

Effective Data Augmentation Approaches to End-to-End Task-Oriented Dialogue

Jun Quan and Deyi Xiong*

School of Computer Science and Technology

Soochow University

Suzhou, China

terryqj0107@gmail.com; dyxiong@suda.edu.cn

Abstract—The training of task-oriented dialogue systems is often confronted with the lack of annotated data. In contrast to previous work which augments training data through expensive crowd-sourcing efforts, we propose four different automatic approaches to data augmentation at both the word and sentence level for end-to-end task-oriented dialogue and conduct an empirical study on their impact. Experimental results on the CamRest676 and KVRET datasets demonstrate that each of the four data augmentation approaches is able to obtain a significant improvement over a strong baseline in terms of Success F_1 score and that the ensemble of the four approaches achieves the state-of-the-art results in the two datasets. In-depth analyses further confirm that our methods adequately increase the diversity of user utterances, which enables the end-to-end model to learn features robustly.

Keywords—effective; data augmentation; end-to-end; task-oriented dialogue; state-of-the-art; robust

I. INTRODUCTION

Task-oriented dialogue systems have evolved from traditional modularized pipeline architectures [1], [2], [3] to recent end-to-end trainable frameworks [4], [5], [6]. However, the major data challenge for both formalisms remains the same: the lack of annotated dialogue dataset in specific tasks or domains. Various slots and values in dialogue utterances need to be manually labeled for use in supervised learning. As the process of manual annotation is time-consuming and expensive, publicly available task-oriented dialogue datasets normally contain only a few thousand dialogues. For data-driven dialogue systems, especially neural dialogue systems which are more data-hungry, insufficient training data will substantially limit their power to learn from data, resulting in poor robustness and performance.

In this paper, we are interested in handling such a data scarce problem via automatic and cheap data augmentation methods. We propose four different data augmentation approaches: synonym substitution and stop-word deletion at the word level, translation and paraphrasing at the sentence level. We only apply these approaches to rephrase user utterances while keeping machine utterances intact on the training data. For user utterances, we leave slots and corresponding values unchanged and reword the remaining parts, keeping the meanings of user utterances as much the same as possible. In doing so, we hope to diversify user utterances so that our dialogue system can learn to deal with language variabilities in a robust way.

*Corresponding author

We use TSCP, an end-to-end dialogue system, recently proposed by [6] to validate the efficacy of our methods. We conduct experiments on two public datasets, CamRest676 and KVRET. The combination of the four data augmentation methods can collectively outperform the basic TSCP model by 4.5 points in terms of F_1 score, the TSCP model with reinforcement learning (RL) by 2.5 points on the CamRest676 dataset. Higher improvements are achieved on the KVRET dataset, 7.8 points and 4.1 points in terms of F_1 over the basic TSCP model and TSCP+RL respectively.

The contributions of the paper are threefold:

- First, we present and empirically investigate four different approaches to data augmentation for end-to-end task-oriented dialogue, which, to the best of our knowledge, is the first attempt in automatic data augmentation for task-oriented dialogue.
- Second, we achieve the state-of-the-art performance on the two datasets with the proposed methods.
- Third, our analyses further display that data augmentation on user utterances is better than augmentation on machine utterances. Details on how the proposed methods improve the performance are also provided.

II. BACKGROUND: END-TO-END TASK-ORIENTED DIALOGUE

Task-oriented dialogue systems that can be trained end-to-end have been studied in recent years as alternatives to traditional pipeline-style dialogue systems. Without loss of generality, we use Sequicity [6] as our baseline system to evaluate our data augmentation methods. It significantly outperforms state-of-the-art pipeline-based methods and obtains a satisfactory entity match rate on out-of-vocabulary (OOV) cases where pipeline-designed competitors totally fail. Sequicity handles both task completion and response generation in a single seq2seq model which can be further optimized with reinforcement learning. It provides a theoretically and aesthetically appealing framework, as it achieves true end-to-end trainability with one single seq2seq model. The key concept introduced in Sequicity is the belief span (bspan), a text span that tracks the dialogue belief states at each turn.

Based on this concept, Sequicity decomposes the task-oriented dialogue problem into the generation of bspans and machine responses in a seq2seq framework. Specifically it decodes in two stages. In the first stage, it generates a bspan to facilitate knowledge base (KB) retrieval. It

then generates a machine utterance in the second stage, conditioned on the knowledge base search result and the bspan from the previous stage. Our work is based on an implementation of the Sequicity as a two-stage copynet (TSCP). In the implementation, CopyNet [7] is used to instantiate Sequicity to allow key words from previous utterances to recur in bspans and generated machine responses.

