Understanding Interpersonal Variations in Word Meanings via Review Target Identification

Daisuke Oba[1], Shoetsu Sato[1], Naoki Yoshinaga[2],
Satoshi Akasaki[1], and Masashi Toyoda[2]

[1] Graduate School of Information Science and Technology, The University of Tokyo
[2] Institute of Industrial Science, The University of Tokyo
{oba, shoetsu, ynaga, akasaki, toyoda}@tkl.iis.u-tokyo.ac.jp

Abstract. When people verbalize what they felt with various sensory functions, they could represent different meanings with the same words or the same meaning with different words; we might mean a different degree of coldness when we say '*this beer is icy cold*,' while we could use different words such as "*yellow*" and "*golden*" to describe the appearance of the same beer. These interpersonal variations in word meanings not only prevent us from smoothly communicating with each other, but also cause troubles when we perform natural language processing tasks with computers. This study proposes a method of capturing interpersonal variations of word meanings by using personalized word embeddings acquired through a task of estimating the target (item) of a given reviews. Specifically, we adopt three methods for effective training of the item classifier; (1) modeling reviewer-specific parameters in a residual network, (2) fine-tuning of reviewer-specific parameters and (3) multi-task learning that estimates various metadata of the target item described in given reviews written by various reviewers. Experimental results with review datasets obtained from ratebeer.com and yelp.com confirmed that the proposed method is effective for estimating the target items. Looking into the acquired personalized word embeddings, we analyzed in detail which words have a strong semantic variation and revealed some trends in semantic variations of the word meanings.

Keywords: semantic variation, personalized word embeddings

1 Introduction

We express what we have sensed with various sensory units as language in different ways, and there exist inevitable semantic variations in the meaning of words because the senses and linguistic abilities of individuals are different. For example, even if we use the word "greasy" or "sour," *how* greasy or *how* sour can differ greatly between individuals. Furthermore, we may describe the appearance of the same beer with different expressions such as "yellow," "golden" and "orange." These semantic variations not only cause problems in communicating with each other in the real world but also delude potential natural language processing (NLP) systems.

In the context of personalization, several studies have attempted to improve the accuracy of NLP models for user-oriented tasks such as sentiment analysis [5], dialogue

systems [12] and machine translation [21], while taking into account the user preferences in the task inputs and outputs. However, all of these studies are carried out 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 variation in the user-generated text (input), but also handles *annotation bias* of the output labels (the deviation of output labels assigned by each annotator) and *selection bias* (the deviation of output labels inherited from the targets chosen by users in sentiment analysis) [5]. The contamination caused by these biases hinders us from understanding the solo impact of semantic variation, which is the target in this study.

The goal of this study is to understand which words have large (or small) interpersonal variations in their meanings (hereafter referred to as *semantic variation* in this study), and to reveal how such semantic variation affects the classification accuracy in tasks with user-generated inputs (*e.g.,* reviews). We thus propose a method for analyzing the degree of personal variations in word meanings by using personalized word embeddings acquired through a review target identification task in which the classifier estimates the target item (*objective* output) from given reviews (*subjective* input) written by various reviewers. This task is free from *annotation bias* because outputs 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 allows us to observe only the impact of semantic variations from acquired personalized word embeddings.

A major challenge in inducing personalized word embeddings is the number of parameters (reviewers), since it is impractical to simultaneously learn personalized word embeddings for thousands of reviewers. We therefore exploit a residual network to effectively obtain personalized word embeddings using reviewer-specific transformation matrices from a small amount of reviews, and apply a fine-tuning to make the training scalable to the number of reviewers. Also, the number of output labels (review targets) causes an issue when building a reliable model due to the difficulty of extreme multi-class classification. We therefore perform multi-task learning with metadata estimation of the target, to stabilize the learning of the model.

In the experiments, we hypothesize that words related to the five senses have inherent semantic variation, and validate this hypothesis. We utilized two large-scale datasets retrieved from ratebeer.com and yelp.com that include a variety of expressions related to the five senses. Using those datasets, we employ the task of identifying the target item and its various metadata from a given review with the reviewer's ID. As a result, our personalized model successfully captured semantic variations and achieved better performance than a reviewer-universal model in both datasets. We then analyzed the acquired personalized word embeddings from three perspectives (frequency, dissemination and polysemy) to reveal which words have large (small) semantic variation.

