MASTER'S THESIS

Modeling Word Meanings by Individuals and Its Applications

（個人が用いる単語の意味のモデル化とその応用）

by

48-186421　Daisuke OBA

in the

Department of Information and Communication Engineering

Graduate School of Information Science and Technology

The University of Tokyo

Supervisor:　Professor Masashi TOYODA

January 30, 2020

Abstract

When people verbalize what they felt with various sensory functions, they represent different meanings (*e.g.*, *temperature range*) with the same word (*e.g.*, *cold*) or the same meaning with different words (*e.g.*, *hazy*, *cloudy*). These interpersonal variations in word meanings not only prevent our smooth communication but also cause troubles when we perform natural language processing (NLP). It is therefore necessary to deepen understanding of these variations in word meanings.

In this thesis, to capture interpersonal semantic variations in word meanings, a method for modeling word meanings by individuals, "personalized word embeddings," is proposed. This method learns personalized word embeddings from an NLP task, distinguishing words used by different people as different words. To prevent meaning-unrelated biases from contaminating word embeddings, review-target identification is adopted as an induction task.

The scalability and stability are major technical issues when conducting the proposed method. As a solution, the scalability is improved by using reviewer-wise fine-tuning of a neural network with residual connection and the stability is also improved by using multi-task learning with target-attribute predictions.

The results of experiments using large-scale review datasets obtained from the RateBeer and Yelp websites confirmed that the proposed method was effective for estimating the target items, and the resulting word embeddings were also effective for solving sentiment analysis and review text personalization tasks. By using the acquired word embeddings, it was possible to extract words with a strong semantic variation and reveal tendencies in semantic variations of the word meanings.

Acknowledgements

I would never think my master's degree could be completed without the support by many people. First of all, I would like to thank my supervisor, Professor Masashi Toyoda. He gave me constructive advice and suggestions whenever I needed. I am also grateful that he have respected my will as much as possible regarding the my research themes. Thanks to him, I have enjoyed my research for two years. I would also like to thank Associate Professor Naoki Yoshinaga. He spent many hours and much energy discussing about my researches and correcting my academic papers, especially before the submission deadline. His critical advice has always made my papers solid. I am also grateful to Professor Masaru Kitsuregawa. He provided the best environment for doing research, especially on computational resources.

I wouold like to thank Assistant Professor Junpei Komiyama and Project Assistant Professor Kazutoshi Umemoto. They gave me a lot of advice on my researches. Especially advice from knowledge in the fields other than mine were very valuable for me.

I also deeply acknowledge the contributions of co-authors of my conference papers, Mr. Shoestu Sato and Mr. Satoshi Akasaki. Since I started my research in this laboratory for the first time, they has provided a lot of advice and suggestion for me. Without them, I could not publish the international conference papers. They not only gave me technical advice for my research, but also taught me hacks to make my student life more comfortable.

I would like to thank all students, researchers, and secretaries in the laboratory. The students were always eager to comment on papers, thesis, and presentation slides. Also, when I was tired just before the submission deadline, I was motivated to see someone struggling for deadlines. And I was very honored to have time to discuss with the best people like you every day. The secretaries were not only perfect for their work, but

also cared about our physical and mental conditions. Thanks to them, two years in the laboratory will be a precious memory for me.

Finally, I would like to thank my family for their great support in continuing my student life. I sincerely thank them for giving me this kind of learning opportunity.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

People express what they have sensed with various sensory units as language in different ways, and semantic variations in the meanings of words inevitably exist because the senses and linguistic abilities of individuals differ. As an example, even if we use the word "*sour*," *how* "*sour*" can differ greatly between individuals. Furthermore, different people may describe the appearance (color) of the same beer with different expressions such as "*yellow*" and "*golden*." These semantic variations in word meanings not only cause problems in our verbal communication but also degrade performance of natural language processing (NLP) systems.

In the context of personalization, several studies have attempted to improve the performance of NLP models in user-oriented tasks such as sentiment analysis [1–3], dialogue systems [4–7], grammatical error correction [8] and machine translation [9–11], taking into account user preferences in regard to the task inputs and outputs. However, all of these studies were based on the settings of estimating *subjective* output from *subjective* input (*e.g.,* estimating a sentiment polarity of the target item from an input review or predicting responses from input utterances in a dialogue system). As a result, the model not

only captures the semantic variations in the user-generated text (input) but also handles *annotation bias* of the output labels (namely, the deviation of output labels assigned by each annotator) [2, 12, 13] and *selection bias* (namely, the deviation of output labels inherited from the targets chosen by users in sentiment analysis) [2]. The contamination caused by these biases hinders the understanding of the solo impact of semantic variations, which is the target in this study.

1.2 Research Goal and Challenges

The goal of this study is to (i) understand which words have large (or small) interpersonal variations in their meanings (hereafter referred to as *semantic variation* in this study) and (ii) reveal how such semantic variation affects the classification accuracy concerning tasks with user-generated inputs (*e.g.,* reviews). A method for analyzing the degree of personal semantic variation in word meanings is thus proposed (Chapter 4). It uses personalized word embeddings acquired through a task called "review-target identification," in which a classifier estimates a target item (*objective* output) from given reviews (*subjective* input) written by various reviewers. This task is free from *annotation bias* because outputs (review target) are automatically determined without annotation. Also, *selection bias* can be suppressed by using a dataset in which the same reviewer evaluates the same target (object) only once, so as not to learn the deviation of output labels caused by the choice of inputs. The resulting model makes it possible to observe only the impact of semantic variations from the acquired personalized word embeddings.

Remaining issues concerning inducing personalized word embeddings are the scalability and stability in learning personalized word embeddings. To make the training scalable in regard to the number of reviewers, a residual network [14] is utilized to (i) obtain personalized word embeddings by using reviewer-specific transformation matrices and biases from a small amount of reviews for each user (§ 4.3.2.1), and (ii) fine-tune these reviewer-specific parameters (§ 4.3.2.4). Also, to make the training via the extreme multi-class classification (*i.e.,* the review-target identification) stable, multi-task learning with

target-attribute predictions is performed during pre-training of the parameters (§ 4.3.2.3). Since the target attributes are likely to be more coarse-grained than the review targets, multi-task learning with the target-attribute predictions makes the training more stable.

In the experiments, it is hypothesized that words related to the five senses especially have inherent semantic variations, and this hypothesis is validated (Chapter 5). Two large-scale datasets retrieved from the RateBeer and Yelp websites including a variety of expressions related to the five senses, are utilized. To confirm the impact of personalized word embeddings obtained by using the proposed method, the datasets were utilized for a certain task: identifying a target item and its attributes from a given review by using the reviewer's ID. As a result, in regard to both datasets, our personalized model successfully captured semantic variation and achieved better performance than a reviewer-universal model (§ 5.3.1). Moreover, the obtained personalized word embeddings were extrinsically evaluated by sentiment analysis and review text personalization. The results of the extrinsic evaluation, in which our model achieved better performances than the other models, demonstrated the capability of the proposed method for suppressing unfavorable biases during the training process (§ 5.3.3). The acquired personalized word embeddings were finally analyzed from three perspectives (frequency, dissemination and polysemy) to reveal which words have large semantic variations (Chapter 6). From the analysis, tendencies that the degree of semantic variation correlates with frequency and dissemination of words were revealed (§ 6.3.1). It was also shown that adjectives and words related to the five senses have large interpersonal differences in their meanings (§ 6.3.2).

1.3 Contributions

The contributions of this thesis are three-fold:

Induction A scalable and stable method for inducing personalized word embeddings without contaminating them with meaning-unrelated biases is proposed. The proposed method induces the word embeddings through review-target identification via reviewer-wise fine-tuning on a neural network with a residual connection and multi-task learning with target-attribute predictions. Effectiveness of the obtained personalized word embeddings on review-target identification itself is confirmed.

Application Usefulness of the obtained personalized word embeddings, not only in the review-target identification task but also in the sentiment analysis task and review text personalization task, is confirmed. The results indicate that the proposed method could capture task-independent word meanings.

Analysis It is shown that meanings of frequent and disseminated words not necessarily agreed by analyzing the obtained word embeddings in terms of three perspectives, which were discussed in previous studies about diachronic and interdomain semantic variations. It is also shown that meanings of adjectives and the words related to the five senses largely differ by individuals.

1.4 Thesis Structure

The structure of this thesis is as follows.

Chapter 2 Basic knowledge related to the proposed method is introduced. Especially, how to represent word meanings with computers, pre-training and fine-tuning of the neural network based models, and the multi-task learning techniques are introduced.

Chapter 3 Related work and how this thesis relates to them are introduced. Especially, studies about personalized natural language processing systems, and interdomain, diachronic and geographical variations in word meanings, and removing unfavorable biases from word embeddings for political correctness are introduced.

Chapter 4 A method for modeling word meanings by individuals, "personalized word embeddings", via review-target identification is proposed. Fine-tuning word embeddings with a residual connection and Multi-task learning with target-attribute predictions is conducted to improve the scalability and stability of the method.

Chapter 5 The impact of the obtained personalized word embeddings on intrinsic and extrinsic tasks is explored using large scale review datasets. Review-target identification, sentiment analysis and review text personalization is adopted as the tasks.

Chapter 6 The obtained personalized word embeddings are analyzed in various perspectives. Mainly, analysis in terms of frequency, dissemination, and polysemy of the words with the newly defined metric, "personal semantic variation", is conducted.

Chapter 7 This thesis are summarized.

Chapter 8 Future directions of the thesis are discussed .

Chapter 2

Preliminary

Our proposed method for modeling word meanings by individuals through review-target identification task uses pre-training of the model based on multi-task learning and reviewer-wise fine-tuning of a part of parameters. In this chapter, firstly, basics on how to represent meanings of words with computer is introduced. Second, pre-training and fine-tuning techniques on neural network models are introduced. Third, techniques of multi-task learning for neural networks are introduced.

2.1 Representation Learning for Word Meanings

Modeling the meanings of words, the most important and fundamental elements in natural language, in a form that can be used by a computer is important for natural language processing systems. Therefore, several researchers have studied on how to represent word meanings with computers.

One of the most widely used approach to representing word meanings is to embed words in a high dimensional vector space. These vectors are, thus, called "*word embeddings*."

A word embedding matrix is defined as follows:

$$\boldsymbol{W}_{emb} \in \mathbb{R}^{d \times |V|} \tag{2.1}$$

where $\boldsymbol{d}$ is the dimension size of the embedding vectors, and $\boldsymbol{V}$ is the vocabularies used in a neural network model. As for the $\boldsymbol{i}$-th word $\boldsymbol{w}_i$ in the vocabulary $\boldsymbol{V}$, its word embedding $\boldsymbol{v}(\boldsymbol{w}_i) \in \mathbb{R}^d$ is computed as

$$\boldsymbol{v}(\boldsymbol{w}_i) = \boldsymbol{W}_{emb}([0, 0, ..., 0, 1, 0, ..., 0]) \tag{2.2}$$

where the last vector is called "*one-hot vector*" for the $\boldsymbol{i}$-th word in the vocabulary, in which only the $\boldsymbol{i}$-th dimension represents 1 and the rest 0 for the $\boldsymbol{i}$-th word in the vocabulary. One-hot vector is the most basic concept of the word meanings in natural language processing systems, but there are some problems. For example, relationships between the words can not be calculated due to its sparseness. On the other hand, with a dense word embedding, the relation between the words can be interpreted or calculated using, for example, the cosine similarity. We thus lookup the $\boldsymbol{i}$-th column vector in $\boldsymbol{W}_{emb}$ when implementing the word embeddings in practice.

In most cases, the word embedding matrix $\boldsymbol{W}_{emb}$ is randomly initialized and then tuned with respect to a loss function $\boldsymbol{L}$ of specific tasks: As for the method for the tuning, there are two main streams; (i) unsupervised method based on distributional hypothesis using raw corpora (§ 2.1.1) and (ii) supervised method exploiting various NLP tasks with their annotated corpora (§ 2.1.2).

