General Paper

Personal Semantic Variations in Word Meanings: Induction, Application, and Analysis

Daisuke Oba[†], Shoetsu Sato[†], Satoshi Akasaki[†],
Naoki Yoshinaga[††] and Masashi Toyoda[††]

When people verbalize what they have felt with different sensory functions, they often represent different meanings such as with *temperature range* using the same word *cold* or the same meaning by using different words (e.g., *hazy* and *cloudy*). These interpersonal variations in word meanings have the effects of not only preventing people from communicating efficiently with each other but also causing troubles in natural language processing (NLP). Accordingly, to highlight interpersonal semantic variations in word meanings, a method for inducing personalized word embeddings is proposed. This method learns word embeddings from an NLP task, distinguishing each word used by different individuals. Review-target identification was adopted as a task to prevent irrelevant biases from contaminating word embeddings. The scalability and stability of inducing personalized word embeddings were improved using a residual network and independent fine-tuning for each individual through multi-task learning along with target-attribute predictions. The results of the experiments using two large scale review datasets confirmed that the proposed method was effective for estimating the target items, and the resulting word embeddings were also effective in solving sentiment analysis. By using the acquired personalized word embeddings, it was possible to reveal tendencies in semantic variations of the word meanings.

Key Words: *Semantic Variation, Personalized Word Embeddings, Residual Network, Multi-Task Learning, Review-Target Identification, Sentiment Analysis*

1 Introduction

People express what they have sensed with various sensory organs as language in different ways, and semantic variations in the meanings of words inevitably exist because the senses and linguistic abilities of individuals differ. For example, if one were to use the word "*sour*," *how* "*sour*" is meant can differ greatly between individuals. Furthermore, different people may describe the appearance (color) of the same beer with different expressions such as "*yellow*" or "*golden*." These semantic variations not only have the potential to cause problems in verbal

[†] Graduate School of Information Science and Technology, The University of Tokyo
[††] Institute of Industrial Science, The University of Tokyo

communication but could also delude the potential 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 (Li, Liu, Jin, Zhao, Yang, and Zhu 2011; Gao, Yoshinaga, Kaji, and Kitsuregawa 2013; Tang, Qin, and Liu 2015; Ebrahimi and Dou 2016), dialogue systems (Li, Galley, Brockett, Spithourakis, Gao, and Dolan 2016; Zhang, Dinan, Urbanek, Szlam, Kiela, and Weston 2018; Gu, Ling, Zhu, and Liu 2019; Madotto, Lin, Wu, and Fung 2019), grammatical error correction (Nadejde and Tetreault 2019), and machine translation (Mirkin and Meunier 2015; Wuebker, Simianer, and DeNero 2018; Michel and Neubig 2018), all of which considered user preferences concerning the task inputs and outputs. However, all these studies were based on the premise of estimating *subjective* output from *subjective* input (e.g., estimating the sentiment polarity of the target item from input review or predicting responses from input utterances in a dialogue system). Consequently, the model not only captures the semantic variations of the user-generated text (input) but also handles the *annotation bias* of the output labels (namely, the deviation of output labels assigned by each annotator) (Gao et al. 2013; Gururangan, Swayamdipta, Levy, Schwartz, Bowman, and Smith 2018; Geva, Goldberg, and Berant 2019) and *selection bias* (namely, the deviation of output labels inherited from the targets chosen by users) (Gao et al. 2013). The contamination caused by these biases hinders the understanding of the solo impact of semantic variations, which is the target of this study.

The goals of this study are to (i) understand which words have large (or small) interpersonal variations in their meanings (hereafter referred to as *semantic variation*) and (ii) reveal how such semantic variation affects the classification accuracy concerning tasks with user-generated inputs (e.g., reviews). To perform such analysis, a method for such analysis into the degree of personal semantic variation in word meanings is thus proposed (§ 3). It uses personalized word embeddings obtained 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. It should be noted that this task is free from *annotation bias* because outputs (review targets) are automatically determined without annotation, along with the suppression of *selection bias* by using a dataset in which the same reviewer evaluates the same target only once, so as not to learn the deviation of output labels caused by choice of inputs. The resulting model makes it possible to observe only the impact of semantic variation from the acquired personalized word embeddings.

Two further remaining issues concerning inducing personalized word embeddings are the scalability and stability in learning personalized word embeddings. To ensure that the training was scalable concerning the number of reviewers, a residual network (He, Zhang, Ren, and Sun 2016) was used to (i) obtain personalized word embeddings by using reviewer-specific transformation matrices and biases from a small amount of reviews for each user (§ 3.2.1) and (ii) fine-tune these reviewer-specific parameters (§ 3.2.4). Moreover, to make the training via the extreme multi-class classification (i.e., the review-target identification) stable, multi-task learning (MTL) with target-attribute predictions was performed during the pre-training of the parameters (§ 3.2.3). As a result of the target attributes being likely to be more coarse-grained than the review targets, MTL using target-attributes made the training more stable.

During the experiments, it was hypothesized that words related to the five senses especially have inherent semantic variations, and this hypothesis was validated (§ 4). Two large-scale datasets retrieved from the RateBeer and Yelp websites, including a variety of expressions related to the five senses, were utilized. To confirm the impact of personalized word embeddings obtained by using the proposed method, the datasets were used for a specific task: identifying a target item and its attributes from a given review, using the reviewer's ID. Consequently and concerning both datasets, the personalized model proposed successfully captured semantic variations and achieved better performance than a reviewer-universal model (§ 4.2.1). Moreover, the personalized word embeddings obtained were extrinsically evaluated by sentiment analysis to assess their usefulness. The results of the extrinsic evaluation evidence that this model achieved higher levels of performance as compared with other models, which demonstrated the capability of the proposed method for suppressing unfavorable biases during the training process (§ 4.2.2). The acquired personalized word embeddings were finally analyzed from three perspectives (frequency, dissemination, and polysemy) to reveal which words have large (or small) semantic variations (§ 4.3).

