Data-oriented Approaches for Improving Neural Dialogue Generation

著者	Akama Reina
学位授与機関	Tohoku University
学位授与番号	11301甲第19932号
URL	http://hdl.handle.net/10097/00134536

Data-oriented Approaches for Improving Neural Dialogue Generation

ニューラル対話応答生成の性能向上のためのデータ駆動アプローチ

Reina Akama

Graduate School of Information Sciences
Tohoku University

This dissertation is submitted for the degree of
Doctor of Information Science

January 2021

Acknowledgements

本研究の遂行にあたり、多くの方々にご協力をいただきました。ここに、心より感謝の意を表します。

主指導教官である乾健太郎教授には、学部4年次での研究室配属から約5年間の研究活動において、終始、熱心なご指導をいただきました。心より感謝申し上げます。
ご多忙のなか、本論文の審査委員をお引き受けくださいました伊藤彰則教授ならびに北村喜文教授に、深く感謝申し上げます。
鈴木潤教授には、とくに博士課程在籍時の研究活動において、多大なるご指導をいただきました。日々の研究活動における的確なご指摘とたくさんの励ましに支えられ、意欲的に研究に取り組むことができました。心より感謝いたします。
横井祥助教には、研究のみならず、研究遂行のための基盤となる諸技能の習得に関しても非常に多くのご助言をいただきました。議論の組み立て方や文章の書き方を一から懇切丁寧にご指導いただきましたこと、心より感謝しております。

本論文を構成する研究の遂行にあたり、豊田中央研究所の徳久良子氏、産業技術総合研究所の渡邉研斗氏、株式会社 Preferred Networksの小林颯介氏にお力添えをいただきました。深く感謝いたします。学会等でいただいた皆様からのご意見や激励は、研究遂行の大きな糧となりました。すべての皆様に感謝いたします。なお、本研究の一部は科学研究費補助金（特別研究員奨励費）の支援を受けたものです。記して感謝いたします。

乾研究室の皆様には、日々の議論を通じてたくさんのご指導ご助言をいただくとともに、ときには人生の先輩や良き友人として様々な相談に乗っていただきました。皆様の暖かい心遣いに支えられ、たのしい研究生活・学生生活を送ることができました。心より感謝いたします。本当にありがとうございました。

最後に、これまでの長きに渡る学生生活を暖かく見守り、研究活動に励む日々の心身の健康を支え続けてくれた両親と妹に、感謝いたします。

博士課程修了後も、しばらくは自然言語処理の研究者として生きていくことになりそうです。今後とも精進して参りますので、これまでの研究活動でお世話になった皆様、これからもどうぞよろしくお願いいたします。

Abstract

The realization of a computer that can talk with others like a human is one of the ideals of artificial intelligence applications that have been considered for a long time.

To date, text generation techniques in natural language processing, including dialogue response generation, have been developing rapidly with the development of deep neural network technology. In general, large-scale and high-quality training data is essential for these deep neural networks-based models to perform optimally.

Toward the improvement of the neural response generation technology through the improvement of their training data, this paper addresses the following studies.

First, we establish the methodologies to acquire large-scale and high-quality dialogue data automatically. Specifically, we discuss two strategies in this thesis: data filtering and data augmentation. Then, through the experiments on the response generation task, we empirically confirm the effectiveness of the proposed methodologies, and demonstrate that high-quality and large-scale training data impact the performance improvement of the neural dialogue response generation model.

Furthermore, we present a methodology for manually constructing new training data for neural response generation models. The proposed method allows collecting high-quality pseudo-dialogue data from crowdworkers in situations where human resources are insufficient.

In addition to these studies, we establish the methodology for modeling the stylistic features of natural language, which is essential for making smooth conversational communication. Specifically, we propose a novel architecture to acquire style-sensitive word vectors independently from semantic or syntactic features of words.

Table of contents

List of figures **vii**

List of tables **ix**

1 Introduction **1**

1.1 Research Issues 2

1.2 Contributions 2

1.3 Thesis Overview 3

2 Dialogue Response Generation with Deep Neural Network **5**

2.1 Neural Dialogue Response Generation Systems 5

2.2 Data-oriented Approaches for Neural Text Generation 6

3 Filtering Noisy Dialogue Corpora by Connectivity and Content Relatedness **8**

3.1 Task Definition: Noisy Data Filtering 10

3.2 Background 10

3.3 Requirements to Utterance Pairs 11

3.3.1 Criteria for Manual Evaluation 11

3.3.2 Observation 12

3.4 Proposed Method 12

3.4.1 Connectivity 13

3.4.2 Content Relatedness 14

3.4.3 Summary 14

3.5 Experiments: Data Scoring 14

3.5.1 Experimental Setup 14

3.5.2 Results and Analysis 17

3.5.3 Discussion on Low Recall Property 18

3.6 Experiments: Response Generation 20

3.6.1 Experimental Setup 20

3.6.2 Results and Analysis 21
3.7 Multilingual Availability 23
3.7.1 Data Scoring 23
3.7.2 Response Generation 24
3.8 Relationship with Evaluation Metric 25
3.9 Conclusion 26

4 Dialogue Data Augmentation by Pairing Single Utterances 27
4.1 Background 28
4.2 Task Setting of Pair Construction 29
4.2.1 Input and Output 29
4.2.2 Difficulties in Dialogue Data 29
4.3 Methodology: Pairing Single Utterances 30
4.3.1 Key Idea 30
4.3.2 Step 1: Template Generation 30
4.3.3 Step 2: Candidate Pair Creation 30
4.3.4 Step 3: Candidate Pair Selection 31
4.4 Experiment 31
4.4.1 Datasets 31
4.4.2 Experimental Setup 32
4.5 Results 33
4.5.1 Paired Data Construction 33
4.5.2 Extrinsic Evaluation: Response Generation 34
4.6 Conclusion 35

5 Corpus construction from crowdworkers imitating target attributes 37
5.1 Societal implementation of Dialogue Systems 38
5.1.1 Issue of Social Isolation 38
5.1.2 Dialogue System as Conversational Partner 39
5.1.3 Human-to-human Communication Supporting System 39
5.2 Methodology: Corpus Construction under Limited Human Resource 40
5.3 Investigations on Data Construction with Imitators 42
5.3.1 Effectiveness of Instructions to Control the Subject of Utterance . . 42
5.3.2 Elderly versus Imitator: Comparison of Collected Utterances under the Instructions 43
5.3.3 Summery 46
5.4 Construction of Japanese QOL-labeled Corpus 46

5.5 Experiments on Reply Candidates Generation 48
5.5.1 Generation-based Response Generation Model 49
5.5.2 Retrieval-based Response Generation Model 49
5.5.3 Experimental Setups 50
5.5.4 Result and Analysis 52
5.5.5 Summary 53
5.6 Conclusion 55

6 Segregation of Word Vector to Style and Semantic Components 56
6.1 Style-sensitive Word Vector 57
6.1.1 Notation 57
6.1.2 Baseline Model (CBOW-NEAR-CTX) 58
6.1.3 Learning Style with Utterance-size Context Window (CBOW-ALL-CTX) 58
6.1.4 Learning the Style and Syntactic/Semantic Separately 59
6.2 Experiments 60
6.2.1 Settings 60
6.2.2 Stylistic Similarity Evaluation 61
6.2.3 Syntactic and Semantic Evaluation 62
6.2.4 Analysis of Trained Word Vectors 63
6.3 Conclusion 65

7 Conclusion 66

References 68

List of Publications 77

List of figures

3.1 *Is the sequence of the two utterances acceptable as a dialogue?* Response acceptability scores are given by humans on the English OpenSubtitles corpus. 9

3.2 Score distributions of our $S_{\mathbf{C}}$, $S_{\mathbf{R}}$, $S_{\mathbf{C+R}}$ across our training data (English). . . 16

3.3 Distributions between human scores and automatically computed scores by each method (English). 17

3.4 Score distributions of our $S_{\mathbf{C}}$, $S_{\mathbf{R}}$, $S_{\mathbf{C+R}}$ across training data (Japanese). . . . 23

3.5 Distributions between human scores and automatically computed scores by each method (Japanese). 24

4.1 Overview of our method. Meaningful pairs $\widetilde{\mathscr{D}}$ are newly created from a small set of utterance–response pairs $\mathscr{D}$ and a set of single utterances $\mathscr{H}$. 28

4.2 Relationship between λ and ROC-AUC score. 33

4.3 Distribution of $\mathrm{Assoc}_{\mathrm{s}}(x,y)$ on $\widetilde{\mathscr{D}}_{\mathrm{cand}}$. 33

5.1 (A) Normal conversation and (B) Conversation including utterance presenting QOL information. 39

5.2 Task instructions for collecting utterances presenting elderly's QOL information. 41

5.3 Attention-check for collecting utterances presenting elderly's QOL information. 41

5.4 Proportion of subject tags of collected utterances. 43

5.5 Task instruction of paired comparison experiment. It was randomly determined which of Reply A and Reply B corresponded to the actual elderly's utterance. In this example, reply A is from Group A (elderly), and reply B is from Group B (imitators). 45

5.6 Result of identification by paired comparison between actual elderly replies and imitator's replies. 46

5.7 Comparison of the contents of replies collected from actual elderly and imitators. 46

5.8 Conditional response generation using QOL-labels. The QOL information (colored blue) is input to the decoder at each step. 49

6.1 Word vector capturing stylistic and syntactic/semantic similarity. 57

6.2 The architecture of CBOW-SEP-CTX. 60

List of tables

3.1 Samples of pairs judged as unacceptable/acceptable in preliminary experiments. Human denotes the average score of five human evaluators on a scale of 1-5. Phrases considered to contribute to connectivity are highlighted . Estimated [topic] of utterance is written in the end of each utterance. 12

3.2 The statistics of our English dataset. 15

3.3 Correlation coefficient between human scores and automatically computed scores (English). 17

3.4 Samples of utterance pairs scored with our method and human judgements (English). The scores of $S_{\mathbf{C}}$ and $S_{\mathbf{R}}$ were normalized by α, β. 18

3.5 Samples of utterance pairs that cause low recall scored with our method and human judgements (English). The scores of $S_{\mathbf{C}}$ and $S_{\mathbf{R}}$ were normalized by α, β. 19

3.6 Comparison of the top and the worst utterance pairs in the training data scored by our method (English). 20

3.7 Evaluation results for generated responses (English; filtered out 50%). **Bold** denotes the best results. The ✗/✔ shows the percentages of the low/high scored responses (i.e., human scores in $[1,3)$ or in $[3,5)$). 21

3.8 Samples of input utterances **(bold)** and responses generated by models trained on filtered data by proposed method $S_{\mathbf{C+R}}$ (Ours). Other responses are generated by models trained on (A) non-filtered data and filtered data by baselines; (B) Csáky et al. (2019)'s SRC method, (C) their TRG method, and (D) Junczys-Dowmunt (2018)'s method, respectively. 22

3.9 Correlation coefficient between human scores and automatically computed scores (Japanese). 24

3.10 Evaluation results for generated responses (Japanese; filtered out 10%). **Bold** denotes the best results. The ✗/✔ shows the percentages of the low/high scored responses (i.e., human scores in $[1,3)$ or in $[3,5)$). 25

4.1 Samples of pairs accepted as $\widetilde{\mathscr{D}}$ (lines 1–8) and rejected (lines 9-10). **Bold** indicates dialogue templates. 34
4.2 Evaluation results of generated responses. 34
4.3 Samples of input utterances and generated responses. 35

5.1 emails from family members used on our preliminary experiments (in Japanese). Underlined indicates that sentences are targeted to reply. 42
5.2 Definition of our subject tags. 43
5.3 Age and gender of Group A (ordinary elderly aged over 65). 44
5.4 Sample of replies created by Group A and Group B to email #1. 44
5.5 Sample of replies created by Group A and Group B to email #2. 44
5.6 Definition of QOL labels (in Japanese). 47
5.7 Sample of our Japanese QOL-labeled corpus. 48
5.8 Number of instances in the database with each QOL label. 51
5.9 Result of human evaluations for QOL communicability and response plausibility. 53
5.10 Samples of response generated by `S2S,W2V,ELMo` on the evaluation set $\mathscr{S}_{10-l12}$. 54
5.11 Samples of response generated by `S2S,W2V,ELMo` on the evaluation set $\mathscr{S}_{100}$. 55

6.1 Results of the quantitative evaluations. 62
6.2 The top similar words for the style-sensitive and syntactic/semantic vectors learned with proposed model, CBOW-SEP-CTX (Japanese). Japanese words are translated into English by the authors. Legend: (translation; impression). **Classical* means wording related to e.g., samurai, ninja. 64
6.3 The top similar words for the style-sensitive and syntactic/semantic vectors learned with proposed model, CBOW-SEP-CTX (English). 65

Chapter 1

Introduction

Dialogue systems have been studied as one of the most important applications of natural language processing (NLP). Dialogue systems research has begun to be studied in earnest around the 1960s. In the early days of the system, the branching conditions and patterns of the conversation and the rules of response content were designed to be manual (Galley et al., 2001; Weizenbaum, 1966). In the 2000s, machine learning-based dialogue systems were actively developed. Instead of following carefully designed rules, the system outputs a response by learning the pattern from an amount of dialogue data and reproducing them (Hasegawa et al., 2013; Ritter et al., 2011). Machine learning techniques have made the systems possible to have various conversations that go beyond rules, and since around this time, dialogue studies on "chit-chat" conversations that do not limit situations or goals have also begun. In 2015, a sophisticated dialogue system was proposed, which is based on a deep neural network (Vinyals and Le, 2015a). Since then, developing deep neural network-based dialogue systems and improving neural response generation techniques have become one of the mainstream of dialogue system research.

Deep neural network-based text generations have brought significant progress in dialogue systems and various text generation fields in NLP, such as machine translation and automated summarization. The sentence generations levering deep neural networks are still one of the hot topics in NLP, and researchers have actively studied novel methodologies to achieve higher performance on various tasks. These methodologies can be divided into two major categories depending on what they are trying to improve: model-oriented approaches and data-oriented approaches. The model-oriented approaches try to improve the performance of text generation models by enhancing their architecture or objective functions. Typical researches include, for example, sequence-to-sequence model (Vinyals and Le, 2015b) and transformer model (Vaswani et al., 2017). These sophisticated models have dramatically improved the sentence generation techniques of NLP. On the other hand, the data-oriented approaches try to improve the performance of the generation models by preparing the best data for training the neural generation models. Previous studies have revealed that the more large-scale and high-quality training data, the higher the performance of the generation model. In the field of neural machine translation, the researchers have actively discussed improving the performance of models through training data improvement, and proposed several practical methodologies, e.g., Back-translation (Sennrich et al., 2016b).

In the field of dialogue response generation, there is a study to improve the performance of response generation by a model-oriented approach. For example, Sordoni et al. (2015) and Li et al. (2016a) are one of the typical examples. On the other hand, data approaches are hardly discussed due to difficulties such as ambiguity of conditions for being dialogue. With these background, in this thesis, we focus on data-oriented approach and discuss the some methodologies to improve neural dialogue generation.

1.1 Research Issues

In this thesis, we address the following research issues:

- **What is the clues to enable augmentation or improvement of dialogue data?** Methodologies for improving the performance of response generation models through the scale and quality of training data have been hampered by difficulties due to the characteristics of dialogue, such as the ambiguity of the criteria for dialogue formulation. We explore the requirements that should be satisfied by a desirable dialogue and provide some criteria for automatically calculating the quality of a dialogue.

- **Methodologies for acquiring desirable resources for training neural response generation models.** We believe that large-scale and high-quality training data make the performance of the response generation model improve. We discuss several methodologies for acquiring such data, including data filtering, data augmentation, and manual data construction.

- **Do the large-scale and high-quality training data improve the performance of neural response generation models?** We empirically reveal the impact of large-scale and high-quality training data on the performance of response generation of neural response generation models.

1.2 Contributions

This thesis makes the following contributions:

- **Establishing the methodologies for dialogue data improvement:** We propose the data filtering methodology to make a large scale-data high-quality by detecting and removing the low-quality utterance pairs. Moreover, we propose the data augmentation methodology to create synthetic utterance pairs from high-quality but small dialogue data.
- **Investigating the impact of training data improvement:** Through the response generation experiments, we demonstrate that large-scale and high-quality training data can improve the performance of neural response generation models.
- **Presenting effective corpus construction methodology:** We propose a practical methodology for manually creating new training data for neural response generation models. The proposed method allows us to collect dialogue data at a scale and quality that can be used for training neural dialogue models even when human resources are limited.
- **Modeling the style of utterances:** Understanding the stylistic features of utterances is one of the essential elements for making smooth conversational communication. To model the stylistic feature of natural languages, we propose an unsupervised methodology to acquire style-sensitive word vectors independently from semantic or syntactic features of words. We demonstrate that our word vectors capture the stylistic similarity between two words successfully.

1.3 Thesis Overview

The rest of this thesis is structured as follows:

- **Chapter 2: Dialogue Response Generation with Deep Neural Network Models.** In this chapter, we introduce deep neural network-based response generation models and briefly summarize the mainstreams of recent research. In addition, we explain the background that large-scale and high-quality learning data are indispensable for the generation model using the deep neural network technology.
- **Chapter 3: Filtering Noisy Dialogue Corpora by Connectivity and Content Relatedness.** In this chapter, we discuss the filtering strategy to acquire large-scale and high-quality training data. We propose a scoring function to detect low-quality utterance-response pairs in training data. We demonstrate that the performance of neural response generation models can be improved by ablating unacceptable utterance pairs.

- **Chapter 4: Dialogue Data Augmentation by Pairing Single Utterances.** As another approach to obtaining high-quality and large-scale training data, in this chapter, we discuss the augmentation strategy. We propose the methodology to augment utterance pairs from existing high-quality but small data.

- **Chapter 5: Corpus construction from crowdworkers imitating target attributions.** In this chapter, we propose a methodology for efficiently creating learning data manually. The proposed method allows collecting data from pseudo target other than the original target by using crowdsourcing under effective instructions.

- **Chapter 6: Segregation of word vector to semantic and style components.** The stylistic feature of utterances is one of the important elements of dialogue. In this chapter, we discuss treating the stylistic feature of utterances as word vectors, independently from the semantic or syntactic features. We introduce a novel task that measures stylistic similarity with new benchmark data and propose an unsupervised methodology to acquire style-sensitive word vectors.

- **Chapter 7: Conclusions.** We summarize our discussion, and present our future direction.

Chapter 2

Dialogue Response Generation with Deep Neural Network

In this chapter, we introduce deep neural network-based response generation models and briefly summarize the mainstreams of recent research. In addition, we explain the background that large-scale and high-quality learning data are indispensable for the generation model using the deep neural network technology.

2.1 Neural Dialogue Response Generation Systems

Research on neural dialogue response generation began with Vinyals and Le (2015a). Vinyals and Le (2015a) naively applied the sequence-to-sequence model Sutskever et al. (2014), which is a novel LSTM-based encoder-decoder model proposed for neural machine translation, to the dialogue response generation settings. Surprisingly, the model produced fluent responses that were good enough to be used as a dialogue system, even though they just prepared a large amount of dialogue data and trained the model on the past utterances as the input and the response as the output. Shang et al. (2015) applied a sequence-to-sequence model with an attention mechanism to dialogues and improved to generate better response over Retrieval-based or statistical machine translation-based response generation. Since then, the inter-sequence response generation model has become the mainstream of neural dialogue response generation, and many subsequent studies have proposed enhancements to this model. Besides, there are also studies using Convolutional Neural Network (CNN) (Mangrulkar et al., 2018), Genera tive Adversarial Network (GAN) (Li et al., 2017a), and Conditional Variational Auto Encoder (CVAE) (Serban et al., 2017; Zhao et al., 2017) as response generation models.

Recently, there has been an increasing number of studies using Transformer (Vaswani et al., 2017) as a response generation model (Csáky et al., 2019).

Dialogue response generation models using deep neural networks can generate quite fluent and acceptable responses to even unknown inputs. However, on the other hand, there are still issues to be addressed in order to achieve more human-like and natural dialogue. For example, one issue is the "diversity" of generated responses. Li et al. (2016a) pointed out that neural response generation models often generate generic responses such as "I don't know" or "OK," i.e., dull responses, to any input. To address this issue, they proposed using Maximum Mutual Information (MMI) between the utterance and the response as the objective function in neural models. Baheti et al. (2018) proposed a response generation method that encourages to generate more content-rich responses while suppressing the generation of dull responses by imposing distributed constraints on the decoder during response generation. In addition, several studies have addressed the "consistency" of generated responses. Sordoni et al. (2015) focused on topic consistency, and proposed a conditional generation method that encodes past information as continuous representation to hidden states of a decoder to generate context-sensitive responses. Akama et al. (2017) focused on style consistency, and presented the training framework to generate stylistically consistent dialogue response, leveraging transfer learning technique. Li et al. (2016b) focused on persona consistency, and proposed a neural model for generating consistent persona-based responses for each speaker by encoding the individual speaker's information as distributed embeddings.

2.2 Data-oriented Approaches for Neural Text Generation

In addition to the dialogue response generation, DNN text generation techniques have been studying in many research areas in NLP, such as machine learning, automatic summarization, and grammatical error correction. In the research area of neural machine translation, which has been leading the other text generation areas for many years, many researchers recognize that **large-scale and high-quality training data** improve the performance of the neural translation models (Edunov et al., 2018; Koehn and Knowles, 2017). Therefore, to date, they have actively discussed methodologies to obtain large-scale and high-quality paralleled translation data. For example, to obtain training data on a **large scale**, Fadaee et al. (2017) proposed a pseudo paralleled translation data creation methodology that leverages language models trained on large amounts of monolingual data, and generated new sentences pairs containing rare words in new synthetically created contexts. Sennrich et al. (2016b) proposed a Back-translation method for paralleled translation data augmentation. They trained the reverse (i.e., target-to-source translation) model and then created pseudo data by pairing the output of the reverse model and the input. This simple but powerful data augmentation methodology achieved significant performance improvement of neural machine translation models. To obtain training data on a **high-quality**, Junczys-Dowmunt (2018) proposed a paralleled translation data filtering method based on the dual conditional cross-entropy computed by pre-trained encoder-decoder model. Regarding the effectiveness in improving training data, several studies have empirically demonstrated that acquiring and utilizing appropriate training data leads to a more significant improvement in translation quality than improving the models themselves. (Edunov et al., 2018; Morishita et al., 2018).

Chapter 3

Filtering Noisy Dialogue Corpora by Connectivity and Content Relatedness

Some million-scale datasets such as movie scripts and social media posts have become available in recent years for building neural dialogue agents (Henderson et al., 2019; Lison and Tiedemann, 2016). Such large-scale datasets can be expected to improve the performance of dialogue response generation models based on deep neural networks (DNNs) since the combination of DNNs and large-scale training datasets has led to considerable performance improvement in many sentence generation tasks (Adiwardana et al., 2020; Koehn and Knowles, 2017; Sennrich and Zhang, 2019).

