Modeling Personal Biases in Language Use by Inducing Personalized Word Embeddings

Daisuke Oba[1], **Naoki Yoshinaga**[2], **Shoetsu Sato**[1], **Satoshi Akasaki**[1], **Masashi Toyoda**[2]
[1]The University of Tokyo, Japan
[2]Institute of Industrial Science, the University of Tokyo, Japan
[1]{oba, shoetsu, akasaki}@tkl.iis.u-tokyo.ac.jp
[2]{ynaga, toyoda}@iis.u-tokyo.ac.jp

Abstract

There exist biases in individual's language use; the same word (*e.g.*, *cool*) is used for expressing different meanings (*e.g.*, *temperature range*) or different words (*e.g.*, *cloudy*, *hazy*) are used for describing the same meaning. In this study, we propose a method of modeling such personal biases in word meanings (hereafter, *semantic variations*) with personalized word embeddings obtained by solving a task on subjective text while regarding words used by different individuals as different words. To prevent personalized word embeddings from being contaminated by other irrelevant biases, we solve a task of identifying a review-target (objective output) from a given review. To stabilize the training of this extreme multi-class classification, we perform a multi-task learning with metadata identification. Experimental results with reviews retrieved from RateBeer confirmed that the obtained personalized word embeddings improved the accuracy of sentiment analysis as well as the target task. Analysis of the obtained personalized word embeddings revealed trends in semantic variations related to frequent and adjective words.

1 Introduction

When we verbalize what we have sensed, there exist inevitable personal biases in word meanings (hereafter, (*personal*) *semantic variations*). For example, when we say "this pizza is greasy," how greasy can vary widely among individuals. When we see the same beer, we may use different words (*e.g.*, *red*, *amber*) to refer its color. The semantic variations will thereby cause problems not only in communicating with each other, but also in building natural language processing (NLP) systems.

Several studies have attempted to personalize models to improve the performance on NLP tasks such as sentiment analysis (Gao et al., 2013) and dialogue systems (Li et al., 2016; Zhang et al., 2018). All of these studies, however, tried to estimate *subjective* output from *subjective* input (*e.g.*, estimating sentiment scores given by reviewers). These personalized models are thereby affected by not only semantic variations in subjective input but also *annotation bias* (deviation of outputs given by the annotators) and *selection bias* (deviation of outputs caused by the deviation of input) (§ 2). This makes it difficult to understand the pure impact of the personal semantic variations.

In this study, aiming at understanding semantic variations and their impact on NLP tasks, we propose a method for modeling personal semantic variations with personalized word embeddings obtained through the review-target identification task. This task estimates the review-target (*objective* output) from a given review (*subjective* input) (§ 3), and is free from *annotation bias* since output labels are given a priori. Also, *selection bias* can be suppressed 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 the choice of inputs. To stabilize the training of this extreme multi-class classification, we apply multi-task learning (MTL) with metadata estimation of the review-target to effectively learn a reliable model (personalized word embeddings).

We validate our hypothesis that words related to the five senses have large semantic variations. We first confirm the impact of personalized word embeddings in the review-target identification using a review dataset obtained from RateBeer, and next evaluate their usefulness in sentiment analysis (§ 4.2). Analysis of the obtained personalized word embeddings on three metrics (frequency, dissemination and polysemy) reveals trends on which words have large semantic variations (§ 4.3).

The contributions of this paper are as follows:

- We established a method to obtain personal semantic variations via multi-task learning on a task with objective outputs (§ 3).
- We categorized personal biases in NLP (§ 2).

- We confirmed the usefulness of personalized word embeddings in review-target identification and sentiment analysis tasks (§ 4.2).
- We revealed trends in personal semantic variations (§ 4.3).

2 Related Work

As discussed in § 1, biases considered by personalization in NLP tasks have three facets: (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)).

Existing studies have modeled (2) and (3) with or without (1) for NLP tasks such as sentiment analysis (Li et al., 2011; Gao et al., 2013; Tang et al., 2015a,b; Chen et al., 2016), 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). However, it is difficult to untangle the different facets of personal biases, there is no study aiming to analyze solely personal semantic variations. Meanwhile, word embeddings induced for a simple NLP task such as sentiment classification conveys less information, which are not suitable for analyzing semantic variations.