The contributions of this paper are three-fold:

- A scalable and stable method for obtaining personalized word embeddings without contaminating them with irrelevant biases is proposed. The proposed method induces personalized word embeddings through a task with objective outputs via effective reviewer-wise fine-tuning on a neural network with a residual connection and MTL with target-attribute predictions.
- The usefulness of the obtained personalized word embeddings not only in the review-target identification task but also in the sentiment analysis task is confirmed.
- The tendencies in the personal semantic variations in terms of three perspectives (frequency, dissemination, and polysemy), which have been discussed in previous studies about diachronic and interdomain semantic variations, are revealed.

2 Related Work

Existing studies on personalization in NLP tasks and analysis of semantic variation of word meanings in terms of diachronic, geographic, domain, and political correctness will initially be discussed in this section. Since existing methods on personalization are mostly aimed at improving accuracy on various tasks, such methods simultaneously model personal variations in word meanings and other irrelevant biases (such as *annotation* and *selection biases*) that contribute to task performances. There are a few studies that try to understand variations in word meanings in terms of time, geography, and domain, which will be reviewed. Finally, differences between interpersonal semantic variations in word meanings and biases related to unfavorable prejudices are then discussed.

As discussed in § 1, in the field of 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 of how annotators label),[1]and (3) *selection bias* of output labels (biases of 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) and (3), particularly in the case of text generation tasks such as machine translation (Mirkin and Meunier 2015; Michel and Neubig 2018; Wuebker et al. 2018) and dialogue systems (Li et al. 2016; Zhang et al. 2018). This is because data-driven approaches without personalization tend to suffer from the writer-dependent diversity of probable outputs. Meanwhile, it is difficult to separate these facets; therefore, to the author's knowledge, *semantic variations* of words among people have not been analyzed independently. It is necessary to be able to eliminate these unfavorable and meaning-unrelated biases to understand the variation of word meanings among individuals.

To quantify the semantic variations of common words among domains, Tredici and Fernández (2017) obtained domain-specific word embeddings using the Skip-gram (Mikolov, Sutskever, Chen, Corrado, and Dean 2013), and they analyzed their 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, inputs and outputs are both defined by the same writer. Ebrahimi and Dou (2016) obtained personal word vectors using the log-bilinear language model (Maas and Ng

[1] It is pointed out that NLP datasets are likely to suffer from *annotation bias* (Geva et al. 2019), whether or not the context of the study is about personalization; models learn to use or rely on this *annotation bias* when task accuracy is optimized (Tsuchiya 2018; Gururangan et al. 2018; Poliak, Naradowsky, Haldar, Rudinger, and Van Durme 2018; Geva et al. 2019).

2010). Their approach also suffers from *annotation bias* because the log-bilinear language model predicts a word according to its previous words; they also use the inputs and outputs defined by the same writer. Consequently, the same word can have dissimilar vectors by person not only because it has different meanings by individuals but also because it just appears with words in different topics.[2] Additionally, their approach cannot be scalable to the number of domains or writers (reviewers in this study), since it simultaneously learns all the domain- or writer-specific parameters.

Semantic variations of word meanings caused by diachronic (Hamilton, Leskovec, and Jurafsky 2016; Rosenfeld and Erk 2018; Jaidka, Chhaya, and Ungar 2018), geographic (Bamman, Dyer, and Smith 2014; Garimella, Mihalcea, and Pennebaker 2016), and interdomain (Tredici and Fernández 2017) differences of text have also been studied. This study analyzes semantic variations of word meanings at the individual level, particularly, as discussed in (Hamilton et al. 2016; Tredici and Fernández 2017), focusing on how semantic variations are correlated with word frequency, dissemination, and polysemy.

Apart from 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 (Bolukbasi, Chang, Zou, Saligrama, and Kalai 2016; Caliskan, Bryson, and Narayanan 2017; Díaz, Johnson, Lazar, Piper, and Gergle 2018; Swinger, De-Arteaga, Heffernan IV, Leiserson, and Kalai 2019; Kaneko and Bollegala 2019). In these studies, "biases" were defined in terms of political correctness, so they differ from biases in personalized word embeddings targeted in this study.

3 Personalized Word Embeddings

The proposed neural network-based model for inducing personalized word embeddings is shown as an overview in Fig. 1. To clarify semantic variations in meanings of individual words, the following approach was taken: personalized word embeddings for each particular person were learned via representation learning in NLP tasks under the assumption that the words used by individuals were different. To implement this approach, two major problems should be solved: (i) what kind of tasks should be used to learn personalized word embeddings and (ii) how to effectively learn them.

[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.

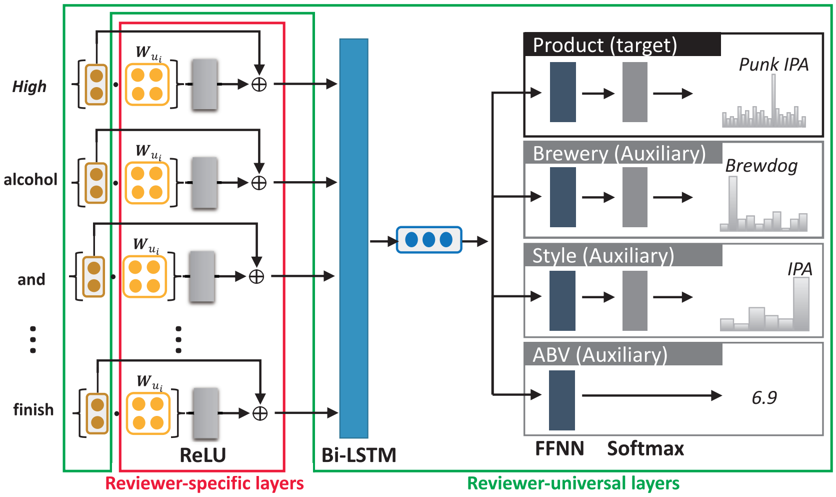


Fig. 1 Overview of the proposed LSTM network with residual connections for inducing personalized word embeddings via review-target identification through MTL with target-attribute predictions.