In contrast to the quantity of the data, the quality of the data has often been problematic. For example, OpenSubtitles (Lison and Tiedemann, 2016; Lison et al., 2018), the most widely used large-scale English dialogue corpus, was constructed by collecting two consecutive lines of movie subtitles under the simplified assumption that one line of a movie subtitle is one utterance and the next line is the next utterance follow it. Inevitably, this corpus includes unacceptable utterance pairs from the viewpoint of a conversational sequence, e.g., caused by scene switching or flashback. Several previous studies have identified such flaws and reported that the corpus is *noisy* (Baheti et al., 2018; Li et al., 2016a; Vinyals and Le, 2015a), where *noisy* refers to unacceptable utterance pairs in this context. Figure 3.1 shows the result of our experimental investigation regarding the acceptability rate of the utterance pairs in the OpenSubtitles corpus.[1] It can be noticed from the figure that only half of the utterance pairs can be considered *acceptable* (i.e., were rated with score 5: Strongly agree or 4: Agree), and

[1] In our experiments, randomly sampled 100 utterance pairs were evaluated by native English speakers. We used Amazon Mechanical Turk (MTurk) to evaluate the data manually. We filtered out unreliable workers by integrating attention checks. We requested five workers to evaluate each pair with a five-point Likert scale (5: Strongly agree to 1: Strongly disagree) Likert (1932) as an answer to the question.

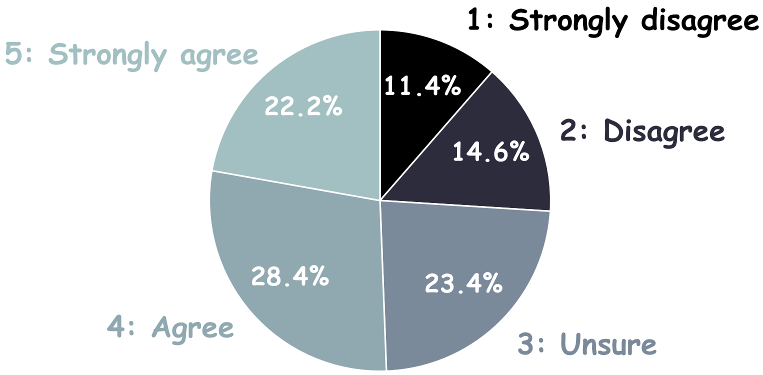


Fig. 3.1 *Is the sequence of the two utterances acceptable as a dialogue?* Response acceptability scores are given by humans on the English OpenSubtitles corpus.

over 25% of utterance pairs are clearly *unacceptable* (i.e., were rated with score 1: Strongly disagree or 2: Disagree) from the human perspective.The samples of unacceptable/acceptable utterance pairs annotated by humans are listed in Table 3.1.

With this situation, a straightforward research question arises, namely, *Can we further improve the performance of neural response generation models by ablating unacceptable utterance pairs from training data?* To the best of our knowledge, no previous study has explicitly focused on this question. Thus, the goal of this paper is to provide an answer to this question. Furthermore, it is not clear whether and how one can effectively discover unacceptable utterance pairs within large-scale training datasets. This study explores a way of constructing a scoring method for filtering *noisy* data filtering to improve the performance of response generation models.

To achieve the set goals, we started with a review of previous arguments about the criteria for identifying appropriate utterances in dialogues and designed our scoring function that is consistent with reflects as much of the community's consensus as possible. In particular, the proposed scoring method estimates the quality of utterance pairs based on the following two aspects: (i) the **connectivity** between source and target utterances and (ii) their **content relatedness** (Section 3.3).

The contributions of this study are the following:

- We propose a scoring method for estimating the quality of utterance pairs in an unsupervised manner (Section 3.4);
- We reveal that our scoring method effectively detects unacceptable utterance pairs, and thus, be appropriate for noisy data filtering (Section 3.5);
- We empirically prove that our proposed data filtering method improves the performance of neural response generation models (Section 3.6); and

- We confirm that our noisy data filtering approach is effective across different languages and dataset sizes (Section 3.7).

3.1 Task Definition: Noisy Data Filtering

Let x be an **utterance** and y be a **response** to x. Then, an **utterance pair** can be denoted as we refer to (x,y). Let $\mathscr{D}$ be a dataset that comprising a set of utterance pairs, $\mathscr{D} = \{(x,y)\}$. Then, the task can be formulated as ablating unacceptable utterance pairs from $\mathscr{D}$ to obtain a less noisy subset $\mathscr{D}' \subseteq \mathscr{D}$, hereinafter referred to as filtered dataset. $\mathscr{D}'$ can then be used to train response generation models. This paper refers to this process as **noisy data filtering**, where *noisy* means unacceptable utterance pairs in this context. Furthermore, we establish a function $S \colon \mathscr{D} \rightarrow \mathbb{R}$ is used to score the degree of *acceptability* of each utterance pair $(x,y) \in \mathscr{D}$.

3.2 Background

Response generation using noisy data. The following two approaches are widely used to address the problem of dialogue response generation noisy dialogue corpora. According to the *model approach*, models are trained while handling noise at the same time. For example, Shang et al. (2018) proposed a method with a calibration framework and demonstrated its effectiveness on a Chinese corpus. According to the *data approach*, training data are pre-processed with the aim of improving their quality before training models. In this study, we take the data approach in light of the success of noisy parallel corpus filtering in machine translation (MT). Additionally, it has become a reasonable strategy to reduce the size of training data since enormous dialogue data has been available. Csáky et al. (2019)'s method is most relevant to our study in that it cleanses dialogue corpora. However, the main goal of their method is to eliminate generic, or boring, responses, whereas the goal of the method proposed here is to eliminate unacceptable utterance pairs. This difference in goals leads to the essential difference in filtering strategies.

Effectiveness of filtering noisy data in neural machine translation. Researchers in the field of neural machine translation (NMT) have recognized that collecting high-quality training data to be equally or even more important than exploring sophisticated model architectures (Junczys-Dowmunt, 2018; Koehn et al., 2018; Morishita et al., 2018). Techniques used in neural response generation and NMT are nearly identical; e.g., sequence-to-sequence models (Sutskever et al., 2014) and Transformers (Vaswani et al., 2017) are often used as base model architectures. We hypothesize that high-quality filtered dialogue data can also improve the performance of dialogue response generators. However, the straightforward application of methods proposed for filtering noisy data in NMT may not work well due to the different nature of NMT and neural response generation tasks. In particular, MT data have one-to-one (ignoring paraphrases) correspondence in source and target sentences, whereas dialogues have many-to-many mappings (Zhao et al., 2017). The experiments presented in this paper provide an answer to whether NMT filtering methods can perform well in dialogue response generation.

3.3 Requirements to Utterance Pairs

In this section, we investigate the requirements that should be satisfied by an acceptable utterance pair.

3.3.1 Criteria for Manual Evaluation

The instructions for manual evaluation provided by the dialogue community explain the key factors for distinguishing acceptable and unacceptable utterance pairs.

In many previous studies, human raters were asked to evaluate the **connectivity** of utterance pairs. For instance, Shang et al. (2015) asked whether a response could be considered as *an appropriate and natural response to the post*. Xing et al. (2017) asked whether *the response can be used as a reply*. Pei and Li (2018) asked whether *the answer is natural* for the question. Other studies have also evaluated the same or similar aspects by using keywords related to the connectivity, such as *semantically appropriate for* (Akama et al., 2017) or *coherent with* (Shen et al., 2017), and *coherence* (Lowe et al., 2017).

Another frequently used metric is **content relatedness**. For instance, Galley et al. (2015) asked human evaluators to evaluate *each response in terms of their relevance to a given utterance*. Li et al. (2016a) asked for the preference of responses *that were more specific to certain utterances*. Ritter et al. (2011) suggested that *an appropriate response should be on the same topic as the utterances*. Several other studies have also focused on evaluating the *relevance* between an utterance and its response (Lowe et al., 2017; Pei and Li, 2018; Xu et al., 2018b).

In summary, the most widely used criteria can be categorized into connectivity and content relatedness of utterance pairs. In fact, these two aspects are considered in the field of sociolinguistics as crucial features of conversation (Sacks, 1989; Sidnell, 2010).

Table 3.1 Samples of pairs judged as unacceptable/acceptable in preliminary experiments. Human denotes the average score of five human evaluators on a scale of 1-5. Phrases considered to contribute to connectivity are highlighted. Estimated `[topic]` of utterance is written in the end of each utterance.

	Utterance	Response	Human
1 :	It'll be like you never left. `[??]`	I painted a white line on the street way over there. `[painting]`	1.4
2 :	You're gonna get us assimilated. `[??]`	Switch to a garlic shampoo. `[??]`	1.8
3 :	I probably asked for too much money. `[money]`	Money's always a problem, isn't it? `[money]`	4.2
4 :	I wonder who I should call back. `[phone]`	They're saying they want to call one of you back. `[phone]`	4.4
5 :	Okay, so where's the rest? `[??]`	Electronically scanned and archived at headquarters but you'll have to speak with them about that. `[work]`	4.4

3.3.2 Observation

Furthermore, we investigated how the two aforementioned aspects can be observed in actual utterance pairs. For this investigation, we use the utterance pairs scored by human raters that were used in our preliminary experiments shown in Figure 3.1. Some examples are shown in Table 3.1.

We observe that typical phrase pair patterns can often be found in utterance pairs with high scores. For example, the pair (*where is* , *at*) in Table 3.1 is one of the typical phrase pair patterns that asks a place and provides an answer to it. Other typical examples include (*why*, *because*) and (*what do you want*, *I want*). In discourse linguistics, such phrase pair patterns are known as the concept of *cohesive devices*. Hereafter, we refer to such a typical phrase pair pattern as **key phrase pair**.

Moreover, in high scored utterance pairs, both an utterance and response are on the **same topic**. For example, in the third example listed in Table 3.1, both the utterance and response mention `[money]`.

3.4 Proposed Method

As per the discussion in the previous section, each acceptable utterance pair should satisfy the following criteria:

- **connectivity** — existence of key phrase pairs

- **content relatedness** — topic commonality

This section presents the proposed scoring functions to assess the degree of satisfying the above two criteria in an unsupervised manner.[2]

3.4.1 Connectivity

Let f and e represent phrases obtained from x and y, respectively. Let $\phi(x,y)$ be a function that returns a set of all possible phrase (n-gram) pairs obtained from the utterance pair (x,y). We can define a finite set of all possible phrase pairs obtained from the entire dialogue data $\mathscr{D}$ as $\overline{\mathscr{P}}_{\mathscr{D}} = \bigcup_{(x,y)\in\mathscr{D}} \phi(x,y)$. Then, let $\mathscr{P}$ represent a set of key phrase pairs (defined in Section 3.3.2). We assume that $\mathscr{P}$ is a subset of $\overline{\mathscr{P}}_{\mathscr{D}}$, i.e., $\mathscr{P} \subseteq \overline{\mathscr{P}}_{\mathscr{D}}$.

To obtain $\mathscr{P}$, we take advantage of a phrase table extraction technique developed in statistical machine translation, e.g., Moses (Koehn et al., 2007). In this task, we require only some phrase pairs that can contribute to the connectivity of an utterance pair (as mentioned in Section 3.3.2), unlike the translation task where the whole sentence must correspond in mutual. Accordingly, in our experiments, we set the null alignment ratio (i.e., probability of no alignment) to 0.5 and extend the phrase extraction algorithm to include only the explicitly corresponding range as phrases in our table.

Then, we define the scoring function $S_{\mathbf{C}}$ to estimate connectivity as:

$$S_{\mathbf{C}}(x,y) := \sum_{(f,e)\in\phi(x,y)\cap\mathscr{P}} \max\bigl(\mathrm{nPMI}(f,e),0\bigr) \cdot \frac{|f|}{|x|} \cdot \frac{|e|}{|y|}, \tag{3.1}$$

where $|\cdot|$ denotes the number of words in the phrase or utterance. To calculate the co-occurrence, we use the normalized pointwise mutual information (nPMI) (Bouma, 2009), which normalizes the value so that low-frequency phrases do not take an extremely large value. Note that we ignore the negative nPMI scores by the $\max(\cdot,0)$ operation because we aim only to consider the positive effect of connectivity. The intuition behind Equation 3.1 is as follows:

- If a phrase pair (f,e) has a high co-occurrence, the association strength of (x,y) including (f,e) might also be high.
- If a phrase f or e occupies almost the entire sentence x or y, (f,e) is a strong indicator of the association of (x,y).

[2] The reason for focusing on an unsupervised approach the lack of data that can provide good supervision for utterance pair evaluation.

3.4.2 Content Relatedness

Let $v(x)$ and $v(y)$ be sentence vector of x and y, respectively. We compute topic commonality of x and y, that is, content relatedness as follows:

$$S_{\mathbf{R}}(x,y) := \max\big(\cos(v(x), v(y)), 0\big). \tag{3.2}$$

Cosine similarity between certain kinds of sentence vectors is known to be a good proxy of the topical relatedness of two sentences (Conneau et al., 2017; Subramanian et al., 2018; Xu et al., 2018a). For the same reasons as Equation 3.1, we ignore the negative cos scores by the $\max(\cdot, 0)$ operation.

3.4.3 Summary

Eventually, combining the above two scoring measures, we propose the following function:

$$S_{\mathbf{C+R}}(x,y) := \alpha S_{\mathbf{C}}(x,y) + \beta S_{\mathbf{R}}(x,y), \tag{3.3}$$

where α, $\beta \in \mathbb{R}_{\geq 0}$ are hyperparameters that weigh the two viewpoints. For our experiments, we fix α and β as follows:

$$\alpha = \frac{1}{\frac{1}{|\mathscr{D}|}\sum\limits_{(x,y)\in\mathscr{D}} S_{\mathbf{C}}(x,y)},\ \beta = \frac{1}{\frac{1}{|\mathscr{D}|}\sum\limits_{(x,y)\in\mathscr{D}} S_{\mathbf{R}}(x,y)}. \tag{3.4}$$

3.5 Experiments: Data Scoring

In this section, we describe our experiments that validate the effectiveness of the proposed scoring method.

3.5.1 Experimental Setup

Dataset. We conducted our experiments on a noisy English dialogue corpus from OpenSubtitles (Lison et al., 2018) containing roughly 441M lines. We automatically obtained dialogue paired-data from the corpus which does not contain speaker annotations on the dialogue turns *following the previous methods*, such as Li et al. (2016a); Vinyals and Le (2015a). Specifically, we extracted the consecutive two lines as an utterance pair based on the assumption that each line corresponds to a full-speaker's turn. As explained at the start of this chapter, it includes many unacceptable utterance pairs. We first applied several rule-based

Table 3.2 The statistics of our English dataset.

Data	# works	# lines	# our pairs
Corpus	446,612	441,452,475	79,621,506
Train	442,433	441,065,310	79,445,453
Valid	200	195,297	90,317
Test	200	191,868	85,736

filtering as rudimentary preprocesses, which are typically used in the related literature. We collected pairs from the dataset in which the length of the utterance and response was 3-25 words each and obtained the dialogue dataset. For counting the number of words, we used SpaCy[3] to tokenize each utterance and response. Some processings were inspired by the technique of noisy-parallel corpus filtering on NMT fields Junczys-Dowmunt (2018). The additional preprocesses that we conducted are as follows:

- Using languid[4], which is a tool that detects the language for given sentences, we removed the utterance pairs judged as any language other than the target language.
- Removed the parrot-back utterance pairs.
- Removed duplicate utterance pairs in order to remove the completely repeated conversational sequences, such as the opening scenes of serial dramas.

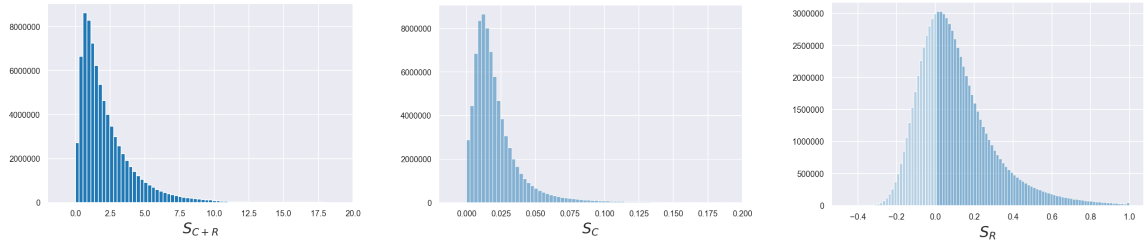
Then, we obtained 79,445,453 utterance pairs as our training data, which excludes our test and validation data. Table 3.2 shows the statistics of our English dataset.

Proposed method: detailed setup. To compute the connectivity $S_{\mathbf{C}}$, we obtained a phrase table on our training data by using Moses (Koehn et al., 2007) with fastAlign (Dyer et al., 2013). We then removed phrase pairs with a low co-occurrence frequency (here, less than 200 times) or composed of the same phrases from the table. As a result, the phrase table included 68,891 phrase pairs, which were used as the key phrase set $\mathscr{P}$ as described in Section 3.4.1.

To compute the content relatedness $S_{\mathbf{R}}$, we created a sentence vector from pre-trained fastText word embeddings (Bojanowski et al., 2017; Mikolov et al., 2018) following Arora et al. (2017)'s method, i.e., using SIF weighting and common component removal. Their method is reported to be useful for computing the relatedness of two given sentences and used in many studies (Baheti et al., 2018; Conneau et al., 2017; Marelli et al., 2014a,b; Subramanian et al., 2018). We learned common components using 30K sentences randomly

[3] https://spacy.io/

[4] https://github.com/saffsd/langid.py



(a) $S_{\mathbf{C+R}}(x,y)$ (b) $S_{\mathbf{C}}(x,y)$ (c) $S_{\mathbf{R}}(x,y)$

Fig. 3.2 Score distributions of our $S_{\mathbf{C}}$, $S_{\mathbf{R}}$, $S_{\mathbf{C+R}}$ across our training data (English).

selected from the training costs appropriately. We then removed the first common component for all sentence vectors.

Figure 3.2 shows the score distributions of our $S_{\mathbf{C}}$, $S_{\mathbf{R}}$, $S_{\mathbf{C+R}}$ across our English training data.

Baselines. For comparison, we prepared the following two baselines:

- Csáky et al. (2019): Entropy-based filtering to remove generic utterances from the training data for promoting less-boring response generation. SRC/TRG indicates that using the entropy of source/target utterances.

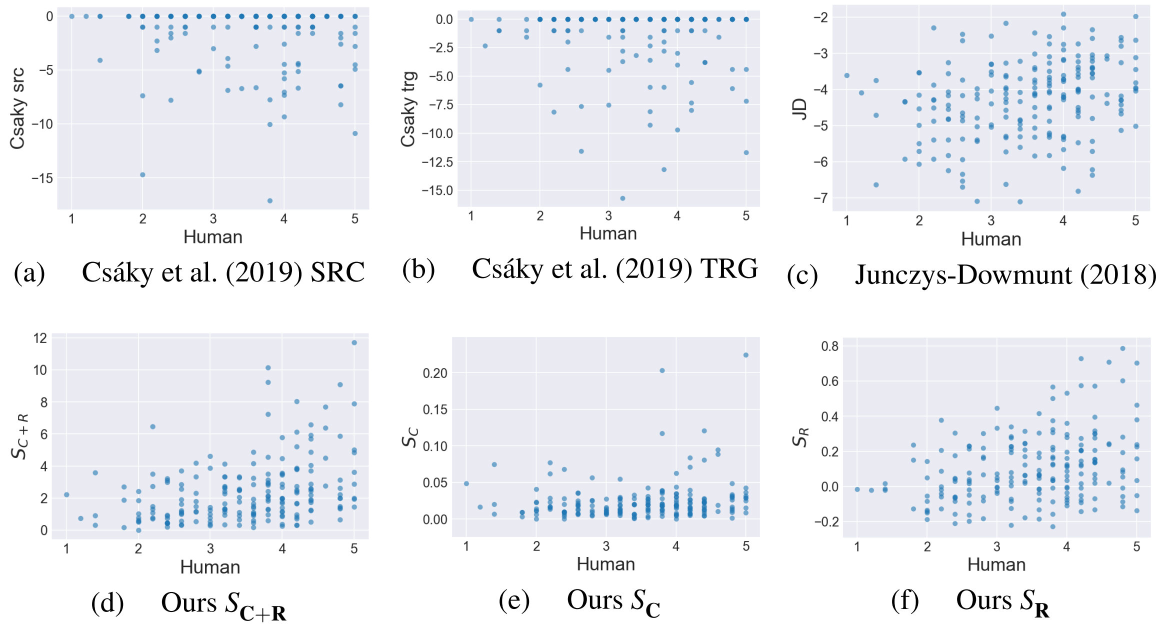
- Junczys-Dowmunt (2018): Filtering for NMT based on the dual conditional cross-entropy computed by a neural encoder-decoder model. It achieved the best performance on the Parallel Corpus Filtering Task at WMT 2018.[5]

Human evaluation. To validate the ability of the proposed method to estimate the quality of utterance pairs, we measured the correlation between its scores and those assigned by humans through crowdsourcing. We used Amazon Mechanical Turk.[6] We randomly extracted 200scored utterance pairs and asked native English-speaking crowdworkers to answer the following question for each pair: *Is the sequence of the two utterances acceptable as a dialogue?* Workers were instructed to provide an answer on a five-point Likert scale (from 5: Strongly agree to 1: Strongly disagree) (Likert, 1932). Unqualified workers were filtered out using attention checks. Eventually, we used the average of the scores provided by five workers as the human score for each pair.

[5] http://www.statmt.org/wmt18/
[6] https://www.mturk.com/

Table 3.3 Correlation coefficient between human scores and automatically computed scores (English).

Scoring method	Spearman's ρ	p-value
Csáky et al. (2019) SRC	−0.1173	9.8×10^{-2}
Csáky et al. (2019) TRG	0.0462	5.2×10^{-1}
Junczys-Dowmunt (2018)	0.2973	1.9×10^{-5}
Ours $S_{\mathbf{C+R}}$	**0.3751**	$\mathbf{4.4 \times 10^{-8}}$
Ours $S_{\mathbf{C}}$ (ablation study)	0.2044	3.7×10^{-3}
Ours $S_{\mathbf{R}}$ (ablation study)	0.3007	1.5×10^{-5}



(a) Csáky et al. (2019) SRC (b) Csáky et al. (2019) TRG (c) Junczys-Dowmunt (2018)

(d) Ours $S_{\mathbf{C+R}}$ (e) Ours $S_{\mathbf{C}}$ (f) Ours $S_{\mathbf{R}}$

Fig. 3.3 Distributions between human scores and automatically computed scores by each method (English).

3.5.2 Results and Analysis

Table 3.3 shows the correlation between human scores and those automatically computed by each method. Among the methods, $S_{\mathbf{C+R}}$ achieved the highest correlation with human scores. Additionally, we also evaluated $S_{\mathbf{C}}$ and $S_{\mathbf{R}}$ as an ablation study of $S_{\mathbf{C+R}}$. We found that both scores were less correlated than $S_{\mathbf{C+R}}$. This result supports the hypothesis that both aspects, namely, connectivity and content relatedness, should be considered when evaluating the quality of utterance pairs.

Figure 3.3 shows the distribution of automatically computed scores corresponding to human scores.As shown in (c), $S_{\mathbf{C+R}}$ rarely overestimates utterance pairs with low human

Table 3.4 Samples of utterance pairs scored with our method and human judgements (English). The scores of $S_{\mathbf{C}}$ and $S_{\mathbf{R}}$ were normalized by α, β.