III. DATA AUGMENTATION APPROACHES

In this section, we elaborate the four data augmentation approaches at both the word and sentence level.

A. Word-Level Data Augmentation

We substitute words with their synonyms and delete stop words so as to produce diversity in user utterances at the word level.

In synonym substitution, we first utilize the NLTK toolkit [8] and WordNet [9], [10] to conduct part-of-speech tagging and synonym retrieval respectively. In order to ensure that the meaning of user utterances does not change semantically, we only allow some specific words to be replaced by their synonyms. Proper nouns (e.g., *Africa*, *America*), qualifiers (e.g., *the*, *a*, *some*, *most*, *every*, *no*), personal pronouns (e.g., *hers*, *herself*, *him*, *himself*), and modal verbs (e.g., *can*, *cannot*, *could*, *couldn't*) should not be replaced as the substitution of them can easily result in inconsistent statements or even semantic changes. For notional verbs (e.g., *want*, *like*, *tell*, *find*), adjectives (e.g., *cheap*, *great*, *delicious*) and nouns (e.g., *food*, *restaurant*, *area*, *south*), we look up their synonyms from WordNet and select the candidate synonyms whose part-of-speech tags are consistent with the corresponding words. For each user utterance, we randomly sample one word that satisfies our substitution rules and randomly select a synonym candidate to replace it. In this way, multiple user utterances can be randomly generated for each original utterance in the training data. These generated utterances will be added to the training data to increase diversity at the word level.

Similarly, we can obtain varieties by deleting stop words in user utterances without changing their meaning. It is common for users to ignore stop words, such as articles, prepositions, adverbs and conjunctions. In order to improve the robustness of the task-oriented dialogue system, and to enable it to pay more attention to the key semantic information in user utterances, we propose to discard these high-frequency stop words in user utterances.

B. Sentence-Level Data Augmentation

For data augmentation at the sentence level, we investigate two different approaches: translation and paraphrasing. These two methods will improve the sentence-level variances, not limited to the presence/absence or variety of some specific words.

We use neural machine translation (NMT) models to translate user utterances into other languages and then use reversed NMT systems to translate the generated translations from other languages back to the original

language. In this paper, we use Google online translation engine as our NMT translation system.

For the sentence-level paraphrasing, we use a seq2seq paraphrase model which contains a bidirectional LSTM encoder and LSTM decoder together with an attention network.¹ The model is trained on a mixed data set consisting of paraphrases from para-nmt-5m, Quora question pairs, SNLI and Semeval [11], [12]. In the decoder part, we can either use a greedy search to generate a single unique paraphrase for each entire user utterance, or generate a plenty of different paraphrases via sampling from a distribution.

C. Implementation Details for the four Data Augmentation Approaches

Synonym substitution: we created four different utterances for each user utterance by randomly replacing words with their synonyms. The created data was combined with the original training data. The size of the augmented data in this way was 5 times as large as that of the original training data.

Stop-word deletion: for this augmentation, we utilized the dictionary of stop words from NLTK toolkit and created only one copy for each user utterance and combined the additional copy with the original data.

Translation: user utterances in original English version data were translated into Chinese, Japanese, French, German via Google Translate, and then translated back to English, thus forming four sets of data expressed in different styles.

Paraphrasing: we generated four sets of dialogue data with the seq2seq-based paraphrase generator.

Assembled Augmentation: we combined all data generated by the four methods above. Together, the size of the assembly augmented data is 14 times as large as that of the original data.

The sizes of mini-batch and vocabulary for each data augmentation approach on the two datasets are shown in Table I, which are chosen according to the performance on the development set.

IV. EXPERIMENTS AND ANALYSES

We conducted extensive experiments and analyses on two datasets to validate the effectiveness of the proposed methods in this section.

A. Datasets and Settings

We used two datasets: CamRest676 [13], [14], [15] and KVRET [5], both of which are manually created by crowd-sourcing workers on the Amazon Mechanical Turk platform by a Wizard-of-Oz method [16]. CamRest676 contains 676 dialogues in the restaurant searching domain while KVRET covers three domains: calendar scheduling, weather information and point of interest (POI) navigation.