3.1 Task for Inducing Personalized Word Embeddings: 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, thus, resulting in the word embeddings being similar or fixed, even if those words were semantically irrelevant. Moreover, as datasets for the task were likely to contain annotation and selection biases, the induced personalized word embeddings may be contaminated with those biases. In consideration of these issues, review-target identification was adopted to induce personalized word embeddings.

The majority of review datasets associate *product name*, *metadata*, *user name*, *rating*, and *time* with each *review text*. The focus of review-target identification is the *product name* which was given *a review text* (Table 1). Compared with conventional tasks, such as sentiment analysis in which *rating* was estimated given *a review text*, the review-target identification is significantly more difficult because of the large number of target classes and therefore requires a model to understand (or distinctively embed) each word in the review. Moreover, since no annotator was involved when labeling output (review target in this case), annotation bias could be excluded. Moreover, since the number of reviews for each review target is one at most per reviewer in most review datasets, selection bias can, therefore, be minimized. Accordingly, the model could learn only the meanings of words through the review-target identification task.

3.2 Method for Inducing Personalized Word Embeddings

As for effective training of personalized word embeddings, mentioned in § 2, the scalability and stability of the training became two major problems because (i) it was necessary to learn embeddings for words amplified by the number of reviewers and (ii) the review-target identification task was an extreme multi-class classification with massive review targets. To solve these problems, a long short-term memory (LSTM) (Hochreiter and Schmidhuber 1997) network with a residual connection similar to ResNet (He et al. 2016) was used to obtain personalized word embeddings by using reviewer-specific transformation parameters (§ 3.2.1). The obtained personalized word embeddings were then fine-tuned in terms of the scalability in accordance with the number of reviewers (§ 3.2.4). After that, the learning of the proposed model was stabilized by applying MTL with target-attribute predictions (§ 3.2.3).

3.2.1 Reviewer-specific layers for personalization

First, the model computed the personalized word embeddings $\boldsymbol{e}_{w_i}^{u_j}$ of each word w_i in input texts via a reviewer-specific matrix $\boldsymbol{W}_{u_j} \in \mathbb{R}^{d \times d}$ and bias vector $\boldsymbol{b}_{u_j} \in \mathbb{R}^{d}$. Concretely, an input word embedding $\boldsymbol{e}_{w_i}$ was transformed to $\boldsymbol{e}_{w_i}^{u_j}$ as follows:

$$\boldsymbol{e}_{w_i}^{u_j} = \mathrm{ReLU}(\boldsymbol{W}_{u_j}\boldsymbol{e}_{w_i} + \boldsymbol{b}_{u_j}) + \boldsymbol{e}_{w_i} \tag{1}$$

where ReLU was a rectified linear unit function. As shown in Eq. (1), a residual network similar to ResNet (He et al. 2016) was used, since semantic variation defined as that from reviewer-universal word embedding. The use of activation functions allowed for non-linear expressions. In this research, rectified linear units (ReLU) were used as the activation function following the structure of ResNet (He et al. 2016). The advantage of using ReLU is that compared with other general activation functions such as sigmoid and hyperbolic tangent (tanh), the computational cost is lower and the vanishing gradients problem (Hochreiter 1991; Bengio, Simard, and Frasconi 1994) (the situation where a deep neural network is unable to propagate gradient from the output back to the layers close to the input) does not occur. The (vector) value of the first term in Eq. (1) was limited to 0 or more by using ReLU that might restrict the expression of the personalized word embeddings. However, since an input word embedding e_{w_i} of the second term in Eq. (1) could represent negative values, its effect on the expression of the personalized word embeddings was limited as a whole. Sharing the reviewer-specific parameters for transformation across words and employing a residual network enabled the model to learn personalized word embeddings even for infrequent words stably.

3.2.2 Reviewer-universal layers

Given the personalized word embedding $\boldsymbol{e}_{w_i}^{u_j}$ of each word w_i in an input text, the model encoded them through LSTM (Hochreiter and Schmidhuber 1997). The LSTM updated the current memory cell $\boldsymbol{c}_t$ and hidden state $\boldsymbol{h}_t$ according to the following equations:

$$\begin{bmatrix} \boldsymbol{i}_t \\ \boldsymbol{f}_t \\ \boldsymbol{o}_t \\ \hat{\boldsymbol{c}}_t \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{bmatrix} \boldsymbol{W}_{\mathrm{LSTM}} \cdot \left[\boldsymbol{h}_{t-1}; \boldsymbol{e}_{w_i}^{u_j}\right] \tag{2}$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \hat{\boldsymbol{c}}_t \tag{3}$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t \odot \tanh\left(\boldsymbol{c}_t\right) \tag{4}$$

where $\boldsymbol{i}_t$, $\boldsymbol{f}_t$, and $\boldsymbol{o}_t$ were the input, forget, and output gate at time step t, respectively. $\boldsymbol{e}_{w_i}$ was an input word embedding at time step t, and $\boldsymbol{W}_{\mathrm{LSTM}}$ was a weight matrix. $\hat{\boldsymbol{c}}_t$ was the current cell state. Operation $\odot$ denoted element-wise multiplication and σ was the logistic sigmoid function. Single-layer bi-directional LSTM (Bi-LSTM) was adopted to use past and future contexts. As the representation of the input text $\boldsymbol{h}$, Bi-LSTM concatenated the outputs from the forward and backward LSTMs:

$$\boldsymbol{h} = \left[\overrightarrow{\boldsymbol{h}_{L-1}}; \overleftarrow{\boldsymbol{h}_0}\right] \tag{5}$$

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

Lastly, a feed-forward layer computed output probability distribution $\hat{\mathbf{y}}$ from the representation $\boldsymbol{h}$ with weight matrix $\boldsymbol{W}_o$ and bias vector $\boldsymbol{b}_o$ as

$$\hat{\mathbf{y}} = \mathrm{softmax}\left(\boldsymbol{W}_o \boldsymbol{h} + \boldsymbol{b}_o\right) \tag{6}$$

3.2.3 Multi-task learning with target-attribute predictions for stable training

Training the model for the target identification task was considered to be unstable because its output space (review targets) was extremely large (more than 50,000 candidates in our datasets). To mitigate this instability, auxiliary tasks that estimate attributes of the target (item) were set up and solved simultaneously with the target identification task (target task) by MTL. The target items and target attributes used in this study are later summarized in Table 1. This approach was motivated by the assumption that the attributes of the review item are more coarse-grained than the review item itself and that understanding those related metadata of the target item would contribute to the accuracy of identifying the review target.