	Utterance	Response	$S_{\mathbf{C}}$	$S_{\mathbf{R}}$	$S_{\mathbf{C+R}}$	Human
1 :	What is the anarchy facing the jail of the sick passion?	Gosh, it's really cold!	0.32	0.00	**0.32**	1.4
2 :	Pushers won't let the junkie go free.	Across 110th Street.	0.00	0.42	**0.42**	2.4
3 :	It started when I was 17.	They'd make a cash drop,	0.63	0.00	**0.63**	2.0
4 :	A big nail should be put in your head	Who are they	0.74	0.00	**0.74**	1.2
5 :	He told me so.	Oh, he did, huh?	2.21	0.00	**2.21**	4.8
6 :	There's a laundry.	Have your clothes dry-cleaned, okay?	0.81	2.89	**3.70**	4.4
7 :	Then if I win, what are you going to do?	When you win?	1.04	7.01	**8.05**	4.2
8 :	But what do you want me to do?	We want you to kick her off the team.	10.20	1.53	**11.72**	5.0

scores but underestimates those with high human scores. The baseline methods presented in (a) and (b) do not show such behavior. This behavior unique to $S_{\mathbf{C+R}}$ is safe for the noisy data filtering task since it can successfully detect lower-quality pairs with high precision. On the other hand, improperly underestimating some acceptable pairs (i.e., low recall) is one downside of $S_{\mathbf{C+R}}$, and we discuss its influences in Section 3.5.3. We emphasize that $S_{\mathbf{C+R}}$ has a desirable property for noisy data filtering in today's situation where a sufficiently large corpus is available; it allows us to obtain a sufficient amount of clean data even if discarding a certain portion of potentially clean data. Interesting future work is to investigate how to improve our methods not to underestimate acceptable pairs while maintaining high precision. It is nearly equivalent to develop an unsupervised approach of dialogue evaluation methods, and thus, this direction is a challenging and essential attempt.

Table 3.4 shows several examples of utterance pairs well-scored by $S_{\mathbf{C}}$, $S_{\mathbf{R}}$, and $S_{\mathbf{C+R}}$. Note that the score ranges differ; e.g., human scores are in $[1,5]$, while $S_{\mathbf{R}}$ is in the range $[0,1]$.Thus, we discuss relative score values; the comparison of absolute score values across the different methods would be meaningless. These examples demonstrate that the complementary contributions of both $S_{\mathbf{C}}$ and $S_{\mathbf{R}}$ allow $S_{\mathbf{C+R}}$ to provide quality estimations close to human judgments.

3.5.3 Discussion on Low Recall Property

What types of pairs cause low recall? Since the proposed method prefers precision over recall, it tends to discard a certain number of acceptable utterance pairs during filtering. To investigate the characteristics of such discarded (yet acceptable) pairs, we analyzed 27 pairs.

Table 3.5 Samples of utterance pairs that cause low recall scored with our method and human judgements (English). The scores of $S_{\mathbf{C}}$ and $S_{\mathbf{R}}$ were normalized by α, β.

	Utterance	Response	$S_{\mathbf{C}}$	$S_{\mathbf{R}}$	$S_{\mathbf{C+R}}$	Human
1 :	What happened to your hand?	Just a scratch.	1.38	0.00	**1.38**	4.8
2 :	But Carcharodontosaurus has the more lethal bite.	This time, the Spinosaurus triumphed.	0.22	0.68	**1.72**	4.0
3 :	I'm right here with you.	Come on, boys.	1.39	0.00	**1.39**	4.0
4 :	What Is It Officer Chan?	Brother Ho, I must leave now,	1.04	0.68	**1.72**	4.4
5 :	Out on the balcony.	You shouldn't have come.	0.30	0.00	**0.30**	4.2

Some examples are listed in Table 3.5. These pairs were selected from those that obtained a human score of 4.0 or above (77 pairs) *and* were among the worst 50% as scored by $S_{\mathbf{C+R}}$ (100 pairs). Consequently, we found two potential issues. One is that human annotators may sometimes easily find the connectivity or the content relatedness for the utterance pairs with the low $S_{\mathbf{C+R}}$ scores. This observation indicates that $S_{\mathbf{C}}$ and $S_{\mathbf{R}}$ are still not perfect for scoring functions, and there remains room for improvement. The possible drawbacks we have already noticed in $S_{\mathbf{C}}$ and $S_{\mathbf{R}}$ are that $S_{\mathbf{C}}$ sometimes fails to capture the connectivity because of the limited coverage by a discrete phrase table-based approach, and $S_{\mathbf{R}}$ is not robust for out-of-vocabulary of word vector. The other case is that the human annotators gave high scores, but we found no connectivity and content relatedness in the utterance pairs. We found that some utterance pairs without any connectivity and content relatedness can be judged as acceptable by the human annotators since they can imagine the underlying context and situation of the utterance pairs using human world knowledge, such as commonsense. We think this is a challenging issue that exceeds our focus in this paper, and thus, remains as future work.

Does our filtering undermine diversity? One might think that our method succeeds in filtering by assigning high scores to generic responses such as dull responses. This concern makes sense since it is known that dialogue systems learned from the training data, including many generic utterances, tend to generate bland responses (Csáky et al., 2019). To answer this interesting question, we confirmed the diversity of utterance pairs with a high score (i.e., remained as training data) and a low score (i.e., removed from training data) in our $S_{\mathbf{C+R}}$ (Table 3.6).As a result, there was no significant difference between them. Therefore, we conclude that the proposed method does not prefer only generic responses and maintains the diversity of data. It is an essential future attempt to improve the quality of dialogue data further (e.g., more diversity) after using the proposed method to remove unacceptable pairs.

Table 3.6 Comparison of the top and the worst utterance pairs in the training data scored by our method (English).

Scored data	Utterance pair			Utterance (source-side)			Response (target-side)		
	len	distinct-1	distinct-2	len	distinct-1	distinct-2	len	distinct-1	distinct-2
Top 50% (remained)	18.06	0.018	0.313	9.05	0.028	0.474	9.02	0.028	0.472
Worst 50% (removed)	17.92	0.019	0.316	8.92	0.030	0.476	9.00	0.030	0.470

3.6 Experiments: Response Generation

This section reports on the effectiveness of the proposed method for filtering noisy data in neural response generation.

3.6.1 Experimental Setup

Training. We obtained the filtered training data $\mathscr{D}'$ by removing utterance pairs with low scores from the original dataset $\mathscr{D}$ (approximately 10% or 50% of total utterance pairs were removed).

As a response generation model, we used a Transformer (Vaswani et al., 2017) based encoder-decoder model implemented in the `fairseq` toolkit (Ott et al., 2019).Transformer has demonstrated high performance in response generation (Dinan et al., 2019) and other NLP tasks. We used ‘`-arch transformer_wmt_en_de_big`’ option with its default configuration, and set the number of maximum training steps to 100K. For token segmentation, we used the byte pair encoding Sennrich et al. (2016c) and set its vocabulary size to 16K. The numbers of parameters in our models were roughly 223M. We trained our models on 8 NVIDIA DGX-1 Tesla V100 GPUs. It took approximately 6 hours for training one model.

Automatic evaluation. Here, we report the following metrics: the average response length in tokens (len), type-token ratio for $\{1,2\}$-grams (distinct-$\{1,2\}$), and and BLEU-1 (Papineni et al., 2002). The latter was used as a reference-based metric; while it is widely used in previous studies (Baheti et al., 2018; Csáky et al., 2019; Zhao et al., 2017), some studies (e.g., Liu et al. (2016)) have reported that BLEU-1 may not be highly correlated with the human evaluation of response generation.

Human evaluation. We evaluated the quality of the generated responses manually. We asked human evaluators recruited via Amazon Mechanical Turk to evaluate responses that

Table 3.7 Evaluation results for generated responses (English; filtered out 50%). **Bold** denotes the best results. The ✗/✔ shows the percentages of the low/high scored responses (i.e., human scores in $[1,3)$ or in $[3,5)$).

Training data	# of pairs	Automatic evaluation				Human evaluation		
		len	distinct-1	distinct-2	BLEU-1	Avg.	✗↓	✔↑
non-filtered	79,445,453	8.44	127/0.030	238/0.064	8.8	3.37	38 %	62 %
Csáky et al. (2019) SRC	40,000,000	7.97	165/0.041	329/0.094	9.1	3.56	25 %	75 %
Csáky et al. (2019) TRG	40,000,000	18.25	213/0.023	591/0.069	5.4	2.85	65 %	35 %
Junczys-Dowmunt (2018)	40,000,000	**8.63**	206/0.048	478/0.125	**9.4**	3.43	32 %	68 %
Ours $S_{\mathbf{C+R}}$	40,000,000	7.13	**345/0.097**	**853/0.278**	**9.4**	3.73	**15 %**	**85 %**
Ours $S_{\mathbf{C}}$ (ablation study)	40,000,000	7.31	201/0.055	466/0.148	9.2	3.69	19 %	81 %
Ours $S_{\mathbf{R}}$ (ablation study)	40,000,000	7.91	270/0.068	662/0.192	**9.4**	**3.76**	20 %	80 %
reference		9.04	1301/0.288	3244/0.807	-	-	-	-

are generated for 100[7] input utterances randomly sampled from the test data. We used the same task setting and protocol as described in Section 3.5.1 to obtain the human scores for each pair. Higher human scores indicate that the better results.

3.6.2 Results and Analysis

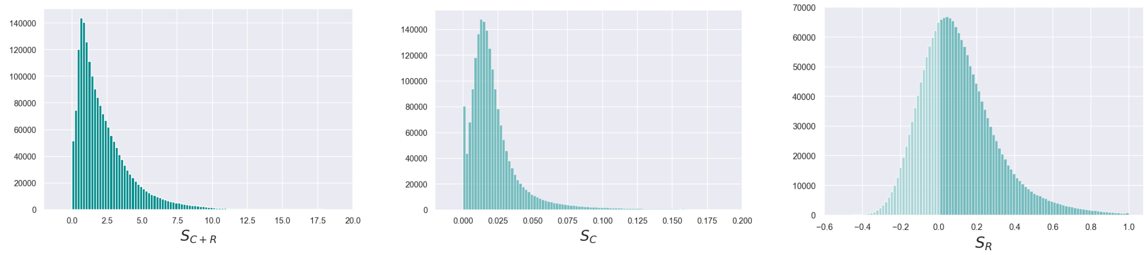
Table 3.7 shows the results of automatic and human evaluations of the generated responses. The model trained on the data filtered using the proposed method $S_{\mathbf{C+R}}$ produced more than three times as many distinct $\{1,2\}$-grams as the model trained on non-filtered data. Furthermore, it outperformed the model trained on non-filtered data in the human evaluation, achieving the highest percentage of acceptable responses of 85%. Additionally, these results of our $S_{\mathbf{C+R}}$ were better than other baselines. To conclude, these experimental results indicate that the proposed scoring method can help generate diverse responses that are judged as acceptable by humans. Samples of the generated responses are shown in Table 3.8.

This experiment provides empirical evidence for supporting our hypothesis that the performance of neural response generation models can be improved by just removing unacceptable utterance pairs from training data, which answers the research question formulated at the start of this paper.

[7] Same size as Shen et al. (2017) and Bao et al. (2020).

Table 3.8 Samples of input utterances **(bold)** and responses generated by models trained on filtered data by proposed method $S_{\mathbf{C+R}}$ (Ours). Other responses are generated by models trained on (A) non-filtered data and filtered data by baselines; (B) Csáky et al. (2019)'s SRC method, (C) their TRG method, and (D) Junczys-Dowmunt (2018)'s method, respectively.

1 :	**Well, I thought we'd wait for Mr. Oliver.**
A :	I don't think so.
B :	I'll be right back.
C :	I'm sorry, but I have to get back to work.
D :	No, I don't think so.
Ours :	Oh, I'm sorry.
2 :	**Are you heading home?**
A :	No, I don't think so.
B :	No, I don't think so.
C :	No, no, no, no, no, no, no, no, no, no, no, no.
D :	No, I don't think so.
Ours :	No, I'm not.
3 :	**Because we want to keep him safe.**
A :	I don't know what you're talking about.
B :	I don't know what you're talking about.
C :	I don't care if he's dead or alive or dead.
D :	I don't know what you're talking about.
Ours :	We need to get him out of here.
4 :	**Quite soon, now, you'll be able to travel.**
A :	I don't think so.
B :	I'll be back in a minute.
C :	I'll be back in a few days, then I'll be back.
D :	I don't want to see you again.
Ours :	No, I'm not going anywhere.
5 :	**Which is where exactly?**
A :	I don't know.
B :	I don't know.
C :	I don't know what you're talking about, but I don't know where you are.
D :	I don't know.
Ours :	I don't know.



(a) $S_{\mathbf{C+R}}(x,y)$ (b) $S_{\mathbf{C}}(x,y)$ (c) $S_{\mathbf{R}}(x,y)$

Fig. 3.4 Score distributions of our $S_{\mathbf{C}}$, $S_{\mathbf{R}}$, $S_{\mathbf{C+R}}$ across training data (Japanese).

3.7 Multilingual Availability

While the proposed method $S_{\mathbf{C+R}}$ was tested on an English corpus, it can potentially work for other languages as well. To demonstrate this, we selected Japanese dialogue data as another case study.The linguistic phenomena in Japanese are quite different from those in English, thus making this experiment to be a good test of the applicability of the proposed method to non-English languages.

Japanese dataset. We prepare the Japanese dialogue data from Japanese OpenSubtitles (Lison et al., 2018) containing roughly 3M lines. We obtain 1,893,477 utterance pairs as our training data, which excludes our test and validation data.

3.7.1 Data Scoring

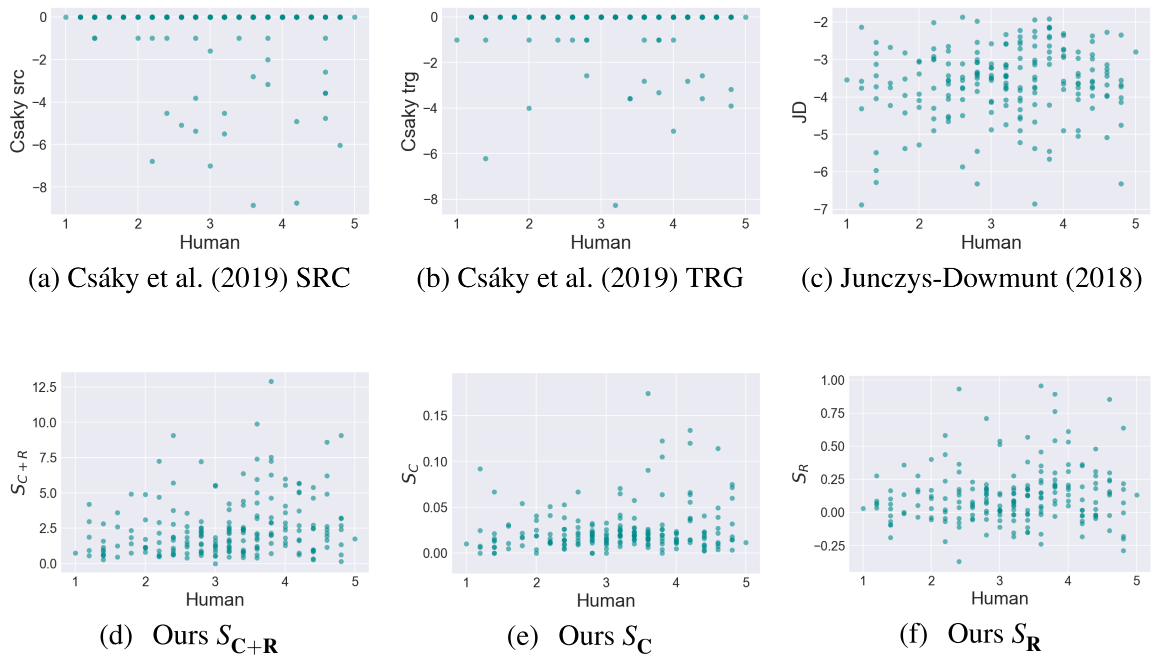
Settings. To compute $S_{\mathbf{C}}$, we defined a low co-occurrence frequency as less than 20, considering the size of the Japanese corpus, and consequently obtained the key phrase pairs $|\mathscr{P}| = 19{,}992$. To compute $S_{\mathbf{R}}$, we used pre-trained fastText (Grave et al., 2018) and learned common components from all sentences in the training data. For human evaluation, we used Yahoo! crowdsourcing[8] to hire native Japanese-speaking workers. The task setting and protocol are the same as those for English (Section 3.5.1), regardless of the crowdsourcing platform. Figure 3.4 shows the score distributions of our $S_{\mathbf{C}}$, $S_{\mathbf{R}}$, $S_{\mathbf{C+R}}$ across our Japanese training data.

Results and analysis. Table 3.9 shows the correlation between human scores and those automatically computed by each method. Our method $S_{\mathbf{C+R}}$ has the highest correlation with human scores, although the overall result is lower than that obtained for the English dataset.

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

Table 3.9 Correlation coefficient between human scores and automatically computed scores (Japanese).

Scoring method	Spearman's ρ	p-value
Csáky et al. (2019) SRC	-0.0553	4.4×10^{-1}
Csáky et al. (2019) TRG	-0.0366	6.1×10^{-1}
Junczys-Dowmunt (2018)	0.1074	1.3×10^{-1}
Ours $S_{\mathbf{C+R}}$	**0.2491**	$\mathbf{3.8 \times 10^{-4}}$
Ours $S_{\mathbf{C}}$ (ablation study)	0.1395	4.9×10^{-2}
Ours $S_{\mathbf{R}}$ (ablation study)	0.1504	3.3×10^{-2}



(a) Csáky et al. (2019) SRC (b) Csáky et al. (2019) TRG (c) Junczys-Dowmunt (2018)

(d) Ours $S_{\mathbf{C+R}}$ (e) Ours $S_{\mathbf{C}}$ (f) Ours $S_{\mathbf{R}}$

Fig. 3.5 Distributions between human scores and automatically computed scores by each method (Japanese).

Figure 3.5 shows the distribution of our $S_{\mathbf{C+R}}$ corresponding to human scores. Similar to the result obtained for English as presented in Figure 3.3 (c), $S_{\mathbf{C+R}}$ rarely overestimates utterance pairs with low human scores but underestimates those with high human scores in Japanese.

3.7.2 Response Generation

Settings. We used the same experimental settings described in Section 3.6.1 for the preparation of filtered data $\mathscr{D}'$ and model training.

Table 3.10 Evaluation results for generated responses (Japanese; filtered out 10%). **Bold** denotes the best results. The ✘/✔ shows the percentages of the low/high scored responses (i.e., human scores in $[1,3)$ or in $[3,5)$).

Training data	# of pairs	Automatic evaluation				Human evaluation		
		len	distinct-1	distinct-2	BLEU-1	Avg.	✘$^{\downarrow}$	✔$^{\uparrow}$
non-filterd	1,893,477	5.91	268/0.091	509/0.207	13.4	3.35	39 %	61 %
Csáky et al. (2019) SRC	1,700,000	5.75	295/0.102	550/0.231	13.2	3.47	37 %	63 %
Csáky et al. (2019) TRG	1,700,000	**7.06**	336/0.095	662/0.219	11.6	3.37	34 %	66 %
Junczys-Dowmunt (2018)	1,700,000	5.31	284/0.107	516/0.240	12.6	3.46	32 %	68 %
Ours $S_{\mathrm{C+R}}$	1,700,000	5.68	**319/0.112**	**582/0.249**	**13.9**	**3.61**	**27 %**	**73 %**
Ours S_{C} (ablation study)	1,700,000	5.51	264/0.096	492/0.218	13.7	3.44	32 %	68 %
Ours S_{R} (ablation study)	1,700,000	5.73	296/0.103	555/0.234	12.5	3.56	30 %	70 %
reference		7.29	750/0.206	1446/0.460	-	-	-	-

Results and analysis. Table 3.10 shows the results of evaluations of the generated responses. The filtered data generated by $S_{\mathrm{C+R}}$ provided the best results in terms of almost all the metrics, including human evaluation. It supports our hypothesis that the proposed method is also suitable for non-English languages.

3.8 Relationship with Evaluation Metric

The proposed method $S_{\mathrm{C+R}}$ maps an utterance pair to a score (scalar value) in terms of the quality of dialogue. That is, formally, our method is similar to the reference-free automatic evaluation metrics for dialogue agents; both of them evaluate the response given an input utterance and also map into a score. Recently, the novel reference-free metrics for evaluating generated responses such as USR (Mehri and Eskenazi, 2020) or MAUDE (Sinha et al., 2020) ware developed. While it is possible to use them as a scoring method for filtering noisy data, in theory, there are some concerns with applying them in practice. One is the difference of the data of interest; since evaluation metrics are intended for responses generated as dialogue, i.e., somewhat valid dialogue data, it is unclear whether they also work for apparently noisy data. Another one is the difference of desired properties; evaluation metrics need to be sensitive to "how good is it?" while the filtering requires to detect "is it a dialogue?" with high accuracy. It would be interesting to investigate the effectiveness of reference-free metrics for noisy dialogue data filtering tasks, and vice versa. We leave these investigations for future work.

In contrast, reference-based metrics require a reference response (i.e., ground truth) when they calculate scores; such metrics include the traditional overlap-based BLEU, ROUGE, METEOR, embedding-based metrics (Liu et al., 2016), and neural network-based RUBER (Tao et al., 2018) and ADEM (Lowe et al., 2017). Thus, these methods cannot straightforwardly be considered as alternatives to the proposed method, which aims at filtering.

3.9 Conclusion

In light of the success of noisy corpus filtering in neural machine translation, we attempted to filter out unacceptable utterance pairs from large dialogue corpora in an unsupervised manner. The proposed scoring method estimates the quality of utterance pairs by focusing on the two crucial aspects of dialogue, namely, the *connectivity* and *content relatedness* of utterance pairs. We demonstrated that our scoring method has a higher correlation with human judgment than recently proposed methods. Furthermore, we provided empirical evidence that our method improves the performance of a response generation model by removing unacceptable utterance pairs from its training data. We hope that this study will facilitate discussions in the dialogue response generation community regarding the issue of filtering noisy corpora.

Chapter 4

Dialogue Data Augmentation by Pairing Single Utterances

In the natural language processing research field, sentence generation technology has been rapidly developed with the help of deep neural network techniques. The recent noteworthy performance improvements of text generation using deep neural networks have been primarily driven by a large amount of high-quality training data. For example, the neural machine translation (NMT) community shares a large amount of high-quality parallel translation data. This is one principal reason why a large number of influential methods and findings have been reported by the community. In contrast, the dialogue community does not have a sufficient amount of high-quality dialogue data for training high-quality deep neural networks. From this background, the goal of this work is to develop a method for efficient dialogue data augmentation to contribute to accelerating progress in the dialogue community. In fact, data augmentation techniques such as back-translation (Sennrich et al., 2016b) are also a central topic in the NMT community, since data augmentation methods have often yielded marked performance improvements (Edunov et al., 2018). However, it is not possible to straightforwardly apply such augmentation methods developed by the NMT community to dialogue data since some essential properties of dialogue data prevent immediate adaptation (see Section 4.2).