2.1.1 Unsupervised Modeling of Word Meanings Based on Distributional Hypothesis

One major approach to learn word embeddings is to use large raw text following the distributional hypothesis [15], in which it is assumed that the meaning of a word depends on the context where it is used. There are several methods that follow this hypothesis. The

most widely used methods are, for example, Skip-gram [16] and CBoW [17]. Although there are differences in the variant methods, they all model which words are likely to co-occur in the same context based on this hypothesis. Since all available documents can be used as it is for the training data, it is easy to prepare a large amount of training data. In fact, documents from Wikipedia and Google News articles are used for learning word embeddings in many studies.

Here, **Skip-gram** used in this study is introduced as an example of the method based on the distributional hypothesis. In the learning process of the Skip-gram word embeddings, from a word, the surrounding context words are predicted. This is based on the assumption that the word with the same meaning should be placed around the same words (context).

Here we suppose that $D = \{w_1, w_2, \cdots, w_N\}$ are the list of words in the given raw corpora and w_i is the target word. In skip-gram, we optimize the parameters (e.g., $\boldsymbol{W}_{emb}$) to predict the context words $\{w_{i-C}, w_{i-C+1}, \cdots, w_{i-1}, w_{i+1}, \cdots, w_{i+C-1}, w_{i+C}\}$ within a window size C from the target word $\boldsymbol{w}_i$ maximizing the loss function L defined as follows:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{\substack{|j|\leq C \\ j\neq 0}} \log p(w_{i+j}|w_i) \tag{2.3}$$

where $\log p(w_{i+j}|w_i)$ is computed as follows:

$$p(w_{i+j}|w_i) = \frac{\exp(\boldsymbol{v}(\boldsymbol{w_i}) \cdot \overline{\boldsymbol{v}}(\boldsymbol{w_{i+j}}))}{\sum_{w\in V}\exp(\boldsymbol{v}(\boldsymbol{w_i}) \cdot \overline{\boldsymbol{v}}(\boldsymbol{w}))} \tag{2.4}$$

where $\boldsymbol{v}(\boldsymbol{w})$ and $\overline{\boldsymbol{v}}(\boldsymbol{w})$ is the vectors of word w when w is a target word and a context word. Eq. (2.4) is called "*softmax*" function in which just a multi-class classification task with $|V|$ classes is conducted.

However, it is too expensive to compute Eq. (2.4) exactly when the vocabulary size $|V|$ is huge. Mikolov et al. [16] thus proposed a fast method called negative sampling that

avoids the exact computations of Eq. (2.4):

$$p(w_{i+j}|w_i) = \sigma(v(w_{i+j}) \cdot v(w_i)) \prod_{k=1}^{V'} \sigma(-v(w_k) \cdot v(w_i)) \tag{2.5}$$

where

$$\sigma(x) = \frac{1}{1 + \exp^{-x}} \tag{2.6}$$

is the logistic sigmoid function, and V' is the words of negative samples for each target word. In general, $|V'|$ is much smaller than $|V|$, resulting the low computational cost.

Finally, the loss function is defined as

$$L = -\frac{1}{N} \sum_{i=1}^{N} \sum_{\substack{|j| \leq C \\ j \neq 0}} (\log \sigma(v(w_{i+j}) \cdot v(w_i)) + \sum_{k=1}^{V'} \sigma(-v(w_k) \cdot v(w_i))) \tag{2.7}$$

The obtained word embeddings capture general semantic similarities between the words, depending on their usage in the text used for training.

2.1.2 Supervised Modeling of Word Meanings for NLP Tasks with Annotated Corpora

The co-occurrence-based word embeddings are widely used to initialize word embedding matrices in various neural network models for many other tasks. Unlike the co-occurrence based methods with raw corpora such as Skip-gram just introduced before, supervised method updates the word embedding matrices by minimizing the loss function of the target downstream tasks along with the other parameters of the model. This is based on the assumption that optimum word meanings depends on each task. For example, in sentiment analysis task classifying sentiment polarity, we want to model meanings of positive words (*e.g.*, "*good*") and negative words(*e.g.*, "*bad*") quite differently, but when parsing sentences, we would like to model the meanings of that adjectives as more

similar embeddings. It should be thus useful if we can model task-dependent characteristics for the target downstream tasks in the word embeddings. In this case, since the word embeddings learned with co-occurrence based method like Skip-gram are often used for initializing embedding matrices, the supervised modeling can be regarded as additional step of the training.

2.2 Pre-training and Fine-tuning of Neural Networks

Assuming source and target tasks or source and target domains (the source and target tasks are the same), and performing different optimizations of the model in a source-to-target order, these optimization are called pre-training and fine-tuning, respectively. Pre-training and fine-tuning, however, has different definitions depending on the conditions; when the source task and the target task are different, a series of optimization is called sequential transfer learning, and when the source domain and the target domain are different, it is called domain adaptation. In this section, sequential transfer learning and domain adaptation are described as pre-training and fine-tuning techniques.

2.2.1 Sequential Transfer Learning

Sequential transfer learning is a variant of transfer learning [18]. Transfer learning aims to transfer the knowledge obtained from a source task to a target task (different from the source task). A typical process of transfer learning is to additionally learn a neural network model pre-trained by the source task by using the target task. This learning process is called sequential transfer learning when the additional training is conducted in a source-to-target order. In addition to the purpose of improving the performance of the target task by using the knowledge associated with the source task, transfer learning is often used in various neural network models when training data of the target task is not available enough.

Training for Source Task (pre-training) Here, the source task used for pre-training of the model should be defined. There are no strict rules about what to use for the source task, but in most cases, language modeling task is used. Language modeling task is to predict the next word or character from the target word in a given document. A variant of a language model: Masked Language Model (MLM), which gained attention with the advent of Bert [19], is also one of the most widely used source tasks in recent years.

Of course, the source task does not need to be a language model. It is determined by consulting the relevance to the target task that the model ultimately wants to solve and the amount of data available. Language models are widely used, since language models do not require the annotation of training data.

Training for Target Task (fine-tuning) The simplest fine-tuning method is to optimize all the parameters of the pre-trained model using the target task. However, if the traning data of the target task is too small or the number of parameters of the model is enormous, over-fitting may occur in the fine-tuning process. Therefore, most of the parameters of the model are fixed, and the fine-tuning step is applied only to some task-specific parameters. In most cases, which parameters to be targeted for fine-tuning are manually determined.

2.2.2 Domain Adaptation

Domain adaptation is a technique that aims to make a model perform better in a target domain whose domain is different from the training data of the source task. Most machine learning techniques assume that the domain of the training data is the same as the domain of the testing data. The coverage of domains in available data, however, is limited in many cases. Domain adaptation was developed to addresses such kind of situations that often occur in real-world applications.

One of the most common domain adaptation methods is to use unlabeled text for adapting features used in the model to the target domain. An example is to map the word embeddings learned from the source domain corpora to the target domain word embedding space [20]. The model, therefore, can handle target domain text on the fly.

2.3 Multi-task Learning

Suppose that the main task that ultimately requires improved performance. Multi-task learning technique give the model kind of induction biases by sharing parameters between multiple related tasks (auxiliary tasks). When performing multi-task learning, several different objective functions of the main and auxiliary tasks are optimized simultaneously. This is essentially different from sequential transfer learning. Multi-task learning is used not only in natural language processing but also in various fields such as computer vision and speech recognition.

Although there are no clear guidelines on what tasks should be used for the auxiliary tasks, many multi-task learning techniques assume that each auxiliary task is related with other auxiliary tasks and main task. In addition, it is assumed that the tasks have labeled data.

There are two main streams of the method fot multi-task learning: (i) hard parameter sharing and (ii) soft parameter sharing of the layers in the model. Hard parameter sharing techniques is the most commonly used method in multi-task learning for neural networks. In hard parameter sharing, the hidden layers of the neural-network based model is shared in all the tasks, except for task-specific layers. On the other hand, in soft-parameter sharing techniques, each task has each model independently. The models are trained by using multiple tasks with constraints that the parameter representations of each model should be close with each other.

Chapter 3

Related Work

Existing studies on personalization in natural language processing (NLP) tasks and analysis of semantic variation of word meanings in terms of diachronic, geographic, domain, and debiasing word embeddings for political correctness are introduced. Since existing methods on personalization are mostly aimed at improving accuracy on various tasks, those methods are simultaneously modeling personal variations in word meanings and other irrelevant biases (such as *annotation biases* and *selection biases*) that will contribute to task performances. Next, a few studies that try to understand variations in word meanings in terms of time, geography, and domain are reviewed. Finally, differences between interpersonal semantic variations in word meanings and biases related to unfavorable prejudices are then discussed.

3.1 Personalization in Natural Language Processing

As discussed in § 1, in NLP, personalization attempts to capture three types of user preferences: (1) *semantic variation* in task inputs (biases in how people use words, *i.e.,* the

target of this study) (2) *annotation bias* of output labels (biases in how annotators label),[1] and (3) *selection bias* of output labels (biases in how people choose perspectives (*e.g.,* review targets) that directly affect outputs (*e.g.,* polarity labels)). As for the history of data-driven approaches to various NLP tasks, existing studies have focused more on (2) or (3) aiming at solving target tasks well, particularly in the case of text-generation tasks such as machine translation [9–11], review generation [23] and dialogue systems [4, 5]. Wang et al, for example, jointly generated a review text and a rating score for each reviewer given reviews written to certain product by other users which is un-reviewed by the target user and formulated this task as opinion recommendation. Tang et al, as an example of classification problems, modeled user- and product-level information for the document sentiment classification. By modeling these features within vector space, they achieved state-of-the-art performance. This is because data-driven approaches without personalization tend to suffer from the writer-dependent diversity of probable outputs. Meanwhile, it is difficult to properly separate these facets; therefore, to the author's knowledge, *semantic variations* of words due to differences between people have not been analyzed independently. To understand variation of word meanings among individuals, it is necessary to be able to eliminate these unfavorable and meaning-unrelated biases.

3.2 Interdomain, Diachronic, and Geographical Semantic Varioations

To quantify the semantic variations of common words among domains, Tredici et al. [24] obtained domain-specific word embeddings by using the Skip-gram [16], and they analyzed obtained word embeddings by using multiple metrics such as frequency. Their approach suffers from *annotation biases* since Skip-gram (or language models in general) attempts to predict words in a sentence given the other words in the sentence; therefore,

[1] It is pointed out that NLP datasets are likely to suffer from *annotation bias* [13], whether or not the context of the study is about personaization; models for NLP tasks learn to use or rely on this *annotation bias* when task accuracy is optimized [12, 13, 21, 22].

inputs and outputs are both defined by the same writer. As a result, the same word can have dissimilar embeddings not only because it has different meanings but also because it just appears with words in different topics. [2] In addition, their approach is not scalable in regard to the number of domains (reviewers in this study) since it simultaneously learns all the domain-specific parameters: the number of parameters explodes in proportion to the number of domains.

Semantic variations of word meanings caused by diachronic [25–27], geographic [28, 29], and interdomain variations [24] have been studied. It is known that meanings of more frequent words are more stable over time [25]. Word frequency may therefore affect semantic variations of words among individuals. Besides frequency aspects, word dissemination has been shown to be predictive of semantic variation in word meanings across the community [24] and of changes in frequency of words over time [30]. And it is that the words meanings increases by being used in diverse contexts [31]. Furthermote, It is known that polysemy of the word change the contextual diversity [31, 32] and word meanings faster over the time [25]. Our study analyzes semantic variations in word meanings at the individual level, in particular, focusing on how semantic variations are correlated with word frequency, dissemination, and polysemy.