The contributions of this paper are three-fold:

- We established an effective and scalable method for obtaining personal word meanings. The method induces personalized word embeddings acquired through tasks

with objective outputs via effective reviewer-wise fine-tuning on a personalized residual network and multi-task learning.
- We confirmed the usefulness of the obtained personalized word embeddings in the review target identification task.
- We found different trends in the obtained personal semantic variations from diachronic and geographical semantic variations observed in previous studies in terms of three perspectives (frequency, dissemination and polysemous).

2 Related Work

In this section, we introduce existing studies on personalization in natural language processing (NLP) tasks and analysis of semantic variation[3] of words.

As discussed in § 1, personalization in NLP attempts to capture three types of user preferences: (1) *semantic variation* in task inputs (biases in how people use words; our target) (2) *annotation bias* of output labels (biases in how annotators label) and (3) *selection bias* of output labels (biases in how people choose perspectives (*e.g.,* review targets) that directly affects outputs (*e.g.,* polarity labels)). In the history of data-driven approaches for various NLP tasks, existing studies have focused more on (2) or (3), particularly in text generation tasks such as machine translation [17, 14, 21] and dialogue systems [12, 22]. This is because data-driven approaches without personalization tend to suffer from the diversity of probable outputs depending on writers. Meanwhile, since it is difficult to properly separate these facets, as far as we know, there is no study aiming to analyze only the semantic variations of words depending on individuals.

To quantify the semantic variation of common words among communities, Tredici et al. [20] obtained community-specific word embeddings by using the Skip-gram [15], and analyzed obtained word embeddings on 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 and therefore both inputs and outputs are defined by the same writer. As a result, the same word can have dissimilar embeddings not only because they have different meanings, but also because they just appear with words in different topics.[4] In addition, their approach is not scalable to the number of communities (reviewers in our case) since it simultaneously learns all the community-specific parameters.

There also exist several attempts in computational linguistics to capture semantic variations of word meanings caused by diachronic [7, 18, 10], geographic [1, 6], or domain [20] variations. In this study, we analyze the semantic variations of meanings of words at the individual level by inducing personalized word embedding, focusing on

[3] Apart from semantic variations, some studies try to find, analyze, or remove biases related to socially unfavorable prejudices (*e.g.,* the association between the words *receptionist* and *female*) from word embeddings [2, 3, 19, 4]. They analyze word "biases" in the sense of political correctness, which are different from biases in personalized word embeddings we targeted.

[4] Let us consider the two user communities of Toyota and Honda cars. Although the meaning of the word "car" used in these two communities is likely to be the same, its embedding obtained by Skip-gram model from two user communities will be different since "car" appears with different sets of words depending on each community.

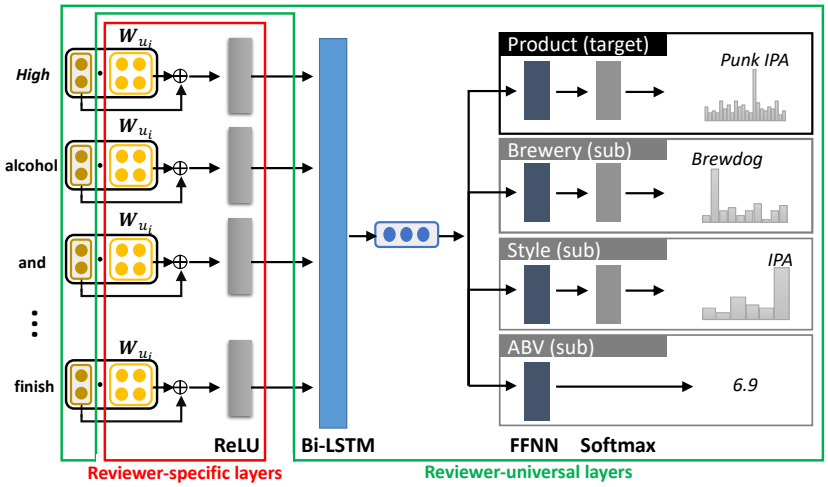


Fig. 1: Overview of our model.

how semantic variations are correlated with word frequency, dissemination, and polysemy as discussed in [7, 20].