In this chapter, therefore, we develop a data creation method of meaningful utterance–response pairs suitable for dialogue response generation. Specifically, our method consists of three steps (see Figure 4.1): (1) generate templates from a set of typical utterance–response pairs, (2) find pairs of matching phrases to fill these templates, and (3) filter out the unreliable candidate pairs to obtain reliable and high-quality utterance–response pairs. The key advantage of our method is that it only requires single utterances, which are already

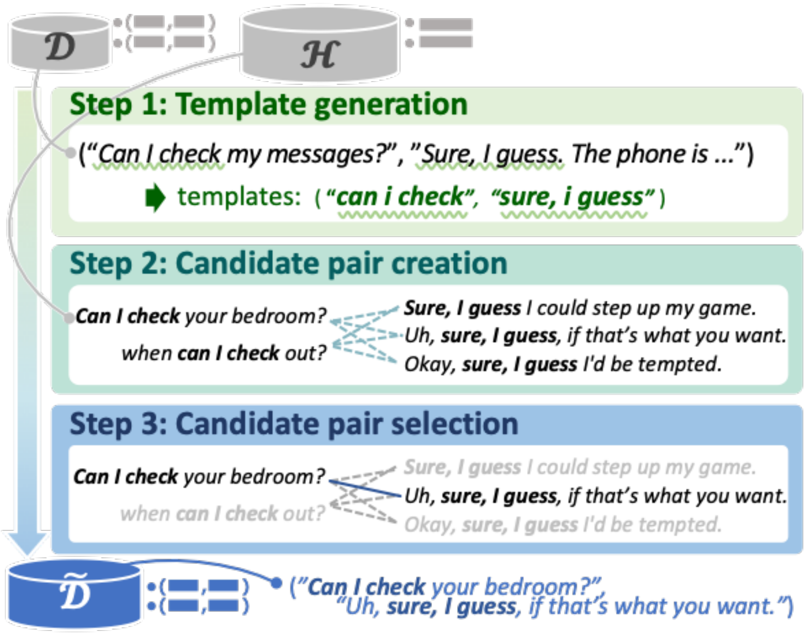

D
H
Step 1: Template generation
("Can I check my messages?", "Sure, I guess. The phone is ...")
templates: ("can i check", "sure, i guess")
Step 2: Candidate pair creation
Can I check your bedroom?
when can I check out?
Sure, I guess I could step up my game.
Uh, sure, I guess, if that's what you want.
Okay, sure, I guess I'd be tempted.
Step 3: Candidate pair selection
Can I check your bedroom?
when can I check out?
Sure, I guess I could step up my game.
Uh, sure, I guess, if that's what you want.
Okay, sure, I guess I'd be tempted.
D̃
("Can I check your bedroom?", "Uh, sure, I guess, if that's what you want.")

Fig. 4.1 Overview of our method. Meaningful pairs $\widetilde{\mathscr{D}}$ are newly created from a small set of utterance–response pairs $\mathscr{D}$ and a set of single utterances $\mathscr{H}$.

available in large quantities. We demonstrate the response generation task in chit-chat using the data obtained by our method.

4.1 Background

The dialogue response generation task is modeled in the same framework as NMT by considering the user utterance as the input sentence and the system response as the output sentence (Shao et al., 2017; Vinyals and Le, 2015a). In other words, we expect that high-quality and rich utterance–response pairs improve performance on the dialogue response generation task in the same way as on the machine translation task. However, presently in dialogue response generation, the methodology for acquiring high-quality and rich utterance–response pairs is hardly discussed.

The current trend of NMT is focusing on researching methodologies of acquiring (or augmenting) rich, high-quality paralleled translation data, such as using a back-translation technique (Sennrich et al., 2016b), more than improving the model architecture itself (Vaswani et al., 2017). This is due to a recent report stating that acquiring and utilizing appropriate training data leads to a more significant improvement in translation quality than improving the model (Edunov et al., 2018; Morishita et al., 2018). We discuss whether the same effect of improved data utilization is also obtainable in the dialogue response sentence generation task.

4.2 Task Setting of Pair Construction

4.2.1 Input and Output

Input. $\mathscr{H}$ denotes a set of single utterances and $\mathscr{D} = \{(x_i, y_i)\}$ denotes a set of pairs consisting of utterance x and its response y_i. Typically, the number of such pairs is small, i.e. $|\mathscr{D}| \ll |\mathscr{H}|$.

Output. Our goal is to construct a large number of utterance–response pairs $\widetilde{\mathscr{D}} \subseteq \mathscr{H} \times \mathscr{H}$ by matching two utterances from the set of single utterances $\mathscr{H}$ based on statistical information obtained from $\mathscr{D}$. Ideally, $|\widetilde{\mathscr{D}}| \gg |\mathscr{D}|$ holds.

4.2.2 Difficulties in Dialogue Data

There are at least two difficulties in data crea for thetion response generation task. The first is data availability. For example, in NMT literature, millions of high-quality parallel translation pairs are freely available as training data for the WMT[1]. Thus, we can train a strong NMT model and generate relatively high-quality data by the back-translation technique. On the other hand, in the dialogue generation task, the most extensive high-quality corpus only offers 30k utterance–response pairs, which renders acquiring such a strong model extremely difficult. Here, we assume high-quality corpus as fully consecutive and meaningful dialogues constructed by human intentionally, such as (Krause et al., 2017; Li et al., 2017b; Zhang et al., 2018). The second is the ambiguity of the task requirement. The dialogue generation task has a one-to-many nature; given a source utterance, multiple responses with diverse meanings can be regarded as appropriate responses. For example, given a source utterance expressing a request (e.g, "*Can you ...?*"), the system can output a response accept, decline, question (e.g., "*Yes*", "*No*", "*Why ...?*") to the order, or some other random responses.

[1] http://www.statmt.org/wmt19/

4.3 Methodology: Pairing Single Utterances

4.3.1 Key Idea

We assume that we have a sufficiently large set of utterances $\mathscr{H}$. Moreover, owing to the one-to-many nature, we expect that a large number of utterances in the same utterance set $\mathscr{H}$ will be acceptable responses to a randomly selected utterance even though they do not have a direct correlation. According to this expectation, we develop a three-step method that automatically extracts and connect two individual utterances as an utterance–response pair (Figure 4.1). Namely, we (1) obtain **dialogue templates** from $\mathscr{D}$, (2) extract candidate utterance–response pairs from $\mathscr{H}$ by applying the obtained templates, and (3) filter out almost all the unreliable candidate pairs to select reliable and possibly high-quality utterance–response pairs.

4.3.2 Step 1: Template Generation

We obtain a typical phrase or word pairs as **dialogue templates** $\bar{\mathscr{P}}$. For example, we extract the typical phrase pair ("*can i check*", "*sure, i guess*") from the utterance–response pair ("*Can I check my messages?*", "*Sure, I guess. The phone is in the back.*"). First, we obtain phrase pairs $\mathscr{P} = \{(f,e)\}$ by using a word/phrase aligner. f and e are extracted from the set of paired sentences $\{(x_i, y_i)\}$, that is, $f \in x$ and $e \in y$. Next, we extract *preferable* phrase pairs $\bar{\mathscr{P}}$ from $\mathscr{P}$ that match the following conditions:

$$\text{NPMI}(f,e) \geq \alpha,\, c(f) > \beta,\, c(e) > \beta, \tag{4.1}$$

where α and β are hyperparameters and $\mathrm{c}(\cdot)$ denotes the frequency of a phrase in $\mathscr{D}$. By imposing the first condition, we extract only strongly co-occurring phrase pairs. $\text{NPMI}(\cdot,\cdot)$ denotes the normalized pointwise mutual information (Bouma, 2009) (NPMI), which is a PMI variant that computes the co-occurrence strength and is not as adversely affected by low-frequency pairs as PMI. The second and third conditions are also imposed to prevent high values from being mistakenly assigned to low-frequency pairs. As a result, we retrieve strongly co-occurring phrase pairs $\bar{\mathscr{P}} \subseteq \mathscr{P}$ from the high-quality seed corpus $\mathscr{D}$.

4.3.3 Step 2: Candidate Pair Creation

We prepare utterance–response pair candidates $\widetilde{\mathscr{D}}_{\text{cand}} \subseteq \mathscr{H} \times \mathscr{H}$ using the dialogue templates $\bar{\mathscr{P}}$. Specifically, we randomly sample utterance candidates that contain the template f and response candidates that contain the corresponding template e of f from the single utterance corpus $\mathscr{H}$. Then, we make pairs of utterance and response candidates using all possible combinations, and obtain utterance–response pair candidates as $\widetilde{\mathscr{D}}_{\text{cand}}$.

4.3.4 Step 3: Candidate Pair Selection

We select the candidate pairs in $\widetilde{\mathscr{D}}_{\text{cand}}$ that connect plausibly and naturally as new pair data $\widetilde{\mathscr{D}} \subseteq \widetilde{\mathscr{D}}_{\text{cand}}$. We define the score $\text{Assoc}_{\text{s}}(x,y)$, which is used to evaluate the connection between utterances x and responses y in $\widetilde{\mathscr{D}}_{\text{cand}}$, by

$$\text{Assoc}_{\text{s}}(x,y) := \sum_{(f,e)\in(x,y)} \text{Assoc}_{\text{p}}(f,e;x,y). \tag{4.2}$$

This function gives a high score to a sentence pair that contains many strongly connected phrase pairs. The strength of connection of a phrase pair (f,e) in a sentence pair (x,y) is computed as

$$\begin{aligned}&\text{Assoc}_{\text{p}}(f,e;x,y) := \\ &\Big(\underbrace{\lambda\cdot\text{NPMI}(f,e)}_{\textbf{(i)}} + \underbrace{(1-\lambda)\frac{|f|+|e|}{Z}}_{\textbf{(ii)}}\Big)\cdot\underbrace{\Big(\frac{|f|}{|x|}+\frac{|e|}{|y|}\Big)}_{\textbf{(iii)}},\end{aligned} \tag{4.3}$$

where $|\cdot|$ is the word length, $\lambda \in [0,1]$ is a hyperparameter, and $Z := \max_{(f,e)\in\mathscr{P}}\{|f|+|e|\}$ is a normalizing constant. An intuitive explanation of the above equations is as follows: (**i**) If a phrase pair (f,e) has strong co-occurrence a high, the association strength of (x,y) including (f,e) might also be high. (**ii**) If a phrase f or e contains a lot of words, such phrase has rich content and information. We assume a phrase pair contains such long phrase has a strong signal to the association. For example, ("*if you think it be*", "*i think*") gives a strong signal rather than ("*you*", "*i*") to the sentence pair include this. (**iii**) If a phrase f or e occupies almost the entire sentence x or y, (f,e) is a strong indicator of the association of (x,y).

4.4 Experiment

4.4.1 Datasets

OpenSubtitles (Lison et al., 2018) is a large corpus of movie subtitles (English data containing roughly 441M lines) that is freely available and has been used as training data in many studies on data-driven dialogue models. The corpus does not contain speaker annotations; thus, we cannot detect the conversation sequence correctly. In our experiments, we sampled 20M utterances with more than five and less than 25 words and used them as the single utterance set $\mathscr{H}$. We also prepared $\widetilde{\mathscr{H}}$ from OpenSubtitles by treating consecutive utterances as utterance–response pairs and sampling 20M pairs.

Although the Cornell Movie Dialogue Corpus (CMDC) (Danescu-Niculescu-Mizil and Lee, 2011) is a much smaller corpus than OpenSubtitles, it has accurate speaker annotations. Hence, we can detect conversation sequences accurately and, thus, we can obtain the less noisy dialogue data. In our experiment, we used a preprocessed version of the CMDC devised by Baheti et al. (2018), which is the evaluation set used in their work and contains 31,487 utterance–response pairs. We sampled 30k pairs from the CMDC to use as our seed corpus $\mathscr{D}$. We also samples 100 pairs from the CMDC as a test set for our generation task.

4.4.2 Experimental Setup

Template generation. To obtain dialogue templates, we used GIZA++ (Brown et al., 1993; Och and Ney, 2003), IBM Model-based statistical machine translation tool, to learn phrase and word alignments on $\mathscr{D}$. Consequently, we obtained 2,094,136 phrase alignments and 178,099 word alignments. We eventually extracted 29,605 alignments as dialogue templates using several selection processes as discussed in Section 4.3.4. We set the hyperparameters $\alpha = 0.8$ and $\beta = 1$ or 3 if (f, e) is a phrase or word pair, respectively.

Candidate pair creation. We obtained up to 400 candidate pairs per template. Consequently, $|\widetilde{\mathscr{D}}_{\mathrm{cand}}|$ was 3,514,296.

Candidate pair selection. To score each candidate, we first tuned the weighting parameter λ of the scoring function (Equation 4.2 in Section 4.3.4) on the development set[2]. We scored all pairs in the development set using $\mathrm{Assoc}_{\mathrm{S}}(x, y)$ and then calculated the probability of correct pairs scoring higher than incorrect pairs in the case of ROC-AUC scoring. Figure 4.2 shows the relationship between λ and ROC-AUC score measured on our development set. We set $\lambda = 0.8$ which maximizes the ROC-AUC score (Fig.4.2).

We scored all the candidate pairs $(x, y) \in \widetilde{\mathscr{D}}_{\mathrm{cand}}$ by $\mathrm{Assoc}_{\mathrm{S}}(x, y)$ and ranked them. Figure 4.3 shows the distribution of $\mathrm{Assoc}_{\mathrm{S}}(x, y)$ scores on $\widetilde{\mathscr{D}}_{\mathrm{cand}}$. Then, we selected the pairs ranked in the top 15% ($\mathrm{Assoc}_{\mathrm{S}} > 1.33$) of the full candidates as new utterance–response pairs $\widetilde{\mathscr{D}}$. Ultimately, our new pair data $|\widetilde{\mathscr{D}}|$ was 379,215.

[2] We prepared 200 pairs; 100 correct pairs sampled from CMDC, and 100 pseudo incorrect pairs which ware sampled single utterances from $\mathscr{H}$ and randomly paired. It is fully disjoint from $\mathscr{H}$ and $\mathscr{D}$.

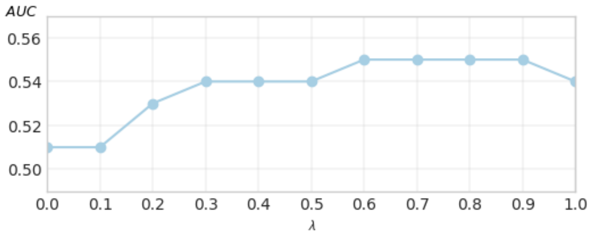


Fig. 4.2 Relationship between λ and ROC-AUC score.

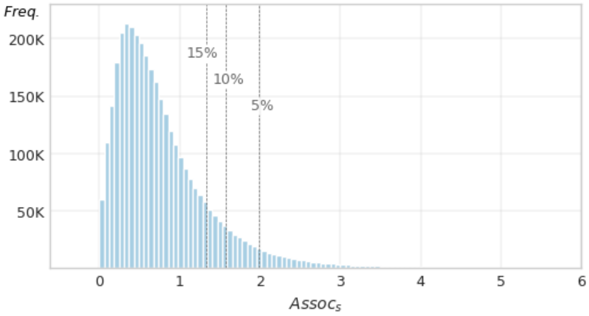


Fig. 4.3 Distribution of $\text{Assoc}_{\text{S}}(x,y)$ on $\widetilde{\mathscr{D}}_{\text{cand}}$.

4.5 Results

4.5.1 Paired Data Construction

Table 4.1 shows actual samples of pair data created with $\text{Assoc}_{\text{S}}(x,y)$. In total, the obtained utterance–response pairs are plausible. In the pair on line 4, for example, the utterance and response are naturally connected, where the pair contains a dialogue template describing a typical situation of someone thinking about whether to agree (e.g., "*sure, i guess*") to a request (e.g., "*can i check*"). Assoc_{S} for this pair was high. In the pair on line 8, however, the connection between the utterance and the response is unnatural despite them containing the same dialogue template as line 4 and Assoc_{S} for this pair was low. There is a tendency that the topic or situation of the conversation is shared between the utterance and the response as a feature of the created pairs. This indicates that our method of connecting each utterance by using selected templates based on some statistics is effective for obtaining naturally paired data. Furthermore, the results show that Assoc_{S} reasonably represented the naturalness of the connection between the utterance and the response, and worked well as a filter.

Table 4.1 Samples of pairs accepted as $\widetilde{\mathscr{D}}$ (lines 1–8) and rejected (lines 9-10). **Bold** indicates dialogue templates.

	Utterance	Response	$\mathrm{Assoc}_\mathrm{s}$
1 :	Glad **to see you, son**.	Good **to see you, dad**.	2.98
2 :	I want to thank you, sir, for giving **me the opportunity to work**.	And **you're welcome** to stay as long as you like.	2.35
3 :	So, raise your hand if you think it was **a Russian** water tentacle.	Uh, it was **Italian**, I think.	1.99
4 :	**Can I check** your bedroom?	Uh, **sure, I guess**, if that's what you want.	1.77
5 :	Now **drop the goddamn gun**.	Where did you **drop it**?	1.75
6 :	I had to **ask him** some questions.	You can ask **him yourself**.	1.62
7 :	Little girl, **are you alone?**	I felt like **I was alone** in the world.	1.58
8 :	**How could I help** you?	You got to start **by trusting** me.	1.54
9 :	You are the world's **worst driver**.	Get in the **passenger** seat.	1.01
10 :	One more question, when **can I check** out?	**Sure, I guess** I could step up my game.	0.35

Table 4.2 Evaluation results of generated responses.

Training data	# of pairs	len	distinct-n		BLEU		
			$n=1$	$n=2$	Precision	Recall	F-value
$\widetilde{\mathscr{H}}$	20,000,000	8.29	229/0.030	**710/0.106**	8.4	7.3	7.8
$\mathscr{D}$	30,000	5.02	83/0.018	176/0.048	**10.5**	5.0	6.8
$\widetilde{\mathscr{D}}$	379,215	**8.90**	**272/0.033**	844/0.103	9.4	**8.5**	**8.9**
reference	921	10.34	1801/0.189	5786/0.608	-	-	-

4.5.2 Extrinsic Evaluation: Response Generation

We also evaluated the obtained utterance–response pairs in terms of their usefulness as a training corpus of neural methods. For this purpose, we trained an LSTM-based sequence-to-sequence model using the `fairseq` toolkit[3]. We prepared three models that were trained on $\widetilde{\mathscr{H}}$, $\mathscr{D}$, and $\widetilde{\mathscr{D}}$.

Table 4.2 shows the evaluation results of the generated responses. We give results in terms of the metrics *distinct*-{1,2}[4] (*dist*), the average response length in words ($\mathrm{len}_\mathrm{avg}$), and the BLEU-1 precision, recall, and F-score (B_p, B_r, and B_F). Here, B_r is obtained by replacing the reference and system output, and B_F is calculated from them.We additionally used B_r

[3] https://github.com/pytorch/fairseq with the default configuration of the '`-arch LSTM`' option.

[4] Following (Baheti et al., 2018; Li et al., 2016a; Xing et al., 2017). The scores are the ratio of types to tokens for {1,2}-grams and indicate the diversity of the responses.

Table 4.3 Samples of input utterances and generated responses.

	Utterance		Response
1 :	**Eve, child! How nice of you to come.**	$\widetilde{\mathscr{H}}$:	I'm sorry, but I'm not going to be here.
		$\mathscr{D}$:	I'm sorry.
		$\widetilde{\mathscr{D}}$:	Oh, it's nice to see you.
2 :	**That sounds good.**	$\widetilde{\mathscr{H}}$:	I don't know what to say.
		$\mathscr{D}$:	I'm going to see you.
		$\widetilde{\mathscr{D}}$:	I'm glad you think so.
3 :	**You coming back?**	$\widetilde{\mathscr{H}}$:	What are you doing here?
		$\mathscr{D}$:	Yeah.
		$\widetilde{\mathscr{D}}$:	I'll be with you in a minute.

and $\mathrm{B_F}$ since the standard BLEU-1 ($\mathrm{B_p}$) is a precision-based evaluation and, thus, a shorter generation tends to misleadingly give a better result. Our $\widetilde{\mathscr{D}}$ outperformed other data in terms of $\mathrm{B_F}$. This indicates that the model trained on our $\widetilde{\mathscr{D}}$ tends to generate better responses with quality closer to the reference data than the model trained on other data. We also observed that $\widetilde{\mathscr{D}}$ achieved a *dist* comparable to $\mathscr{H}$. Moreover, $\mathrm{len_{avg}}$ of $\widetilde{\mathscr{D}}$ was 8.90, which was the closest to the reference of 10.34. Incidentally, $\mathscr{D}$ had a rather low $\mathrm{len_{avg}}$, which was less than half of the reference. This is the main reason why $\mathscr{D}$ had a relatively high $\mathrm{B_p}$ but a much lower $\mathrm{B_r}$. In summary, $\widetilde{\mathscr{D}}$ constructed by our method obtained consistently better results for most of the metrics.

Table 4.3 shows examples of the responses to the given utterances generated by the model. We confirmed that $\widetilde{\mathscr{D}}$ mostly generated fluent and contextually reasonable responses. This indicates that our method can generate meaningful utterance–response pairs for training neural dialogue models.

4.6 Conclusion

In this chapter, we discussed an automatic dialogue data construction method in order to prepare the desirable resource for training deep neural network-based response generation models. We proposed a method that can automatically obtain potentially high-quality utterance–response pairs. Our basic strategy was to link two single utterances extracted from a large utterance set if such a pair is reliable and can potentially be a high-quality utterance–response pair as evaluated by certain statistics. We evaluated the set of obtained pairs qualitatively and experimentally demonstrated via a dialogue generation task that the proposed method obtaines useful pairs as a training corpus of neural methods. The series of discussions in this chapter aimed at establishing dialogue augmentation methodologies have provided new insights and directions for obtaining large-scale, high-quality dialogue data and, ultimately, for improving the performance of neural dialogue response generation.

Chapter 5

Corpus construction from crowdworkers imitating target attributes

With the advancements of dialogue response generation technology, in recent years, it has become relatively easy to build a dialogue system with good performance if we can obtain a large amount of high-quality dialogue data. These advances encourage real-world applications of dialogue systems in society. Social demand for dialogue systems has been increasing year by year. One of the most common possible problems with using dialogue systems in the real world is the lack of resources available as training data for building dialogue systems. When building a unique dialogue system that suits the various demands of the real world, one of the most common possible problems is the lack of resources as training data. If you want to develop a dialogue system and there is no training data for it, you have to start by creating resources. In the NLP field, one of the practical choices to construct new resources is crowdsourcing (Callison-Burch et al., 2015). Using crowdsourcing allows us to manually create large amounts of high-quality data in a relatively short amount of time and at a low cost. As crowdworkers, there are many workers with various attributes (e.g., age, nationality, gender) all over the world, while the number of workers with each of these attributes is unevenly distributed. Therefore, if you need to collect data from workers with specific attributes, it may happen that workers are not available in the required quantities.

In this chapter, we propose a methodology for efficiently collecting dialogue data using crowdsourcing under the lack of human resource. The proposed method allows collecting data from pseudo target over the original target by under our practical instructions. We empirically confirm the validity of our method. Furthermore, we newly construct dialogue corpus with our method and demonstrate building a dialogue system using the corpus.

As an example of the real-world application of dialogue systems, in this work, we focus on developing the system to address the elderly's social isolation. It is one of the biggest demands in today's society. For this purpose, several previous works have been developed dialogue systems focusing on making conversation with the elderly on behalf of humans to alleviate their loneliness (Lala et al., 2017; Sidner et al., 2013). However, in these cases, the fundamental problem of the elderly's lack of communication with others in society still remains. As one possible approach to solve this fundamental problem, it is considered that we apply dialogue response generation technology to a communication support system for human-to-human, i.e., the elderly and others, conversation. We consider a system that supports the elderly in their textual communication such as email, specifically, a system that provides the elderly with some reply candidates for an email. We believe that such a system will be useful in reducing the burden of writing replies for the elderly, and in promoting communication that is beneficial to the elderly, such as proactively informing their physical and mental health to others.