Computational linguists have utilized word embeddings to capture semantic variations of words caused by diachronic (Hamilton et al., 2016; Szymanski, 2017; Rosenfeld and Erk, 2018; Jaidka et al., 2018), geographic (Bamman et al., 2014; Garimella et al., 2016) or domain (Tredici and Fernández, 2017) differences. In these studies, they have mainly discussed relationships between semantic variations of words and their frequency, dissemination (the number of users), or polysemy of the words. Hamilton et al. (2016) report that the meanings of more frequent words are more stable over time, and the meanings of polysemous words are likely to change over time since polysemous words appear in diverse contexts (Winter et al., 2014; Bréal, 1897). Tredici and Fernández (2017) report that the meanings of words used by more people are more stable. In this study, we analyze the personal semantic variations by inducing personalized word embeddings, mainly focusing on how frequent, disseminated or polysemous words are biased, following these studies.

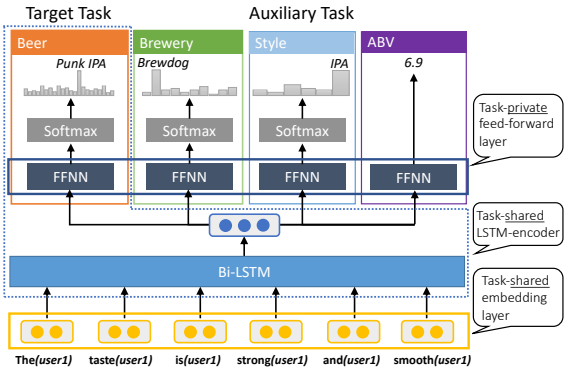


Figure 1: The overview of our model.

3 Personalized Word Embeddings

This section describes our neural network-based model (Figure 1) designed for inducing personalized word embeddings via review-target identification. This model estimates the review-target from a given review.

Model Overview: The whole process is as follows. First, a given review, represented as a sequence of words, is transformed to a sequence of their word embeddings via an embedding layer. Here, our model regards words written by different reviewers as different words for personalization. Next, we apply bi-directional long-short term memory (Bi-LSTM) (Gers et al., 1999) to the sequence of word embeddings and use the concatenation of outputs from the forward and backward LSTMs as a review representation. Finally, a feed-forward layer computes an output probability distribution from the encoded representation of the review.

Multi-task Learning (MTL): The extremely large number of labels (review-targets) makes it difficult to stably train the target identification model. To mitigate this, we jointly train auxiliary tasks that estimate the metadata of the review-target along with the target task. This approach assumes that understanding metadata contributes the performance of the target identification. Concretely, our MTL model contains a task-shared embedding layer, a task-shared LSTM-encoder, and task-private feed-forward layers similarly to (Dong et al., 2015; Luong et al., 2016). In our experiments, these task-private layers consist of three layers for classification and one layer for regression (Figure 1). In the classification tasks, the model computes log probability over target labels as the output and cross-entropy is used as the loss

function. In the regression task, the output is the metadata itself represented as a scalar value and squared error is used as the loss function.

Here, multi-task learning raises a new problem. In auxiliary tasks, since the same reviewer can select the same label multiple times, the personalized word embeddings trained through the multi-task learning may implicitly include the *selection bias* of the output labels depending on the reviewers. Therefore, to exclude those irrelevant biases from the personalized embeddings, we introduce personalized bias terms to feed-forward layers of each task. These bias terms are fixed to the prior distributions of outputs in the training set depending on reviewers so that they absorb selection biases instead of personalized word embeddings.

4 Experiments

We first evaluate the effect of personalization in the target identification task. Next, to confirm the usefulness of the obtained personalized embeddings, we exploit them to solve a sentiment analysis task for extrinsic evaluation. Finally, we analyze the degree and tendencies of semantic variations captured by the obtained personalized word embeddings.

