Generating Stylistically Consistent Dialog Responses with Transfer Learning

Reina Akama[†1], **Kazuaki Inada**[†2], **Naoya Inoue**[†3], **Sosuke Kobayashi**[‡4], **Kentaro Inui**[†*5]
[†]Graduate School of Information Sciences, Tohoku University
[‡]Preferred Networks, Inc.
[*]RIKEN Center for Advanced Intelligence Project
{[1]reina.a,[2]kazuaki.inada,[3]naoya-i,[5]inui}@ecei.tohoku.ac.jp,
[4]sosk@preferred.jp

Abstract

We propose a novel, data-driven, and stylistically consistent dialog response-generation system. To create a user-friendly system, it is crucial to make generated responses not only appropriate but also stylistically consistent. For leaning both the properties effectively, our proposed framework has two training stages inspired by transfer learning. First, we train the model to generate appropriate responses, and then we ensure that the responses have a specific style. Experimental results demonstrate that the proposed method produces stylistically consistent responses while maintaining the appropriateness of the responses learned in a general domain.

1 Introduction

Dialog response-generation (DRG) systems have been studied extensively over the decades. In the recent years, there is a growing interest in a data-driven DRG system leveraging a large amount of conversational texts in social networking services (SNSs) such as Twitter. One successful approach is to train a neural sequence-to-sequence model (seq2seq) (Sutskever et al., 2014) on conversations from SNS, where the seq2seq learns a mapping from a user utterance to the appropriate response with recurrent neural networks (RNNs) (Shang et al., 2015; Li et al., 2016; Serban et al., 2017).

The prior data-driven approach, however, suffers a major drawback. Since a wide variety of users, ranging from young to old, and from female to male, participates in an SNS, the learned responses are not guaranteed to be stylistically consistent (e.g. in terms of politeness). For example, we show actual responses generated by a vanilla seq2seq-based DRG system trained on Twitter conversations in Table 1. The responses have inconsistent style as if the system had multiple personalities; the responses are friendly (1b), polite (2b), or feminine (3b). To create a user-friendly DRG system, it is crucial to generate responses that are stylistically consistent.

1a)	*User*:	こんにちは (konnichiha . / Hello.)
1b)	*System*:	今日も元気いっぱいでいこうな (kyou mo genki ippai de iko u na . / Yo, man! How ya doing today?)
2a)	*User*:	今日は雪です (kyou ha yuki desu . / It is snowing today.)
2b)	*System*:	雪降ってるんですか (yuki ga hu tte iru n desu ka . / Is it?)
3a)	*User*:	寒いです (samui desu . / It's cold.)
3b)	*System*:	よしよし手をつないであげますわ (yoshi yoshi , te wo tsunai de age masu wa . / Oh, dear. I'm gonna hold your lovely hands.)

Table 1: Example of stylistically inconsistent responses generated by a vanilla seq2seq trained on our Twitter dialog corpus.

This paper proposes a novel, data-driven, and stylistically consistent DRG model. A vanilla seq2seq cannot generate stylistically consistent responses without training on a large amount of stylistically consistent dialog responses, which is prohibitively costly. To address this issue, we apply transfer learning, namely transferring knowledge about response generation in a general domain into the task of stylistically consistent response generation.

In the literature, little attention has been paid to the stylistic consistency of the generated responses. There are some previous attempts on transforming a style of dialog utterances into a specified one (Walker et al., 2012; Miyazaki et al., 2015). Their approaches assume that original utterances are given by a separate independent system and need some manual annotations. Li et al.

Proceedings of the The 8th International Joint Conference on Natural Language Processing, pages 408–412,
Taipei, Taiwan, November 27 – December 1, 2017

(2016) aim for response generation with a consistent "persona" by conditioning a seq2seq on a specific Twitter user ID embedding. However, their work focuses on the consistency in the content of the generated responses and they did not directly evaluate the stylistic consistency of their system.

This is the first study that focuses on building a stylistically consistent end-to-end and data-driven DRG model and empirically evaluates the stylistic consistency of generated responses in single-turn dialogs. Our experiments demonstrate that the proposed method produces stylistically consistent responses while maintaining the appropriateness of responses learned from a general domain.

2 Related Work

The literature includes some prior studies that aim for transforming a style of dialog utterances into a specified one (Walker et al., 2012; Miyazaki et al., 2015). Walker et al. (2012) extract rules representing characters from their annotated movie subtitle corpora. Miyazaki et al. (2015) propose a method of converting utterances using rewriting rules automatically derived from a Twitter corpus. These approaches have a fundamental problem to need some manual annotations, which is a main issue to be solved in this work. We propose an end-to-end and data-driven framework which addresses both response generation and stylistic transformation.