Specifically, independent feed-forward layers were added and used to compute outputs from shared sentence representation $\boldsymbol{h}$ defined by Eq. (5) for each auxiliary task (Fig. 1). As shown in Table 1, three types of auxiliary tasks were assumed: (1) multi-class classification (the same as the target task), (2) multi-label classification, and (3) regression. Multi-task learning under a loss that sums up individual losses for the target and auxiliary tasks was performed. Cross-entropy loss was used for multi-class classification, a summation of cross-entropy loss of each class was used for multi-label classification, and mean-square loss was used for regression.

3.2.4 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 limitations. Therefore, the model was first pre-trained by using all the training data without personalization. Fine-tuning was then applied only to the parameters in reviewer-specific layers (Fig. 1) by training reviewer-independent models based on the reviews written by each reviewer while keeping reviewer-universal layers (Fig. 1) fixed.

More specifically, in the pre-training, the model used reviewer-universal parameters $\boldsymbol{W}$ and $\boldsymbol{b}$ (instead of $\boldsymbol{W}_{u_j}$ and $\boldsymbol{b}_{u_j}$) for Eq. (1). It then initialized the reviewer-specific parameters $\boldsymbol{W}_{u_j}$ and $\boldsymbol{b}_{u_j}$ in reviewer-specific layers by using $\boldsymbol{W}$ and $\boldsymbol{b}$. Finally, the initialized parameters $\boldsymbol{W}_{u_j}$ and $\boldsymbol{b}_{u_j}$ were fine-tuned per reviewer using the reviews written by the reviewer. This approach makes the model scalable even to a large number of reviewers. Note that all the reviewer-universal parameters in the reviewer-universal layers were fixed at the time of reviewer-wise fine-tuning.

Furthermore, all the parameters in reviewer-universal and reviewer-specific layers were subjected to MTL only during the pre-training without personalization. Reviewer-specific parameters $\boldsymbol{W}_{u_j}$ and $\boldsymbol{b}_{u_j}$ of the pre-trained model were then fine-tuned while the target task only was optimized. This fixing stops the model introducing *selection bias* into the personalized embeddings; otherwise, the prior output distribution of the auxiliary tasks by individuals could be implicitly learned.

4 Experiments

The target identification task was evaluated first using two review datasets to confirm the effectiveness of the personalized word embeddings induced by the method. If the model can successfully solve this objective task more accurately than the reviewer-universal model obtained by only pre-training the proposed reviewer-specific model, it is considered that those personalized word embeddings would capture the personal semantic variations of input words. Next, to verify the usefulness of the personalized word embeddings concerning not only an intrinsic task but also an extrinsic task and to confirm whether the proposed method could remove biases unrelated to the meanings, the proposed model was applied to solve sentiment analysis task. Personal semantic variation of each word was then defined, and the degree and tendencies of the semantic variation in the obtained personalized word embeddings were analyzed from the same perspectives as discussed in previous studies on semantic variations in word meanings.

4.1 Settings

Datasets Datasets containing reviews of beer and services related to foods were adopted for evaluating the proposed method, since there were a variety of expressions that describe what people have sensed with various sensory units in these domains of the datasets. The RateBeer dataset, which included a variety of beers, was extracted from ratebeer.com[3] (McAuley and Leskovec 2013). Written by reviewers who posted at least 100 reviews, 2,695,615 reviews about 109,912 types of beer were selected. The Yelp dataset, which includes a diverse range of services, was derived from yelp.com.[4] 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. Consequently, 426,816 reviews of 56,574 services (restaurants or foods) written by 2,414 reviewers in total were extracted. 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**.

Target and Auxiliary Tasks Table 1 summarizes the settings of the target and auxiliary tasks. The target task took a review, and estimated target beer for the RateBeer dataset or services from the Yelp dataset reviewed therein. Regarding the target attributes for MTL, **style** with 89 types and **brewery** with 6,870 types were chosen for multi-class classification, and **alcohol by volume (ABV)** was chosen for regression in the experiments with the RateBeer dataset. As for the Yelp dataset, **location** with 19 types was used for multi-class classification, and **category** with 683 types was used for multi-label classification.

Sentiment Analysis Task The settings of the sentiment analysis task are also summarized in Table 1. As for the sentiment analysis task, **ratings** of given reviews annotated by the reviewers were used for regression. The ratings are integers and range from 1 to 20 in the RateBeer dataset

[3] https://www.ratebeer.com
[4] https://www.yelp.com/dataset

Table 1 Summary of task settings for inducing and evaluating personalized word embeddings

(a) RateBeer dataset

Tasks	Input	Output	Type	Loss
Induction and intrinsic evaluation				
(Target) review-target identification	review text	beer	classification	cross entropy
(Auxiliary) target-attribute prediction	review text	style	classification	cross entropy
	review text	brewery	classification	cross entropy
	review text	ABV	regression	mean square error
Extrinsic evaluation				
Sentiment analysis	review text	rating	regression	mean square error

(b) Yelp dataset

Tasks	Input	Output	Type	Loss
Induction and intrinsic evaluation				
(Target) review-target identification	review text	service	classification	cross entropy
(Auxiliary) target-attribute prediction	review text	location	classification	cross entropy
	review text	category	multi-label classification	binary cross entropy
Extrinsic evaluation				
Sentiment analysis	review text	rating	regression	mean square error

and from 1 to 5 in the Yelp dataset. This task was solved as a regression task since it is natural to treat the fine-grained ratings as continuous values.