3 Personalized Word Embeddings

In this section, we describe our neural network-based model for inducing personalized word embeddings via review target identification (Fig. 1). Our model is designed to identify the target item from a given review with the reviewer's ID. A major challenge in inducing personalized word embeddings is the number of parameters. We therefore exploit a residual network to effectively obtain personalized word embeddings using reviewer-specific transformation matrices and apply a fine-tuning for the scalability to the number of reviewers. Also, the number of output labels makes building a reliable model challenging due to the difficulty of extreme multi-class classification. We therefore perform multi-task learning to stabilize the learning of the model.

3.1 Reviewer-specific Layers for Personalization

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

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

where ReLU is a rectified linear unit function. As shown in Eq. (1), we employ a Residual Network (ResNet) [8] since semantic variation is namely the variation from the reviewer-universal word embedding. By sharing the reviewer-specific parameters for transformation across words and employing ResNet, we aimed for the model to stably learn personalized word embeddings even for infrequent words.

3.2 Reviewer-universal Layers

Given the personalized word embedding $\boldsymbol{e}_{w_i}^{u_j}$ of each word w_i in an input text, our model encodes them through Long short-term Memory (LSTM) [9]. LSTM updates the current memory cell $\boldsymbol{c}_t$ and the hidden state $\boldsymbol{h}_t$ following the equations below:

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

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

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

where $\boldsymbol{i}_t$, $\boldsymbol{f}_t$, and $\boldsymbol{o}_t$ are the input, forget, and output gate at time step t, respectively. $\boldsymbol{e}_{w_i}$ is the input word embedding at time step t, and $\boldsymbol{W}_{\text{LSTM}}$ is a weight matrix. $\hat{\boldsymbol{c}}_t$ is the current cell state. The operation $\odot$ denotes element-wise multiplication and σ is the logistic sigmoid function. We adopt single-layer Bi-directional LSTM (Bi-LSTM) to utilize the past and the future context. As the representation of the input text $\boldsymbol{h}$, Bi-LSTM concatenates the outputs from the forward and the backward LSTM:

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

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

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

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

3.3 Multi-task Learning of Target Attribute Predictions for Stable Training

We consider that training our model for the target identification task can be unstable because its output space (review targets) is extremely large (more than 50,000 candidates). To mitigate this problem, we set up auxiliary tasks that estimate metadata of the target item and solve them simultaneously with the target identification task (target task) by multi-task learning. This idea is motivated by the hypothesis that understanding related metadata of the target item contributes to the accuracy of target identification.

Specifically, we add independent feed-forward layers to compute outputs from the shared sentence representation $\boldsymbol{h}$ defined by Eq. (5) for each auxiliary task (Fig. 1). We assume three types of auxiliary tasks: (1) multi-class classification (same as the target task), (2) multi-label classification, and (3) regression. We perform the multi-task learning under a loss that sums up individual losses for the target and auxiliary tasks. We adopt cross-entropy loss for multi-class classification, a summation of cross-entropy loss of each class for multi-label classification and mean-square loss for regression.

3.4 Training

Considering the case where the number of reviewers is enormous, it is impractical to simultaneously train the reviewer-specific parameters of all reviewers due to memory limitation. Therefore, we first pre-train the model using all the training data without personalization, and then we apply fine-tuning only to reviewer-specific parameters by training independent models from the reviews written by each reviewer.

In this pre-training, the model uses reviewer-universal parameters $\boldsymbol{W}$ and $\boldsymbol{b}$ (instead of $\boldsymbol{W}_{u_j}$ and $\boldsymbol{b}_{u_j}$) in Eq. (1), and then initializes the reviewer-specific parameters $\boldsymbol{W}_{u_j}$ and $\boldsymbol{b}_{u_j}$ by them. This method makes our model scalable even to a large number of reviewers. We fix all the reviewer-universal parameters at the time of fine-tuning.

Furthermore, we perform multi-task learning only during the pre-training without personalization. We then fine-tune reviewer-specific parameters $\boldsymbol{W}_{u_j}, \boldsymbol{b}_{u_j}$ of the pre-trained model while only optimizing the target task. This enables the model to prevent the personalized embeddings from containing the *selection bias*, otherwise the prior output distribution of the auxiliary tasks by individuals can be implicitly learned.

4 Experiments

We first evaluate the target identification task using two review datasets to confirm the effectiveness of the personalized word embeddings induced by our method. If our model can successfully solve this objective task better than the reviewer-universal model obtained by the pre-taining of our reviewer-specific model, it is considered that those personalized word embeddings capture the personal semantic variation. We then analyze the degree and tendencies of the semantic variation in the obtained word embeddings.