3.3 Debiasing of Word Embeddings in Terms of Political Correctness

Apart from personal semantic variations, biases related to socially unfavorable prejudices (*e.g.,* the association between the words *receptionist* and *female*) have been identified, analyzed, and removed from word embeddings [33–37]. For example, Bolukbasi et al. [33],

[2] As for two user groups, one of Toyota cars and one of Honda cars, although the meaning of the word "*car*" used in these two groups is likely to be the same, its embedding obtained by the Skip-gram model from the two user groups will differ since "*car*" appears with different sets of words according to each group.

showed that even word embeddings trained on Google news articles wrote by professional journalists have gender stereotypes[3] in their configuration of the vector space (*e.g.,* some occupation words such as *homemaker* and *nurse* are close to the word *she* than to *he*). He argued that such unfavorable biases could be amplified in a wide variety of applications. They proposed a learning algorithm of debiased word embeddings with a constraint that the gender-neutral words are equally close to each element of gender-specific words, while keeping the performance of benchmark test, analogy task. Some studies have addressed Age biases in word embeddings as a typical discrimination different from gender biases. Diaz et al. [35], revealed a age-related biases in a large number of word embeddings and sentiment analysis tools. Acoording to their analysis, sentences with the word *young* are 66% more likely to be scored positively than sentences with *old*. It is also shown that the above bias is alleviated by data pre-processing with simple word replacement.

In these studies, word "biases" were defined in terms of political correctness, so they differ from biases in personalized word embeddings targeted in this study, such as *annotation bias* and *selection bias*. In addition, compared to these political correctness studies that have a correct (desired) answer in their embedding configuration depend on their attributes, there is no correct answer on personalized word embeddings handled in this study, so a constraint using such gender-neutral words cannot be used. This is also a challenge of this study.

[3]Biases that are widely spread and held among people. [33]

Chapter 4

Induction of Interpersonal Semantic Variations in Word Meanings

4.1 Overview

To clarify interpersonal semantic variations in meanings of individual words, the following straightforward approach is taken: word embeddings for each individual person (personalized word embeddings) are learnt via representation learning in an NLP task under the assumption that the words used by individuals are different. To implement this approach, two major problems need to be solved: (i) what kind of tasks should be used to learn personalized word embeddings (§ 4.2) and (ii) how to effectively learn them (§ 4.3).

4.2 Induction Task: Review-target Identification

As for the task of learning personalized word embeddings, if the task is too simple, a distinction between words may not be required, and the resulting word embeddings may be similar or fixed even if those words are semantically irrelevant. In addition, datasets for the task are likely to contain *annotation* and *selection biases*, the induced personalized

product	Asahi Super Dry
category	Alcoholic beverage
user	Daniel C.
time	Oct 11, 2019
rating	4/5
text	*I came to Tokyo for the first time in two years, so I tried this stuff that I often drank at that time. The same brilliant yellow color. Clean and dry aftertaste and slightly sweet. It doesn't have any surprising features, but I like it because it goes well with any dish.*

word embeddings therefore will be contaminated with those biases. In consideration of these issues, review-target identification, in which the review target is estimated given a review text, is adopted as the task of inducing personalized word embeddings.

For a more specific explanation, the following review data is used as an example. Most review datasets usually have *product name or id*, its *metadata or id* (category in this review), *user name or id*, *rating*, *time*, and *review text* for each instance as this example has. The target of review-target identification is *product name or id* (Asahi Super Dry in this review) given a review text. Compared to conventional tasks such as sentiment analysis in which *rating* is estimated given a review text, the review-target identification is highly more difficult due to the large number of classes, so it is necessary to distinguish and understand each word at the time of inducing them. In addition, the fact that no annotator other than writer him/herself is involved when labeling output (review target in this case) allows us to minimize *annotation biases*. Moreover, the number of each review target is one at most in most review datasets. We can therefore suppress *selection biases*, which is the deviation of output label for each person. Due to these characteristics, the model can learn only the meanings of words through the review-target identification task.

4.3 Induction Method

In this section, the methods for inducing personalized word embeddings by using review-target identification are described. First, our main approach and the technical issues that occur when conducting it, *the scalability* and *stability*, is outlined using a simple method as an example in § 4.3.1. Next, a scalable and stable method to address those technical issues is proposed in § 4.3.2.

4.3.1 Approach and Issues

In order to clarify interpersonal semantic variations in word meanings, an approach of computing word embeddings by individuals (personalized word embeddings) by solving the review-target identification task is taken. More concretely, the approach is conducted as follows.

First, a given review of review target $c \in C$, represented as a sequence of words S, is transformed to a sequence of their word embeddings. Here, words written by different reviewers are regarded as different words, *e.g.*, a word $w_i \in V$ used by reviewer $u_j \in U$ and a word $w_i \in V$ used by reviewer $u_k \in U$ are distinguished. As a result, embeddings of the word w_i for each person are also defined as different parameters: $\boldsymbol{e}_{w_i}^{u_j}$ and $\boldsymbol{e}_{w_i}^{u_k}$. The set of review targets, vocabularies, and users are denoted as C, V, and U respectively, and embedding matrix is denoted as $\boldsymbol{W}_{emb} \in \mathbb{R}^{(|U|\times|V|)\times d}$, where d is the dimension size of word embeddings. Next, an neural network model is applied to the sequence of word embeddings and the outputs from the network is used as a semantic representation of a given review. Finally, an output probability distribution of review targets is computed. Through the learning process, embedding matrix $\boldsymbol{W}_{emb}$ along with other parameters is updated to maximize the conditional probability $p(c|S)$ of target c conditioned on a given text S.

However, two technical issues arise when performing this approach straightforwardly. The first one is *the scalability* to the number of reviewers. With the above approach,

as you can see from the explanation, the number of parameters in embedding matrix $\boldsymbol{W}_{emb} \in \mathbb{R}^{(U \times V) \times d}$ increases linearly with the number of reviewers. Since there is a limit to the computational resources, learning may be impossible if the number of target reviewers is increased. The second one is *the stability* in learning process. review-target estimation task has favorable characteristics for eliminating meaning-unrelated biases in learning process, as decribed in § 4.2, but it is an extreme multi-class classification problem. Compared to widely used tasks such as sentiment classification where classifying positive and negative class ($|C| = 2$,) the number of classes of review targets ranges from thousands to tens of thousands in most datasets. Therefore, learning can not be carried out stably. To induce personalized word embeddings through review-target identification, it is necessary to overcome those issues.

4.3.2 Proposed Method

As for effective training of personalized word embeddings, mentioned in § 4.3.1 and § 3, the scalability and stability of the training become two major problems because (i) it is necessary to learn embeddings for words amplified by the number of reviewers, and (ii) the review-target identification task is an extreme multi-class classification with massive review targets. In this section, to solve with these problems, a scalable and stable method for inducing personalized word embeddings is proposed.

In the method, a Long short-term memory (LSTM) [38] network with a residual connection [14] is utilized to obtain personalized word embeddings by using reviewer-specific transformation parameters (§ 4.3.2.1). The personalized word embeddings are obtained by fine-tuning these reviewer-specific parameters per reviewer after pre-training all the parameters (including reviewer-specific transformation parameters) in terms of the scalability in accordance with the number of reviewers (§ 4.3.2.4). Learning of the proposed model is also stabilized by applying multi-task learning with target-attribute predictions when pretraining all the parameters (§ 4.3.2.3).

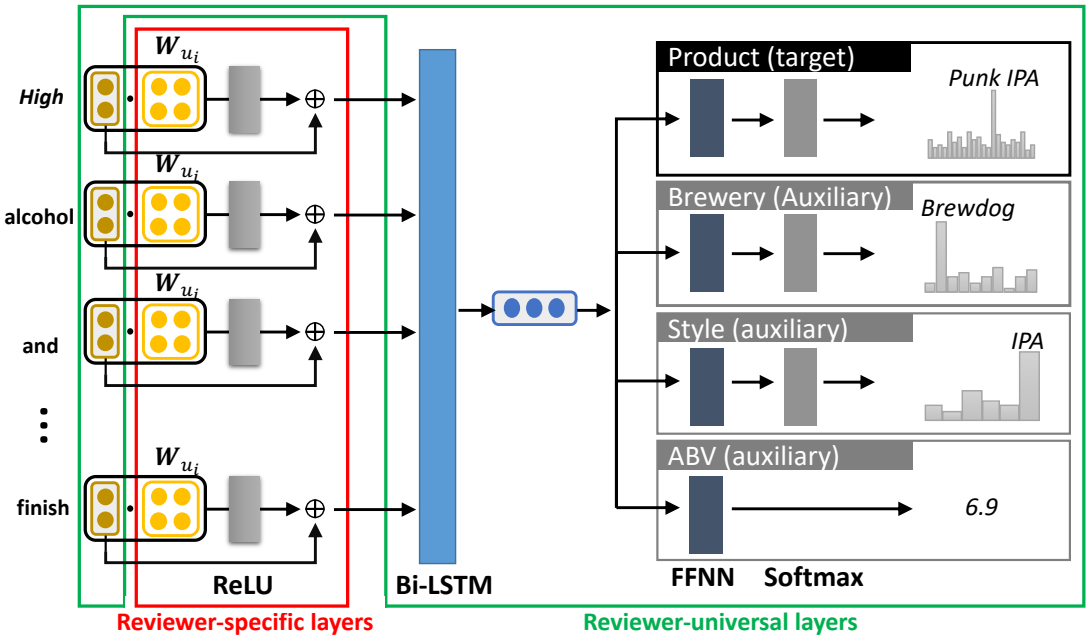


FIGURE 4.1: Overview of the proposed model.
Overview of the proposed LSTM network with residual connection for inducing personalized word embeddings via review-target identification through multi-task learning with target-attribute predictions.

The proposed neural network based model with a residual connection for inducing personalized word embeddings is overviewed in Fig. 4.1. The model consists of reviewer-specific layers and reviewer-universal layers.

4.3.2.1 Reviewer-specific Layers for Personalization

In the layers, the model computes the personalized word embeddings $\boldsymbol{e}_{w_i}^{u_j}$ of each word $w_i \in V$ for each user $u_j \in U$ in input text via a reviewer-specific matrix $\boldsymbol{W}_{u_j} \in \mathbb{R}^{d \times d}$ and bias vector $\boldsymbol{b}_{u_j} \in \mathbb{R}^d$. Here, the set of vocabularies and users is denoted by V and U. Concretely, an input reviewer-universal word embedding $\boldsymbol{e}_{w_i}$ is transformed to a personalized word embedding $\boldsymbol{e}_{w_i}^{u_j}$ as follows:

$$\boldsymbol{e}_{w_i}^{u_j} = \mathrm{ReLU}(\boldsymbol{W}_{u_j}\boldsymbol{e}_{w_i} + \boldsymbol{b}_{u_j}) + \boldsymbol{e}_{w_i} \tag{4.1}$$

where ReLU is a rectified linear unit function.

Compared to the method to increase parameters for $|V|$ word embeddings per person introduced as an example in § 4.3.1, the proposed method increases only a few parameters for $|d|$ ($<< |V|$) word embeddings per person. This representation with a few parameters enhances the scalability of the training. Moreover, As shown in Eq. (4.1), a residual connection inspired by a residual network (ResNet) [14] is used, since semantic variation defined as that from reviewer-universal word embedding. Sharing the reviewer-specific parameters for transformation, $\boldsymbol{W}_{u_j}$ and $\boldsymbol{b}_{u_j}$, across words and employing a residual connection enables the model to stably learn personalized word embeddings even for infrequent words.

4.3.2.2 Reviewer-universal Layers

In this layer, reviewer-universal word embeddings, common elements that constitute personalized word embeddings for each individual, are given to reviewer-specific layers. Subsequently, review-target classification is performed on the later layers.