Throughout all the tasks, accuracy was used for classification, and root mean square loss (RMSE) was used for regression tasks. For multi-label classification, micro-F1 score was used.

Models and Hyperparameters As for the target item and attribute identification tasks, the proposed model (described in § 3) was evaluated in terms of four different settings.[5] The differences of the models were (1) whether fine-tuning for personalization was applied and (2) whether the model was trained through MTL before the fine-tuning. Table 2 lists the major hyperparameters. The embedding layer was initialized by Skip-gram embeddings (Mikolov et al. 2013) pre-trained using review texts of training and validation sets of each dataset. The vocabulary for each dataset includes all the words that appeared ten times or more in the dataset. For optimization, the models were trained up to 100 epochs with Adam (Kingma and Ba 2015), and the model at the epoch with the best results in the target task on the development set was selected as the test model.

As for the sentiment analysis task for extrinsically evaluating the obtained personalized word

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

Table 2 Hyperparameters of the proposed model

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

embeddings, another set of models (with the same architecture and hyperparameters) was trained as review-target identification models in Fig. 1 (except that they have only one feed-forward layer for the target sentiment regression task). 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.

4.2 Results

4.2.1 Inducing and evaluating personalized word embeddings by review-target identification

Table 3 lists the 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, whereas 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 the reviewer-specific parameters were updated only on the target task by using fine-tuning. This means that the improvements in the performance on auxiliary tasks were obtained purely by the semantic variations captured by reviewer-specific parameters.

Moreover, the model with fine-tuning can successfully solve this objective task more accurately than the reviewer-universal model. This indicates the validity of the proposed method to represent the personalized word embeddings.

Table 3 Results on the review-target identification task using the RateBeer dataset and Yelp dataset

(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

Accuracy or 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).

4.2.2 Evaluating personalized word embeddings by sentiment analysis

Table 4 lists the 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 results obtained by the proposed model with both multi-task and personalization outperformed those of the other models. These results indicate that the personalized word embeddings obtained by the proposed method are useful not only for an intrinsic task used to obtain them but also for an extrinsic task. In other words, the proposed method can model task-independent personal semantic variations as personalized word embeddings.

Moreover, 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 using the sentiment analysis task. To compare with the personalized word embeddings obtained by the proposed model using review-target identification, personalized word embeddings were also obtained using auxiliary tasks considered to be affected by selection bias (because the same output label appears multiple times in a person's training data). Table 5 shows that the model with the word embeddings obtained by the proposed method achieved the best performance. The results suggest that the proposed method can suppress the meaning-unrelated biases and obtain task-independent word meanings.

4.2.3 Impact of the number of reviews for personalization

The impact of the number of reviews for personalization when solving the review-target identification problem was investigated. The reviewers were first grouped into several bins according to their number of reviews. Classification accuracies for reviews written by the reviewers in the same bin were then evaluated. Classification accuracy of the target task was plotted against

Table 4 Results of sentiment analysis: embedding layers are kept fixed to those of the corresponding models in Table 3

		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

RMSE marked with (i) + was significantly better than the model without multi-task and personalization on the RateBeer dataset ($p < 0.05$ assessed by z-test), and (ii) # was significantly better than the model without multi-task and personalization and the model with only personalization on the Yelp dataset ($p < 0.05$ assessed by z-test).

Table 5 Comparison of sentiment analysis results for different tasks used for personalization with the RateBeer dataset and the Yelp dataset

(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

Embedding layers are kept fixed after personalization on each task. The proposed target identification task is beer and service in each dataset, respectively.

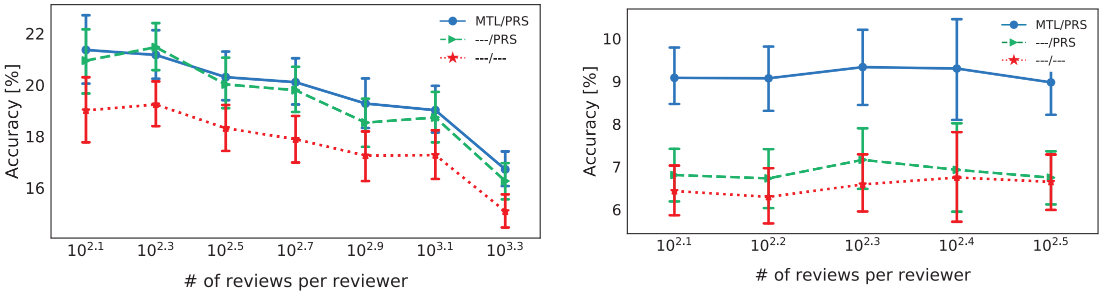
the number of reviews per reviewer in Fig. 2. For example, the plots (and error bars) for $10^{2.3}$ represent the accuracy (variation) of the target identification for reviews written by each reviewer with n reviews ($10^{2.1} \leq n < 10^{2.3}$).

Contrary to expectations, for the RateBeer dataset (Fig. 2 (a)), all models obtained lower accuracies as the number of reviews increased. However, as for the Yelp dataset (Fig. 2 (b)), the performance of the models did not deteriorate as the 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 all reviews, whereas the top-10% frequent restaurants in the Yelp dataset only accounted 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 experienced reviewers. The model without MTL and personalization 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.

4.3 Analysis of Personalized Word Embeddings

The personalized word embeddings were analyzed to determine what kind of personal biases existed in each word. Here, to remove the influences of low-frequent words, only words used by 30% or more reviewers (excluding stop words) were targeted.

Personal semantic variation[6] of a word w_i was first defined to determine how the repre-



(a) RateBeer dataset

(b) Yelp dataset

Fig. 2 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.