5.1 Societal implementation of Dialogue Systems

5.1.1 Issue of Social Isolation

As the global population ages, the isolation of the elderly from society is a particularly serious issue. According to a Japanese Cabinet Office survey, the proportion of the elderly (i.e., aged over 65) who live alone or as a couple is 56.9%, which is higher than those who live with their children (39%) (内閣府, 2017). In addition, a survey of people over 60 years of age on how often they interact with others, not only face-to-face but also via email and telephone, found that 7.5% of men and 4.9% of women who live alone talk with others less than once a week (内閣府, 2015).

In the field of gerontology and geriatrics medicine, the researchers have been studying the health condition of the elderly and the ideal form of care, and are focusing on the concept of Quality of Life (QOL). The QOL is a scale for evaluating the health condition of the elderly many-sided from the physical, mental, and social sides. Some studies reported that the care suitable for the elderly could be realized by evaluating not only a physical condition but also the feeling and situation of the elderly by grasping the health condition of the elderly by QOL. (Hellstrm and IR, 2001; Vaarama, 2009).

徳久 et al. (2019) analyzed how family members and the elderly communicate with each other via email in such a way as to express QOL information (Refer to as QOL expressive speech). According to their report, 85.7% (3,574utterances 3,064utterances) of elderly's responses to emails from family members were family-centered ones (e.g., Figure 5.1 (A)), and only 6.4% (3,574utterances 229utterances) were elderly-centered ones (Figure 5.1 (B)).

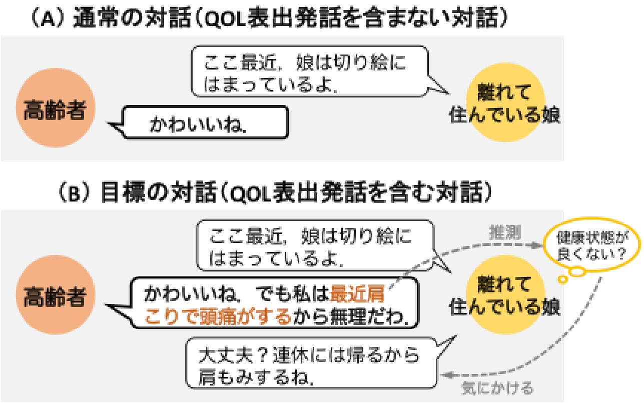

（A）通常の対話（QOL表出発話を含まない対話）
高齢者
ここ最近，娘は切り絵にはまっているよ．
離れて住んでいる娘
かわいいね．
（B）目標の対話（QOL表出発話を含む対話）
高齢者
ここ最近，娘は切り絵にはまっているよ．
推測
健康状態が良くない？
離れて住んでいる娘
かわいいね．でも私は最近肩こりで頭痛がするから無理だわ．
大丈夫？連休には帰るから肩もみするね．
気にかける

Fig. 5.1 (A) Normal conversation and (B) Conversation including utterance presenting QOL information.

5.1.2 Dialogue System as Conversational Partner

Under these social backgrounds, dialogue systems that can be used as a talking partner for the elderly have been actively developed. For example, Lala et al. (2017); Shitaoka et al. (2017) proposed attentive listening agents that listens to the elderly and Sidner et al. (2013) proposed always-on system that can make conversation and play table games with the elderly to reduce their isolation. Such dialogue systems may be able to alleviate the loneliness of the elderly by interacting with them instead of people, however, they cannot solve the fundamental problem of the elderly's lack of communication with others in society.

5.1.3 Human-to-human Communication Supporting System

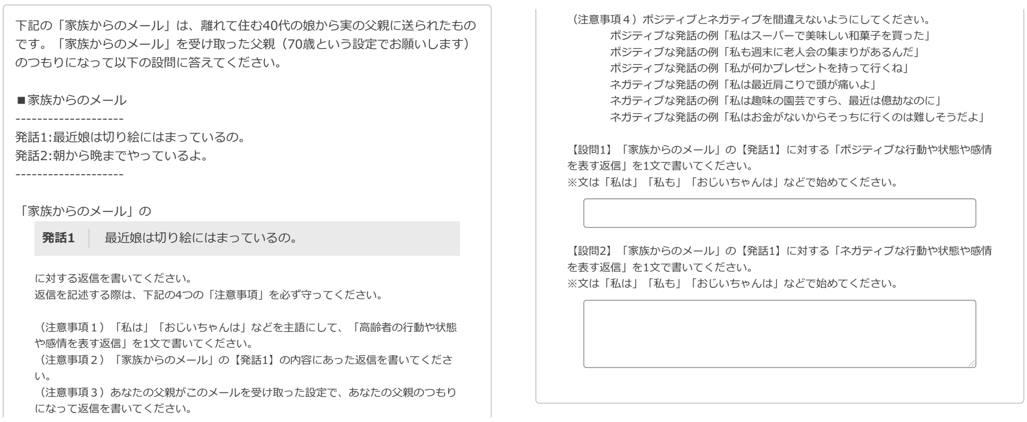
Our goal is to fundamentally solve the lack of communication with others in the elderly. For this goal, we consider developing a system that supports the elderly in their textual communications such as email. Specifically, in daily communication between the elderly and their family members who live apart from them, the system supports the generation of replies that naturally communicate the elderly's own QOL to their family members. We expect that the system will be useful in reducing the burden of writing replies for the elderly, and in promoting communication that is beneficial to the elderly, such as proactively informing their physical and mental health, i.e., QOL information, to others. Figure 5.1 shows the examples of the elderly's textual communications with their family members. Figure 5.1 (A) "かわいいね" is an appropriate response, but it does not convey the QOL of the elderly person to her daughter. On the other hand, in Figure 5.1 (B), "かわいいね．でも私は最近肩こりで頭痛がするから無理だわ" is an appropriate response and conveying the QOL information of the elderly. Then this reply elicited the daughter's utterance "大丈夫？連休には帰るから肩もみするね" that cares about elderly's health. We aim to create a reply assistance system that stimulates the elderly by presenting QOL-expressed utterances, such as those in Figure 5.1(B), to them as hints for their responses and induces QOL-expressed utterances from them that would not be described without the support of the system.

In this work, assuming to develop such a support system for the elderly, we try to generate the elderly's reply using the current dialogue response generation techniques. Specifically, we construct a new dialogue corpus consisting of the system's ideal input and output (i.e., the input is a family's message, and the output is an elderly's reply with their QOL information), and then generate the elderly's reply candidates with the constructed corpus. Through a series of experiments, we demonstrate that our proposed method can efficiently construct dialogue data even when human resources are limited, and that the dialogue response generation technologies in current NLP fields generate appropriate responses as the reply candidates that convey specific QOL information.

5.2 Methodology: Corpus Construction under Limited Human Resource

As a method for collecting large amounts of dialogue data, crowdsourcing has been employed in many studies (Callison-Burch et al., 2015; He et al., 2018; Zhang et al., 2018). However, depending on the features of the data to be collected, there may be restrictions on the available crowd workers, and it may not be possible to reserve a sufficient amount of human resources. In this paper, we discuss the data construction method with crowdsourcing by "imitators" and propose an effective task instruction for this purpose. The "imitator" refers to a worker who does not have a specific attribute but acts as if he or she does have it. To effectively collect data that strongly reflects the attributes to be imitated, we ask the crowdworkers who work as imitators through the task instructions to (i) provide concrete settings to imitate attributes and (ii) create data with the first-person subject.

Even for our purpose, it is not easy to collect a large amount of dialogue data from the elderly through crowdsourcing because there are not enough elderly crowdworkers. Therefore, we consider asking non-elderly crowdworkers to imitate the elderly and to create data by giving them the following instructions. As the (i) above, we instruct crowdworkers aged 40-59, imitators, to create data as if they were their fathers. We expect that the workers will imitate the elderly with more precision by imagining a concrete goal: their own fathers.



page 1/2

下記の「家族からのメール」は、離れて住む40代の娘から実の父親に送られたものです。「家族からのメール」を受け取った父親（70歳という設定でお願いします）のつもりになって以下の設問に答えてください。

■家族からのメール

発話1:最近娘は切り絵にはまっているの。
発話2:朝から晩までやっているよ。

「家族からのメール」の

発話1 最近娘は切り絵にはまっているの。

に対する返信を書いてください。
返信を記述する際は、下記の4つの「注意事項」を必ず守ってください。

（注意事項１）「私は」「おじいちゃんは」などを主語にして、「高齢者の行動や状態や感情を表す返信」を1文で書いてください。
（注意事項２）「家族からのメール」の【発話1】の内容にあった返信を書いてください。
（注意事項３）あなたの父親がこのメールを受け取った設定で、あなたの父親のつもりになって返信を書いてください。

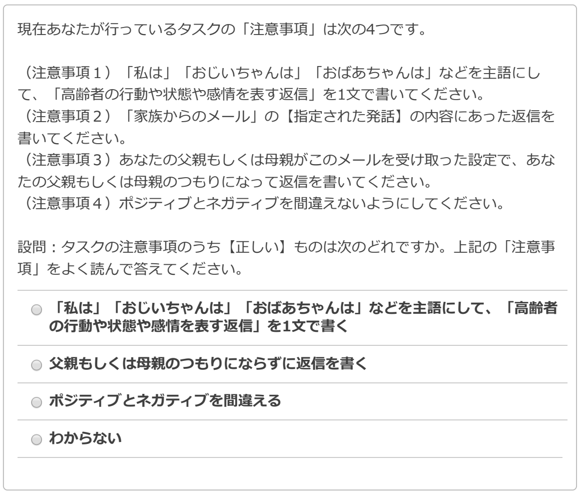
page 2/2

（注意事項４）ポジティブとネガティブを間違えないようにしてください。
ポジティブな発話の例「私はスーパーで美味しい和菓子を買った」
ポジティブな発話の例「私も週末に老人会の集まりがあるんだ」
ポジティブな発話の例「私が何かプレゼントを持って行くね」
ネガティブな発話の例「私は最近肩こりで頭が痛いよ」
ネガティブな発話の例「私は趣味の園芸ですら、最近は億劫なのに」
ネガティブな発話の例「私はお金がないからそっちに行くのは難しそうだよ」

【設問1】「家族からのメール」の【発話1】に対する「ポジティブな行動や状態や感情を表す返信」を1文で書いてください。
※文は「私は」「私も」「おじいちゃんは」などで始めてください。

【設問2】「家族からのメール」の【発話1】に対する「ネガティブな行動や状態や感情を表す返信」を1文で書いてください。
※文は「私は」「私も」「おじいちゃんは」などで始めてください。

Fig. 5.2 Task instructions for collecting utterances presenting elderly's QOL information.



現在あなたが行っているタスクの「注意事項」は次の4つです。

（注意事項１）「私は」「おじいちゃんは」「おばあちゃんは」などを主語にして、「高齢者の行動や状態や感情を表す返信」を1文で書いてください。
（注意事項２）「家族からのメール」の【指定された発話】の内容にあった返信を書いてください。
（注意事項３）あなたの父親もしくは母親がこのメールを受け取った設定で、あなたの父親もしくは母親のつもりになって返信を書いてください。
（注意事項４）ポジティブとネガティブを間違えないようにしてください。

設問：タスクの注意事項のうち【正しい】ものは次のどれですか。上記の「注意事項」をよく読んで答えてください。

○ **「私は」「おじいちゃんは」「おばあちゃんは」などを主語にして、「高齢者の行動や状態や感情を表す返信」を1文で書く**
○ **父親もしくは母親のつもりにならずに返信を書く**
○ **ポジティブとネガティブを間違える**
○ **わからない**

Fig. 5.3 Attention-check for collecting utterances presenting elderly's QOL information.

As the (ii) above, we instruct imitators to write replies describing the actions, states, and sentiments of the elderly, using "I am ..." or "Grandpa is ..." as the subject. 徳久 et al. (2019) reported that utterances with such actions and states are more likely to convey QOL information. As described in Section 5.1.1, the elderly's replies to emails from family members, in the case where there is no precise control, tend to mention family members. In other words, the replies that are mainly about the elderly themselves are hardly written. Under these situations, the instruction that forces the response's subject can be critical for effectively collecting utterances, including the elderly's QOL information. Eventually, we collect data from the imitator by these instructions shown in Figure 5.2 including the two above, and some attention checks (Figure 5.3). We believe that the proposed method is not

Table 5.1 emails from family members used on our preliminary experiments (in Japanese). Underlined indicates that sentences are targeted to reply.

#	All messages on email and reply target.
1:	最近娘は切り絵にはまっているの．朝から晩までやっているよ．
2:	やっと金曜日！明日は結婚記念日だから，娘がポテトを揚げるのを手伝ってくれて，みんなでケーキも作ったよ．
3:	天龍川で川下り．お宿はプレイルームもあるところでなかなか良かったよ
4:	久しぶりのうなぎ！美味しかった！
5:	2階に久々にプラレールの大作ができた！お母さんが買ってくれたトミカの道路も活躍しているよ．
6:	今日からラジオ体操が始まるよ．今年は子ども会の会長だから，毎朝子どもたちの前でお手本役をやらなきゃいけないんだ．頑張るね.

limited to the communication support for the elderly that is assumed in this study, but is expected to apply to all cases where there is no available data for a specific application.

5.3 Investigations on Data Construction with Imitators

5.3.1 Effectiveness of Instructions to Control the Subject of Utterance

To verify the proposed instructions are practical in collecting elderly subjective dialogue data from imitators, we comparably analyze the subject of the utterances in data collected with the instructions (ours) and without the instructions (徳久 et al., 2019).

Following the utterance collection experiment without the instructions by 徳久 et al. (2019), we asked crowdworkers aged 40-59 years to write utterances as replies to the six emails shown in Table 5.1 with the instructions shown in Figure 5.2, and then collected 4,717 pseudo elderly’ s utterances from 386 workers for analysis.[1] The collected utterances were assigned the subject tags, shown in table 5.2, by the two annotators. Figure 5.4 shows the ratio of subject tags for collected utterances by the two settings. Without the instruction, the percentage of utterances with the Elderly tag was the subject was 6.4% (229 out of 3,574 utterances). In contrast, with the instructions, this percentage increased to 89.2% (4,207 out of 4,717 utterances). This result demonstrated that our instructions allow us to effectively collect the elderly subjective utterances, including such as the elderly’s behaviors, statuses, or sentiments.

[1] We collected utterances from a total of 400 workers (200 men, 200 women) and excluded 18 workers (13 men, 5 women; 4.5% of the total) from our analysis because they wrote unethical contents.

Table 5.2 Definition of our subject tags.

Tag	Definition	Example (in Japanese)
Elderly	The subject of the act or state described in the utterance is elderly or elderly event.	でも最近肩がこるから私には無理だわ
Family	The subject of the act or state described in the utterance is family or family event.	かわいいね
FamEld	The subject of the act or state described in the utterance is both family and elderly.	今度一緒にやろう
NOTAG	The Dialogue Act of the utterance is Greeting, Auto-Positive, or Thanking.	お疲れ様，ありがとう
Other	The subject of the act or state described in the utterance is neither family nor elderly.	友達の娘さんも切り絵が上手よ

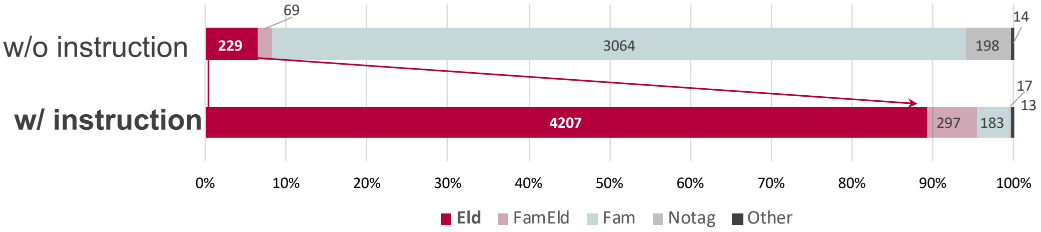


Fig. 5.4 Proportion of subject tags of collected utterances.

5.3.2 Elderly versus Imitator: Comparison of Collected Utterances under the Instructions

We confirm that the imitators can simulate the elderly’s utterances including their behaviors (actions), statuses, or sentiments, and create high-quality dialogue data with the instructions. Specifically, we demonstrate that (1) the dialogue data created by the imitators is indistinguishable from the data created by the elderly, and (2) there is no difference between them in the breakdown of contents of utterances. For our investigations, we prepared the two types of dialogue data created by the following two groups.

Group A: It consists of 10 workers who are ordinary elderly aged over 65. Table 5.3 shows the details of the workers. The workers created positive and negative responses to each of the 6 emails shown in 5.1 by writing on paper in their own houses. We collected 12 replies per worker; eventually, we obtained a total of 120 utterances.

Group B: It consists of 10 workers (3 men, 7 women) who are crowdworkers aged 40-59, i.e., imitators. We randomly sampled them from workers in Section 5.3.1 and used their created replies; eventually, we obtained a total of 120 utterances.

Table 5.3 Age and gender of Group A (ordinary elderly aged over 65).

	age 65-69	age 70-79	age 80-89	Total
Men	0	2	1	3
Women	3	2	2	7

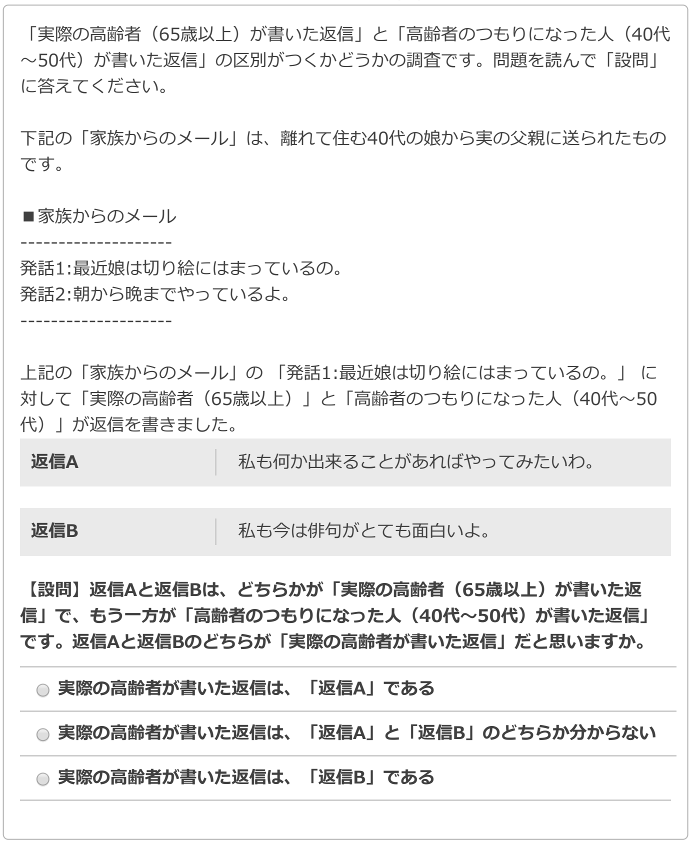
Table 5.4 Sample of replies created by Group A and Group B to email #1.

Replies created by Group A (Elderly aged over 65)
・私は吹矢を健康の為にやっているよ.
・私も切り絵は，小さい時大好きだったのよ.
・私も何か出来ることがあればやってみたいわ.
・私はかたが痛いので出来ないよ.
・私は最近病気が進んで入院中．切り絵どころではナイヨ.
・私は根気がないから長く続きそうにないわ.
Replies created by Group B（Crowdworkers aged 40-59 and who imitated elderly）
・私は，若いころ切り絵に没頭していたよ.
・私も切り絵が好きですよ.
・私も今は俳句がとても面白いよ.
・おじいちゃんは指先が思うように動かせないからできないな.
・私は腰が痛くて何もしたくないな.
・おじいちゃんはそんな細かい事はもう難しいかなあ.

Table 5.5 Sample of replies created by Group A and Group B to email #2.

Replies created by Group A (Elderly aged over 65)
・私も，あなたと一緒に結婚記念日には一緒にケーキを作った事を思い出したわ.
・私もシャンパンでも買って来よう.
・私は6月が結婚記念日なので今年はケーキでも作ろうかな.
・私も参加したかったよ.
・私はそちらに行けないからお祝いできなくてごめんね.
・私は老いて，皆を記念日に招んであげられないことが悲しい.
Replies created by Group B（Crowdworkers aged 40-59 and who imitated elderly）
・私も，久しぶりにケーキ作りしてみようかしら.
・私も一緒に食べたいなぁ.
・私もお祝いにお寿司を作ってあげるよ.
・私も一緒にお祝いできなくて残念です.
・私は体調が悪くて何もしてあげられないわ.
・おばあちゃんもプレゼントを送ってやりたいけど，お金が無くてごめんね.

The sample of replies created by Group A and B are listed in Table 5.4 and Table 5.5.



「実際の高齢者（65歳以上）が書いた返信」と「高齢者のつもりになった人（40代～50代）が書いた返信」の区別がつくかどうかの調査です。問題を読んで「設問」に答えてください。

下記の「家族からのメール」は、離れて住む40代の娘から実の父親に送られたものです。

■家族からのメール

発話1:最近娘は切り絵にはまっているの。
発話2:朝から晩までやっているよ。

上記の「家族からのメール」の 「発話1:最近娘は切り絵にはまっているの。」 に対して「実際の高齢者（65歳以上）」と「高齢者のつもりになった人（40代～50代）」が返信を書きました。

返信A	私も何か出来ることがあればやってみたいわ。
返信B	私も今は俳句がとても面白いよ。

【設問】返信Aと返信Bは、どちらかが「実際の高齢者（65歳以上）が書いた返信」で、もう一方が「高齢者のつもりになった人（40代～50代）が書いた返信」です。返信Aと返信Bのどちらが「実際の高齢者が書いた返信」だと思いますか。

○ **実際の高齢者が書いた返信は、「返信A」である**

○ **実際の高齢者が書いた返信は、「返信A」と「返信B」のどちらか分からない**

○ **実際の高齢者が書いた返信は、「返信B」である**

Fig. 5.5 Task instruction of paired comparison experiment. It was randomly determined which of Reply A and Reply B corresponded to the actual elderly's utterance. In this example, reply A is from Group A (elderly), and reply B is from Group B (imitators).

First, we confirmed that the dialogue data created by the imitators is indistinguishable from the data created by the elderly. We paired the collected replies from Group A and Group B one-on-one, and then we give these to the evaluators and asked to judge which utterance was written by the elderly, i.e., Grop A (Figure 5.5). Here, the evaluators were allowed to answer with "unsure." Figure 5.6 shows the result of the paired comparison. The "correct" indicates the number of evaluators who could select which was the elderly's reply, the "incorrect" indicates the number of evaluators who could not. The percentage of responses in which the evaluators succeeded in selecting the elderly's utterances correctly was 39.2%, which was not significantly different from the percentage of responses in which they could not succeed. Besides, the answer "unsure" occupied 24% of the total. These results show that it is difficult to distinguish the imitators' utterances under the instructions from the elderly's utterances. It indicates that we can collect the dialogue data from imitators in place of the elderly.