4.1 Settings

Data For training and intrinsic evaluation, we use a huge review dataset about beers constructed from RateBeer[1] (McAuley and Leskovec, 2013). It contains 2,924,163 reviews about 110,369 types of beers with various metadata (*e.g.,* brewery name, style, rating, etc.) written by 29,265 reviewers. From this dataset, we extracted 527,157 reviews about 83,049 types of beers written by the top-100 reviewers who wrote the most reviews, to guarantee enough data size per reviewer. After that, we randomly divided these reviews into training (421,725), development (52,716), and test sets (52,716) in the ratio of 8:1:1. We refer to this dataset as **RateBeer dataset**.

Tasks Our target task takes a beer review and estimates the target beer reviewed in it. Regarding the metadata estimated in multi-task learning (MTL), we chose **style** with 89 types and **brewery** with 6,208 types for classification tasks and **alcohol by volume (ABV)** for a regression task. Note that these metadata are objective and our MTL is free from annotation bias.

[1] https://www.ratebeer.com/

# Layers of Bi-LSTM	1
Dimensions of LSTM output	200
Dimensions of word embeddings	200
Dropout rate	0.5
Mini-batch size	400
Initial learning rate	0.005
Vocabulary size (w/o personalization)	23,556
Vocabulary size (w/ personalization)	469,346

Table 1: Hyperparameters of our model.

In the sentiment analysis task, we estimate the **ratings** of given reviews annotated by the reviewers. The ratings are integers and range from 1 to 20. Here, we solve this task as a regression task since it is natural to treat the fine-grained rating as continuous values.

Models and Hyperparameters In the review-target and its metadata identification tasks, we compare our model described in § 3 with five models with different settings.[2] Their differences are, (1) whether the model is trained through MTL, (2) whether personalization is applied to the embeddings, and (3) whether personalization is applied to the bias term in the output layers. When MTL is not employed, multiple models are independently trained by tasks without sharing layers.

Table 1 shows major hyperparameters. We initialize the embedding layer by pretrained skip-gram embeddings (Mikolov et al., 2013) induced from the training set of RateBeer dataset. The vocabulary is defined by all the words that appeared more than or equal to 10 times in the training set, and the top-100 reviewers have used at least once. For optimization, we train the models up to 100 epochs with Adam (Kingma and Ba, 2015) and select the one at the epoch with the best results on the development set.[3]

In the sentiment analysis task for extrinsic evaluation of the obtained personalized word embeddings, we train another set of models with the same architecture and hyperparameters as the review-target identification models in Figure 1 except that they have only one feed-forward layer for the target regression task. The embedding layers of the models are kept fixed after initialized by the word embeddings extracted from the corresponding review-target identification models with the same settings of personalization and MTL.

[2] All of our models were implemented by PyTorch (https://pytorch.org/) in the version of 0.4.0.

[3] Regarding MTL, we select the model at the epoch with the best results in the target task.

Model MTL	personalized emb.	bias	Target task beer [Acc.(%)]	Auxiliary tasks brewery [Acc.(%)]	style [Acc.(%)]	ABV(%) [RMSE]
			2.99	**8.70**	**46.60**	1.437
	✓		3.32	7.88	44.52	1.462
✓			3.81	8.03	44.12	**1.425**
✓	✓		4.14	7.41	43.74	1.467
✓	✓	✓	**4.47**	7.83	43.93	1.478
baseline			0.03	0.69	5.46	2.284

Table 2: Results on the review-target and its metadata identification.

Model MTL	personalized emb.	bias	sentiment rating [RMSE]
			1.452
	✓		1.406
✓			1.447
✓	✓		1.381
✓	✓	✓	**1.377**
baseline			2.903

Table 3: Results on the sentiment analysis: embedding layers are kept fixed to those of the corresponding models in Table 2.

4.2 Results

Table 2 shows the accuracies on the three classification tasks (product, style, and brewery) and RMSE on the regression task (ABV) through the test sets. We can see two insights from the results: (1) In the target task, the model adopted all the methods outperformed others, (2) In the auxiliary tasks, MTL and personalization had no effect.

As for the first one, since the identification of the review-target requires both detailed understandings of all the related metadata and capturing biases of word meanings, our proposed method considering both elements achieved the best performance as a natural consequence. The second one is not surprising since the metadata estimated in the auxiliary tasks are weakly related to each other. Thus multi-task learning and personalization did not contribute to the improvement of these auxiliary tasks.