First, given the transformed personalized word embedding $\boldsymbol{e}_{w_i}^{u_j}$ of each word w_i in an input text, the model encodes them through Long short-term memory (LSTM) [38]. LSTM updates current memory cell $\boldsymbol{c}_t$ and hidden state $\boldsymbol{h}_t$ according to the following equations:

$$\begin{bmatrix} \boldsymbol{i}_t \\ \boldsymbol{f}_t \\ \boldsymbol{o}_t \\ \hat{\boldsymbol{c}}_t \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{bmatrix} \boldsymbol{W}_{\text{LSTM}} \cdot \left[\boldsymbol{h}_{t-1}; \boldsymbol{e}_{w_i}^{u_j}\right] \tag{4.2}$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \hat{\boldsymbol{c}}_t \tag{4.3}$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t \odot \tanh\left(\boldsymbol{c}_t\right) \tag{4.4}$$

where $\boldsymbol{i}_t$, $\boldsymbol{f}_t$, and $\boldsymbol{o}_t$ are the input, forget, and output gate at time step t, respectively. $\boldsymbol{e}_{w_i}^{u_j}$ is an input personalized word embedding at time step t, and $\boldsymbol{W}_{\text{LSTM}}$ is a weight matrix of the LSTM network. $\hat{\boldsymbol{c}}_t$ is the current cell state. Operation $\odot$ denotes element-wise

multiplication and σ is the logistic sigmoid function. In the proposed model, single-layer Bi-directional LSTM (BI-LSTM), a variant of LSTM, is adopted as the encoding layer to utilize past and future context. LSTM is a special kind of recurrent neural networks (RNNs), in which LSTM uses gate functions at each time step to retain sequential information and thus keeps the long-range dependencies of input sequences. As a successor, Bi-LSTM was proposed to capture not only past context at time step t but also future context.

Second, as the representation of the input text $\boldsymbol{h}$, BI-LSTM concatenates the outputs from the forward and backward LSTMs as follows:

$$\boldsymbol{h} = \left[\overrightarrow{\boldsymbol{h}_{L-1}}; \overleftarrow{\boldsymbol{h}_0}\right] \tag{4.5}$$

Here, L denotes the length of the input text, and $\overrightarrow{\boldsymbol{h}_{L-1}}$ and $\overleftarrow{\boldsymbol{h}_0}$ denote the outputs from the forward and backward LSTM at the last time step, respectively.

Lastly, a feed-forward layer computes output-probability distribution $\hat{\mathbf{y}}$ of review targets from the representation $\boldsymbol{h}$ as follows:

$$\hat{\mathbf{y}} = \mathrm{softmax}\left(\boldsymbol{W}_o \boldsymbol{h} + \boldsymbol{b}_o\right) \tag{4.6}$$

where $\boldsymbol{W}_o$ is the weight matrix and $\boldsymbol{b}_o$ is the bias vector for the last layer.

4.3.2.3 Multi-task Learning with Target-Attribute Predictions for Stable Training

As mentioned in § 4.3.1, training the model for the target identification task is considered to be unstable because its output space (review targets) is extremely large (thousands or tens of thousands more). To mitigate this instability, auxiliary tasks that estimate attributes of the target item were set up and solved simultaneously with the review-target identification task (target task) by multi-task learning. This approach is motivated from the assumption that understanding those related metadata of the target item will contribute to the accuracy of identifying the review target. Moreover, supervision can be increased

by solving relatively easy sub-problems; attributes of the review item are more coarse-grained than the review item itself.

Specifically, independent feed-forward layers are added and used to compute output probabilities for each auxiliary task from shared sentence representation $\boldsymbol{h}$ defined by Eq. (4.5) (Fig. 4.1) as follows:

$$\hat{\mathbf{y}_{\mathbf{o_k}}} = \mathrm{softmax}\left(\boldsymbol{W}_{o_k}\boldsymbol{h} + \boldsymbol{b}_{o_k}\right) \tag{4.7}$$

where o_k denotes one of the tasks used for multi-task learning and $\hat{\mathbf{y}_{\mathbf{o_k}}}$ denotes the output probability of a task. In this thesis, three types of auxiliary tasks are assumed: (i) multi-class classification (the same as the target task), (ii) multi-label classification, and (ii) regression. Cross-entropy loss is used for multi-class classification, a summation of cross-entropy loss of each class is used for multi-label classification and mean-square loss for regression. Finally, optimization of parameters under a loss that sums up individual losses for the target and auxiliary tasks is performed.

All reviewer-universal parameters are subjected to multi-task learning only during the pre-training without personalization. The model, therefore, uses reviewer-universal parameters $\boldsymbol{W}$ and $\boldsymbol{b}$ (instead of $\boldsymbol{W}_{u_j}$ and $\boldsymbol{b}_{u_j}$) for Eq. (4.1), and it then initializes the reviewer-specific parameters $\boldsymbol{W}_{u_j}$ and $\boldsymbol{b}_{u_j}$ by using $\boldsymbol{W}$ and $\boldsymbol{b}$ in fine-tuning step (§ § 4.3.2.4).

4.3.2.4 Fine-tuning for Scalable Training

Under the assumption that the number of reviewers is enormous, it is impractical to simultaneously train the reviewer-specific parameters of all the reviewers due to memory limitation (even if the number of parameters per person could be reduced by our model configuration). Therefore, fine-tuning of pre-trained parameters (§ § 4.3.2.3) is then applied only to reviewer-specific parameters by training independent models on the basis of the reviews written by each reviewer.

In the fine-tuning step, reviewer-specific parameters $\boldsymbol{W}_{u_j}$ and $\boldsymbol{b}_{u_j}$ of the pre-trained model are tuned while the target task (review-target identification task) only is optimized. Concretely, the reviewer-specific parameters $\boldsymbol{W}_{u_j}$ and $\boldsymbol{b}_{u_j}$ are initialized by using $\boldsymbol{W}$ and $\boldsymbol{b}$

pretrained in § 4.3.2.3. All the other parameters (reviewer-universal parameters) are fixed at the time of fine-tuning. This approach makes the model scalable even to a large number of reviewers. In addition, this fixing stops the model introducing *selection bias* into the personalized embeddings; otherwise, as discussed in § 4.2, the prior output distribution of the auxiliary tasks for individual person can be implicitly learned.

Chapter 5

Evaluation

5.1 Overview

In this chapter, evaluation of the personalized word embeddings by the review-target identification task using two review datasets are conducted. If the model can successfully solve this objective task more accurately than the reviewer-universal model obtained by pre-taining of the proposed reviewer-specific model, it is considered that those personalized word embeddings capture the interpersonal semantic variations of input words. Next, to verify the usefulness of the personalized word embeddings in regard to not only an intrinsic task but also an extrinsic task, the proposed model is applied to solve sentiment analysis task and review text personalization task. Lastly, a summary of the evaluation results is described.

5.2 Settings

In this section, settings of evaluating the personalized word embeddings obtained by the proposed method are detailed. First, two datasets commonly used throughout the

TABLE 5.1: Dataset statistics.

datasets	RateBeer dataset		Yelp dataset	
# reviews		2,695,615		426,816
# reviewers		3,670		2,414
# review targets	beer	109,912	service	56,574
# target attributes	style	89	location	19
	brewery	6,870	category	683
	ABV	20		

evaluation are summarized. Next, the task details are described. Lastly, the model and implementation details for each task is described.

5.2.1 Datasets

Two datasets containing reviews of beer and services related to foods were adopted for evaluating the proposed method, since there are a variety of expressions that describe what people have sensed with various sensory units in these domains of the datasets. Table 5.1 sumarizes statistics of the two datasets.

The RateBeer dataset, which includes reviews of a variety of beers, was extracted from RateBeer[1] [39]. Written by reviewers who posted at least 100 reviews, 2,695,615 reviews about 109,912 types of beer were selected. All the extracted reviews are containing the metadata about "style", "brewery", and "alcohol by volume (ABV)". The Yelp dataset, which includes reviews of a diverse range of services, was derived from yelp.com.[2] The selected reviews were (1) those containing location metadata, (2) those falling under either the "food" or "restaurant" categories, and (3) those written by a reviewer who posted at least 100 reviews. As a result, 426,816 reviews of 56,574 services (restaurants or foods) written by 2,414 reviewers in total were extracted.

[1] https://www.ratebeer.com
[2] https://www.yelp.com/dataset

These two datasets were randomly divided into training, development, and testing sets with the ratio of 8:1:1. Hereafter, the former is referred to as **RateBeer dataset** and the latter as **Yelp dataset**. Both datasets are used for target task and sentiment analysis.

Regarding review text personalization, we further extracted the data of the top-100 reviewer with the most reviews in each of training, development, and testing sets of RateBeer datasets, and created 1,266,958 (training), 19,382 (development), and 19,432 (testing) pairs of sentences referring to the same review-target. At that time, sentences exceeding 50 words were excluded. Yelp dataset may contain the reviews mentioning the foods offered by the service while also mentioning the service too. Therefore, in the review text personalization, the RateBeer dataset only is used, where each review seems to refer to an object of the same scale (beer).

5.2.2 Tasks

Three tasks are used to evaluate the personalized word embeddings. First, (i) review-target identification task is utilized as the target task. Next, (ii) sentiment analysis and (iii) review text personalization task are utilized to show the application possibility of the induced personalized word embeddings to extrinsic NLP tasks. Task information is summarized in Table 5.2.

Review-target Identification Target task takes a review as an input and estimates target beer for RateBeer dataset or services for Yelp dataset reviewed in it. Regarding the target attributes for multi-task learning (MTL), style and brewery were chosen for multi-class classification, and alcohol by volume (ABV) was chosen for regression in the experiments with RateBeer dataset. As for MTL with the Yelp dataset, location was used for multi-class classification, and category was used for multi-label classification.

Sentiment Analysis The sentiment analysis task also takes a review as an input and estimate ratings of given reviews annotated by the reviewers themselves. The ratings are integers and range from 1 to 20 in RateBeer dataset and from 1 to 5 in Yelp dataset. This

TABLE 5.2: Overview of task information.

(A) **RateBeer dataset**

task	output	type	metric
review-target identification	beer	classification	accuracy
target-attribute prediction	style	classification	accuracy
	brewery	classification	accuracy
	ABV	regression	RMSE
sentiment analysis	rating	regression	RMSE
review text personalization	sentence	text generation	BLEU

(B) **Yelp dataset**

task	output	type	metric
review-target identification	service	classification	accuracy
target-attribute prediction	location	classification	accuracy
	category	multi-label classification	micro-F1
sentiment analysis	rating	regression	RMSE

task was solved as a regression task since it is natural to treat the fine-grained ratings as continuous values the discrete classes.

Review Text Personalization As for the review text personalization task, the model basically takes a review and id of a reviewer (target-reviewer) as an input, and estimates the sentence the target-reviewer wrote. This task was solved as text generation task.

Throughout all the tasks, accuracy was used for metrics of classification and root mean square loss (RMSE) was used for metric of regression tasks. For multi-label classification, micro-F1 score that is a weighted average of precision and recall was used as evaluation metric. As for the text generation, BLEU score [40], an automatic evaluation index for machine translation, was used as the metric.

TABLE 5.3: Hyperparameters in review-target identification and sentiment analysis.

Model		Optimization	
Dimensions of hidden layer	200	Dropout rate	0.2
Dimensions of word embeddings	200	Algorithm	Adam
Vocabulary size (RateBeer dataset)	59,653	Learning rate	0.001
Vocabulary size (Yelp dataset)	42,412	Batch size	200

5.2.3 Models and Hyperparameters

Model configuration, baseline, and hyperparameters are explained for each task. The hyperparameters in the sentiment analysis task are the same as those in the review-target identification task.

5.2.3.1 Review-target Identification

Models As for the target item and attribute identification tasks, the proposed model (described in 4.3.2) was evaluated in terms of four different settings.[3] The differences of the models are (1) whether fine-tuning for personalization is applied and (2) whether the model is trained through MTL before the fine-tuning.