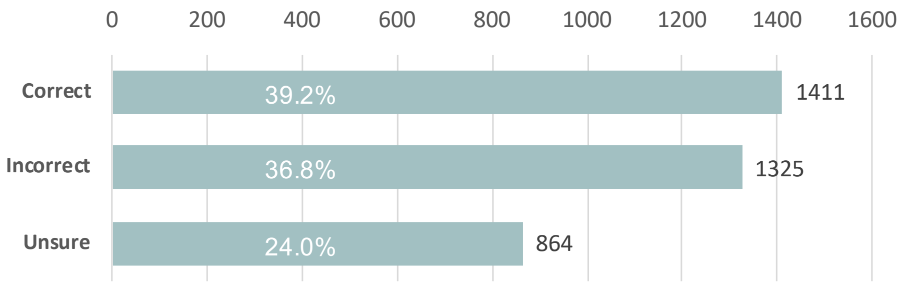


Fig. 5.6 Result of identification by paired comparison between actual elderly replies and imitator's replies.

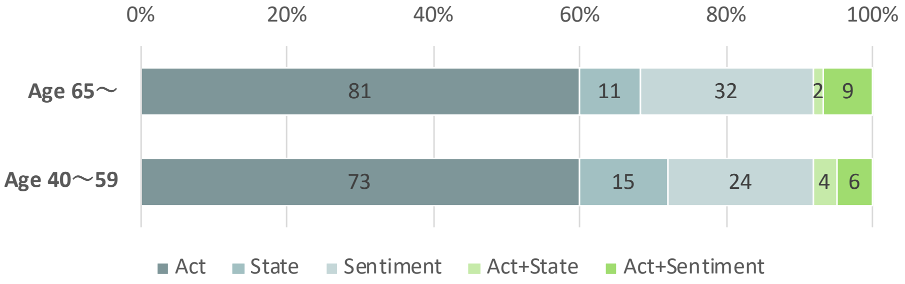


Fig. 5.7 Comparison of the contents of replies collected from actual elderly and imitators.

Next, we confirm that there is no difference between the dialogue data created by the elderly and imitators in the breakdown of contents of utterances. For the utterances collected from each group, we divided them into five fields based on the contents of what the utterance includes, i.e., the elderly's actions, states, and sentiments. Figure 5.7 shows the propotion of the contents. It demonstrates that there were no significant differences in the proportions of the contents of utterances between the two dialogue data.

5.3.3 Summery

To conclude these investigations, we confirmed that the proposed method allows us to collect dialogue data with comparable quality and properties as original from imitators.

5.4 Construction of Japanese QOL-labeled Corpus

Following the method described in Section 5.3, we constructed a new corpus, named Japanese QOL-labeled Corpus. Specifically, we took the following procedure.

Table 5.6 Definition of QOL labels (in Japanese).

#	QOL label	Definition
1	生活活動力 (positive)	高齢者が身の回りのことをひとりでできるか（移動，買い物，洗濯，食事の支度など）－できる
2	生活活動力 (negative)	高齢者が身の回りのことをひとりでできるか（移動，買い物，洗濯，食事の支度など）－できない
3	健康満足感 (positive)	高齢者の健康状態－健康状態が良い
4	健康満足感 (negative)	高齢者の健康状態－健康状態が悪い
5	人的サポート満足感 (positive)	高齢者の人付き合い－人付き合いがある
6	人的サポート満足感 (negative)	高齢者の人付き合い－人付き合いがない
7	経済的ゆとり満足感 (positive)	高齢者の金銭的な余裕－余裕がある
8	経済的ゆとり満足感 (negative)	高齢者の金銭的な余裕－余裕がない
9	精神的健康 (positive)	高齢者のさみしさ・無力さ－寂しくない・無力と感じない
10	精神的健康 (negative)	高齢者のさみしさ・無力さ－寂しい・無力と感じる
11	精神的活力 (positive)	高齢者の趣味や生きがい－趣味や生きがいがある
12	精神的活力 (negative)	高齢者の趣味や生きがい－趣味や生きがいがない

Step 1. Collecting family members' messages: We collected family members' messages from crowdworkers aged 40 to 59 who lived apart from their parents. They were instructed to write their messages in 2 or 3 sentences as if they were writing an email to their parents. Our preliminary experiment revealed a bias in the topics of the collected messages, depending on when we did the crowdsourcing. Therefore, to avoid this bias and collect conversations on various topics across seasons and situations, we specified the topic of email to workers in advance, e.g., children's school arts festival, daily chores, cherry blossom viewing, or autumn leaves hunting. The workers wrote messages following the given topics.

Step 2. Collecting elderly's reply: We collected the elderly's replies to family member's messages obtained by Step 1 from imitators who are crowdworkers aged 40 to 59 following the method described in Section 5.3. The instruction to the imitators and task format are shown in Figure 5.2. Since our preliminary experiment revealed that workers tend to write only positive replies without this instruction, in order to balance the polarity of the responses we collect, we asked the crowdworkers to write both positive and negative replies to each message. We filtered out unqualified workers using attention-checks (one example are shown in Figure 5.3).

Step 3. Annotating QOL-labels: Three crowdworkers manually assigned the plausible QOL labels shown in Table 5.6 to each elderly's reply collected in Step 2. We defined these QOL labels based on 太田 et al. (2001)'s six kinds of QOL (i.e., life

Table 5.7 Sample of our Japanese QOL-labeled corpus.

#1	Family's message :	子どもが野菜が苦手だから小さくみじん切りにしてチャーハンに入れて食べさせているよ.
	Elderly's reply :	私も料理をしているぞ.
	QOL label :	生活活動力 (positive), 健康満足感 (positive)
#2	Family's message :	子どもたちもお代わりをして，あっという間になくなりました.
	Elderly's reply :	私は最近食欲がないよ.
	QOL label :	健康満足感 (negative), 精神的健康 (negative)
#3	Family's message :	発表は昨日学部掲示板に張られていて，合格したの.
	Elderly's reply :	合格祝いにおじいちゃんが何か買ってやるか.
	QOL label :	経済的ゆとり満足感 (positive)
#4	Family's message :	サークルのみんなと相談してるんだけどなかなか意見がまとまらないんだ.
	Elderly's reply :	私は最近人と話すのも億劫で嫌だな.
	QOL label :	人的サポート満足感 (negative)
#5	Family's message :	日ごと雨だと外にも出られないね.
	Elderly's reply :	私は毎日畑仕事をしているよ.
	QOL label :	生活活動力 (positive), 精神的活力 (positive)

activity capacity, health satisfaction, human support satisfaction, economical comfort satisfaction, mental health, and mental vitality). We allowed workers to annotate multiple types of QOL labels to a single reply.

Step 4. Selecting QOL-labels: Among the QOL labels assigned to the elderly's replies in Step 3, we employ only those with the agreement of at least 2 of the workers. We do not include in the dialogue corpus the responses for which no label was assigned. Eventually, our QOL-labeled dialogue corpus was constructed with $81,228$ QOL labels for $52,079$ replies (in average, 1.6 labels per one reply).

Table 5.7 shows examples of dialogue data included in the Japanese QOL-labeled Corpus. We obtained "Family's message" in Step 1, "elderly's reply" in Step 2, and "QOL label" in Steps 3 and 4, respectively.

5.5 Experiments on Reply Candidates Generation

In this section, we attempt to build a model to generate utterances containing QOL information using the QOL-labeled dialogue corpus created in chapter 5.4. We assume to use these utterances as candidate replies in a reply assistance system for the elderly. Therefore, these utterances should be appropriate utterances that the elderly would naturally want to select as

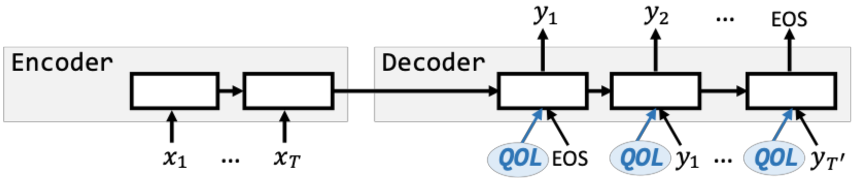


Fig. 5.8 Conditional response generation using QOL-labels. The QOL information (colored blue) is input to the decoder at each step.

their response and should allow others to read certain QOL information from them. We will build models with several response generation techniques and our corpus, and then empirically confirm that they generate plausible responses that convey certain QOL information for a given message.

5.5.1 Generation-based Response Generation Model

As a high-performance machine learning-based response generation technique, sequence to sequence (seq2seq) model (Sutskever et al., 2014; Vinyals and Le, 2015b) has been highlighted in recent years. Many studies reported that the seq2seq models generate plausible and fluent responses. For example, Li et al. (2016b) proposed a method with a seq2seq model to consistently generate personalized responses by inputting distributed embeddings of speaker information to its decoder during training and generation. Let $X = (x_1, \dots, x_T)$ denote input message and $Y = (y_1, \dots, y_{T'})$ denote output response. Motivated by the Li et al. (2016b)'s method, we generate utterances to include particular QOL information by maximizing the following prediction probability:

$$p(Y|X,q) = \prod_{t=1}^{T'} p(y_t|X, y_{<t}, q), \tag{5.1}$$

where, T, T' are word length of input utterance and output response, respectively. The decoder predicts the next word y_t by using the previous output $y_{<t}$ and the given QOL label q at each timestep. We refer to this model as S2S model. Figure 5.8 shows the overview of S2S model.

5.5.2 Retrieval-based Response Generation Model

In addition to generation-based methods, a retrieval-based method is one of the other possible choices to build response generation models (Isbell et al., 2000; Ritter et al., 2011; Sordoni et al., 2015). Retrieval-based response generation can be formulated as the problem of selecting the most suitable response $\widetilde{Y}$ from the instance-database $\mathscr{D}$ for a given utterance X. In this work, we find the pair $(\widetilde{X},\widetilde{Y}) \in \mathscr{D}$ with the highest similarity between input message X and $\widetilde{X}$, and then output $\widetilde{Y}$ as the response.

$$\widetilde{Y} = \underset{(\widetilde{X},\widetilde{Y})\in\mathscr{D}_q}{\arg\max}\,\mathrm{sim}\big(v(X), v(\widetilde{X})\big). \tag{5.2}$$

Where, $\mathscr{D}_q$ denote instance-database that consists of only dialogue data with specific QOL-label q. We prepared 12 types of $\mathscr{D}_q$ for 12-types of QOL labels (5.6). $v(X)$ is sentence vector of X. The function $\mathrm{sim}(\cdot,\cdot)$ compute the similarity between two sentences vectors; we used the cosine similarity. To obtain a vector representation of a sentence, we used two methods: the one is word2vec (Mikolov et al., 2013a), one of the most standard methods to obtain word embeddings in today's NLP research. The other is ELMo (Peters et al., 2018), one of the latest methods effective to obtain contextualized word embeddings. We refer to these models as `W2V` model and `ELMo` model, respectively.

5.5.3 Experimental Setups

We verify that the models built with our QOL-labeled corpus generate plausible responses that convey certain QOL information for a given message, through the response generation experiments and manual evaluation for the results.

Dataset Preparation

We used 90% of the QOL-labeled corpus as our training set and 10% for our test set.

Setup for `S2S` Model

We projected the QOL label onto a 12-dimensionalul binary vector as input to the decoder. The seq2seq encoder and decoder were 2-layer LSTMs (Hochreiter and Schmidhuber, 1997) with 512-dimensional hidden layers and 512-dimensional embedding layers. The number of maximum training epochs was 100. We used Adam (Kingma and Ba, 2015) for parameter optimization.

Setup for `W2V` Model, `ELMo` Model

For the `W2V` model, we used pre-trained 300-dimensional Japanese word2vec embeddings[2]. For the `ELMo` model, we obtained 512-dimensional embeddings by training the ELMo on

[2] https://github.com/Kyubyong/wordvectors

Table 5.8 Number of instances in the database with each QOL label.

#	QOL label q	$\lvert\mathscr{D}_q\rvert$
1	生活活動力 (positive)	14,002
2	生活活動力 (negative)	3,970
3	健康満足感 (positive)	11,316
4	健康満足感 (negative)	15,228
5	人的サポート満足感 (positive)	3,419
6	人的サポート満足感 (negative)	1,761
7	経済的ゆとり満足感 (positive)	5,258
8	経済的ゆとり満足感 (negative)	3,064
9	精神的健康 (positive)	4,593
10	精神的健康 (negative)	5,456
11	精神的活力 (positive)	6,788
12	精神的活力 (negative)	2,335

Japanese Wikipedia data. In both models, we created sentence embeddings by averaging all words' embeddings included in the sentence. We prepared the instance-databases D_q from our training set. Table 5.8 shows the size of each database corresponding to each kind of QOL label. The pair $(\widetilde{X},\widetilde{Y})$ such that multiple QOL labels q_i, q_j are assigned is included in the all database corresponding to assigned labels: $(\widetilde{X},\widetilde{Y}) \in \mathscr{D}_{q_i}, (\widetilde{X},\widetilde{Y}) \in \mathscr{D}_{q_j}$.

Evaluation datasets Preparation

To evaluate the generated responses, we prepared the two types of evaluation datasets from our test set. The one is $\mathscr{S}_{10-l12}$ that includes 120 pairs consisting of (utterance, QOL label), which are created by randomly sampling 10 utterances from our test set and then attaching 12 different QOL labels to each utterance. We use $\mathscr{S}_{10-l12}$ to evaluate the models in terms of whether they can generate various responses that convey different QOL information for a single utterance. The other is $\mathscr{S}_{100}$ that includes 100 triples consisting of (utterance, response, QOL label), which are created by randomly sampling 100 utterances from our test set. We use $\mathscr{S}_{100}$ to evaluate the models in terms of whether they can generate human-like responses while including certain QOL information in naturally possible situations.

Human Evaluation Settings

The quality of the generated responses for our evaluation datasets was manually evaluated. We asked native Japanese-speakers via Yahoo! crowdsourcing to evaluate the responses in terms of the following three points:

- Point (1): The response expresses/implies the specified QOL information

- Point (2): The response expresses/implies the polarity of specified QOL information
- Point (3): The response is plausible as a reply to a given message.

For all points, the evaluators were given a message and a generated response. For point (1), the evaluators were asked to answer the question *What QOL states could you read from the response?* by selecting the best and second-best ones from among 13 options, which consist of the 12 types of QOL states defined in Table 5.6 and "*Unsure*." For point (2), the evaluators were asked to answer the question *Which polarities of QOL states could you read from the response?* by selecting one from {*Positive*, *Negative*, *Unsure*}. For point (3), the evaluators were asked to answer the question *Are given utterances and responses plausible for dialogue?* by selecting one from {*Yes*, *No*, *Unsure*}. For each generated response, five evaluators answered each question.

5.5.4 Result and Analysis

Result of Human Evaluation

Table 5.9 shows the result of human evaluations. The (1)QOL on the table indicates a percentage of that the evaluators' answer for the above evaluation point (1) is the same as the QOL label that the model was specified to generate. Here, @1 is the percentage of agreement with the QOL label that evaluators selected as the best, and @2 is the agreement with their second-best. In $\mathscr{S}_{10-l12}$, that is under the setting to forcibly generate responses for conveying all 12 types of QOL even where are some contextual unnaturalness, only at most half of the cases succeeded in conveying the specified QOL. On the other hand, in $\mathscr{S}_{100}$, the models succeeded with the high percentages in conveying the specified QOL via generated responses. The (2)Pos/Neg on the table indicates a percentage of that evaluators' answer for the above evaluation point (2) is the same (✔) or not (✘) as the QOL label's polarity (i.e., positive or negative) that the model was specified to generate. All models succeeded with the high accuracies in conveying the polarity of specified QOL via generated responses. The (3)Plausibility on the table indicates a percentage of that generated response is plausible (✔) or not (✘) as a reply to a given message. The results show that the `S2S-QOL` model and `ELMo-QOL` model generated the plausible responses at over 60%. `W2V-QOL` generated a smaller percentage of plausible responses than the other two models.

Qualitative Analysis of Generated Responses

The samples of generated responses for our evaluation dataset $\mathscr{S}_{10-l12}$ are shown in Table 5.10. The right side of the table shows the results of the human evaluation for the

Table 5.9 Result of human evaluations for QOL communicability and response plausibility.

Evaluation set	Model	(1) QOL		(2) Pos/Neg		(3) Plausiblity	
		@1	@2	✔	✘	✔	✘
$\mathscr{S}_{10-l12}$	S2S	0.31 (188)	0.43 (256)	0.87 (524)	0.05 (27)	0.61 (364)	0.35 (210)
	W2V	0.38 (227)	0.51 (308)	0.93 (558)	0.03 (17)	0.51 (305)	0.42 (249)
	ELMo	0.37 (222)	0.50 (302)	0.94 (566)	0.03 (17)	0.60 (361)	0.35 (211)
$\mathscr{S}_{100}$	S2S	0.61 (304)	0.70 (352)	0.91 (456)	0.03 (13)	0.60 (302)	0.34 (169)
	W2V	0.62 (312)	0.72 (358)	0.90 (451)	0.04 (22)	0.61 (304)	0.33 (163)
	ELMo	0.62 (308)	0.70 (352)	0.92 (459)	0.03 (13)	0.68 (338)	0.26 (128)

generated response. These responses were generated for a single input message while conditioning it with 12 different QOL labels. First of all, we qualitatively confirmed that the S2S model generated sufficiently fluent Japanese. It means that the corpus we constructed was of sufficient size and quality to use for training a neural response generation model. Moreover, we confirmed that the conditional response generation methods using QOL labels generated various responses that convey the desired specific QOL for the same input. For example, for the same input, “着くとすぐに本を読んでいるよ,” the models generated responses regarding the purchase of a book when the QOL labels related to financial comfort were specified, while the models generated responses regarding one’s interest or hobby, e.g., reading books, when the QOL labels related to mental vitality were specified. Regarding polarity, the positive QOL labels tended to generate positive terms such as “今度,” “〜する,” or “〜してあげる,” in contrast, the negative QOL labels tended to generate negative terms such as “億劫,” “面倒,” or “無駄.”

Table 5.11 shows the generated responses by the models on our evaluation dataset $\mathscr{S}_{100}$. We qualitatively confirmed that the model generated fluent and plausible responses reflecting the given QOL label. These responses are desirable as candidate responses for the communication support system we assume to develop.

5.5.5 Summary

To summarize, in our response generation experiments, we have confirmed that both the generation-based response generation model and the retrieval-based response generation models generate appropriate responses that convey a specific QOL method. These results indicate that the corpus created by the proposed manually corpus construction method can be used to build dialogue systems with current NLP technologies.

Table 5.10 Samples of response generated by S2S,W2V,ELMo on the evaluation set $\mathscr{S}_{10-l12}$.

Input message: 着くとすぐに本を読んでいるよ				
QOL label / Output response	(1) QOL@1@2		(2)Pos/Neg	(3)Plausiblity
#1 **QOL label : 生活活動力 (positive)**				
S2S: 今度わしが見に行くよ	1/5	3/5	5/5	3/5
W2V: わしも新聞を読むのだけが楽しみだよ.	0/5	1/5	2/5	5/5
ELMo: わしも新聞を読むのだけが楽しみだよ.	0/5	0/5	1/5	4/5
#2 **QOL label : 生活活動力 (negative)**				
S2S: わしも見に行ってみたいが，今週では車に乗るとすぐに疲れちゃうから，もう乗るのが億劫になるよ	0/5	0/5	5/5	4/5
W2V: そうね．それはいいことだわ．お婆ちゃんは具合が悪いから絵本を読んでやりたいけどできないわごめんね.	1/5	1/5	5/5	5/5
ELMo: 雪はうれしいが，足が悪いから外に出られない	0/5	1/5	5/5	1/5
#3 **QOL label : 健康満足感 (positive)**				
S2S: 今度本を読んであげよう	0/5	0/5	4/5	3/5
W2V: お母さんは毎日ジョギングで鍛えてるわ.	3/5	4/5	5/5	3/5
ELMo: 私も雪でも散歩しているぞ	4/5	5/5	5/5	0/5
#4 **QOL label : 健康満足感 (negative)**				
S2S: 私は新聞を読むのも億劫だよ	1/5	2/5	5/5	5/5
W2V: 足腰が悪くて，七夕飾りどころじゃないのよ.	5/5	5/5	5/5	1/5
ELMo: 私は目が悪くなって本が読めないよ.	4/5	5/5	5/5	5/5
#5 **QOL label : 人的サポート満足感 (positive)**				
S2S: 今度教えてあげるね	2/5	4/5	2/5	1/5
W2V: 私もこの間お友達と映画を見に行ったわ	3/5	3/5	5/5	1/5
ELMo: 私もテニスして汗かいてるよ.	0/5	1/5	5/5	3/5
#6 **QOL label : 人的サポート満足感 (negative)**				
S2S: わしも見に来て欲しいんだがな	0/5	0/5	3/5	1/5
W2V: 私は誰にも会いたくない	4/5	4/5	5/5	2/5
ELMo: 私は一人だから寂しい	2/5	2/5	5/5	1/5
#7 **QOL label : 経済的ゆとり満足感 (positive)**				
S2S: 今度本を買ってそちらに行くよ	3/5	3/5	4/5	5/5
W2V: 今度本を買って持っていくよ.	2/5	3/5	5/5	4/5
ELMo: 今度本を買って持っていくよ.	4/5	4/5	5/5	5/5
#8 **QOL label : 経済的ゆとり満足感 (negative)**				
S2S: 今度本を買ってあげたいけど，私にはお金がないからそっちに行くのは難しいな	4/5	4/5	5/5	5/5
W2V: おばあちゃん今お金がないんだよ	5/5	5/5	5/5	0/5
ELMo: おばあちゃんあちこち痛いけど湿布買うお金無いわ.	1/5	4/5	5/5	0/5
#9 **QOL label : 精神的健康 (positive)**				
S2S: 今度私が教えてあげるよ	0/5	0/5	5/5	4/5
W2V: ワシも婆さんもワクワクしているよ.	2/5	3/5	5/5	1/5
ELMo: また私がプレゼントするよ.	0/5	0/5	5/5	4/5
#10 **QOL label : 精神的健康 (negative)**				
S2S: 私にも見せてもらいたい	0/5	0/5	5/5	5/5
W2V: 私は毎日に疲れてるわ.	1/5	3/5	5/5	1/5
ELMo: 俺が一緒に読んでやれたらいいんだけどな.	1/5	1/5	2/5	4/5
#11 **QOL label : 精神的活力 (positive)**				
S2S: 今度わしが見に行くよ	0/5	1/5	5/5	2/5
W2V: わしも新聞を読むのだけが楽しみだよ.	2/5	2/5	4/5	3/5
ELMo: わしも新聞を読むのだけが楽しみだよ.	2/5	2/5	3/5	5/5
#12 **QOL label : 精神的活力 (negative)**				
S2S: 私には興味がないな	2/5	3/5	4/5	5/5
W2V: 本にはわしは興味ないからなあ.	3/5	3/5	5/5	5/5
ELMo: 私は外に出かけることも少なくなったよ.	3/5	3/5	5/5	1/5

Table 5.11 Samples of response generated by `S2S`,`W2V`,`ELMo` on the evaluation set $\mathscr{S}_{100}$.