Transfer learning is a machine learning technique effective for many NLP tasks (Pan and Yang, 2010), which effectively trains a machine learning model by transferring knowledge about a general domain into a target domain. By applying transfer learning to a stylistically consistent DRG system, once we build a DRG system without stylistic consistency, it is easy to change its style by adding a small stylistically consistent corpus.

3 Generating Stylistically Consistent Responses with Transfer Learning

Given an utterance style, our goal is to create a DRG system that can keep producing utterances with the specified style. Inspired by the success of the data-driven approach, one can prepare a corpus of conversations for every possible style and feed it to a seq2seq. However, in order to obtain a stylistically consistent DRG system through a vanilla seq2seq, this method requires millions of training instances for each target style, which is prohibitively expensive.

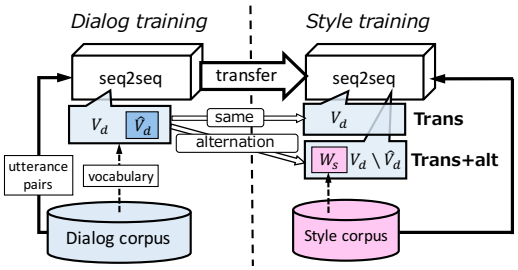

Figure 1: Overview of our framework. Trans shares the same vocabulary in dialog and style training. Trans+alt alters a vocabulary to consider expressions specific in a style corpus.

3.1 General Framework

To address this issue, we propose a novel two-staged training framework for building a stylistically consistent response-generation model, as depicted in Figure 1. The core idea is as follows. A stylistically consistent response generation model has to learn at least three aspects of responses: *content, fluency* and *style*. We hypothesize that learning the content and fluency requires a large amount of training instances, whereas learning the style requires far less training instances. Inspired by transfer learning, our training strategy is divided into two steps (henceforth, dialog training and style training). We thus assume two types of corpora as training data: (i) *dialog corpus*, a large conversational corpus *without* stylistic consistency, and (ii) *style corpus*, a small conversational corpus *with* stylistic consistency.

In our experiments, we used single-turn dialog (i.e. utterance pairs) as training data. However, the proposed framework is applicable to multi-turn dialog as long as we have a response-generation model.

First, in dialog training, we train a seq2seq response generation model on a dialog corpus to learn the content and fluency of responses, namely how to generate responses without taking into account stylistic consistency. Seq2seq (Sutskever et al., 2014) is an RNN-based approach effective for response generation (Shang et al., 2015; Li et al., 2016; Serban et al., 2017). A typical seq2seq has a vocabulary of a limited size (e.g., tens of thousands) to speed up the training process, reduce the computational memory usage and prevent overfitting. Following a typical practice in seq2seq, we use the top N_d most frequent words in a dialog corpus as the seq2seq vocabulary (henceforth, V_d) and convert the other infrequent words into a spe-

cial symbol UNK.

Second, in style training, we use a style corpus to fine-tune the seq2seq that is already trained with a dialog corpus to ensure that the generated responses are stylistically consistent. The model inherits all the model parameters (i.e. word embeddings and weight matrices in the RNNs) and the seq2seq vocabulary (henceforth, V_s) from dialog training.

Note that obtaining a large-scale dialog corpus is relatively easy. For example, one can simply use the whole Twitter conversations as a dialog corpus. As for a style corpus, one can extract utterances made by a specific character from a movie script as in (Walker et al., 2012). One advantage of our framework is that once a large-scale dialog corpus is obtained and dialog training is done, it is easy to change its style by adding a small style corpus.

3.2 Vocabulary in Transferred Model

When applying transfer learning to a seq2seq, creating a vocabulary for style training is not a trivial job. In our experiments, we explore two strategies. In Trans, our first strategy, we simply use the same vocabulary for dialog training and style training (i.e. $V_s = V_d$). However, this may not be a good strategy, because the top N_d most frequent words in a dialog corpus do not necessarily include words that are potentially useful for generating stylistically consistent responses.