Baseline In the review-target identification and each of attribute prediction tasks, the baseline model selects the class (classification) that appears most frequently in the training data or average value (regression) of training data is adopted.

Hyperparameters Table 5.3 lists major hyperparameters. The embedding layer was initialized by Skip-gram embeddings [16] pretrained using review text of training and validation sets of each dataset. The vocabulary for each dataset includes all the words that appeared 10 times or more in the dataset. For optimization, the models were trained up to 100 epochs with Adam [41], and the model at the epoch with the best results in the target task on the development set was selected as the test model.

[3] All models were implemented by using PyTorch (https://pytorch.org/) version 1.2.0.

5.2.3.2 Sentiment Analysis

Models As for the sentiment analysis task for extrinsically evaluating the obtained personalized word embeddings, another set of models with the same architecture and hyperparameters was trained as review-target identification models in Fig. 4.1 (except that they have only one feed-forward layer for the regression of sentiment ratings.) The embedding layers of the models are kept fixed after being initialized by the personalized word embeddings obtained from the corresponding review-target identification models with the same settings of personalization and MTL.

Baseline In the sentiment analysis task, the model that selects average value in the training data as the estimation value is adopted.

5.2.3.3 Review Text Personalization

Models In the review text personalization task, another task for extrinsically evaluating the obtained personalized word embeddings, a token representing the target-reviewer (writer of the target sentence) is always connected to the beginning of the input sentence. This token is converted into a vector representation with the same size as that of a personalized word embedding[4]. Two models based on 1-layer bidirectional RNN encoder-decoder model with attention architecture [43] are proposed. The first one is (1) the model in which the embedding layer of the encoder is replaced with the personalized word embeddings obtained by the proposed method. The second one is (2) the model in which the embedding layers of both the encoder and decoder are replaced with the personalized word embeddings. Both models can take the semantic representation of the input sentence expressed with personalized word embeddings as an input. As for the non-personalized embedding layers, the Skip-gram embeddings are used that are also used for the initialization of the review-target identification model. Throughout the optimization, the embedding layers of encoder and decoder are kept fixed.

[4]It is inspired by the implementation of the Google Neural Machine Translation Model [42].

TABLE 5.4: Hyperparameters in review text personalization.

Model		Optimization	
Dimensions of hidden layer (Encoder)	128	Dropout rate	0.5
Dimensions of hidden layer (Decoder)	128	Algorithm	Adam
Dimensions of word embeddings	200	Learning rate	0.003
Dimensions of reviewer token embeddings	200	Batch size	128
Vocabulary size	59,653	Teacher forcing ratio	0.5

Baselines As for the baseline models, a token representation of the target-reviewer is also connected to the beginning of the input sentence. There are two baselines for confirming the impact of personalized word embeddings in review text personalization task. The first one is (1) the model whose embedding layers of the encoder and decoder was the Skip-gram embeddings (not the personalized word embeddings). This model is set as a baseline that can not know the personalized word meanings of the input sentence. As for the second one, (2) the model that copy the input sentence and paste as the output sentence is adopted. This is set to confirm the difficulty of review text personalization when neither writer of input nor writer of output is known.

Hyperparameters Table 5.4 shows the major hyperparameters. The vocabulary set is the same as that of the review-target identification models. For optimization, the models were trained up to 50 epochs with Adam [41], and the model at the epoch with the best BLEU score on the development set was selected as the testing model.

5.3 Results

In this section, the evaluation results for each task are shown. As for the review-target identification, evaluation on the number of reviews used for personalization was also conducted.

TABLE 5.5: Results of the review-target identification task using the RateBeer dataset and Yelp dataset. Accuracy and RMSE marked with $**$ or $*$ was significantly higher than that of the other models ($p < 0.01$ or $0.01 < p \leq 0.05$ assessed by paired t-test for accuracy and z-test for RMSE).

(A) **RateBeer dataset**

model		target task	auxiliary tasks		
multi-task	personalize	product [Acc.(%)]	brewery [Acc.(%)]	style [Acc.(%)]	ABV [RMSE]
		15.76	n/a	n/a	n/a
	✓	16.71	n/a	n/a	n/a
✓		16.18	(19.83)	(49.26)	(1.415)
✓	✓	**17.53****	(**20.64****)	(**50.07****)	(**1.399***)
baseline		0.08	1.51	6.19	2.321

(B) **Yelp dataset**

model		target task	auxiliary tasks	
multi-task	personalize	service [Acc.(%)]	location [Acc.(%)]	category [Micro F1]
		6.50	n/a	n/a
	✓	6.83	n/a	n/a
✓		8.15	(70.61)	(**0.567**)
✓	✓	**9.11****	(**83.02****)	(0.563)
baseline		0.05	27.00	0.315

5.3.1 Evaluating Personalized Word Embeddings by Review-target Identification

Table 5.5 lists results of the review-target identification task using the two datasets. It can be inferred from these results that (1) as for the target task, the model with both MTL and personalization outperformed the others and (2) personalization also improved the performance of auxiliary tasks.

The model without personalization assumes that the same words written by different

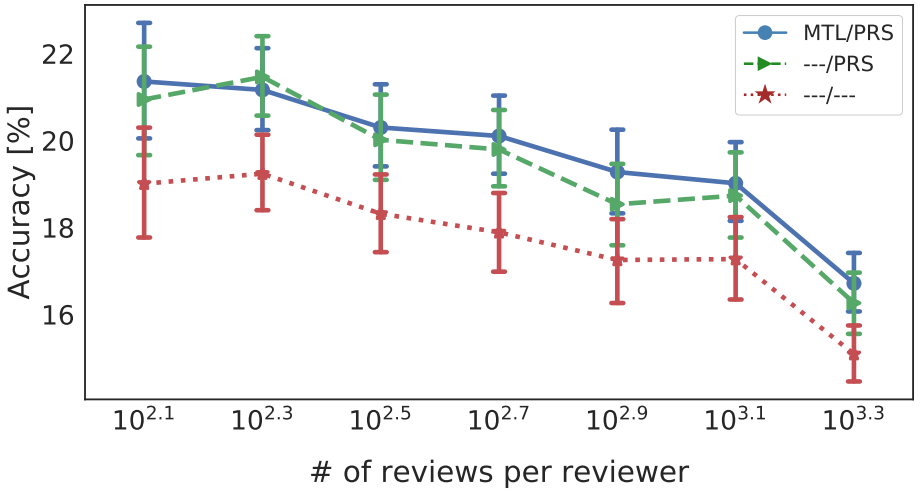
reviewers have the same meanings, while the model with personalization distinguishes them. The improvement by personalization on the target task with objective outputs partly supports the fact that the same words written by different reviewers have different meanings, even though they are in the same domain (beer, restaurant, and food). Simultaneously solving the auxiliary tasks that estimate attributes of the target item guided the model to understand the target item from various perspectives, like part-of-speech tags of words.

It should be mentioned here that only the reviewer-specific parameters were updated on the target task by using fine-tuning. This means that the improvements of the performance on auxiliary tasks were obtained purely by the semantic variations captured by reviewer-specific parameters.

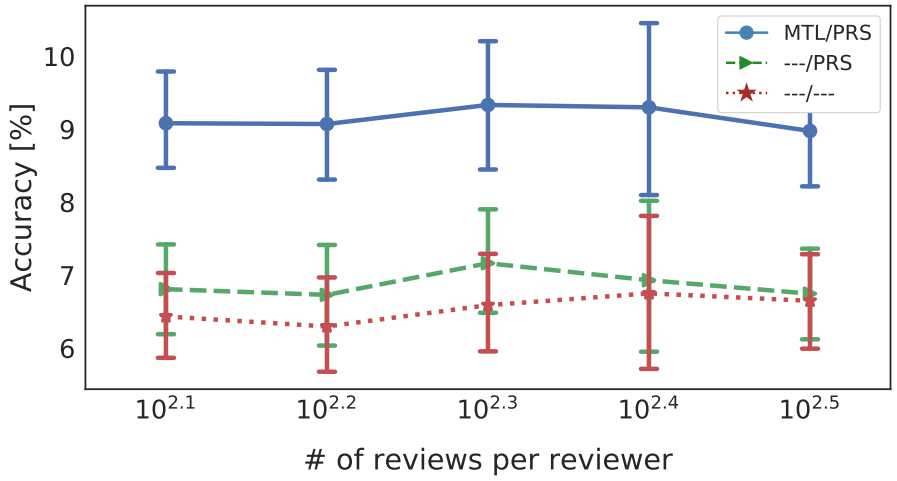
5.3.2 Impact of the Number of Reviews for Personalization

The impact of number of reviews for personalization when solving the review-target identification problem was investigated. The reviewers were first grouped into several bins according to number of reviews. Classification accuracies for reviews written by the reviewers in the same bin were then evaluated. Classification accuracy of the target task is plotted against number of reviews per reviewer in Fig. 5.1. For example, the plots (and error bars) for $10^{2.3}$ represent the accuracy (variation) of the target identification for reviews written by each reviewer with n reviews ($10^{2.1} \leq n < 10^{2.3}$).

Contrary to our expectation, as for the RateBeer dataset (Fig. 5.1 (a)), all models obtained lower accuracies as number of reviews increased. On the contrary, as for the Yelp dataset (Fig. 5.1 (b)), the performance of the models did not deteriorate as number of reviews increased. We consider that this difference is due to the biases of frequencies in the review targets. Since the RateBeer dataset is heavily skewed, the top-10% frequent beers account for 74.3% of the entire reviews, while the top-10% frequent restaurants in the Yelp dataset account for 48.0% of the reviews. Therefore, it is more difficult to estimate infrequent targets in the RateBeer dataset, and such reviews tend to be written by



(A) **RateBeer dataset**



(B) **Yelp dataset**

FIGURE 5.1: Accuracies of target identification task against the number of reviews per reviewer. In the legend, **MTL** and **PRS** stands for multi-task learning and personalization.

the experienced reviewers. Although the model without MTL and personalization also obtained slightly lower accuracies, even in the case of the Yelp dataset, the model with both MTL and personalization successfully exploited the increased reviews and obtained higher accuracies.

TABLE 5.6: Results of sentiment analysis: embedding layers are kept fixed to those of the corresponding models in Table 5.5. RMSE marked with (i) + is significantly better than the model without multi-task and personalization on the RateBeer dataset ($p < 0.05$ assessed by z-test), and (ii) # is significantly better than the model without multi-task and personalization and the model with personalization on the Yelp dataset ($p < 0.05$ assessed by z-test).

		RateBeer dataset	Yelp dataset
model		sentiment analysis	
multi-task	personalize	rating [RMSE]	
		1.729	0.683
	✓	1.645	0.665
✓		1.726	0.655
✓	✓	**1.622**$^{+}$	**0.631**$^{\#}$
baseline		3.239	1.046

5.3.3 Evaluating Personalized Word Embeddings by Sentiment Analysis

Table 5.6 lists results of the extrinsic evaluation of the obtained personalized word embeddings on the sentiment analysis task. Similar to the results of the review-target identification task, the result obtained by the proposed model with both multi-task and personalization outperformed those of the other models. These results confirm that the personalized word embeddings trained through the proposed method successfully learned task-independent personal semantic variations in word meanings. In other words, it is even helpful for solving tasks other than the review-target identification used to obtain the personalized word embeddings.

In addition, to confirm whether the personalized word embeddings obtained by the proposed method could remove the biases unrelated to the meanings, the performances of models with different tasks used for personalization were compared by using sentiment analysis. To compare with the personalized word embeddings obtained by the proposed model via the proposed task: review-target identification, personalized word embeddings were also obtained by using auxiliary tasks considered to be affected by *selection bias*

TABLE 5.7: Comparison of sentiment analysis results for different tasks used for personalization with the RateBeer dataset and the Yelp dataset. Embedding layers are kept fixed after personalization on each task. Proposed target identification task is beer and service in each dataset respectively.

