银行啊，我在哪去了？还是要用哪个银行。(What bank is it? Where will I go? Which bank to use?) [HGFU] 我也是六天。(I am also six days.) [**HGFU$^+$**] 顺丰么？(Is it expressed by Shun Feng Express?) [**KP-S2S**] 我建议你去工行 (I suggest you to go to the Industrial and Commercial Bank of China.)
这是光驱接口，得用支架 (This is a CD driver interface. It needs a bracket.)	[MMI-bidi] *我的也是光驱。(Mine is also a CD driver.) [CMHAM] 这个是接口么 (Is this interface?) [MrRNN] *谢谢大佬，我还以为是个好光驱啊。(Thank you, big guy. I thought it was a good CD driver.) [HGFU] *这个是什么光驱？我是这的 (What kind of CD drive is this? I'm this one.) [**HGFU$^+$**] 这个是什么光驱接口？我是这的 (What kind of CD drive is this? I'm this one.) [**KP-S2S**] 你用的什么支架啊，我在网上买？(What kind of interface do you use? I will buy it online.)

Fig. 3. Examples of generated responses. Inappropriate responses are labeled with *, and underlines indicate semantic association relationships captured by KP-S2S.

"go to the Industrial and Commercial Bank of China" and "open an bank account", "buy it online" and "needs a interface", "search in Baidu" and "how to match the card" . It indicates that KP-S2S can successfully learn semantic association from the training data, and such association relations can be seen as knowledge to some extend, as shown in Case 1, 2 and 4. It seems that this kind of knowledge implicitly represented in KP-S2S model bring a significant performance improvement in terms of response appropriateness and informativeness. In contrast, baseline models fail to capture such kind of semantic association.

4 Related Works

Lots of work in chitchat focus on learning response generation models from large scale human-to-human dialogue corpus within a seq2seq framework [10, 14, 3, 9, 7, 4, 17, 12, 20, 19, 13].

Tao et al. [15] tried to extract different semantic aspects from a message and the whole response is expected to focus on only a few words in each semantic aspect. Semantic aspects are calculated by projecting hidden states of the encoder with k different parameter matrices. However, the problem that irrelevant parts interfere with the association modeling between messages and relevant parts in responses still exists. Serban et al. [8] tried to model high-level semantic association between coarse words(e.g. entities) from messages and responses respectively. Coarse words are different from key-phrase as they are restricted to be specific categories of words, including nouns, manually selected verbs and technical entities in utterances. Moreover, as coarse words can come from both

relevant and irrelevant part in the response, the problem that irrelevant parts interfere with the association modeling between message and relevant parts still exists. Mou et al. [7] and Yao et al. [18] proposed to introduce cue words to the model, they use PMI scores to model the semantic association between message words and response words. However, irrelevant parts still interfere with the association modeling between message and relevant parts. Only 53% of the predicted cue words based on PMI are appropriate to respond given message.

Many work [5, 11] attempted to calculate quality scores for message-response pairs to promote contribution of high-quality instances to the training. These scores are calculated and utilized at utterance level. Lei et al. [2] using reinforcement learning to put more weights on informative words, however, it is for task-oriented dialogue systems rather than open-domain dialogue.

5 Conclusion

In this paper, we showed that sometimes only a part of response has semantic association relationships with the message. We built a key-phrase annotation model to reveal semantic association in message-response pairs. These annotation results are applicable for other data-driven conversation models. Further, We proposed a key-phrase aware end2end neural response generation model (KP-S2S) that can effectively capture semantic association between messages and relevant parts in responses. Experimental results showed that KP-S2S can generate more appropriate and informative responses than state-of-the-art baseline models. In addition, simple use of key-phrase information in training data can bring performance improvement for a cue-word based response generation model in previous works [18].

References

1. Jiao, Z., Sun, S., Sun, K.: Chinese lexical analysis with deep bi-gru-crf network. arXiv preprint arXiv:1807.01882 (2018), https://arxiv.org/abs/1807.01882
2. Lei, W., Jin, X., Kan, M.Y., Ren, Z., He, X., Yin, D.: Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). vol. 1, pp. 1437–1447 (2018)
3. Li, J., Galley, M., Brockett, C., Gao, J., Dolan, B.: 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. pp. 110–119 (2016)
4. Li, J., Galley, M., Brockett, C., Spithourakis, G., Gao, J., Dolan, B.: A persona-based neural conversation model. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). vol. 1, pp. 994–1003 (2016)
5. Lison, P., Bibauw, S.: Not all dialogues are created equal: Instance weighting for neural conversational models. In: Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue. pp. 384–394 (2017)

6. Liu, C.W., Lowe, R., Serban, I., Noseworthy, M., Charlin, L., Pineau, J.: 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. pp. 2122–2132 (2016)
7. Mou, L., Song, Y., Yan, R., Li, G., Zhang, L., Jin, Z.: Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. pp. 3349–3358 (2016)
8. Serban, I.V., Klinger, T., Tesauro, G., Talamadupula, K., Zhou, B., Bengio, Y., Courville, A.C.: Multiresolution recurrent neural networks: An application to dialogue response generation. In: AAAI. pp. 3288–3294 (2017)
9. Serban, I.V., Sordoni, A., Bengio, Y., Courville, A.C., Pineau, J.: Building end-to-end dialogue systems using generative hierarchical neural network models. In: AAAI. vol. 16, pp. 3776–3784 (2016)
10. Shang, L., Lu, Z., Li, H.: 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). vol. 1, pp. 1577–1586 (2015)
11. Shang, M., Fu, Z., Peng, N., Feng, Y., Zhao, D., Yan, R.: Learning to converse with noisy data: Generation with calibration. In: IJCAI. pp. 4338–4344 (2018)
12. Shao, Y., Gouws, S., Britz, D., Goldie, A., Strope, B., Kurzweil, R.: Generating high-quality and informative conversation responses with sequence-to-sequence models. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2210–2219 (2017)
13. Song, Y., Yan, R., Feng, Y., Zhang, Y., Zhao, D., Zhang, M.: Towards a neural conversation model with diversity net using determinantal point processes. In: AAAI (2018)
14. Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., Nie, J.Y., Gao, J., Dolan, B.: 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. pp. 196–205 (2015)
15. Tao, C., Gao, S., Shang, M., Wu, W., Zhao, D., Yan, R.: Get the point of my utterance! learning towards effective responses with multi-head attention mechanism. In: IJCAI. pp. 4418–4424 (2018)
16. Wu, X., Zhou, J., Sun, Y., Liu, Z., Yu, D., Wu, H., Wang, H.: Generalization of words for chinese dependency parsing. Proc. IWPT'13 pp. 73–81 (2013)
17. Xing, C., Wu, W., Wu, Y., Liu, J., Huang, Y., Zhou, M., Ma, W.Y.: Topic aware neural response generation. In: AAAI. vol. 17, pp. 3351–3357 (2017)
18. Yao, L., Zhang, Y., Feng, Y., Zhao, D., Yan, R.: Towards implicit content-introducing for generative short-text conversation systems. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2190–2199 (2017)
19. Zhang, R., Guo, J., Fan, Y., Lan, Y., Xu, J., Cheng, X.: Learning to control the specificity in neural response generation. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). vol. 1, pp. 1108–1117 (2018)
20. Zhou, G., Luo, P., Cao, R., Lin, F., Chen, B., He, Q.: Mechanism-aware neural machine for dialogue response generation. In: AAAI. pp. 3400–3407 (2017)