For TSCP, the dimensionality d for both hidden states and word embeddings was set to 50. Vocabulary size V was 800 for CamRest676 and 1400 for KVRET. The

¹<https://github.com/vsuthichai/paraphraser>

Table I
THE SIZES OF MINI-BATCH AND VOCABULARY FOR THE FOUR DATA AUGMENTATION APPROACHES.

	CamRest676		KVRET	
	Batch size	Vocab size	Batch size	Vocab size
Synonym Substitution	64	800	32	1800
Stop-Word Deletion	32	800	32	1400
Translation	100	800	32	1800
Paraphrasing	64	800	64	1800
Assembled Augmentation	64	800	256	1800

Table II
EXPERIMENT RESULTS ON CAMREST676 AND KVRET.

	CamRest676	KVRET
	Success F ₁	
Results from [6]		
TSCP	0.834	0.774
TSCP + RL	0.854	0.811
Our implementation		
TSCP	0.832	0.815
TSCP + RL	0.858	0.831
Results obtained by data augmentation		
Translation	0.869	0.842
Paraphrasing	0.869	0.841
Synonym Substitution	0.871	0.833
Stop-Word Deletion	0.856	0.831
Assembled Augmentation	0.879	0.852
Machine Utterance Augmentation (synonym substitution)	0.775	-
User + Machine Utterance Augmentation (translation)	0.822	-

size of mini-batch for both datasets was set 32. The model was trained with the Adam optimizer [17], with a learning rate of 0.003 and a decay parameter of 0.5. We used a learning rate of 0.0001 and decay of 0.8 for the subsequent reinforcement learning process. We used beam search strategy with a beam size of 10 on CamRest676 and greedy search strategy on KVRET. Early stopping was also performed to improve the training efficiency.

We used the Success F_1 score as the automatic metric for dialogue evaluation. The Success F_1 score estimates both the precision and recall of requested slots being successfully answered in dialogues, a metric proposed by [6].

B. Results

Table II shows the experiment results on the two datasets, from which we have three findings. First, the results demonstrate that all the proposed data augmentation methods contribute to the significant improvements in F_1 over the basic TSCP model. Except for the stop-word deletion method, all other methods perform better than the RL-enhanced TSCP. Second, the sentence-level augmentation methods are better than the word-level methods in most cases as the former provide more variances for user utterances. Third, the assembled augmentation, which combines all data generated by the four data augmentation methods, achieve the new state-of-the-art performance on the two datasets, with more than 2 points higher than the RL-enhanced TSCP model in terms of F_1 score.

C. Effect of Augmentation on Machine Utterances

At each turn in a dialogue from the two datasets, a user utterance triggers some special requests and a machine

response utterance provides answers to these requests. In our previous experiments, we performed data augmentation only on user utterances. In order to study the effect of data augmentation on machine utterances, we further carried out two experiments. One is to generate both user and machine utterances with the translation augmentation method. The other is to create copies only for machine utterances with synonym substitution. Both experiments were carried out on the CamRest676 dataset.

Results are displayed at the bottom of Table II. It is clear to observe that machine utterance augmentation seriously deteriorates the performance. The reason for this may be that data augmentation introduces both variance and noise. The variance and noise in user utterances can prevent the system from over-sensitivity [18], thus making the system more robust. However, the variance and noise in machine utterances will distract the system. This resonates with the back translation that uses real target sentences and translated source sentences, widely used for seq2seq-based neural machine translation [19].

D. Analysis

We took a deep look into the data to investigate how the proposed data augmentation methods improve the Success F_1 score that computes both the precision and recall of requested slots being correctly answered.

The precision and recall in F_1 can be formulated as follows:

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (1)$$

Table III
ANALYSIS ON THE PRECISION AND RECALL.

	CamRest676					
	Success F ₁	Precision	Recall	TP	FP	FN
baseline	0.832	0.885	0.786	422	55	115
Assembled Augmentation	0.879	0.883	0.875	467	62	67

Table IV
DIALOGUE EXAMPLES OF THE BASELINE AND DATA AUGMENTATION.