	Input (Message, QOL labels)		Output
#1	この前スーパーで並んでいてね	S2S :	私もよくスーパーマーケット行く
	QOL : 生活活動力 (positive)	W2V :	おじいちゃんもよくスーパーマーケット行く
		ELMo :	おじいちゃんもよくスーパーマーケット行く
#2	昨日の日曜日，小学校の運動会だったの	S2S :	私は趣味の散歩ですら最近は億劫なのに
	QOL : 健康満足感 (negative)	W2V :	私は入院中で動けなかったよ
		ELMo :	私は入院中で動けなかったよ
#3	行かせてやりたいけどちょっと無理だな	S2S :	私にはお金がないから助けてあげるのは難しいな
	QOL : 経済的ゆとり満足感 (negative)	W2V :	私はお金がないからそんなこと言ってこないでよ
		ELMo :	私はお金がないからそんなこと言ってこないでよ
#4	子どもの偏食がすごいのよ	S2S :	私は何でも美味しく感じるよ
	QOL : 健康満足感 (positive), 精神的健康 (positive)	W2V :	私は何でも食べられるけどね、歯も丈夫だし
		ELMo :	私は何でも食べられるけどね、歯も丈夫だし
#5	なので今日は子供の服を買いに行って来たよ	S2S :	私の家は散らかり放題だわ
	QOL : 生活活動力 (negative), 健康満足感 (negative)	W2V :	私は買い物に出かけるのが面倒で、もう何年もデパートには行っていないわ.
		ELMo :	私は買い物に出かけるのが面倒で、もう何年もデパートには行っていないわ.

5.6 Conclusion

In this chapter, we discussed a practical methodology for manually generating new dialogue data as training data for building a dialogue system. We proposed a method to collect pseudo-dialogue data from a large number of crowdworkers, i.e., imitators, who are instructed to imitate the behavior of the original targets. We demonstrated the validity of the proposed method by qualitatively confirmed that the data generated by the imitators was of the same quality as the original data. When we consider the social implementation of dialogue systems, one of the most possible confronting issues is the lack of dialogue data for system development. Our proposed method is advantageous for such cases. In this work, as one example of social implementation of dialogue systems, we focused on textual communication support for the elderly who are isolated in society, and then we considered developing a system that provides the elderly with reply candidates when they create replies to emails. Through a series of experiments assuming the development of a textual communication supporting system for the elderly, we demonstrated that our proposed method could efficiently construct dialogue data even when human resources are limited. Furthermore, we confirmed that the dialogue data we constructed was a scale and quality used as training data for current deep neural response generation models.

Chapter 6

Segregation of Word Vector to Style and Semantic Components

Analyzing and generating natural language texts requires the capturing of two important aspects of language: *what is said* and *how it is said*. In the literature, much more attention has been paid to studies on *what is said*. However, recently, capturing *how it is said*, such as stylistic variations, has also proven to be useful for natural language processing tasks such as classification, analysis, and generation (Niu and Carpuat, 2017; Pavlick and Tetreault, 2016; Wang et al., 2017).

In this chapter, we studies the stylistic variations of words in the context of the representation learning of words. The lack of subjective or objective definitions is a major difficulty in studying style (Xu, 2017). Previous attempts have been made to define a selected aspect of the notion of style (e.g., politeness) (Flekova et al., 2016; Mairesse and Walker, 2007; Niu et al., 2017; Pavlick and Nenkova, 2015; Preotiuc-Pietro et al., 2016; Sennrich et al., 2016a); however, it is not straightforward to create strict guidelines for identifying the stylistic profile of a given text. The systematic evaluations of style-sensitive word representations and the learning of style-sensitive word representations in a supervised manner are hampered by this. In addition, there is another trend of research forward controlling style-sensitive utterance generation without defining the style dimensions (Akama et al., 2017; Li et al., 2016b); however, this line of research considers style to be something associated with a given specific character, i.e., a persona, and does not aim to capture the stylistic variation space.

The contributions of this work are three-fold.

- We propose a novel architecture that acquires style-sensitive word vectors (Figure 6.1) in an unsupervised manner.

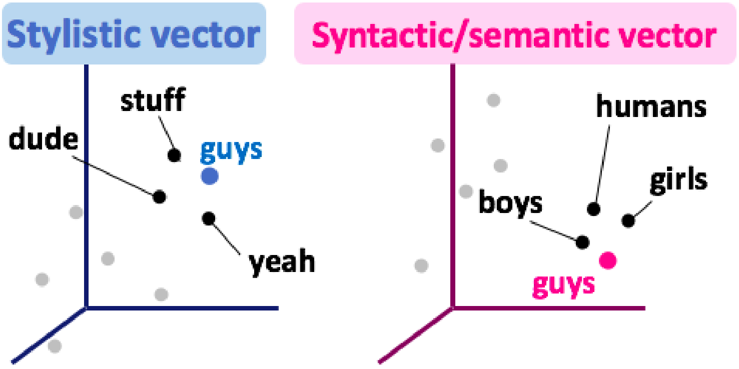


Fig. 6.1 Word vector capturing stylistic and syntactic/semantic similarity.

- We construct a novel dataset for style, which consists of pairs of style-sensitive words with each pair scored according to its stylistic similarity.
- We demonstrate that our word vectors capture the stylistic similarity between two words successfully.

6.1 Style-sensitive Word Vector

The key idea is to extend the continuous bag of words (CBOW) (Mikolov et al., 2013a) by distinguishing nearby contexts and wider contexts under the assumption that a style persists throughout every single utterance in a dialog. We elaborate on it in this section.

6.1.1 Notation

Let w_t denote the target word (token) in the corpora and $\mathscr{U}_t = \{w_1, \ldots, w_{t-1}, w_t, w_{t+1}, \ldots, w_{|\mathscr{U}_t|}\}$ denote the utterance (word sequence) including w_t. Here, w_{t+d} or $w_{t-d} \in \mathscr{U}_t$ is a context word of w_t (e.g., w_{t+1} is the context word next to w_t), where $d \in \mathbb{N}_{>0}$ is the distance between the context words and the target word w_t.

For each word (token) w, bold face v_w and $\tilde{v}_w$ denote the vector of w and the vector predicting the word w. Let $\mathscr{V}$ denote the vocabulary.

6.1.2 Baseline Model (CBOW-NEAR-CTX)

First, we give an overview of CBOW, which is our baseline model. CBOW predicts the target word w_t given nearby context words in a window with width δ:

$$\mathscr{C}_{w_t}^{\text{near}} := \{w_{t \pm d} \in \mathscr{U}_t \mid 1 \leq d \leq \delta\} \tag{6.1}$$

The set $\mathscr{C}_{w_t}^{\text{near}}$ contains in total at most 2δ words, including δ words to the left and δ words to the right of a target word. Specifically, we train the word vectors $\tilde{v}_{w_t}$ and v_c ($c \in \mathscr{C}_{w_t}^{\text{near}}$) by maximizing the following prediction probability:

$$P(w_t | \mathscr{C}_{w_t}^{\text{near}}) \propto \exp\left(\tilde{v}_{w_t} \cdot \frac{1}{|\mathscr{C}_{w_t}^{\text{near}}|} \sum_{c \in \mathscr{C}_{w_t}^{\text{near}}} v_c\right). \tag{6.2}$$

The CBOW captures both semantic and syntactic word similarity through the training using nearby context words. We refer to this form of CBOW as CBOW-NEAR-CTX. Note that, in the implementation of Mikolov et al. (2013b), the window width δ is sampled from a uniform distribution; however, in this work, we fixed δ for simplicity. Hereafter, throughout our experiments, we turn off the random resizing of δ.

6.1.3 Learning Style with Utterance-size Context Window (CBOW-ALL-CTX)

CBOW is designed to learn the semantic and syntactic aspects of words from their nearby context (Mikolov et al., 2013b). However, an interesting problem is determining the location where the stylistic aspects of words can be captured. To address this problem, we start with the assumption that a style persists throughout each single utterance in a dialog, that is, the stylistic profile of a word in an utterance must be consistent with other words in the same utterance. Based on this assumption, we propose extending CBOW to use all the words in an utterance as context,

$$\mathscr{C}_{w_t}^{\text{all}} := \{w_{t \pm d} \in \mathscr{U}_t \mid 1 \leq d\}, \tag{6.3}$$

instead of only the nearby words. Namely, we expand the context window from a fixed width to the entire utterance. This training strategy is expected to lead to learned word vectors that are more sensitive to style rather than to other aspects. We refer to this version as CBOW-ALL-CTX.

6.1.4 Learning the Style and Syntactic/Semantic Separately

To learn the stylistic aspect more exclusively, we further extended the learning strategy.

Distant-context Model (CBOW-DIST-CTX)

First, remember that using nearby context is effective for learning word vectors that capture semantic and syntactic similarities. However, this means that using the nearby context can lead the word vectors to capture some aspects other than style. Therefore, as the first extension, we propose excluding the *nearby* context $\mathscr{C}_{w_t}^{\text{near}}$ from *all* the context $\mathscr{C}_{w_t}^{\text{all}}$. In other words, we use the *distant* context words only:

$$\mathscr{C}_{w_t}^{\text{dist}} := \mathscr{C}_{w_t}^{\text{all}} \setminus \mathscr{C}_{w_t}^{\text{near}} = \{w_{t \pm d} \in \mathscr{U}_t \mid \delta < d\}. \tag{6.4}$$

We expect that training with this type of context will lead to word vectors containing the style-sensitive information only. We refer to this method as CBOW-DIST-CTX.

Separate Subspace Model (CBOW-SEP-CTX)

As the second extension to distill off aspects other than style, we use both *nearby* and *all* contexts ($\mathscr{C}_{w_t}^{\text{near}}$ and $\mathscr{C}_{w_t}^{\text{all}}$). As Figure 6.2 shows, both the vector v_w and $\tilde{v}_w$ of each word $w \in \mathscr{V}$ are divided into two vectors:

$$v_w = x_w \oplus y_w, \ \ \tilde{v}_w = \tilde{x}_w \oplus \tilde{y}_w, \tag{6.5}$$

where $\oplus$ denotes vector concatenation. Vectors x_w and $\tilde{x}_w$ indicate the style-sensitive part of v_w and $\tilde{v}_w$ respectively. Vectors y_w and $\tilde{y}_w$ indicate the syntactic/semantic-sensitive part of v_w and $\tilde{v}_w$ respectively. For training, when the context words are near the target word ($\mathscr{C}_{w_t}^{\text{near}}$), we update both the style-sensitive vectors ($\tilde{x}_{w_t}$, x_c) and the syntactic/semantic-sensitive vectors ($\tilde{y}_{w_t}$, y_c), i.e., $\tilde{v}_{w_t}$, v_c. Conversely, when the context words are far from the target word ($\mathscr{C}_{w_t}^{\text{dist}}$), we only update the style-sensitive vectors ($\tilde{x}_{w_t}$, x_c). Formally, the prediction probability is calculated as follows:

$$P_1(w_t | \mathscr{C}_{w_t}^{\text{near}}) \propto \exp\left(\tilde{v}_{w_t} \cdot \frac{1}{|\mathscr{C}_{w_t}^{\text{near}}|} \sum_{c \in \mathscr{C}_{w_t}^{\text{near}}} v_c\right), \tag{6.6}$$

$$P_2(w_t | \mathscr{C}_{w_t}^{\text{dist}}) \propto \exp\left(\tilde{x}_{w_t} \cdot \frac{1}{|\mathscr{C}_{w_t}^{\text{dist}}|} \sum_{c \in \mathscr{C}_{w_t}^{\text{dist}}} x_c\right). \tag{6.7}$$

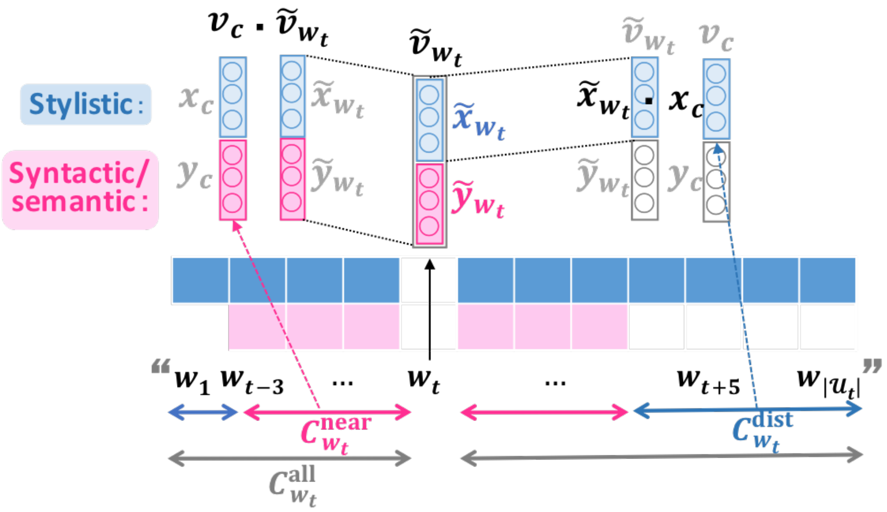


Fig. 6.2 The architecture of CBOW-SEP-CTX.

At the time of learning, two prediction probabilities (loss functions) are alternately computed, and the word vectors are updated. We refer to this method using the two-fold contexts separately as the CBOW-SEP-CTX.

6.2 Experiments

We investigated which word vectors capture the stylistic, syntactic, and semantic similarities.

6.2.1 Settings

Training and Test Corpus We collected Japanese fictional stories from the Web to construct the dataset. The dataset contains approximately 30M utterances of fictional characters. We separated the data into a 99%–1% split for training and testing. In Japanese, the function words at the end of the sentence often exhibit style (e.g., *desu+wa*, *desu+ze*[1];) therefore, we used an existing lexicon of multi-word functional expressions (Miyazaki et al., 2015). Overall, the vocabulary size $|\mathcal{V}|$ was 100K.

Hyperparameters We chose the dimensions of both the style-sensitive and the syntactic/semantic-sensitive vectors to be 300, and the dimensions of the baseline CBOWs were 300. The learning rate was adjusted individually for each part in $\{x_w, y_w, \tilde{x}_w, \tilde{y}_w\}$ such that “the product of the learning rate and the expectation of the number of updates” was a fixed constant. We

[1] These words mean the verb *be* in English.

ran the optimizer with its default settings from the implementation of Mikolov et al. (2013a). The training stopped after 10 epochs. We fixed the nearby window width to $\delta = 5$.

6.2.2 Stylistic Similarity Evaluation

Data Construction

To verify that our models capture the stylistic similarity, we evaluated our style-sensitive vector x_{w_t} by comparing to other word vectors on a novel artificial task matching human stylistic similarity judgments. For this evaluation, we constructed a novel dataset with human judgments on the stylistic similarity between word pairs by performing the following two steps. First, we collected only style-sensitive words from the test corpus because some words are strongly associated with stylistic aspects (Kinsui, 2003; Teshigawara and Kinsui, 2011) and, therefore, annotating random words for stylistic similarity is inefficient. We asked crowdsourced workers to select style-sensitive words in utterances. Specifically, for the crowdsourced task of picking "style-sensitive" words, we provided workers with a word-segmented utterance and asked them to pick words that they expected to be altered within different situational contexts (e.g., characters, moods, purposes, and the background cultures of the speaker and listener.). Then, we randomly sampled $1,000$ word pairs from the selected words and asked 15 workers to rate each of the pairs on five scales (from -2: "*The style of the pair is different*" to $+2$: "*The style of the pair is similar*"), inspired by the syntactic/semantic similarity dataset (Finkelstein et al., 2002; Gerz et al., 2016). Finally, we picked only word pairs featuring clear worker agreement in which more than 10 annotators rated the pair with the same sign, which consisted of random pairs of highly agreeing style-sensitive words. Consequently, we obtained 399 word pairs with similarity scores. To our knowledge, this is the first study that created an evaluation dataset[2]to measure the lexical stylistic similarity.

In the task of selecting style-sensitive words, the pairwise inter-annotator agreement was moderate (Cohen's kappa κ is 0.51). In the rating task, the pairwise inter-annotator agreement for two classes ($\{-2,-1\}$ or $\{+1,+2\}$) was fair (Cohen's kappa κ is 0.23). These statistics suggest that, at least in Japanese, native speakers share a sense of style-sensitivity of words and stylistic similarity between style-sensitive words.

Stylistic Sensitivity

We used this evaluation dataset to compute the Spearman rank correlation (ρ_{style}) between the cosine similarity scores between the learned word vectors $\cos(v_w, v_{w'})$ and the human

[2] https://jqk09a.github.io/style-sensitive-word-vectors/

Table 6.1 Results of the quantitative evaluations.

Model	ρ_{style}	ρ_{sem}	SYNTAXACC @5	SYNTAXACC @10
CBOW-NEAR-CTX	12.1	27.8	86.3	85.2
CBOW-ALL-CTX	36.6	24.0	85.3	84.1
CBOW-DIST-CTX	**56.1**	15.9	59.4	58.8
CBOW-SEP-CTX				
x (Stylistic)	**51.3**	**28.9**	68.3	66.2
y (Syntactic/semantic)	9.6	18.1	**88.0**	**87.0**

judgements. Table 6.1 shows the results on its left side. First, our proposed model, CBOW-ALL-CTX outperformed the baseline CBOW-NEAR-CTX. Furthermore, the x of CBOW-DIST-CTX and CBOW-SEP-CTX demonstrated better correlations for stylistic similarity judgments ($\rho_{style} = 56.1$ and 51.3, respectively). Even though the x of CBOW-SEP-CTX was trained with the same context window as CBOW-ALL-CTX, the style-sensitivity was boosted by introducing joint training with the near context. CBOW-DIST-CTX, which uses only the distant context, slightly outperforms CBOW-SEP-CTX. These results indicate the effectiveness of training using a wider context window.

6.2.3 Syntactic and Semantic Evaluation

We further investigated the properties of each model using the following criterion: (1) the model's ability to capture the syntactic aspect was assessed through a task predicting part of speech (POS) and (2) the model's ability to capture the semantic aspect was assessed through a task calculating the correlation with human judgments for semantic similarity.

Syntactic Sensitivity

First, we tested the ability to capture syntactic similarity of each model by checking whether the POS of each word was the same as the POS of a neighboring word in the vector space. Specifically, we calculated SYNTAXACC@N defined as follows:

$$\frac{1}{|\mathcal{V}|N} \sum_{w \in \mathcal{V}} \sum_{w' \in \mathcal{N}(w)} \mathbb{I}[\mathrm{POS}(w) = \mathrm{POS}(w')], \tag{6.8}$$

where $\mathbb{I}[\text{condition}] = 1$ if the condition is true and $\mathbb{I}[\text{conditon}] = 0$ otherwise, the function $\mathrm{POS}(w)$ returns the actual POS tag of the word w, and $\mathcal{N}(w)$ denotes the set of the N top similar words $\{w'\}$ to w w.r.t. $\cos(v_w, v_{w'})$ in each vector space.

Table 6.1 shows SYNTAXACC@N with $N = 5$ and 10. For both N, the y (the syntactic/semantic part) of CBOW-NEAR-CTX, CBOW-ALL-CTX and CBOW-SEP-CTX achieved similarly good. Interestingly, even though the x of CBOW-SEP-CTX used the same context as that of CBOW-ALL-CTX, the syntactic sensitivity of x was suppressed. We speculate that the syntactic sensitivity was distilled off by the other part of the CBOW-SEP-CTX vector, i.e., y learned using only the *near* context, which captured more syntactic information. In the next section, we analyze CBOW-SEP-CTX for the different characteristics of x and y.

Semantic and Topical Sensitivities

To test the model's ability to capture the semantic similarity, we also measured correlations with the Japanese Word Similarity Dataset (JWSD) (Sakaizawa and Komachi, 2018), which consists of 4,000 Japanese word pairs annotated with semantic similarity scores by human workers. For each model, we calculate and show the Spearman rank correlation score (ρ_{sem}) between the cosine similarity score $\cos(v_w, v_{w'})$ and the human judgements on JWSD in Table 6.1[3]. CBOW-DIST-CTX has the lowest score ($\rho_{sem} = 15.9$); however, surprisingly, the stylistic vector x_{w_t} has the highest score ($\rho_{sem} = 28.9$), while both vectors have a high ρ_{style}. This result indicates that the proposed stylistic vector x_{w_t} captures not only the stylistic similarity but also the captures semantic similarity, contrary to our expectations (ideally, we want the stylistic vector to capture only the stylistic similarity). We speculate that this is because not only the *style* but also the *topic* is often consistent in single utterances. For example, "サンタ (Santa Clause)" and "トナカイ (reindeer)" are topically relevant words and these words tend to appear in a single utterance. Therefore, stylistic vectors $\{x_w\}$ using all the context words in an utterance also capture the topic relatedness. In addition, JWSD contains topic-related word pairs and synonym pairs; therefore the word vectors that capture the topic similarity have higher ρ_{sem}. We will discuss this point in the next section.

6.2.4 Analysis of Trained Word Vectors

Finally, to further understand what types of features our CBOW-SEP-CTX model acquired, we show some words[4] with the four most similar words in Table 6.2. The top side of

[3] Note that the low performance of our baseline ($\rho_{sem} = 27.8$ for CBOW-NEAR-CTX) is unsurprising comparing to English baselines (cf., Taguchi et al. (2017)).

[4] We arbitrarily selected style-sensitive words from our stylistic similarity evaluation dataset.

Table 6.2 The top similar words for the style-sensitive and syntactic/semantic vectors learned with proposed model, CBOW-SEP-CTX (Japanese). Japanese words are translated into English by the authors. Legend: (translation; impression). **Classical* means wording related to e.g., samurai, ninja.

	The top similar words $\{w'\}$ to w w.r.t. cosine similarity			
	俺 (I; male, colloquial)	拙者 (I; classical*)	かしら (wonder; female)	サンタ (Santa Clause; shortened)
Stylistic $\cos(x_w, x_{w'})$	おまえ (you; colloquial, rough)	でござる (be; classical)	わね (QUESTION; female)	サンタクロース (Santa Clause; -)
	あいつ (he/she; colloquial, rough)	ござる (be; classical)	ないわね (not; female)	トナカイ (reindeer; -)
	ねーよ (not; colloquial, rough, male)	ござるよ (be; classical)	わ (SENTENCE-FINAL; female)	クリスマス (Christmas; -)
	俺 (I; male, colloquial)	拙者 (I; classical)	かしら (wonder; female)	サンタ (Santa Clause; shortened)
Syntactic/ Semantic $\cos(y_w, y_{w'})$	僕 (I; male, childish)	僕 (I; male, childish)	かな (wonder; childish)	お客 (customer; little polite)
	あたし (I; female, childish)	俺 (I; male, colloquial)	でしょうか (wonder; fomal)	プロデューサー (producer; -)
	私 (I; formal)	私 (I; formal)	かしらね (wonder; female)	メイド (maid; -)

Table 6.2 (for stylistic vector x) shows the results. We found that the Japanese word "拙者 (I; classical)" is similar to "ござる (be; classical)" or words containing it (the second column of Table 6.2). The result looks reasonable, because words such as "拙者 (I; classical)" and "ござる (be; classical)" are typically used by Japanese *Samurai* or *Ninja*. We can see that the vectors captured the similarity of these words, which are stylistically consistent across syntactic and semantic varieties. Conversely, the bottom side of the table (for the syntactic/semantic vector y) shows that the word "拙者 (I; classical)" is similar to the personal pronoun (e.g., "僕 (I; male, childish)"). We further confirmed that 15 the top similar words are also personal pronouns (even though they are not shown due to space limitations). These results indicate that the proposed CBOW-SEP-CTX model jointly learns two different types of lexical similarities, i.e., the stylistic and syntactic/semantic similarities in the different parts of the vectors. However, our stylistic vector also captured the topic similarity, such as "サンタ (Santa Clause)" and "トナカイ (reindeer)" (the fourth column of Table 6.2). Therefore, there is still room for improvement in capturing the stylistic similarity.