[6] Unlike the definitions of the semantic variation in existing studies (Hamilton et al. 2016; Garimella et al. 2016; Tredici and Fernández 2017) that measure the degree of change in the meaning of a word that occurs by diachronic or domain difference of text, personal semantic variation measures how much meanings of a word

sentations of the word differ for each individual as

$$\frac{1}{|\mathrm{U}(w_i)|}\sum_{u_j \in \mathrm{U}(w_i)}(1-\cos(\boldsymbol{e}_{w_i}^{u_j}, \overline{\boldsymbol{e}}_{w_i})) \tag{7}$$

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.

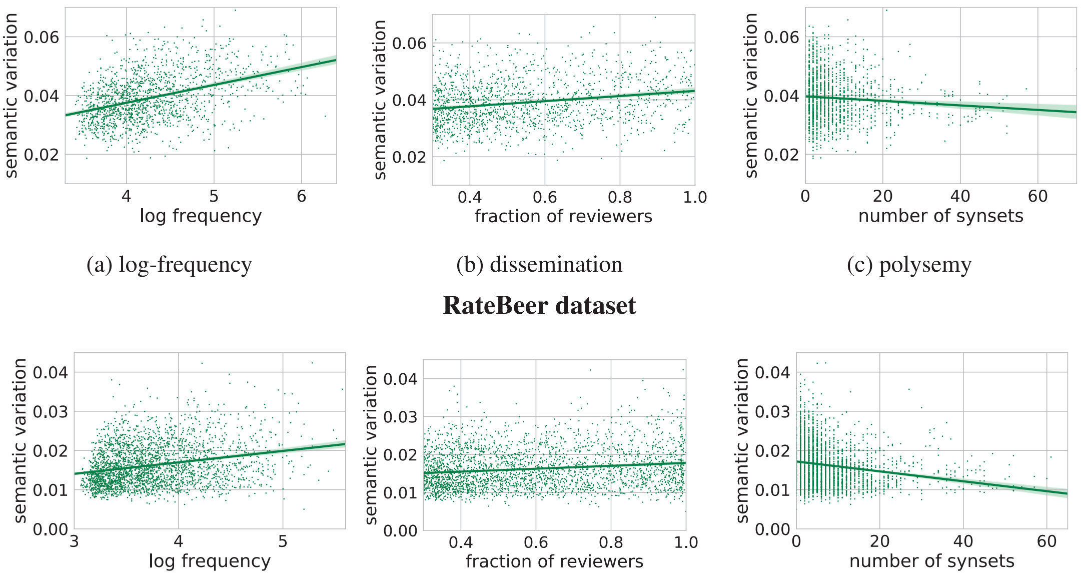
Three perspectives were focused, namely **frequency**, **dissemination**, and **polysemy**, that have been discussed in the studies on semantic variations of words caused by diachronic or domain differences of text used to obtain them (Hamilton et al. 2016; Garimella et al. 2016; Tredici and Fernández 2017) (§ 2). Fig. 3 shows semantic variations against the three metrics. Each x-axis corresponds to log frequency of the word ((a) and (d)), the ratio of the reviewers who used the word ((b) and (e)), and the number of synsets found in WordNet (Miller 1995) ver. 3.0 ((c) and (f)), respectively.

Interestingly, in contrast to the reports by (Hamilton et al. 2016) on diachronic semantic variations but consistently to the reports by (Tredici and Fernández 2017) on interdomain semantic variations, semantic variations correlate highly with frequency and dissemination but poorly with polysemy in these results. This tendency of interpersonal semantic variations can be explained as follows. In the datasets used in these 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 (e.g., as introduced in § 1, even if how "*sour*" differs by individuals, the meaning itself does not change). This is because those words are still used only in skewed contexts related to food and drink where word senses do not fluctuate significantly.

Table 6 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 (20% 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**$_1$ (vision), **hearing**$_2$ (audition), **taste**$_3$ (gustation), **smell**$_4$ (olfaction), and **touch**$_5$ (somatosensation). From the results, in the

defined by individuals diverge.



(a) log-frequency (b) dissemination (c) polysemy

RateBeer dataset

(d) log-frequency (e) dissemination (f) polysemy

Yelp dataset

Fig. 3 Personal semantic variations computed from personal word embeddings for the same words on the two datasets, the RateBeer and the Yelp dataset. Their Pearson coefficient correlations are (a) 0.40, (b) 0.22, (c) −0.08, (d) 0.25, (e) 0.16, (f) −0.19. The trendlines show 95% confidence intervals obtained from kernel regressions.

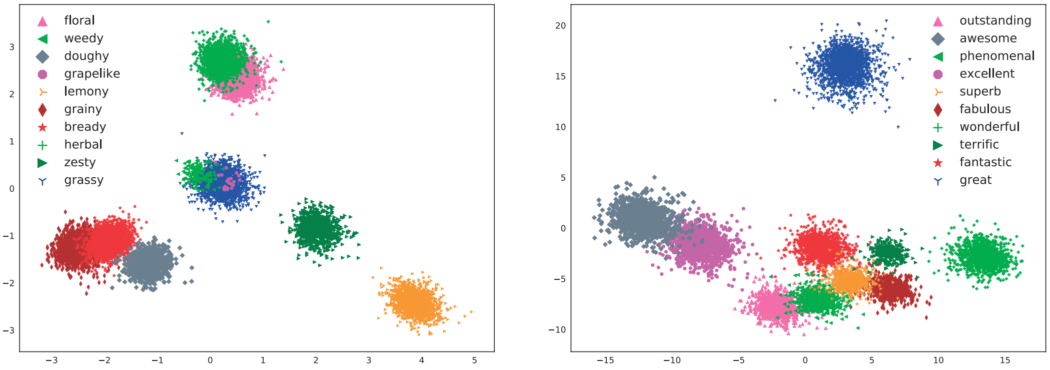
top-50 words in the RateBeer dataset, more words represented each sense (except hearing) than the bottom-50 words. Differing from this, the list of top-50 words in the Yelp dataset included fewer words related to the five senses than the RateBeer dataset; however, many adjectives that could apply to various domains (e.g., "*great*," and "*excellent*") were 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.