To remedy this, in Trans+alt, our second strategy, we alter the seq2seq vocabulary V_s that is used for dialog training before style training (henceforth, *vocabulary alternation*). Let W_s be a set of the top N_s most frequent words which are only included in a style corpus, and $\hat{V}_d$ be a set of the top N_s most *infrequent* words in V_d. Instead of simply setting $V_s = V_d$, we set $V_s = V_d \setminus \hat{V}_d \cup W_s$ (i.e. the top $N_d - N_s$ most frequent words in a dialog corpus plus the N_s most frequent words in a style corpus). Intuitively, we basically employ V_d as V_s, but replace infrequent words from the dialog corpus with frequent words in the style corpus.

4 Experiments

4.1 Setups

Datasets A summary of our corpora is shown in Table 2. The dialog corpus contains approximately 3.7 M Japanese utterance pairs extracted from tweet-reply chains on Twitter.[1] The style corpus contains pairs of utterances extracted from subtitles of a Japanese TV show. We prepare two instances of style corpus: (i) *Tetsuko corpus*, where all the responses are made by Tetusko Kuroyanagi, an elderly, polite, and female TV personality, and (ii) *Oja corpus*, where all the responses are made by Ojarumaru, a five-year-old kids' cartoon character who uses classical Japanese expressions (e.g., see Appendix A, Table 5). We use 95 % of the corpora for training and 5 % for validation.

	Utterance pairs	Vocabulary	Overlap (%)
Dialog corpus	3,688,162	591,880	-
Tetsuko corpus	12,564	12,102	7,230 (59.7)
Oja corpus	1,476	2,137	1,532 (71.7)

Table 2: Dialog corpus and style corpora used in our experiments. 'Overlap' represents the number of overlap words between a dialog corpus and style corpus.

Model	Transfer learning	Dialog corpus	Style corpus	Vocabulary alternation
Base		✓		
Mix		✓	✓	✓
Trans	✓	✓	✓	
Trans+alt	✓	✓	✓	✓

Table 3: Four models for our experiments

Baselines We prepare two baseline models without transfer learning: (i) Base, a vanilla seq2seq trained on the dialog corpus only, and (ii) Mix, a vanilla seq2seq trained on the mixture of a dialog corpus and style corpus, where Mix is applied to the vocabulary alternation as well as Trans+alt. We summarize the baseline models and proposed models in Table 3.

Settings All the four models use the following settings. The seq2seq encoder and decoder are two-layer LSTMs with 2048 units using 1024-dimensional word embeddings. The vocabulary size (N_d) is $25,000$. We train the models using Adam (Kingma and Ba, 2015) with mini-batch size 64 and performed early stopping for perplexity using the validation data. For the vocabulary alternation in Mix and Trans+alt, we use $N_s = 1,000$ for Tetsuko corpus and $N_s = 500$ for Oja corpus.[2]

4.2 Evaluation Method

Conventionally, DRG systems are evaluated through reference-based evaluation (e.g. BLEU

[1] Noisy sentences (e.g., URLs) are filtered out.

[2] We decide the values of Ns with reference to Table 2.

Style		Base	Mix	Trans	Trans+alt
Tetsuko	AR	49.0	50.0	53.0	$60.0^{\dagger B,\dagger M}$
	SC	19.5	22.5	$30.5^{\ddagger B,\dagger M}$	$34.0^{\ddagger B,\ddagger M}$
Oja	AR	50.0	46.5	49.0	$57.0^{\ddagger B,\ddagger M,\dagger T}$
	SC	2.5	26.5	$72.0^{\ddagger B,\ddagger M}$	$82.5^{\ddagger B,\ddagger M,\dagger T}$

Table 4: Results of human evaluation. AR and SC denotes the appropriateness of response and stylistic consistency, respectively. Superscripts $^{\dagger B,\dagger M,\dagger T}$ (and $^{\ddagger B,\ddagger M,\ddagger T}$) indicate the statistical significance against Base, Mix and Trans (sign test, $p < 0.05$ for †, $p < 0.01$ for ‡), respectively.

(Papineni et al., 2002; Sordoni et al., 2015; Li et al., 2016)) or subjective human evaluation (Walker et al., 2012; Vinyals and Le, 2015; Shang et al., 2015). We employ human evaluation, because human evaluation captures more stylistic consistency than reference-based evaluation, and word-overlap similarity metrics such as BLEU correlates weakly with human judgments (Liu et al., 2016).

For query utterances, we randomly extract 50 sentences from Twitter. There is no overlap between these query utterances, and the training and validation data. Each model generates four responses from one query utterance by beam sampling (beam width 3) using four different random seeds. Therefore, the total number of responses generated by one model is 200.