Baseline	Data augmentation
User: Can you tell me if there is a moderately priced restaurant serving spanish food anywhere in town? Machine: la tasca serves spanish food. User: What is the address, phone number, and area? Machine: la tasca's phone number is 01223 464630. User: That is it. Thank you. Machine: I'm sorry , there are no restaurants serving spanish food. User: Thanks again! bye. Machine: Thank you for using the cambridge restaurant system.	User: Can you tell me if there is a moderately priced restaurant serving spanish food anywhere in town? Machine: la tasca serves spanish food in the moderate price range. Would you like their location ? User: What is the address, phone number, and area? Machine: Their address is 14 -16 bridge street and their phone number is 01223 464630. User: That is it. Thank you. Machine: Have a nice day. User: Thanks again! bye. Machine: Have a nice day.
User: I would like a restaurant in the north part of town that serves australasian food Machine: There are no restaurants in the north part of town. Would you like a different type of food ? User: How about asian oriental food? Machine: dojo noodle bar serves asian oriental food. User: Yes, I would like their address and phone number. Machine: dojo noodle bar is located at 40210 millers yard city centre and their phone number is 01223 363471. User: Thank you goodbye. Machine: I'm sorry , but dojo noodle bar is a nice day.	User: I would like a restaurant in the north part of town that serves australasian food. Machine: I'm sorry , there are no restaurants serving australasian food. Would you like me to find a different restaurant? User: How about asian oriental food? Machine: saigon city serves asian oriental food User: Yes, I would like their address and phone number. Machine: Their address is 169 high street chesterton chesterton. Their phone number is 01223 356555. User: Thank you goodbye. Machine: Have a nice day !

$$precision = \frac{TP}{TP + FP} \quad (2)$$

$$recall = \frac{TP}{TP + FN} \quad (3)$$

where TP denotes the number of requested slots that are correctly predicted and do exist in real machine responses, FN the number of slots that exist in real responses but not answered at all, FP the number of slots being predicted but not present in real responses.

We provide the values for the precision, recall, TP, FN and FP in Table III for the assembled augmentation. Obviously, our method can significantly improve the recall by nearly 9 points while keeping the precision basically the same as the baseline. The reason behind the improvement of the recall is that the proposed methods substantially increases TP and decreases FN. This is because the diversity in user utterances created by data augmentation helps the dialogue system recognize more requested slots and further allows the decoder to answer these slots in machine responses. Without data augmentation, some slots are just not detected at all in the baseline (thus a higher FN).

E. Dialogue Samples

Table IV shows some dialogue examples generated by the model with or without data augmentation. The dialogues on the left side of the table is generated by the baseline model, while on the right side is the examples generated by the model with assembled data augmentation.

Obviously, the model after our data augmentation is more robust to understand the user utterances and can produce more appropriate machine responses.

V. RELATED WORK

Data augmentation has achieved great success in various tasks including computer vision [20], speech recognition [21] and text classification [22], but is explored in a very limited way for the natural language understanding (NLU) module of traditional pipeline systems of task-oriented dialogue. [23] propose to augment data for the NLU module by adding noise to one single user utterance without considering its relation with other utterances. [24] introduce a technique to expand the limited in-domain data for a new spoken language understanding task. [25] propose a data-augmentation framework to model relations between utterances of the same semantic frame in the training data. Other researchers present methods for gathering dialogue data through crowd-sourcing, e.g., via talking to myself [26] or MultiWOZ [27]. Different from our methods, these methods either focus solely on the NLU module or rely on expensive human efforts.

VI. CONCLUSION AND FUTURE WORK

In this paper, we have presented four different effective methods of data augmentation for end-to-end task-oriented dialogue systems at both the word and sentence level. Empirical study on two public datasets CamRest676 and KVRET shows that data augmentation can prevent the dialogue system from the omission of key information in

user utterances and significantly improve the F_1 score via effectively solving the problem of data scarcity.

In the future, we intend to apply our data augmentation methods on more datasets and to explore some other efficient ways to increase the diversity of machine responses as well.

ACKNOWLEDGMENT

The present research was supported by the National Natural Science Foundation of China (Grant No. 61622209 and 61861130364). We would like to thank the three anonymous reviewers for their insightful comments.