Here, for English readers, we also report a result for English. We trained another CBOW-SEP-CTX model on an English fan-fiction dataset that was collected from the Web[5]. The English result (Table 6.3) also shows an example of the performance of our model on another language.

[5] https://www.fanfiction.net/

Table 6.3 The top similar words for the style-sensitive and syntactic/semantic vectors learned with proposed model, CBOW-SEP-CTX (English).

	The top similar words $\{w'\}$ to w w.r.t. cosine similarity			
	shit	hi	guys	ninja
Stylistic $\cos(x_w, x_{w'})$	fuckin fuck goddamn	hello bye hiya	stuff guy bunch	shinobi genin konoha
	shit	hi	guys	ninja
Syntactic/Semantic $\cos(y_w, y_{w'})$	shitty crappy sucky	goodbye goodnight good-bye	boys humans girls	shinobi pirate soldier

6.3 Conclusion

In this chapter, we presented the unsupervised learning of style-sensitive word vectors, which extends CBOW by distinguishing nearby contexts and wider contexts. We created a novel dataset for style, where the stylistic similarity between word pairs was scored by human. Our experiment demonstrated that our method leads word vectors to distinguish the stylistic aspect and other semantic or syntactic aspects.

Chapter 7

Conclusion

Toward the improvement of the neural response generation technology through the improvement of their training data, in this thesis, we have addressed the following three research issues:

- **What is the clues to enable augmentation or improvement of dialogue data?** From the series of investigations and experiments, we found the connectivity and relatedness of utterances are possible to use as criteria for automatically calculating the quality of dialogue. We hope that these insights will facilitate discussions on data-oriented approaches for improving neural dialogue response generation.

- **Methodologies for acquiring desirable resources for training neural response generation models.** We proposed several methodologies for acquiring large-scale and high-quality training data, including data filtering, data augmentation, and manual data construction.

- **Do the large-scale and high-quality training data improve the performance of neural response generation models?** In the dialogue response generation task, as in other neural text generation tasks, we empirically confirmed that large-scale and high-quality training data improves the performance of neural response generation models.

The key contribution of this thesis can be summarized as follows:

- **Establishing the methodologies for dialogue data improvement:** We proposed the data filtering methodology to make a large scale-data high-quality by detecting and removing the low-quality utterance pairs. Moreover, we proposed the data augmentation methodology to create synthetic utterance pairs from high-quality but small dialogue data.

- **Investigating the impact of training data improvement:** Through the response generation experiments, we demonstrated that large-scale and high-quality training data can improve the performance of neural response generation models.

- **Presenting effective corpus construction methodology:** We proposed a practical methodology for manually creating new training data for neural response generation models. We empirically confirmed that the proposed method collected dialogue data at a scale and quality that can be used for training neural dialogue models even when human resources are limited.

- **Modeling the style of utterances:** We introduce a novel task and new benchmark data for measuring stylistic similarity of words and proposed an unsupervised methodology to acquire style-sensitive word vectors independently from semantic or syntactic features of words. We demonstrated that our word vectors capture the stylistic similarity between two words successfully.

References

Adiwardana, D., Luong, M.-T., So, D. R., Hall, J., Fiedel, N., Thoppilan, R., Yang, Z., Kulshreshtha, A., Nemade, G., Lu Quoc, Y., and Le, V. (2020). Towards a Human-like Open-Domain Chatbot. In *aiXiv preprint arXiv:2001.09977*.

Akama, R., Inada, K., Inoue, N., Kobayashi, S., and Inui, K. (2017). Generating Stylistically Consistent Dialog Responses with Transfer Learning. In *Proceedings of the 8th International Joint Conference on Natural Language Processing (IJCNLP)*, volume 2, pages 408–412.

Arora, S., Liang, Y., and Ma, T. (2017). A Simple but Tough-to-Beat Baseline for Sentence Embeddings. In *5th International Conference on Learning Representations (ICLR)*.

Baheti, A., Ritter, A., Li, J., and Dolan, B. (2018). Generating More Interesting Responses in Neural Conversation Models with Distributional Constraints. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3970–3980.

Bao, S., He, H., Wang, F., Wu, H., and Wang, H. (2020). PLATO: Pre-trained Dialogue Generation Model with Discrete Latent Variable. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 85–96.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics (TACL)*, 5:135–146.

Bouma, G. (2009). Normalized (Pointwise) Mutual Information in Collocation Extraction. In *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology (GSCL)*, pages 31–40.

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., and Mercer, R. L. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311.

Callison-Burch, C., Ungar, L., and Pavlick, E. (2015). Crowdsourcing for nlp. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL): Tutorial Abstracts*, pages 2–3.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., and Bordes, A. (2017). Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 670–680.

Csáky, R., Purgai, P., and Recski, G. (2019). Improving Neural Conversational Models with Entropy-Based Data Filtering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5650–5669.

Danescu-Niculescu-Mizil, C. and Lee, L. (2011). Chameleons in Imagined Conversations: A New Approach to Understanding Coordination of Linguistic Style in Dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87.

Dinan, E., Roller, S., Shuster, K., Fan, A., Auli, M., and Weston, J. (2019). Wizard of Wikipedia: Knowledge-Powered Conversational Agents. In *7th International Conference on Learning Representations (ICLR)*.

Dyer, C., Chahuneau, V., and Smith, N. A. (2013). A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 644–648.

Edunov, S., Ott, M., Auli, M., and Grangier, D. (2018). Understanding Back-Translation at Scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 489–500.

Fadaee, M., Bisazza, A., and Monz, C. (2017). Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 567–573.

Finkelstein, L., Gabrilovich, E., Matians, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. (2002). Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131.

Flekova, L., PreoŢiuc-Pietro, D., and Ungar, L. (2016). Exploring stylistic variation with age and income on twitter. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 313–319.

Galley, M., Brockett, C., Sordoni, A., Ji, Y., Auli, M., Quirk, C., Mitchell, M., Gao, J., and Dolan, B. (2015). deltaBLEU: A Discriminative Metric for Generation Tasks with Intrinsically Diverse Targets. 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)*, volume 2, pages 445–450.

Galley, M., Fosler-Lussier, E., and Potamianos, A. (2001). Hybrid natural language generation for spoken dialogue systems. In *Proceedings of the 7th European Conference on Speech Communication and Technology (EUROSPEECH–01)*, pages 1735–1738.

Gerz, D., Vulić, I., Hill, F., Reichart, R., and Korhonen, A. (2016). Simverb-3500: A large-scale evaluation set of verb similarity. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2173–2182.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning Word Vectors for 157 Languages. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)*, pages 3483–3487.

Hasegawa, T., Kaji, N., Yoshinaga, N., and Toyoda, M. (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.

He, H., Chen, D., Balakrishnan, A., and Liang, P. (2018). Decoupling strategy and generation in negotiation dialogues. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2333–2343.

Hellstrm, Y. and IR, H. (2001). Perspectives of elderly people receiving home help on health care and quality of life. In *Health and Social Care in the Community*, pages 61–71.

Henderson, M., Budzianowski, P., Casanueva, I., Coope, S., Gerz, D., Kumar, G., Mrkši´c, N. M., Spithourakis, G., Su, P.-H., Vuli´c, I. V., and Wen, T.-H. (2019). A Repository of Conversational Datasets. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 1–10.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Isbell, C. L., Kearns, M., Kormann, D., Singh, S., and Stone, P. (2000). Cobot in lambdamoo: A social statistics agent. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence (AAAI)*, pages 36–41.

Junczys-Dowmunt, M. (2018). Dual Conditional Cross-Entropy Filtering of Noisy Parallel Corpora. In *Proceedings of the 3rd Conference on Machine Translation: Shared Task Papers (WMT)*, pages 888–895.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *The Third International Conference on Learning Representations (ICLR)*.

Kinsui, S. (2003). *Vaacharu nihongo: yakuwari-go no nazo (In Japanese)*. Tokyo, Japan: Iwanami.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Mit, C. M., Zens, R., Aachen, R., Dyer, C., Bojar, O., and Cornell, E. H. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion (ACL) Volume Proceedings of the Demo and Poster Sessions*, pages 177–180.

Koehn, P., Khayrallah, H., Heafield, K., and Forcada, M. L. (2018). Findings of the WMT 2018 Shared Task on Parallel Corpus Filtering. In *Proceedings of the 3rd Conference on Machine Translation: Shared Task Papers (WMT)*, pages 726–739.

Koehn, P. and Knowles, R. (2017). Six Challenges for Neural Machine Translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39.

Krause, B., Damonte, M., Dobre, M., Duma, D., Fainberg, J., Fancellu, F., Kahembwe, E., Cheng, J., and Webber, B. (2017). Edina: Building an Open Domain Socialbot with Self-dialogues. *1st Proceedings of Alexa Prize*.

Lala, D., Milhorat, P., Inoue, K., Ishida, M., Takanashi, K., and Kawahara, T. (2017). Attentive listening system with backchanneling, response generation and flexible turn-taking. In *Proceedings of the 18th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 127–136.

Li, J., Galley, M., Brockett, C., Gao, J., and Dolan, B. (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)*, pages 110–119.

Li, J., Galley, M., Brockett, C., Spithourakis, G., Gao, J., and Dolan, B. (2016b). A Persona-Based Neural Conversation Model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 1, pages 994–1003.

Li, J., Monroe, W., Shi, T., Jean, S., Ritter, A., and Jurafsky, D. (2017a). Adversarial Learning for Neural Dialogue Generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2157–2169.

Li, Y., Su, H., Shen, X., Li, W., Cao, Z., and Niu, S. (2017b). DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings of the 8th International Joint Conference on Natural Language Processing (IJCNLP)*, volume 1, pages 986–995.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*, 22(140):1–55.

Lison, P. and Tiedemann, J. (2016). OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*, pages 923–929.

Lison, P., Tiedemann, J., and Kouylekov, M. (2018). OpenSubtitles2018: Statistical Rescoring of Sentence Alignments in Large, Noisy Parallel Corpora. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)*, pages 1742–1748.

Liu, C.-W., Lowe, R., Serban, I. V., Noseworthy, M., Charlin, L., and Pineau, J. (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 (EMNLP)*, pages 2122–2132.

Lowe, R., Noseworthy, M., Serban, I. V., Angelard-Gontier, N., Bengio, Y., and Pineau, J. (2017). Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 1, pages 1116–1126.

Mairesse, F. and Walker, M. (2007). Personage: Personality generation for dialogue. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 496–503.

Mangrulkar, S., Shrivastava, S., Thenkanidiyoor, V., and Aroor Dinesh, D. (2018). A context-aware convolutional natural language generation model for dialogue systems. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue (SIG-DIAL)*, pages 191–200.

Marelli, M., Bentivogli, L., Baroni, M., Bernardi, R., Menini, S., and Zamparelli, R. (2014a). SemEval-2014 Task 1: Evaluation of Compositional Distributional Semantic Models on Full Sentences through Semantic Relatedness and Textual Entailment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval)*, pages 1–8.

Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., and Zamparelli, R. (2014b). A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, pages 216–223.

Mehri, S. and Eskenazi, M. (2020). USR: An Unsupervised and Reference Free Evaluation Metric for Dialog Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 681–707.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. In *the International Conference on Learning Representations (ICLR) Workshop*.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., and Joulin, A. (2018). Advances in Pre-Training Distributed Word Representations. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)*, pages 52–55.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *The 26th Annual Conference on Neural Information Processing Systems*, pages 3111–3119.

Miyazaki, C., Hirano, T., Higashinaka, R., Makino, T., and Matsuo, Y. (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*, pages 307–314.

Morishita, M., Suzuki, J., and Nagata, M. (2018). NTT’ s Neural Machine Translation Systems for WMT 2018. In *Proceedings of the 3rd Conference on Machine Translation: Shared Task Papers (WMT)*, pages 461–466.

Niu, X. and Carpuat, M. (2017). Discovering stylistic variations in distributional vector space models via lexical paraphrases. In *Proceedings of the Workshop on Stylistic Variation at the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 20–27.

Niu, X., Martindale, M., and Carpuat, M. (2017). A study of style in machine translation: Controlling the formality of machine translation output. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2804–2809.

Och, F. J. and Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations) (NAACL-HLT)*, pages 48–53.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318.

Pavlick, E. and Nenkova, A. (2015). Inducing lexical style properties for paraphrase and genre differentiation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 218–224.

Pavlick, E. and Tetreault, J. (2016). An empirical analysis of formality in online communication. *Transactions of the Association of Computational Linguistics*, 4:61–74.

Pei, J. and Li, C. (2018). S2SPMN: A Simple and Effective Framework for Response Generation with Relevant Information. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 745–750.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 2227–2237.

Preotiuc-Pietro, D., Xu, W., and Ungar, L. H. (2016). Discovering user attribute stylistic differences via paraphrasing. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, pages 3030–3037.

Ritter, A., Cherry, C., and Dolan, W. B. (2011). Data-Driven Response Generation in Social Media. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 583–593.

Sacks, H. (1989). Lecture One: Rules of Conversational Sequence. *Human Studies*, 12(3/4):217–233.

Sakaizawa, Y. and Komachi, M. (2018). Construction of a japanese word similarity dataset. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, pages 948–951.

Sennrich, R., Haddow, B., and Birch, A. (2016a). Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40.

Sennrich, R., Haddow, B., and Birch, A. (2016b). Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 86–96.

Sennrich, R., Haddow, B., and Birch, A. (2016c). Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 1, pages 1715–1725.

Sennrich, R. and Zhang, B. (2019). Revisiting Low-Resource Neural Machine Translation: A Case Study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 211–221.

Serban, I. V., Sordoni, A., Lowe, R., Charlin, L., Pineau, J., Courville, A., and Bengio, Y. (2017). A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 3295–3301.

Shang, L., Lu, Z., and Li, H. (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)*, volume 1, pages 1577–1586.

Shang, M., Fu, Z., Peng, N., Feng, Y., Zhao, D., and Yan, R. (2018). Learning to Converse with Noisy Data: Generation with Calibration. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*, pages 4338–4344.

Shao, Y., Gouws, S., Britz, D., Goldie, A., Strope, B., and Kurzweil, R. (2017). 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 (EMNLP)*, pages 2210–2219.

Shen, X., Su, H., Li, Y., Li, W., Niu, S., Zhao, Y., Aizawa, A., and Long, G. (2017). A Conditional Variational Framework for Dialog Generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 2, pages 504–509.

Shitaoka, K., Tokuhisa, R., Yoshimura, T., Hoshino, H., and Watanabe, N. (2017). Active listening system for a conversation robot (in Japanese). *Journal of Natural Language Processing*, 24(1):3–47.

Sidnell, J. (2010). *Conversation Analysis: An Introduction.* Language in Society. John Wiley & Sons.

Sidner, C., Bickmore, T., Rich, C., Barry, B., Ring, L., Behrooz, M., and Shayganfar, M. (2013). Demonstration of an always-on companion for islated older adults. In *Proceedings of the 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 148–150.

Sinha, K., Parthasarathi, P., Wang, J., Lowe, R., Hamilton, W. L., and Pineau, J. (2020). Learning an Unreferenced Metric for Online Dialogue Evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2430–2441. Association for Computational Linguistics.

Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., Nie, J.-Y., Gao, J., and Dolan, B. (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)*, pages 196–205.

Subramanian, S., Trischler, A., Bengio, Y., and Pal, C. J. (2018). Learning General Purpose Distributed Sentence Representations via Large Scale Multi-task Learning. In *6th International Conference on Learning Representations (ICLR)*.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems 27 (NIPS)*, pages 3104–3112.

Taguchi, Y., Tamori, H., Hitomi, Y., Nishitoba, J., and Kikuta, K. (2017). Learning Japanese word distributional representation considering of synonyms (in Japanese). Technical Report 17, The Asahi Shimbun Company, Retrieva Inc.

Tao, C., Mou, L., Zhao, D., and Yan, R. (2018). RUBER: An Unsupervised Method for Automatic Evaluation of Open-Domain Dialog Systems. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI-18)*.

Teshigawara, M. and Kinsui, S. (2011). Modern Japanese 'role language' (yakuwarigo): fictionalised orality in Japanese literature and popular culture. *Sociolinguistic Studies*, 5(1):37.

Vaarama, M. (2009). Care-related quality of life in old age. In *European Journal of Aging*, pages 113–125.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is All you Need. In *Advances in Neural Information Processing Systems 30 (NIPS)*, pages 5998–6008.

Vinyals, O. and Le, Q. (2015a). A Neural Conversational Model. In *Proceedings of the 31st International Conference on Machine Learning (ICML) Deep Learning Workshop*.

Vinyals, O. and Le, Q. (2015b). A neural conversational model. In *International Conference on Machine Learning (ICML) Deep Learning Workshop*.

Wang, D., Jojic, N., Brockett, C., and Nyberg, E. (2017). Steering output style and topic in neural response generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2140–2150.

Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9(1):36–45.

Xing, C., Wu, W., Wu, Y., Liu, J., Huang, Y., Zhou, M., and Ma, W.-Y. (2017). Topic Aware Neural Response Generation. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pages 3351–3357.

Xu, W. (2017). From shakespeare to twitter: What are language styles all about? In *Proceedings of the Workshop on Stylistic Variation at the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1–9.

Xu, X., Dušek, O., Konstas, I., and Rieser, V. (2018a). Better Conversations by Modeling, Filtering, and Optimizing for Coherence and Diversity. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3981–3991.

Xu, Z., Jiang, N., Liu, B., Rong, W., Wu, B., Wang, B., Wang, Z., and Wang, X. (2018b). LSDSCC: a Large Scale Domain-Specific Conversational Corpus for Response Generation with Diversity Oriented Evaluation Metrics. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, volume 1, pages 2070–2080.

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., and Weston, J. (2018). Personalizing Dialogue Agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 1, pages 2204–2213.

Zhao, T., Zhao, R., and Eskenazi, M. (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)*, pages 654–664.

内閣府 (2015). 平成27年度版高齢社会白書 (全体版).

内閣府 (2017). 平成29年度版高齢社会白書 (全体版).

太田, 芳賀, 長田, 田中, 前田, 嶽崎, 関, 大山, 中西, and 石川 (2001). 地域高齢者のための QOL 質問表の開発と評価. In 日本公衆衛生雑誌, pages 258–267.

徳久, 寺嶌, and 乾 (2019). 高齢者と家族とのコミュニケーションの質の向上に向けて：高齢者の Quality of Life 表出発話の分析. 情報処理学会論文誌, 60(2):1–8.

List of Publications

Journal Papers (Refereed)

1. Reina Akama, Ryoko Tokuhisa, Kentaro Inui. Generating Candidate Responses for Supporting Human-to-human Communication of Quality of Life (in Japanese). In Journal of Natural Language Processing, Volume 26, Number 3, pp.579-612, September 2019.

International Conference Papers (Refereed)

1. Reina Akama, Sho Yokoi, Jun Suzuki and Kentaro Inui. Filtering Noisy Dialogue Corpora by Connectivity and Content Relatedness. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020), pp.941-958, November 2020.

2. Reina Akama, Kento Watanabe, Sho Yokoi, Sosuke Kobayashi, Kentaro Inui. Unsupervised Learning of Style-sensitive Word Vectors. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018), Volume 2 pp.572-578, July 2018.

3. Reina Akama, Kazuaki Inada, Naoya Inoue, Sosuke Kobayashi, Kentaro Inui. Generating Stylistically Consistent Dialog Responses with Transfer Learning. In Proceedings of the 8th International Joint Conference on Natural Language Processing (IJCNLP 2017), Volume 2 pp.408-412, November 2017.

Awards

1. The 34th Annual Conference of the Japanese Society for Artificial Intelligence (JSAI 2020) Student Incentive Award, June 2020.
2. The 34th Annual Conference of the Japanese Society for Artificial Intelligence (JSAI 2020) Award, June 2020.
3. The 26th Annual Meeting of the Association for Natural Language Processing Grand Prize, March 2020.
4. The 14th Symposium of Young Researcher Association for NLP Studies (YANS) Encouragement Award, August 2019.
5. The 26th Annual Meeting of the Association for Natural Language Processing Young Researcher Encouragement Award, March 2019.
6. The 26th Annual Meeting of the Association for Natural Language Processing Best Poster Presentation Award, March 2019.
7. The 1st Dialogue System Live Competition Excellence Award (The 3rd place), November 2018.
8. The 32th Annual Conference of the Japanese Society for Artificial Intelligence (JSAI 2018) Award, June 2018.
9. The 5th CWRU-Tohoku Joint Workshop Best Poster Presentation Award, August 2018.
10. The 81st Special Interest Group on Spoken Language Understanding and Dialogue Processing (SIG-SLUD) 8th Dialogue Symposium Young Researcher Encouragement Award, October 2017.
11. The 12th Symposium of Young Researcher Association for NLP Studies (YANS) Encouragement Award, September 2017.

Other Publications (Not refereed)

1. 赤間怜奈, 横井祥, 鈴木潤, 乾健太郎. ニューラル対話応答生成のための言語非依存な低品質対話データフィルタリング法の提案と分析. 人工知能学会第34回全国大会（JSAI 2020）, オンライン, 2020年6月.

2. 赤間怜奈, 鈴木潤, 横井祥, 乾健太郎. 句の呼応と話題の一貫性に着目した低品質対話データの教師なしフィルタリング. 言語処理学会第26回年次大会（NLP 2020）, pp.1507-1510, オンライン, 2020年3月.

3. 赤間怜奈, 武藤由依, 鈴木潤, 乾健太郎. 独立発話の繋ぎ合わせによる発話-応答ペアの獲得. 言語処理学会第25回年次大会（NLP 2019）, pp.1153-1156, 名古屋, 2019年3月.

4. 赤間怜奈, 徳久良子, 乾健太郎. Quality of Life 情報の伝達補助を目的とする対話応答生成. 人工知能学会音声・言語理解と対話処理研究会第84回研究会第9回対話システムシンポジウム（SLUD-84）, 早稲田, 2018年11月.

5. 赤間怜奈, 渡邉研斗, 横井祥, 小林颯介, 乾健太郎. スタイルの類似性を捉えた単語ベクトルの教師なし学習. 人工知能学会第32回全国大会（JSAI 2018）, 鹿児島, 2018年6月.

6. 赤間怜奈, 横井祥, 渡邉研斗, 田然, 小林颯介, 乾健太郎. サンプリング戦略に基づく単語ベクトルの意味成分とスタイル成分の分離. 言語処理学会第24回年次大会（NLP 2018）, pp.718-721, 岡山, 2018年3月.

7. 赤間怜奈, 横井祥, 渡邉研斗, 乾健太郎. 発話における語の文体ベクトルの半教師あり学習. 人工知能学会音声・言語理解と対話処理研究会第81回研究会第8回対話システムシンポジウム（SLUD-81）, 早稲田, 2017年10月.

8. 赤間怜奈, 渡邉研斗, 横井祥, 乾健太郎. 発話スタイル空間の教師なし学習およびスタイル制御可能な対話システムの実現. NLP若手の会第12回シンポジウム（YANS 2017）, 那覇, 2017年9月.

9. 赤間怜奈, 稲田和明, 小林颯介, 佐藤祥多, 乾健太郎. 転移学習を用いた対話応答のスタイル制御. 言語処理学会第23回年次大会（NLP 2017）, pp.338-341, 筑波, 2017年3月.