We use Yahoo! Crowd Sourcing[3] for human evaluation. Given (i) a query utterance Q, (ii) the response R generated by a model, and (iii) a style description S that the model are trying to generate with, the workers are independently asked to answer the following two questions:

- Whether R is a grammatically and semantically appropriate response to Q.
- Whether the style of R matches S.

Note that the workers were given only a style description S with several example utterances but not the specific name of individual target character. Each response is judged by five workers, who do not know which model generated each response. The final answer is determined by majority vote.

4.3 Results

Table 4 shows the percentage of responses judged as 'appropriate response' (AR) and 'stylistically consistent' (SC).[4] Base indicates that the dialog corpus's style only matches 19.5% in Tetsuko corpus and 2.5% in Oja corpus. Trans and Trans+alt, the proposed transfer learning frameworks, successfully generate more stylistically consistent responses while maintaining the appropriateness of generated responses learned from a dialog corpus (Base v.s. Trans, Trans+alt). In addition, transfer learning is more effective than simply mixing a dialog corpus and style corpus (Mix v.s. Trans+alt). Recall that the difference between Mix and Trans+alt is whether transfer learning is applied, namely the two models are trained on the same corpora and seq2seq vocabulary. Furthermore, the vocabulary alternation in style training (see Sec. 3.2) helps to make the generated responses more stylistically consistent (Trans v.s. Trans+alt). Overall, the improvement on Oja corpus is more salient than that on Tetsuko corpus. We attribute this to the fact that Tetsuko corpus is closer to the original dialog corpus.

Moreover, Trans+alt improves the appropriateness of the responses. On both style corpora, Trans+alt achieves the best result among the four models. Table 6 in Appendix B and Table 7 in Appendix C show actual responses generated by Trans+alt. By analyzing the generated responses, we find that inappropriate responses such as dull responses (e.g., I don't know) and Internet slangs are relatively fewer, even though we did not make any special treatment. We attribute this to the fact that Trans+alt is additionally trained on less noisy real conversations (i.e., TV subtitles) with a better vocabulary, where new frequent words in a less noisy style corpus are pushed.

The overall results support our assumption that style training requires far less training data than dialog training (see Sec. 3.1). We speculate that styles of utterances are characterized by a smaller variety of lexical choices such as sentence-final auxiliary expressions and personal pronouns. For Japanese, in fact, it is shown that sentence-final auxiliary expressions are an important factor for changing the character of a dialog system (Miyazaki et al., 2015).

[3] https://crowdsourcing.yahoo.co.jp/

[4] The percentage of judgments agreed by the workers, where the number of votes to *yes* or *no* is more than 3, is 72.4 (AR) and 70.1 (SC) on Tetsuko corpus, and 68.0 (AR) and 82.0 (SC) on Oja corpus.

5 Conclusion

We have presented a novel end-to-end framework to build a stylistically consistent dialog response-generation system, leveraging transfer learning. We have demonstrated that we are able to produce stylistically consistent responses by transfer learning while maintaining the appropriateness of responses learned from a general domain. The proposed framework allows us to train a response generation model on a large-scale, and easily-obtainable dialog corpus without stylistic consistency and then on a small-scale stylistically consistent corpus. This is the first work to focus on creating a stylistically consistent end-to-end DRG system and evaluating stylistic consistency in neural dialog response generation studies.

In future work, we plan to improve style training so that it can learn only the style of responses. We will assign a weight indicating the degree of stylistic peculiarity to each word in a style corpus, which controls the aggressiveness of style training. Another future work includes exploring an effective way of creating a style corpus, e.g. automatically collecting polite utterances from a large Twitter corpus with a filter, or generating stylistically consistent responses with a smaller or even no specific style corpus.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 15H01702.

References

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *The International Conference on Learning Representations (ICLR)*.

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

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. 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*, pages 2122–2132. Association for Computational Linguistics.

Chiaki Miyazaki, Toru Hirano, Ryuichiro Higashinaka, Toshiro Makino, and Yoshihiro Matsuo. 2015. Automatic conversion of sentence-end expressions for utterance characterization of dialogue systems. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation (PACLIC)*, pages 307–314.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

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

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Thirty-First AAAI Conference on Artificial Intelligence*, pages 3295–3301.

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. Association for Computational Linguistics.

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*, pages 196–205. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. In *International Conference on Machine Learning(ICML) Deep Learning Workshop 2015*.

Marilyn A Walker, Grace I Lin, and Jennifer Sawyer. 2012. An annotated corpus of film dialogue for learning and characterizing character style. In *LREC*, pages 1373–1378.