(A) **RateBeer dataset**

multi-task	personalization task	sentiment analysis rating [RMSE]
	style	1.668
✓	style	1.657
	brewery	1.634
✓	brewery	1.633
	ABV	1.679
✓	ABV	1.678
	beer	1.645
✓	beer	**1.622**
baseline		3.239

(B) **Yelp dataset**

multi-task	personalization task	sentiment analysis rating [RMSE]
	location	0.650
✓	location	0.647
	category	0.662
✓	category	0.658
	service	0.665
✓	service	**0.631**
baseline		1.046

(because the same output label appears multiple times in an individual person's training data).

Table 5.7 shows that the embeddings obtained by the proposed method achieved the best performances. That result suggests that the proposed method can suppress the meaning-unrelated biases and obtain task-independent word meanings.

TABLE 5.8: Results of review text personalization in RateBeer dataset.

personalization encoder	personalization decoder	BLEU
		17.65
✓		17.84
✓	✓	**17.87**
baseline		3.87

5.3.4 Evaluating Personalized Word Embeddings by Review Text Personalization

Table 5.8 shows the results of the extrinsic evaluation of the obtained personalized word embeddings on review text personalization task. Baseline represents a model that copies the input sentence.

First of all, from the results of the baseline, it can be seen that the review description varies from person to person even for the same review target, and review text personalization task is difficult when neither the writer of the input sentence nor the writer of the output sentence.

Next, when the embedding layer of the encoder is changed to personalized word embeddings, the performance was improved. This result suggests that in order for the model to understand descriptions that varies by person, the meanings of input words should be represented depending on individual reviewers.

In addition, replacing the embedding layers of both the encoder and decoder with the obtained personalized word embeddings outperformed the others. The result indicates that, in order to generate sentences that are more personalized to the target-reviwer, it is necessary to pay attention to the word meanings the target-reviewer has. There is no much difference, however, between the model with the personalized word embeddings only in encoder and the model with the personalized word embeddings in both encoder

and decoder. It is considered that, if the input sentence can be represented by personalized word meanings, the reviewer token representing the target reviewer may be a strong enough to decode personalized sentence.

5.3.5 Summary of Evaluation Results

In evaluation on the review-target identification (§ 5.3.1), the proposed the model with personalized word embeddings induced by our method with multi-task learning and fine-tuning outperformed the other models. The result indicated that the personalized word meanings can be successfully learned by the proposed model. In exploration of the impact of the number of reviews for personalization (§ 5.3.2), It was shown that models obtained lower accuracies as the number of reviews increased in skewed dataset. In exploration of application of the obtained word embeedings, performence on the sentiment analysis and review text personalization were improved by using the personalized word embeddings (§ 5.3.3, § 5.3.4). The results also indicated that the personalized word embeddings obtained by the proposed method could remove task dependent biases.

Chapter 6

Analysis

6.1 Overview

In this chapter, personalized word embeddings obtained by the proposed method are analyzed in two ways: (i) on the correlation between real-world values and the composition of the word embedding space (§ 6.2) and (ii) with the uniquely defined metric: personal semantic variation (§ 6.3). As for the former, it is confirmed whether the reviewers who mentioned “bitter“ for the similar bitter beers have similar word embeddings of “bitter”. As for the latter, the degree and tendencies of the semantic variation in the obtained personalized word embeddings are analyzed from the same perspectives as discussed in previous studies on interdomain and diachronic semantic variations in word meanings.

6.2 Correlation between Acquired Personalized Word Embeddings and Real-world Values

In this section, whether the personalized word embeddings obtained by the proposed method can express the personal preference that can be actually observed in the embedding space was evaluated. Thirty reviewers with the most reviews in the RateBeer dataset were targeted. In addition, the International Bittering Units scale (IBU), which represents the bitterness of beer were utilized as values that can actually be observed. Furthermore, "*bitter*" was defined as the word representing IBU. As a basic policy, the median IBU value of beer with "*bitter*" written in the reviews was calculated by individuals, and cosine distance of words "*bitter*" of all combinations of pairs of reviewers from 30 people was also calculated. Correlation coefficient between differences of the median on IBU and the cosine distance of word "*bitter*" were were calculated for evaluation.

Since IBU values do not exist in the RateBeer dataset originally, data for the 100 most-popular beers that IBU can be acquired on the web were prepared. Reviews containing adverbs such as "*very*" and "*moderately*" just before the target word "*bitter*" were excluded from the calculation of the median of IBU value. As a comparison with the personalized word embeddings obtained by the proposed method, word embeddings personalized through auxiliary tasks were prepared. All the models are pretrained with multi-task learning.

The results are listed in 6.1. In terms of the value of the correlation coefficient in itself, personalized word embeddings of "*bitter*" obtained through the target identification task seem to be better overall than the other embeddings. However, the p-value shows that there is no significant difference even from the uncorrelated case. This result suggests that learning might not be sufficient to drastically change the embedding configuration. In addition, the fact that all reviewers do not necessarily express bitterness as "*bitter*" and that documents using "*bitter*" with an excluded adverb are also used for learning personalized embeddings may also affect the results of weak correlation. Improvement and invention of effective analysis method of this part is a future work.

TABLE 6.1: Correlation between difference of the median of international bitterness unit (IBU) of the review target and cosine distance between personalized word embeddings of *bitter* of all the combinations of the thirty reviewers.

personalization task	spearman's rank correlation coefficient	kendall's rank correlation coefficient	pearson correlation coefficient
style	-0.1007	-0.1497	-0.1558
brewery	-0.0649	-0.0970	-0.1244
ABV	0.0140	**0.0225**	-0.0178
beer	**0.0143**	0.0209	**0.0107**

6.3 Personal Semantic Variation in Word Meanings

In this section, the obtained personalized word embeddings were analyzed to determine what kind of personal biases exist in each word. Proposed metric, personal semantic variation is first defined and its relatedness with three perspectives are shown.

Personal semantic variation[1] of a word w_i is first defined to determine how the representations of the word differ for each individual as

$$\frac{1}{|\mathrm{U}(w_i)|} \sum_{u_j \in \mathrm{U}(w_i)} (1 - \cos(\boldsymbol{e}_{w_i}^{u_j}, \overline{\boldsymbol{e}}_{w_i})) \tag{6.1}$$

where $\boldsymbol{e}_{w_i}^{u_j}$ is the personalized word embedding to w_i of a reviewer u_j, $\overline{\boldsymbol{e}}_{w_i}$ is the average of $\boldsymbol{e}_{w_i}^{u_j}$ for $\mathrm{U}(w_i)$, and $\mathrm{U}(w_i)$ is the set of the reviewers who used the word w_i at least once in training data. Here, to remove the influences of low-frequent words, only words used by 30% or more reviewers (excluding stop words) were targeted.

[1] Unlike the definition of the semantic variation in existing studies [24], which measure the degree of change from a domain to a domain of a word meaning, personal semantic variation measures how much a number of meanings of a word defined by individuals are diverged.

6.3.1 Analysis on Three Perspectives

Three perspectives are focused on: **frequency**, **dissemination**, and **polysemy**, which have been discussed in the studies on semantic variations caused by diachronic or domain differences of text [24, 25, 29].

Frequency In this thesis, frequency of each word was computed as base 10-logarithm of frequency counted across training, validation, testing sets of the review-target identification task (Table 5.1):

$$\mathrm{Frequency(w_i)} = \log_{10} \mathrm{N_{w_i}} \tag{6.2}$$

where N_{w_i} denotes the count of word w_i in the dataset.

dissemination Dissemination of each word is calculated as the ratio of the reviewer who used the word in the dataset:

$$\mathrm{Dissemination(w_i)} = \mathrm{U_{w_i}/U} \tag{6.3}$$

where U_{w_i} is the number of users who used the word w_i and U is the total number of users in the dataset.

Polysemy In this thesis, polysemy of each word is represented as the number of synsets found in WordNet [44] ver. 3.0. In WordNet, meanings of each word is annotated and recorded as a database.

Fig. 6.1, Fig. 6.2, and Fig. 6.3 show semantic variations against the three metrics. Each x-axis corresponds to frequency (Fig. 6.1), dissemination (Fig. 6.2), and polysemy (Fig. 6.3), respectively. Interestingly, in contrast to the reports by [25] on diachronic semantic variations but consistently to reports by [24] on interdomain semantic variations, semantic variations correlate highly with frequency and dissemination but poorly with polysemy in our results. This tendency of interpersonal semantic variations can be explained as follows. In the datasets used in our experiments, words related to the five senses, such as "*soft*" and "*creamy*," frequently appear, and their usage depends on feelings and experiences by individuals. Therefore, their meanings show high semantic variations. As

for polysemy, although the semantic variations might change the degree or nuance of the word sense, they do not change its synset. This is because those words are still used only in skewed contexts related to food and drink where word senses do not fluctuate significantly.

6.3.2 Example Study

Table 6.2 lists the top-50 (and bottom-50) words with the largest (and smallest) semantic variations. As can be seen from the tables, the list of the top-50 words contains many more adjectives (50% and 38% on the RateBeer and Yelp dataset, respectively) than the list of the bottom-50 words (22% and 14% on the RateBeer and Yelp dataset) , which are likely to be used to represent individual feelings that depend on the five senses.

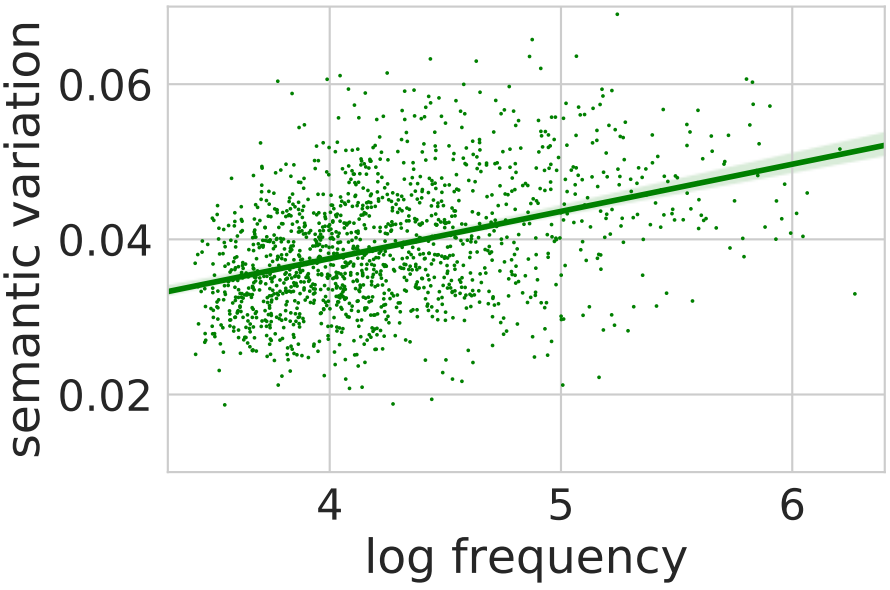
To determine what kind of words have large semantic variations, the adjectives of the top-50 (and bottom-50) were classified by the five senses, which are **sight** (vision), **hearing** (audition), **taste** (gustation), **smell** (olfaction), and **touch** (somatosensation). From the results, as for the RateBeer dataset, in the top-50 words, more words are representing each sense (except hearing) than the bottom-50 words. On the contrary, the list of words on the Yelp dataset include less words related to the five senses than the RateBeer dataset; however, many adjectives that could be applicable to various domains (*e.g.,* "*great*," and "*excellent*") are included. This result may be due to the domain size and the lack of reviews detailing specific products in the restaurant reviews contained in the Yelp dataset; reviews in RateBeer dataset describe each product (beer), while Yelp dataset, which consists of reviews related to product (food), talks about services that provide products such as restaurants.