4.1 Settings

Dataset We adopt review datasets of beer and services related to foods for evaluation, since there are a variety of expressions that describe what we have sensed with various sensory units in these domains. RateBeer dataset is extracted from ratebeer.com[5] [13] that includes a variety of beers. We selected 2,695,615 reviews about 109,912 types of beers written by reviewers who posted at least 100 reviews. Yelp dataset is derived from yelp.com[6] that includes a diverse range of services. We selected reviews that (1) have location metadata, (2) fall under either the "food" or "restaurant" categories, and (3) are written by a reviewer who posted at least 100 reviews. As a result, we extracted 426,816 reviews of 56,574 services (restaurants or foods, in this study) written by 2,414 reviewers in total. We randomly divided these two datasets into training, development, and testing sets with the ratio of 8:1:1. In the rest of this paper, we refer the former as **RateBeer dataset** and the latter as **Yelp dataset**.

Auxiliary Tasks Regarding the metadata for multi-task learning (MTL), we chose **style** and **brewery** for multi-class classification and **alcohol by volume (ABV)** for regression in the experiments with RateBeer dataset. As for the Yelp dataset, we used **location** for multi-class classification and **category** for multi-label classification.

[5] `https://www.ratebeer.com`
[6] `https://www.yelp.com/dataset`

Table 1: Hyperparameters of our model.

Model		Optimization	
Dimensions of hidden layer	200	Dropout rate	0.2
Dimensions of word embeddings	200	Algorithm	Adam
Vocabulary size (Ratebeer dataset)	100,288	Learning rate	0.0005
Vocabulary size (Yelp dataset)	98,465	Batch size	200

Table 2: Results on the product identification task on RateBeer dataset. Accuracy and RMSE marked with $**$ or $*$ was significantly better than 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).

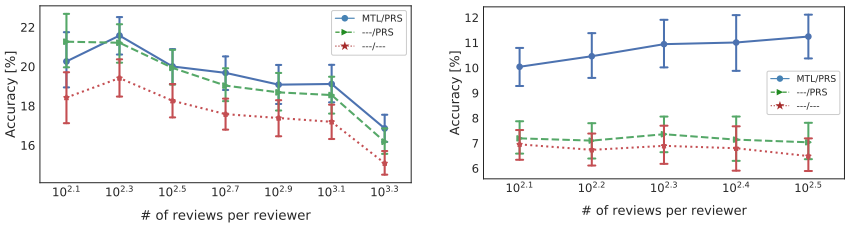
model		target task	auxiliary tasks		
multi-task	personalize	product [Acc.(%)]	brewery [Acc.(%)]	style [Acc.(%)]	ABV [RMSE]
		15.74	n/a	n/a	n/a
	✓	16.69	n/a	n/a	n/a
✓		16.16	(19.98)	(49.00)	(1.428)
✓	✓	**17.56****	(**20.81****)	(**49.78****)	(**1.406***)
baseline		0.08	1.51	6.19	2.321

Table 3: Results on the service identification task on Yelp dataset. Accuracy marked with $**$ was significantly better than the others ($p < 0.01$ assessed by paired t-test).

model		target task	auxiliary tasks	
multi-task	personalize	service [Acc.(%)]	location [Acc.(%)]	category [Micro F1]
		6.75	n/a	n/a
	✓	7.15	n/a	n/a
✓		9.71	(70.33)	(**0.578**)
✓	✓	**10.72****	(**83.14****)	(0.577)
baseline		0.05	27.00	0.315

Models and Hyperparameters In the target item and metadata identification tasks, we compare our model described in § 3 with four different settings.[7] Their differences are, (1) whether the fine-tuning for personalization is applied and (2) whether the model is trained through MTL before the fine-tuning. Table 1 shows major hyperparameters. We initialize the embedding layer by Skip-gram embeddings [15] pretrained from each of the original datasets, containing all the reviews in RateBeer and Yelp datasets, respectively. The vocabulary for each dataset includes all the words that appeared 10 times or more in the dataset. For optimization, we trained the models up to 100 epochs with Adam [11] and selected the model at the epoch with the best results in the target task on the development set as the test model.

[7] We implemented all the models using PyTorch (`https://pytorch.org/`) version 0.4.0.



(a) RateBeer dataset (b) Yelp dataset

Fig. 2: Accuracies in target identification task against the number of parameters per reviewer. In the legend, **MTL** and **PRS** stands for multi-task learning and personalization.