REFERENCES

- [1] A. I. Rudnicky, E. Thayer, P. Constantinides, C. Tchou, R. Shern, K. Lenzo, W. Xu, and A. Oh, "Creating natural dialogs in the carnegie mellon communicator system," in *Sixth European Conference on Speech Communication and Technology*, 1999.
- [2] V. Zue, S. Seneff, J. R. Glass, J. Polifroni, C. Pao, T. J. Hazen, and L. Hetherington, "Juplter: a telephone-based conversational interface for weather information," *IEEE Transactions on speech and audio processing*, vol. 8, no. 1, pp. 85–96, 2000.
- [3] V. W. Zue and J. R. Glass, "Conversational interfaces: Advances and challenges," *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1166–1180, 2000.
- [4] M. Eric and C. D. Manning, "A copy-augmented sequence-to-sequence architecture gives good performance on task-oriented dialogue," *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 468–473, 2017.
- [5] M. Eric, L. Krishnan, F. Charette, and C. D. Manning, "Key-value retrieval networks for task-oriented dialogue," in *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, 2017, pp. 37–49.
- [6] W. Lei, X. Jin, M.-Y. Kan, Z. Ren, X. He, and D. Yin, "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, 2018, pp. 1437–1447.
- [7] J. Gu, Z. Lu, H. Li, and V. O. Li, "Incorporating copying mechanism in sequence-to-sequence learning," *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 1631–1640, 2016.
- [8] S. Bird and E. Loper, "Nltk: the natural language toolkit," in *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*. Association for Computational Linguistics, 2004, p. 31.
- [9] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller, "Introduction to wordnet: An on-line lexical database," *International journal of lexicography*, vol. 3, no. 4, pp. 235–244, 1990.
- [10] L. F. D'Haro, S. Kim, and R. E. Banchs, "A robust spoken q&a system with scarce in-domain resources," in *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE, 2015, pp. 47–53.
- [11] J. M. John Wieting and K. Gimpel, "Learning paraphrastic sentence embeddings from back-translated bitext," in *Proceedings of Empirical Methods in Natural Language Processing*, 2017.
- [12] J. Wieting and K. Gimpel, "Paranmt-50m: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations," *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, pp. 451–462, 2018.
- [13] T.-H. Wen, Y. Miao, P. Blunsom, and S. Young, "Latent intention dialogue models," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 3732–3741.
- [14] T.-H. Wen, M. Gasic, N. Mrksic, L. M. Rojas-Barahona, P.-H. Su, S. Ultes, D. Vandyke, and S. Young, "Conditional generation and snapshot learning in neural dialogue systems," *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2153–2162, 2016.
- [15] T.-H. Wen, D. Vandyke, N. Mrksic, M. Gasic, L. M. Rojas-Barahona, P.-H. Su, S. Ultes, and S. Young, "A network-based end-to-end trainable task-oriented dialogue system," *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 438–449, 2016.
- [16] J. F. Kelley, "An iterative design methodology for user-friendly natural language office information applications," *ACM Transactions on Information Systems (TOIS)*, vol. 2, no. 1, pp. 26–41, 1984.
- [17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Proceedings of the 3rd International Conference for Learning Representations, San Diego*, 2015.
- [18] T. Niu and M. Bansal, "Adversarial over-sensitivity and over-stability strategies for dialogue models," *arXiv preprint arXiv:1809.02079*, 2018.
- [19] R. Sennrich, B. Haddow, and A. Birch, "Improving neural machine translation models with monolingual data," *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 86–96, 2016.
- [20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
- [21] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates *et al.*, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.
- [22] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Advances in neural information processing systems*, 2015, pp. 649–657.
- [23] G. Kurata, B. Xiang, and B. Zhou, "Labeled data generation with encoder-decoder lstm for semantic slot filling," in *INTERSPEECH*, 2016, pp. 725–729.
- [24] S. Jalalvand, A. Ljolje, and S. Bangalore, "Automatic data expansion for customer-care spoken language understanding," *arXiv preprint arXiv:1810.00670*, 2018.

- [25] Y. Hou, Y. Liu, W. Che, and T. Liu, “Sequence-to-sequence data augmentation for dialogue language understanding,” in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 1234–1245.
- [26] J. Fainberg, B. Krause, M. Dobre, M. Damonte, E. Kahembwe, D. Duma, B. Webber, and F. Fancellu, “Talking to myself: self-dialogues as data for conversational agents,” *arXiv preprint arXiv:1809.06641*, 2018.
- [27] P. Budzianowski, T.-H. Wen, B.-H. Tseng, I. Casanueva, S. Ultes, O. Ramadan, and M. Gašić, “Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling,” *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 5016–5026, 2018.