6.4 Visualizing Personalized Word Embeddings

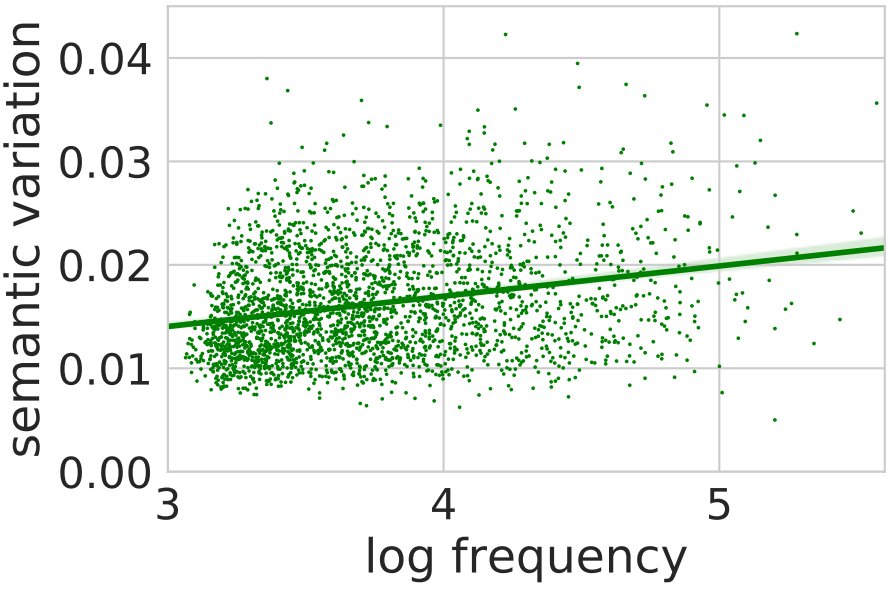
Whether some words get confused was also analyzed. The adjective words "*grassy*" and "*great*" with large semantic variations in each dataset were used as an example. Personalized word embeddings were visualized using Principal Component Analysis (PCA), with the nine adjective words closest to the target words in the universal embedding space in Fig. 6.4. As can be seen, clusters of "*grainy*," "*bready*," and "*doughy*" in the RateBeer dataset and "*awesome*" and "*excellent*" in the Yelp dataset are mixed each other, suggesting that words representing the same meaning may differ for each individuals.

TABLE 6.2: The list of top-50 (and bottom-50) words with the largest (and the smallest) semantic variation in the RateBeer and Yelp datasets. Adjectives are boldfaced.

	top-50	**bottom-50**
RateBeer dataset	**deep** grass **grassy lingering soapy toasty bready** tobacco underneath pours **pleasing** ery **medium** mildly **subtle** underlying hints dough lots **subdued sharp** mainly ark **updated tangy** resin **bright** hue **flowery fairly good rich upfront nice** crisp **dusty** toffee **creamy** kind citrus zest **citrusy** profile presence hay **earthy** aromas dominated toast **doughy**	dogfish batch reminds **course** needs bells cask rye **hot** ask honey unlike reminded raspberry **canned** packs liquor hand **barley** stone rogue maple never horse line rice bourbon minute **belgium** raspberries dog heat bomb **mexican triple** rock difference **scottish** coconut ton **burning** dead **organic** bock brewing dubbel **pink** missing becoming champagne
Yelp dataset	**great fantastic excellent superb amazing awesome phenomenal tasty delish good delicious yummy** sides sauce **nice incredible**t flatbread entrees **outstanding wonderful** appetizers desserts **fabulous** ambiance chicken atmosphere rice salmon ambience **flavorful** patio sauces risotto dishes sausage chorizo went items garlic sandwiches veggies cabbage decor **ordered** asparagus pistachio sandwich stopped restaurant calamari	note nearly aside easily eye single possibly almost together mark exact warning **major alone** even lack zero **opposite** wish somehow saving **short** changing **apart** practically yet thus ends replaced part deciding handful thumbs hardly desired rather except **enough** c favor meaning none hearing via meant reading b ups **biggest** iron



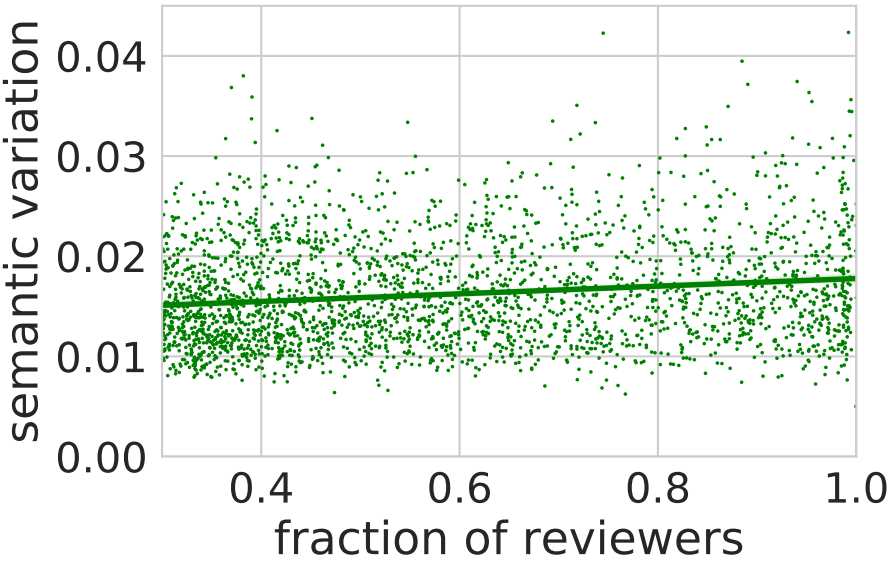
(A) log-frequency at RateBeer dataset



(B) log-frequency at Yelp dataset

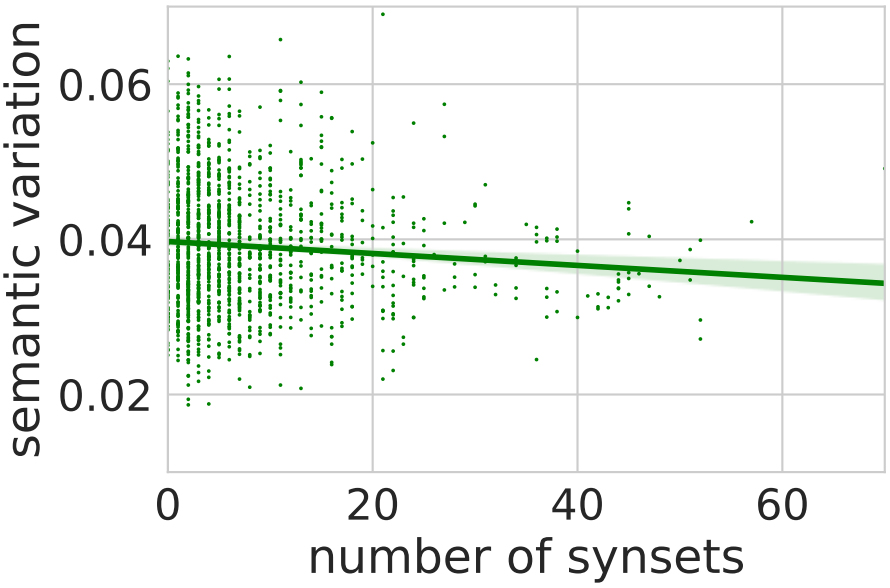
FIGURE 6.1: Relationship between personal semantic variations and frequency computed from personalized word embeddings of the same words on the two datasets, RateBeer and Yelp dataset. Their Pearson coefficient correlations are (A) 0.40 and(B) 0.25. The trendlines show 95% confidence intervals obtained from kernel regressions.

(A) dissemination at RateBeer dataset

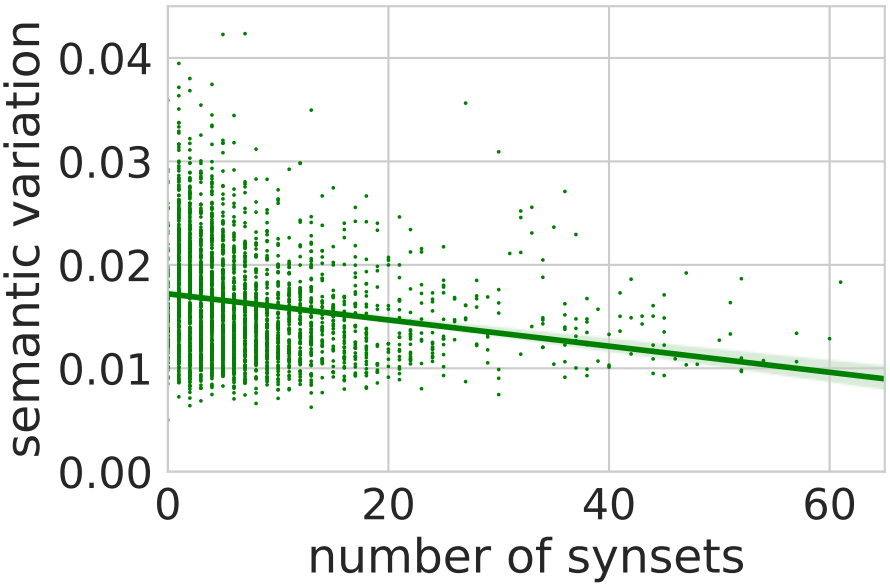


(B) dissemination at Yelp dataset

FIGURE 6.2: Relationship between personal semantic variations and dissemination computed from personalized word embeddings of the same words on the two datasets, RateBeer and Yelp dataset. Their Pearson coefficient correlations are (A) 0.22 and (B) 0.16. The trendlines show 95% confidence intervals obtained from kernel regressions.

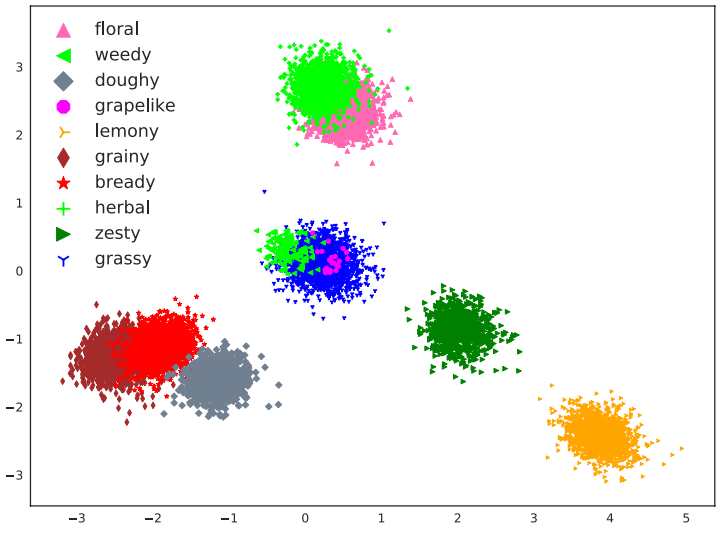


(A) polysemy at RateBeer dataset

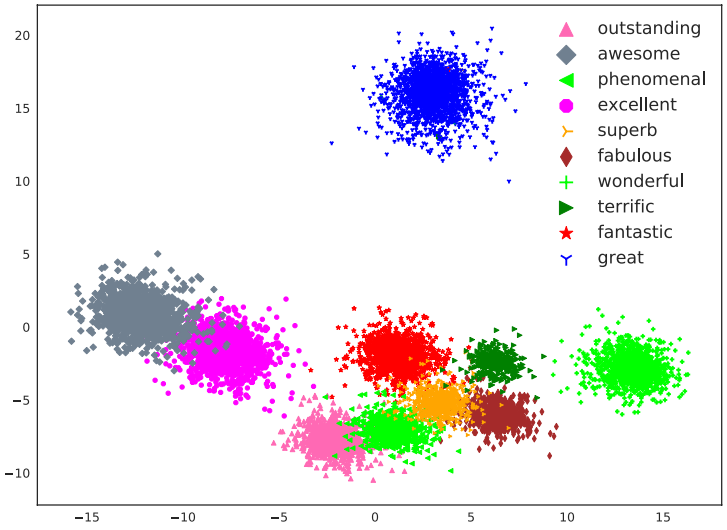


(B) polysemy at Yelp dataset

FIGURE 6.3: Relationship between personal semantic variations and polysemy computed from personalized word embeddings of the same words on the two datasets, RateBeer and Yelp dataset. Their Pearson coefficient correlations are (A) -0.08 and (B) -0.19. The trendlines show 95% confidence intervals obtained from kernel regressions.