4.2 Overall Results

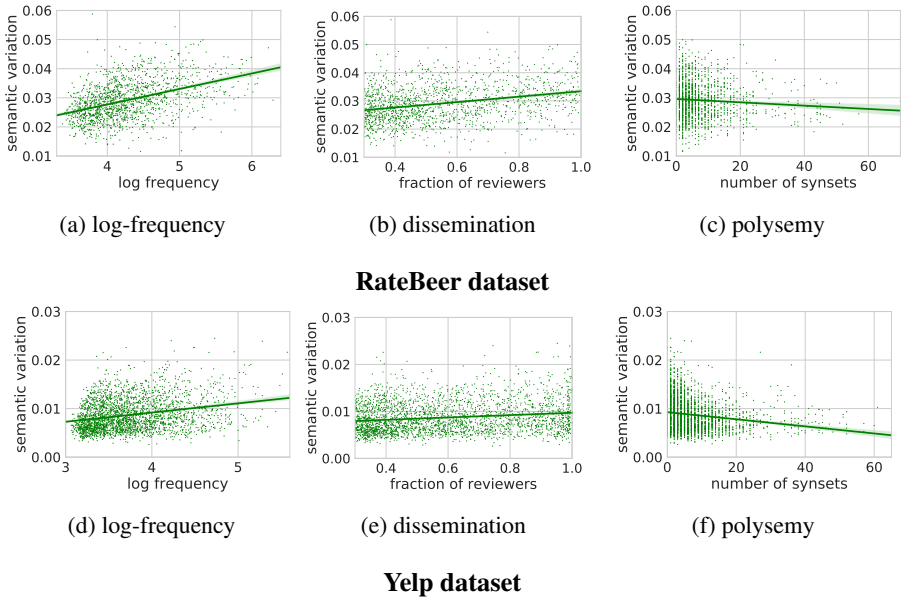
Table 2 and Table 3 show the results on the two datasets. We gain two insights from the results: (1) in the target task, the model with both MTL and personalization outperformed the others, (2) personalization also improves the auxiliary tasks.

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 or restaurant). Simultaneously solving the auxiliary tasks that estimate metadata of the target item guided the model to understand the target item from various perspectives, like part-of-speech tags of words.

We should mention that only the reviewer-specific parameters are updated for the target task in fine-tuning. This means that the improvements on auxiliary tasks were obtained purely by the semantic variations captured by reviewer-specific parameters.

Impact of the number of reviews for personalization We investigated the impact of the number of reviews for personalization when we solved the review target identification. We first grouped the reviewers into several bins according to the number of reviews, and then evaluated the classification accuracies for reviews written by the reviewers in the same bin. Fig. 2 shows the classification accuracy of the target task plotted against the number of reviews per reviewer; 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 review n ($10^{2.1} \leq n < 10^{2.3}$).

Contrary to our expectation, in (a) RateBeer dataset, all of the models obtained lower accuracies as the number of reviews increased. On the other hand, in (b) Yelp dataset, only the model employing MTL and personalization obtained higher accuracies as they increased. We consider that this difference came from the biases of frequencies in review targets. Since RateBeer dataset is heavily skewed, where the top-10% frequent beers account for 74.3% of the entire reviews, while the top-10% frequent restaurants in Yelp dataset account for 48.0% of the reviews. Therefore, it is more difficult to estimate infrequent targets in RateBeer dataset and such reviews tend to be written by experienced reviewers. Although the model without MTL and personalization also obtained



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

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

Fig. 3: Personal semantic variations of the words on the two datasets. Their Pearson coefficient correlations are (a) 0.43, (b) 0.29, (c) -0.07, (d) 0.27, (e) 0.16, (f) -0.19, respectively. The trendlines show 95% confidence intervals from kernel regressions.

slightly lower accuracies even in Yelp dataset, the model with both MTL and personalization successfully exploited the increased reviews and obtained higher accuracies.

4.3 Analysis

In this section, we analyze the obtained personalized word embeddings to see what kind of personal biases exist in each word. Here, we target only the words used by 30% or more reviewers (excluding stopwords) to remove the influences of low frequent words.

We first define the **personal semantic variation**[8] of a word w_i, to determine how the representations of the word are different by individuals, as:

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

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.