Whether or not some words got confused was also analyzed. The adjective words "*grassy*" and "*great*" with large semantic variations in each dataset were used as examples. 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. 4. As can be seen, clusters of "*grainy*," "*bready*," and "*doughy*" in the RateBeer dataset and "*awesome*" and "*excellent*" in the Yelp dataset were mixed each others, suggesting that words representing the same meaning may differ for each individual. Moreover, personalized word embeddings, for some words, overlapped those for multiple other words. For example, the personalized word

Table 6 The list of top-50 (and bottom-50) words with the largest (and the smallest) semantic variation in the RateBeer and Yelp datasets

	top-50	bottom-50
RateBeer dataset	**deep** grass **grassy**^{3,4} **lingering** **soapy**^{3,4} **toasty**^{3,4} **bready**^{3,4} tobacco underneath pours **pleasing** ery **medium** mildly **subtle** underlying hints dough lots **subdued** **sharp**^{3,5} mainly ark **updated** **tangy**^{3} resin **bright**^{1} hue **flowery**^{4} **fairly** **good** **rich** **upfront** **nice** crisp **dusty**^{1} toffee **creamy**^{5} kind citrus zest **citrusy**^{3,4} profile presence hay **earthy**^{3,4} aromas dominated toast **doughy**^{3,4}	dogfish batch reminds course needs bells cask rye **hot**^{3,5} ask honey unlike reminded raspberry **canned** packs liquor hand **barley**^{3} stone rogue maple never horse line rice bourbon minute **belgium** raspberries dog heat bomb **mexican** **triple** rock difference **scottish** coconut ton **burning**^{5} dead **organic** bock brewing dubbel **pink**^{1} missing becoming champagne
Yelp dataset	**great** **fantastic** **excellent** **superb** **amazing** **awesome** **phenomenal** **tasty**^{3} **delish**^{3} **good** **delicious**^{3} **yummy**^{3} sides sauce **nice** **incredible**t flatbread entrees **outstanding** **wonderful** appetizers desserts **fabulous** ambiance chicken atmosphere rice salmon ambience **flavorful**^{3,4} 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

Adjectives are boldfaced and classified into the five senses: **sight**$_1$ (vision), **hearing**$_2$ (audition), **taste**$_3$ (gustation), **smell**$_4$ (olfaction), and **touch**$_5$ (somatosensation).



(a) RateBeer dataset (b) Yelp dataset

Fig. 4 Two-dimensional representation of the words, ***grassy*** and ***great*** in the two datasets, respectively, with the words closest to them in the universal embedding space.

embeddings for “*bready*” overlapped the embeddings for “*grainy*” and those for “*doughy*” in the RateBeer dataset. This demonstrates that the same word could have different meanings by individuals.

5 Conclusions

Interpersonal variations in word meanings were focused on, and a hypothesis that words related to the five senses have inevitable personal semantic variations was explored. To verify this, a novel method for obtaining semantic variations by inducing personalized word embeddings through a task with objective outputs was proposed (§ 3). Experiments using large-scale review datasets from the RateBeer and Yelp websites showed that the combination of MTL and personalization improved the performance of the review-target identification (§ 4.2.1). Experiments on sentiment analysis proved that personalized word embeddings obtained by the proposed method are useful extrinsically (§ 4.2.2). This finding shows that the proposed method can capture interpersonal variations of word meanings. The analysis showed that adjectives and words related to the five senses have large interpersonal semantic variations (§ 4.3).

As for future studies, relationships between semantic variations and demographic factors, such as gender and age of the reviewers, which are inevitable for expressing individuality, will be analyzed. In addition to the review text, methodologies for acquiring personal semantic variations of word meanings from social media texts like Twitter will be studied.

Acknowledgement

Part of this study was presented at the 20th International Conference on Computational Linguistics and Intelligent Text Processing from 7th to 13th of April 2019 in La Rochelle, France (Oba, Sato, Yoshinaga, Akasaki, and Toyoda 2019a) and also at the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics from 2nd to 7th of June 2019 in Minneapolis, USA (Oba, Yoshinaga, Sato, Akasaki, and Toyoda 2019b). We thank Satoshi Tohda for proofreading the early draft of our paper. This work was partially supported by Commissioned Research (201) of the National Institute of Information and Communications Technology of Japan, JST CREST Grant Number JPMJCR19A4, Japan, and NII CRIS Contract Research 2019.

Reference

Bamman, D., Dyer, C., and Smith, N. A. (2014). "Distributed Representations of Geographically Situated Language." In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, pp. 828–834.

Bengio, Y., Simard, P., and Frasconi, P. (1994). "Learning Long-Term Dependencies with Gradient Descent is Difficult." *IEEE Transactions on Neural Networks*, **5** (2), pp. 157–166.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). "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)*, pp. 4349–4357.

Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). "Semantics Derived Automatically from Language Corpora Contain Human-like Biases." *Science*, **356** (6334), pp. 183–186.

Díaz, M., Johnson, I., Lazar, A., Piper, A. M., and Gergle, D. (2018). "Addressing Age-Related Bias in Sentiment Analysis." In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI 2018)*, p. 412.

Ebrahimi, J. and Dou, D. (2016). "Personalized Semantic Word Vectors." In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pp. 1925–1928.

Gao, W., Yoshinaga, N., Kaji, N., and Kitsuregawa, M. (2013). "Modeling User Leniency and Product Popularity for Sentiment Classification." In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013)*, pp. 1107–1111.

Garimella, A., Mihalcea, R., and Pennebaker, J. (2016). "Identifying Cross-Cultural Differences in Word Usage." In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers (COLING 2016)*, pp. 674–683.