(A) **RateBeer dataset**



(B) **Yelp dataset**

FIGURE 6.4: Two-dimensional representations of the words, ***grassy*** and ***great*** in the two datasets, respectively, with the words closest to them in the universal embedding space.

Chapter 7

Conclusion

In this thesis, interpersonal variations in word meanings were focused on, and which words have largely different meanings by individuals was explored. To verify this, a novel method for modeling word meanings by individuals, "personalized word embeddings," through a task with objective outputs was proposed (Chapter 4). As a task for inducing personalized word embeddings, in order to suppress meaning-unrelated biases from contaminating word embeddings, the review-target identification is adopted (§ 4.2). To improve the scalability to the number of reviewer are improved by using reviewer-wise fine-tuning of the proposed model with residual connection (§ 4.3.2.4). The stability of the method in inducing personalized word embeddings was also improved by using multi-task learning with target-attribute prediction when pre-traing model parameters (§ 4.3.2.3).

Experiments using large-scale review datasets from the RateBeer and Yelp websites was conducted (Chapter 5). Experimental results showed that the combination of multi-task learning and personalization of word embeddings improved the performance of the review-target identification (§ 5.3.1). At the same time, it was shown that the word meanings of each individual persons could be stably learned by the proposed method. The scalability was also indicated by the fact that the word embeddings for thousands of reviewers were obtained.

Experiments using sentiment analysis and review text personalization task was also conducted to examine the applications of the acquired personalized word embeddings (§ 5.3.3, § 5.3.4). The results showed that the personalized word embeddings are effective not only in the review-target identification itself but also extrinsic NLP tasks. The results also indicated that the task-independent semantic representations of words can be obtained by the proposed method.

The obtained personalized word embeddings were then analyzed (Chapter 6). First, the metric to calculate how the personalized word embeddings differ by individuals was newly defined (§ 6.3). Analysis in terms frequency, dissemination, and polysemy of words, which have been discussed in the interdomain and diachronic semantic variation, showed that frequent and widely used words have strong semantic variations (§ 6.3.1). Further analysis revealed that words related to the five senses and adjectives had strong semantic variations (§ 6.3.2).

Chapter 8

Future Work

In this thesis, a scalable and stable method for modeling the word meanings for each individual person using the task of estimating the review target given a review text was proposed, and what words have strong individual differences are clarified. It is also showed the acquired personalized word embeddings are applicable to sentiment analysis and review text personalization tasks. In this section, future work are described based on this study.

Personalization with Documents Other than Review Documents

In this study, we performed experiments in the review document domain. This is because the task of estimating the review target is an appropriate situation so that meaning-unrelated biases can be excluded. However, there are many texts written by individuals other than review documents such as social media text. The proposed method needs to be applicable to a wide range of domains in order to improve language communication between individuals in such domains and to improve the performance of NLP tasks dealing with such domain text. Carefully selecting labels that do not induce *selection or annotation biases*, the validity of the selected labels should be examined through intrinsic and extrinsic evaluations as in the experiments in this study. For example, in Twitter data, named entities referred to by each tweet can be considered as a label.

Phrase Level Personalization

In this study, we looked for differences in meaning at the word level. However, individuals may have their own meanings even at the phrase level, so personlization of semantic representations at the phrase level should be explored. In this case, it will be a technical issue to recognize phrases when words that make up the phrase of their order is wrong. It will happen especially in the documents written by the second language learner.

Generalizing existing text transformation tasks

Quality of language communication are degraded by differences in the linguistic abilities of text writers and readers and the performance of computers to process languages. There exist studies trying to remove the language barrier such as text simplification and grammatical error correction. However, these are designed as tasks that transfer input text into text that is natural for a set of readers, and do not consider the linguistic abilities and senses of the individual writer or the individual reader in most cases. Therefore, they cannot completely eliminate the language barrier between individuals, as this thesis focuses on. In the case of review text personalization explored in this thesis, the problem described above can be solved because the language used by different person is treated as different language. In addition, existing rewriting tasks can be generalized by formulating them as translation tasks in the same language.

At that time, it is first necessary to build a model that can perform sufficient text personalization. Considering the performance using the BLEU, the model in this study can still improve performance. Regarding training data, in addition to **Personalization with Documents Other than Review Documents** described above, it is also necessary to avoid data sparseness. Text generation tasks require a large amount of data in source and target side (per individual person in this case), in general. Therefore, an experiment on the minimum required amount of personal text data should be conducted, and a data augmentation should be considered such as grouping people who have documents that are less than the obtained threshold. Furthermore, a method to deal with individuals who do not exist in the learning data is also required. Method like a technique that conduct domain adaptation at the instance level on the fly used in the field of machine translation should be explored as a solution.

Bibliography

[1] Fangtao Li, Nathan Liu, Hongwei Jin, Kai Zhao, QiangYang, and Xiaoyan Zhu. Incorporating reviewer and product information for review rating prediction. In *Proceedings of the 22nd international joint conference on Artificial Intelligence (IJCAI 2011)*, pages 1820–1825, 2011.

[2] Wenliang Gao, Naoki Yoshinaga, Nobuhiro Kaji, and Masaru Kitsuregawa. Modeling user leniency and product popularity for sentiment classification. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing (IJCNLP 2013)*, pages 1107–1111, 2013.

[3] Duyu Tang, Bing Qin, and Ting Liu. Learning semantic representations of users and products for document level sentiment classification. 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 2015)*, pages 1014–1023, 2015.

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

[5] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 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 2018)*, pages 2204–2213, 2018.

[6] Jia-Chen Gu, Zhen-Hua Ling, Xiaodan Zhu, and Quan Liu. Dually interactive matching network for personalized response selection in retrieval-based chatbots. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019)*, pages 1845–1854, 2019.

[7] Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, and Pascale Fung. Personalizing dialogue agents via meta-learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pages 5454–5459, July 2019.

[8] Maria Nadejde and Joel Tetreault. Personalizing grammatical error correction: Adaptation to proficiency level and l1. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 27–33, 2019.

[9] Shachar Mirkin and Jean-Luc Meunier. Personalized machine translation: Predicting translational preferences. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*, pages 2019–2025, 2015.

[10] Joern Wuebker, Patrick Simianer, and John DeNero. Compact personalized models for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*, pages 881–886, 2018.

[11] Paul Michel and Graham Neubig. Extreme adaptation for personalized neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, pages 312–318, 2018.

[12] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2018)*, pages 107–112, 2018.

[13] Mor Geva, Yoav Goldberg, and Jonathan Berant. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding

datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019)*, pages 1161–1166, 2019.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR 2016)*, pages 770–778, 2016.

[15] Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

[16] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of Advances in neural information processing systems 26 (NIPS 2013)*, pages 3111–3119, 2013.

[17] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proceedings of the First International Conference on Learning Representations (ICLR)*, 2013.

[18] Sebastian Ruder. *Neural transfer learning for natural language processing*. PhD thesis, NUI Galway, 2019.

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2019)*, pages 4171–4186. Association for Computational Linguistics, June 2019.

[20] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, 2018.

[21] Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. In

*Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics (*SEM 2018)*, pages 180–191, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[22] Masatoshi Tsuchiya. Performance impact caused by hidden bias of training data for recognizing textual entailment. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, May 2018.

[23] Zhongqing Wang and Yue Zhang. Opinion recommendation using a neural model. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pages 1626–1637, September 2017.

[24] Marco Del Tredici and Raquel Fernández. Semantic variation in online communities of practice. In *Proceedings of 12th International Conference on Computational Semantics (IWCS 2017)*, 2017.

[25] William L. Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pages 1489–1501, 2016.

[26] Alex Rosenfeld and Katrin Erk. Deep neural models of semantic shift. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2018)*, pages 474–484, 2018.

[27] Kokil Jaidka, Niyati Chhaya, and Lyle Ungar. Diachronic degradation of language models: Insights from social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, pages 195–200, 2018.

[28] David Bamman, Chris Dyer, and Noah A. Smith. Distributed representations of geographically situated language. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, pages 828–834, 2014.

[29] Aparna Garimella, Rada Mihalcea, and James Pennebaker. Identifying cross-cultural differences in word usage. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers (COLING 2016)*, pages 674–683, 2016.

[30] Eduardo G Altmann, Janet B Pierrehumbert, and Adilson E Motter. Niche as a determinant of word fate in online groups. *PloS one*, 6(5):e19009, 2011.

[31] BODO winter, Graham Thompson, and Matthias Urban. Cognitive factors motivating the evolution of word meanings: Evidence from corpora, behavioral data and encyclopedic network structure. In *Proceedings of the 10th International Conference on the Evolution of Language (EVOLANG 2014)*, pages 353–360, 2014.

[32] Michel Bréal. The history of words. *The beginnings of semantics: Essays, lectures and reviews*, pages 152–175, 1897.

[33] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of Advances in neural information processing systems 29 (NIPS 2016)*, pages 4349–4357, 2016.

[34] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

[35] Mark Díaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI 2018)*, page 412, 2018.

[36] Nathaniel Swinger, Maria De-Arteaga, Neil Thomas Heffernan IV, Mark DM Leiserson, and Adam Tauman Kalai. What are the biases in my word embedding? In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES 2019)*, pages 305–311, 2019.

[37] Masahiro Kaneko and Danushka Bollegala. Gender-preserving debiasing for pre-trained word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pages 1641–1650, July 2019.

[38] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[39] Julian McAuley and Jure Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems (RecSys 2013)*, pages 165–172, 2013.

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

[41] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations (ICLR 2015)*, 2015.

[42] Melvin Johnson, Mike Schuster, Quoc Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernand a ViÃ © gas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 2017.

[43] Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. Multi-task sequence to sequence learning. 2016.

[44] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

Publications

International conferences (referred)

- Daisuke Oba, Shoetsu Sato, Naoki Yoshinaga, Satoshi Akasaki, Masashi Toyoda. Understanding Interpersonal Variations in Word Meanings via Review Target Identification. In proceedings of the 20th International Conference on Computational Linguistics and Lntelligent Text Processing (CICLing2019), No.129, April, 2019
- Daisuke Oba, Naoki Yoshinaga, Shoetsu Sato, Satoshi Akasaki, Masashi Toyoda. Modeling Personal Biases in Language Use by Inducing Personalized Word Embeddings. In proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT2019) , pp.2102-2108, June, 2019

Domestic conferences (non-referred)

- 大葉大輔，吉永直樹，赤崎智，豊田正史. 五感に基づく言語表現における個人のバイアスとその補正. NLP 若手の会 第 13 回シンポジウム, August, 2018
- 大葉大輔，佐藤翔悦，赤崎智，吉永直樹，豊田正史. 人の言語使用における単語の意味の揺らぎの解明に向けて. 言語処理学会第 25 回年次大会，March, 2019

- 大葉大輔，佐藤翔悦，吉永直樹，赤崎智，豊田正史. Understanding Interpersonal Variations in Word Meanings via Review Target Identification. 東京大学音声・言語・コミュニケーション研究会, December, 2019