[8] Unlike the definition of the semantic variation of the existing studies [20], which measures the degree of change from a point to a point of a word meaning, personal semantic variation measures how much a number of meanings of a word defined by individuals are diverged.

Table 4: The list of top-50 words with the largest (and the smallest) semantic variation on the RateBeer dataset and Yelp dataset. Adjectives are boldfaced.

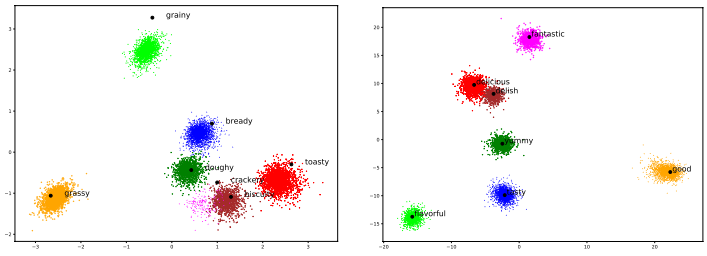
	top-50	**bottom-50**
RateBeer dataset	ery **bready** ark **slight floral toasty tangy updated citrusy soft deep** mainly **grassy** aroma **doughy dissipating** grass ot **great earthy** smell **toasted** somewhat **roasty soapy** perfume **flowery lingering musty** citrus **malty** background malt present hue **minimal** earth **foamy faint dark medium clean nice** copper hay bread herbs **chewy** complexity toast reddish	reminds cask batch oil reminded beyond **canned conditioned double** abv hope horse oats rye brewery blueberry blueberries maple bells **old** cork shame dogfish become dog hand **plastic** course remind christmas cross rogue **extreme organic fat** lost words islands etc growler **hot** heat stout alcohol unibroue pass nitro **longer** scotch **rare**
Yelp dataset	**tasty fantastic great awesome delish excellent yummy delicious good amazing phenomenal superb** asparagus risotto **flavorful** calamari salmon **creamy** chicken got veggies **incredible** ordered scallops sides **outstanding** sausage flatbread shrimp eggplant patio ambiance sandwich **wonderful** desserts **salty** gnocchi **fabulous** quesadilla atmosphere bacon mussels sauce vegetables restaurant broth **grilled** mushrooms ravioli decor food	easily note possibly almost nearly warning aside **opposite alone** even **needless** saving yet mark thus wish apart thankfully **straight possible** iron **short** eye period thumbs **old** deciding **major** zero meaning **exact** replaced fully somehow **single** de key **personal** desired hence pressed rock exactly ups keeping hoping whole meant seeing test hardly

Here, we focus on three perspectives: **frequency**, **dissemination**, and **polysemy** which have been discussed in the studies of semantic variations caused by diachronic or geographical differences of text [7, 6, 20] (§ 2). Fig. 3 shows the semantic variations against the three metrics. Each of the x-axes 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 [16] ((c) and (f)), respectively.

Interestingly, in contrast to the reports by [7] and [20], semantic variations correlate highly with frequency and dissemination, and 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 five senses such as "*soft*" and "*creamy*" frequently appear and their usage depend on feelings and experiences by individuals. Therefore, they 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.

Table 4 shows the top-50 words with the largest (and smallest) semantic variations. As can be seen from the tables, the list of top-50 words contains much more adjectives compared with the list of bottom-50 , which are likely to be used to represent individual feelings that depend on the five senses.

To see in detail what kind of word have large semantic variation, we classify the adjectives of the top-50 (and bottom-50) by the five senses, which are **sight** (vision), **hearing** (audition), **taste** (gustation), **smell** (olfaction), and **touch** (somatosensation). From the results, on the RateBeer dataset, there were more words representing each sense except hearing in the top-50 words compared with the bottom-50. On the other hand, the list of words on Yelp dataset include less words related to the five senses than the RateBeer dataset, but there are many adjectives that could be applicable to various domains (*e.g.,* "*tasty*," and "*excellent*"). This may be due to the domain size of Yelp dataset and the lack of reviews detailing the specific products in the restaurant reviews.



(a) RateBeer dataset (b) Yelp dataset

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

We also analyze the relationships between words to confirm that there are words that gets confused. We use the word "*bready*" and "*tasty*" with the highest semantic variation in each dataset. We visualized the personalized word embeddings using Principal Component Analysis (PCA), with six words closest to the target words in the universal embedding space in Fig. 4. As can be seen, clusters of "*cracky*," "*doughy*," and "*biscuity*" are mixed each other, suggesting that words representing the same meaning may differ by individuals.