Geva, M., Goldberg, Y., and Berant, J. (2019). "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)*, pp. 1161–1166.

Gu, J.-C., Ling, Z.-H., Zhu, X., and Liu, Q. (2019). "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)*, pp. 1845–1854.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S., and Smith, N. A. (2018).

"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)*, pp. 107–112.

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016). "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change." In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pp. 1489–1501.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep Residual Learning for Image Recognition." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, pp. 770–778.

Hochreiter, S. (1991). "Untersuchungen zu Dynamischen Neuronalen Netzen. Diploma thesis, Institut für Informatik, Lehrstuhl Prof. Brauer, Technische Universität München.".

Hochreiter, S. and Schmidhuber, J. (1997). "Long Short-term Memory." *Neural Computation*, **9** (8), pp. 1735–1780.

Jaidka, K., Chhaya, N., and Ungar, L. (2018). "Diachronic Degradation of Language Models: Insights from Social Media." In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, pp. 195–200.

Kaneko, M. and Bollegala, D. (2019). "Gender-preserving Debiasing for Pre-trained Word Embeddings." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pp. 1641–1650.

Kingma, D. P. and Ba, J. (2015). "Adam: A Method for Stochastic Optimization." *3rd International Conference on Learning Representations (ICLR 2015).*

Li, F., Liu, N., Jin, H., Zhao, K., Yang, Q., and Zhu, X. (2011). "Incorporating Reviewer and Product Information for Review Rating Prediction." In *Proceedings of the 22nd international joint conference on Artificial Intelligence (IJCAI 2011)*, pp. 1820–1825.

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

Maas, A. L. and Ng, A. Y. (2010). "A Probabilistic Model for Semantic Word Vectors." In *Proceedings of Workshop on Deep Learning and Unsupervised Feature Learning*, pp. 1–8.

Madotto, A., Lin, Z., Wu, C.-S., and Fung, P. (2019). "Personalizing Dialogue Agents via Meta-Learning." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pp. 5454–5459.

McAuley, J. and Leskovec, J. (2013). "Hidden Factors And Hidden Topics: Understanding Rating Dimensions with Review Text." In *Proceedings of the 7th ACM conference on Recommender* *systems (RecSys 2013)*, pp. 165–172.

Michel, P. and Neubig, G. (2018). "Extreme Adaptation for Personalized Neural Machine Translation." In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, pp. 312–318.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). "Distributed Representations of Words and Phrases and Their Compositionality." In *Proceedings of Advances in Neural Information Processing Systems 26 (NIPS 2013)*, pp. 3111–3119.

Miller, G. A. (1995). "WordNet: A Lexical Database for English." *Communications of the ACM*, **38** (11), pp. 39–41.

Mirkin, S. and Meunier, J.-L. (2015). "Personalized Machine Translation: Predicting Translational Preferences." In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*, pp. 2019–2025.

Nadejde, M. and Tetreault, J. (2019). "Personalizing Grammatical Error Correction: Adaptation to Proficiency Level and L1." In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pp. 27–33.

Oba, D., Sato, S., Yoshinaga, N., Akasaki, S., and Toyoda, M. (2019a). "Understanding Interpersonal Variations in Word Meanings via Review Target Identification." In *Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2019)*, No. 129.

Oba, D., Yoshinaga, N., Sato, S., Akasaki, S., and Toyoda, M. (2019b). "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-HLT 2019)*, pp. 2102–2108.

Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R., and Van Durme, B. (2018). "Hypothesis Only Baselines in Natural Language Inference." In *Proceedings of the 7th Joint Conference on Lexical and Computational Semantics (*SEM 2018)*, pp. 180–191.

Rosenfeld, A. and Erk, K. (2018). "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-HLT 2018)*, pp. 474–484.

Swinger, N., De-Arteaga, M., Heffernan IV, N. T., Leiserson, M. D., and Kalai, A. T. (2019). "What are the Biases in My Word Embedding?" In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES 2019)*, pp. 305–311.

Tang, D., Qin, B., and Liu, T. (2015). "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)*, pp. 1014–1023.

Tredici, M. D. and Fernández, R. (2017). "Semantic Variation in Online Communities of Practice." In *Proceedings of 12th International Conference on Computational Semantics (IWCS 2017)*.

Tsuchiya, M. (2018). "Performance Impact Caused by Hidden Bias of Training Data for Recognizing Textual Entailment." In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, pp. 1506–1511.

Wuebker, J., Simianer, P., and DeNero, J. (2018). "Compact Personalized Models for Neural Machine Translation." In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*, pp. 881–886.

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., and Weston, J. (2018). "Personalizing Dialogue Agents: I have a Dog, Do You Have Pets Too?" In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, pp. 2204–2213.

Daisuke Oba: He received his M.S. degree from the University of Tokyo in 2020. He is currently a doctoral student at the University of Tokyo. His research interests include natural language processing (NLP), in particular, personalized NLP.

Shoetsu Sato: A doctoral student at the University of Tokyo, majoring in dialogue and domain adaptation. He received a master's degree from the university in 2017.

Satoshi Akasaki: He received his master's degree from the University of Tokyo, Graduate School of Information Science and Technology in 2018. He is currently a doctoral student of the same university and JSPS Research Fellow (DC1). He conducts research on information extraction and dialogue systems.

Naoki Yoshinaga: He received his Ph.D. from the University of Tokyo in 2005. He has been JSPS Research Fellow (DC1, PD) from 2002 to 2008, and Associate Professor at the University of Tokyo since 2016. His research interests include computational linguistics and machine learning, in particular natural language processing in the wild.

Masashi Toyoda: He received his Ph.D. degree from the Tokyo Institute of Technology in 1999. He worked as Associate Professor at the University of

Tokyo from 2006 to 2018, and has been professor since 2018. His research interests include analysis and interactive visualization of Web, social media, and IoT data.

(Received December 1, 2019)
(Revised February 29, 2020)
(Accepted April 9, 2020)