Finally, Table 3 shows the results of the sentiment analysis task for extrinsic evaluation. Similarly to the review-target identification, the model with both MTL and personalization performed the best. The personalization of output bias term also slightly improved RMSE. These results confirm that the personalized word embeddings trained through our methods successfully learned task-independent personal semantic variations. In other words, they were helpful even for solving tasks other than the review-target identification.

4.3 Analysis

In this section, we analyze the personalized word embeddings extracted from the best model with MTL and personalization to confirm what kind of personal biases exist in each word. Here, we target on only the words used by more than or equal to 30% of the reviewers excluding stopwords to remove the influences of low frequent words.

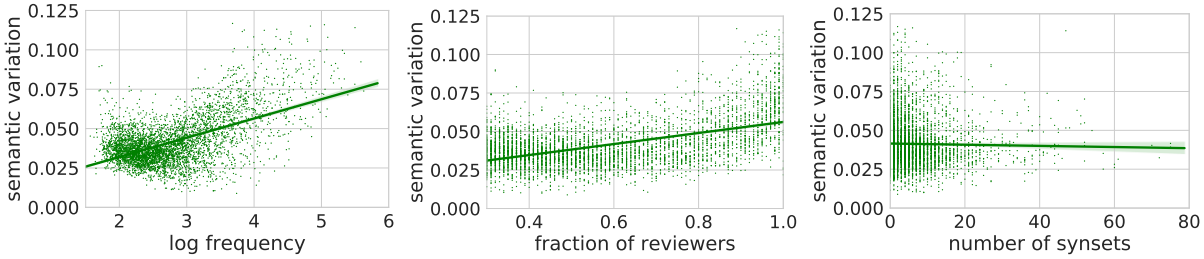
We first define the **personal semantic variations** of word w, to determine how the representations of the word are different by individuals, as:

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

where $\mathrm{U}(w_i)$ is the set of the reviewers who used the word w_i in the training set, $\boldsymbol{e}_{w_i}^{u_j}$ is the word embedding of w_i personalized to reviewer u_j, and $\overline{\boldsymbol{e}}_{w_i}$ is the average of $\boldsymbol{e}_{w_i}^{u_j}$ for $u_j \in \mathrm{U}(w_i)$.

Here, we focus on the three factors, **frequency**, **dissemination**, and **polysemy** which have been studied on semantic variations caused by diachronic, geographical or domain differences of text (see § 2). Figure 2 shows the semantic variations of words against the degree of the three metrics. The x-axes correspond to (a) log frequency of the word, (b) the ratio of the reviewers who used the word, and (c) the number of synsets found in WordNet (Miller, 1995) ver. 3.0, respectively.

Interestingly, in contrast to the results reported in previous studies (Hamilton et al., 2016; Tredici and Fernández, 2017), personal semantic variations correlate highly with frequency and dissemination, and poorly with polysemy in our results. This tendency can be explained as follows: In the dataset used in our experiments, words related to five senses such as "*mild*," "*dry*" and "*soapy*" frequently appear and their usages depend on the feelings and experiences of each individual. Therefore, these words show high semantic variations. Regarding polysemy, although the semantic variations acquired through our method 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 highly skewed contexts related to beer where word senses and their meanings do not significantly fluctuate.



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

Figure 2: Personal semantic variations: The Pearson coefficient correlations are (a) **0.55**, (b) **0.51**, (c) -0.02, respectively. The trendlines show 95% confidence intervals from bootstrapped kernel regressions.

top-50
surprisingly, **nice**, quite, **light**, **pleasant**, actually, though, buttery, **grassy**, really, **bready**, **dusty**, **fruity**, **decent**, **mild**, rather, little, **toffee**, **earthy**, **woody**, **subtle**, **nutty**, **strange**, even, still, **dry**, **tasty**, maybe, **medium**, bit, **soapy**, **interesting**, somewhat, malt, **pretty**, brewery, character, **solid**, lovely, **floral**, **herbal**, **grainy**, **big**, yet, nose, fruit, fairly, aroma, **good**, almost, **metallic**
bottom-50
lasted, primary, system, **secondary**, **personal**, test, acquired, ii, **greater**, standout, roof, england, flow, scored, purchase, partly, colorado, spare, rocks, ounce, se, jug, source, shipping, fullness, denmark, center, **diminished**, greatly, met, spirits, burns, comments, **surrounded**, scores, expectations, carmel, crew, die, **annual**, laces, reading, consumed, handpump, **disappeared**, suits, husks, duck, rise, meal, hall

Table 4: Top-50 words with the largest (and smallest) semantic variations. Adjectives are boldfaced.