5 Conclusions

In this study, we focused on interpersonal variations in word meanings, and explored a hypothesis that words related to the five senses have inevitable personal semantic variations. To verify this, we proposed a novel method for obtaining semantic variation using personalized word embeddings induced through a task with objective outputs. Experiments using large-scale review datasets from ratebeer.com and yelp.com showed that the combination of multi-task learning and personalization improved the performance of the review target identification, which means that our method could capture interpersonal variations of word meanings. Our analysis showed that words related to the five senses have large interpersonal semantic variations.

For future studies, besides factors we worked on this study such as frequency, we plan to analyze relationships between semantic variations and demographic factors of the reviewers such as gender and age which are inevitable for expressing individuals.

We will release the experimental codes for the academic and industrial communities at http://www.tkl.iis.u-tokyo.ac.jp/˜oba/cicling-19/ to facilitate the reproducibility of our results and their use in various application contexts.

Acknowledgements

This work was partially supported by Commissioned Research (201) of the National Institute of Information and Communications Technology of Japan.

References

1. Bamman, D., Dyer, C., Smith, N.A.: Distributed representations of geographically situated language. In: Proc. ACL 2018 (Short Papers). pp. 828–834 (2014)
2. Bolukbasi, T., Chang, K.W., Zou, J.Y., Saligrama, V., Kalai, A.T.: Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In: NIPS 2016. pp. 4349–4357 (2016)
3. Caliskan, A., Bryson, J.J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. Science **356**(6334), 183–186 (2017)
4. Díaz, M., Johnson, I., Lazar, A., Piper, A.M., Gergle, D.: Addressing age-related bias in sentiment analysis. In: Proc. CHI 2018. p. 412. ACM (2018)
5. Gao, W., Yoshinaga, N., Kaji, N., Kitsuregawa, M.: Modeling user leniency and product popularity for sentiment classification. In: Proc. IJCNLP 2013. pp. 1107–1111 (2013)
6. Garimella, A., Mihalcea, R., Pennebaker, J.: Identifying cross-cultural differences in word usage. In: Proc. COLING 2016. pp. 674–683 (2016)
7. Hamilton, W.L., Leskovec, J., Jurafsky, D.: Diachronic word embeddings reveal statistical laws of semantic change. In: Proc. ACL 2016. pp. 1489–1501 (2016)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. CVPR 2016. pp. 770–778 (2016)
9. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
10. Jaidka, K., Chhaya, N., Ungar, L.: Diachronic degradation of language models: Insights from social media. In: Proc. ACL 2018 (Short Papers). pp. 195–200 (2018)
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. ICLR 2015 (2015)
12. Li, J., Galley, M., Brockett, C., Spithourakis, G., Gao, J., Dolan, B.: A persona-based neural conversation model. In: Proc. ACL 2016. pp. 994–1003 (2016)
13. McAuley, J., Leskovec, J.: Hidden factors and hidden topics: understanding rating dimensions with review text. In: Proceedings of the 7th ACM conference on Recommender systems. pp. 165–172 (2013)
14. Michel, P., Neubig, G.: Extreme adaptation for personalized neural machine translation. In: Proc. ACL 2018 (Short Papers). pp. 312–318 (2018)
15. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: NIPS 2013. pp. 3111–3119 (2013)
16. Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM **38**(11), 39–41 (1995)
17. Mirkin, S., Meunier, J.L.: Personalized machine translation: Predicting translational preferences. In: Proc. EMNLP 2015. pp. 2019–2025 (2015)
18. Rosenfeld, A., Erk, K.: Deep neural models of semantic shift. In: Proc. NAACL 2018. pp. 474–484 (2018)
19. Swinger, N., De-Arteaga, M., Heffernan, I., Thomas, N., Leiserson, M.D., Kalai, A.T.: What are the biases in my word embedding? arXiv:1812.08769 (2018)
20. Tredici, M.D., Fernández, R.: Semantic variation in online communities of practice. In: proc. IWCS 2017 (2017)
21. Wuebker, J., Simianer, P., DeNero, J.: Compact personalized models for neural machine translation. In: Proc. EMNLP 2018. pp. 881–886 (2018)
22. Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., Weston, J.: Personalizing dialogue agents: I have a dog, do you have pets too? In: Proc. ACL 2018. pp. 2204–2213 (2018)