Table 4 shows the top-50 words with the largest (and smallest) semantic variations. As can be seen from the table, the top-50 words contain much more adjectives (58%) compared with the bottom-50 ones (16%), which are likely to be used to represent our feelings depending on the five senses.

To see more precisely what kind of words have large semantic variations, we manually classify the adjectives of the top-50 (and bottom-50) by the five senses. 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 ones.

Finally, we analyze the relationships between words beyond the analysis focusing on the single word. We visualized the obtained personalized word embeddings of the word "*mild*" and some closest words in the embedding space as an example in Figure 3. From the results, intersection of the clusters (*e.g.*, "*grainy*" and "*grassy*") means that the same meaning can be represented in different ways by individuals.

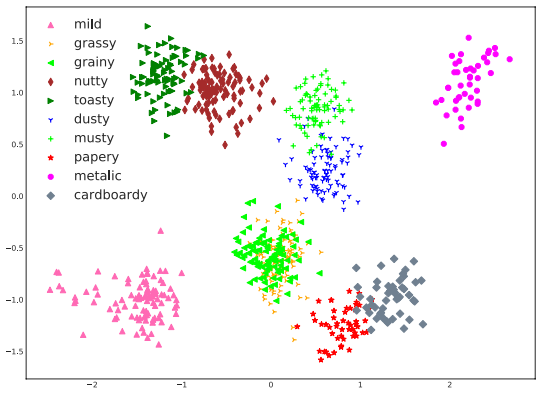


Figure 3: Two-dimensional visualization of the word "*mild*" with some closest words.

5 Conclusions

In this study, we proposed a method of modeling personal semantic variations with personalized word embeddings induced through the review-target identification task. The experimental results on the large-scale beer review dataset showed that personalized word embeddings obtained by multi-task learning with metadata identification improved the accuracy of sentiment analysis as well as the target task. Our analysis revealed that words related to the five senses and adjectives have large semantic variations.

We plan to analyze relationships between semantic variations and user factors of writers who used the target words such as age and gender. We will develop a generic method of inducing personalized word embeddings for any subjective text.

Acknowledgements

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

References

David Bamman, Chris Dyer, and Noah A. Smith. 2014. Distributed representations of geographically situated language. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014, Short Papers)*, pages 828–834.

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

Huimin Chen, Maosong Sun, Cunchao Tu, Yankai Lin, and Zhiyuan Liu. 2016. Neural sentiment classification with user and product attention. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)*, pages 1650–1659.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2015)*, pages 1723–1732.

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

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

Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. 1999. Learning to forget: Continual prediction with LSTM. In *Proceedings of the ninth International Conference on Artificial Neural Networks (ICANN 1999)*, pages 850–855.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 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)*, pages 1489–1501.

Kokil Jaidka, Niyati Chhaya, and Lyle Ungar. 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, Short Papers)*, pages 195–200.

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

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

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

Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence to sequence learning. *International Conference on Learning Representations (ICLR 2016)*.

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

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

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

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

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

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

Terrence Szymanski. 2017. Temporal word analogies: Identifying lexical replacement with diachronic word embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL, Short Papers)*, pages 448–453.

Duyu Tang, Bing Qin, and Ting Liu. 2015a. 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 2015)*, pages 1014–1023.

Duyu Tang, Bing Qin, Ting Liu, and Yuekui Yang. 2015b. User modeling with neural network for review rating prediction. In *Proceedings of 24th International Joint Conference on Artificial Intelligence (IJCAI 2015)*.

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

Bodo Winter, Graham Thompson, and Matthias Urban. 2014. Cognitive factors motivating the evolution of word meanings: Evidence from corpora, behavioral data and encyclopedic network structure. In *Evolution of Language: Proceedings of the tenth International Conference (EVOLANG10)*, pages 353–360.

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

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