Unsupervised Learning of Style-sensitive Word Vectors

Reina Akama[*1], **Kento Watanabe**[†2], **Sho Yokoi**[*‡3], **Sosuke Kobayashi**[§4], **Kentaro Inui**[*‡5]
[*]Graduate School of Information Sciences, Tohoku University
[†]National Institute of Advanced Industrial Science and Technology (AIST)
[§]Preferred Networks, Inc.
[‡]RIKEN Center for Advanced Intelligence Project
{[1]reina.a,[3]yokoi,[5]inui}@ecei.tohoku.ac.jp,
[2]kento.watanabe@aist.go.jp,[4]sosk@preferred.jp

Abstract

This paper presents the first study aimed at capturing stylistic similarity between words in an unsupervised manner. We propose extending the continuous bag of words (CBOW) model (Mikolov et al., 2013a) to learn style-sensitive word vectors using a wider context window under the assumption that the style of all the words in an utterance is consistent. In addition, we introduce a novel task to predict lexical stylistic similarity and to create a benchmark dataset for this task. Our experiment with this dataset supports our assumption and demonstrates that the proposed extensions contribute to the acquisition of style-sensitive word embeddings.

1 Introduction

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

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

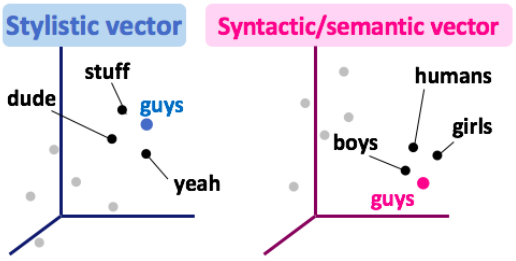

Figure 1: Word vector capturing stylistic and syntactic/semantic similarity.

The contributions of this paper are three-fold. (1) We propose a novel architecture that acquires style-sensitive word vectors (Figure 1) in an unsupervised manner. (2) We construct a novel dataset for style, which consists of pairs of style-sensitive words with each pair scored according to its stylistic similarity. (3) We demonstrate that our word vectors capture the stylistic similarity between two words successfully. In addition, our training script and dataset are available on https://jqk09a.github.io/style-sensitive-word-vectors/.

2 Style-sensitive Word Vector

The key idea is to extend the continuous bag of words (CBOW) (Mikolov et al., 2013a) by distin-

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 572–578
Melbourne, Australia, July 15 - 20, 2018.

guishing nearby contexts and wider contexts under the assumption that a style persists throughout every single utterance in a dialog. We elaborate on it in this section.

2.1 Notation

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

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

2.2 Baseline Model (CBOW-NEAR-CTX)

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

$$\mathcal{C}^{\text{near}}_{w_t} := \{w_{t\pm d} \in \mathcal{U}_t \mid 1 \leq d \leq \delta\} \quad (1)$$

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

$$P(w_t|\mathcal{C}^{\text{near}}_{w_t}) \propto \exp\left(\tilde{\boldsymbol{v}}_{w_t} \cdot \frac{1}{|\mathcal{C}^{\text{near}}_{w_t}|} \sum_{c \in \mathcal{C}^{\text{near}}_{w_t}} \boldsymbol{v}_c\right). \quad (2)$$

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

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

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

$$\mathcal{C}^{\text{all}}_{w_t} := \{w_{t\pm d} \in \mathcal{U}_t \mid 1 \leq d\}, \quad (3)$$

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

2.4 Learning the Style and Syntactic/Semantic Separately

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

Distant-context Model (CBOW-DIST-CTX)

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

$$\mathcal{C}^{\text{dist}}_{w_t} := \mathcal{C}^{\text{all}}_{w_t} \setminus \mathcal{C}^{\text{near}}_{w_t} = \{w_{t\pm d} \in \mathcal{U}_t \mid \delta < d\}. \quad (4)$$

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

Separate Subspace Model (CBOW-SEP-CTX)

As the second extension to distill off aspects other than style, we use both *nearby* and *all* contexts ($\mathcal{C}^{\text{near}}_{w_t}$ and $\mathcal{C}^{\text{all}}_{w_t}$). As Figure 2 shows, both the vector $\boldsymbol{v}_w$ and $\tilde{\boldsymbol{v}}_w$ of each word $w \in \mathcal{V}$ are divided into two vectors:

$$\boldsymbol{v}_w = \boldsymbol{x}_w \oplus \boldsymbol{y}_w, \quad \tilde{\boldsymbol{v}}_w = \tilde{\boldsymbol{x}}_w \oplus \tilde{\boldsymbol{y}}_w, \quad (5)$$

where $\oplus$ denotes vector concatenation. Vectors $\boldsymbol{x}_w$ and $\tilde{\boldsymbol{x}}_w$ indicate the style-sensitive part of $\boldsymbol{v}_w$ and $\tilde{\boldsymbol{v}}_w$ respectively. Vectors $\boldsymbol{y}_w$ and $\tilde{\boldsymbol{y}}_w$ indicate the syntactic/semantic-sensitive part of $\boldsymbol{v}_w$ and $\tilde{\boldsymbol{v}}_w$ respectively. For training, when the context words are near the target word ($\mathcal{C}^{\text{near}}_{w_t}$), we update both the style-sensitive vectors ($\tilde{\boldsymbol{x}}_{w_t}$, $\boldsymbol{x}_c$) and the syntactic/semantic-sensitive vectors ($\tilde{\boldsymbol{y}}_{w_t}$, $\boldsymbol{y}_c$), i.e., $\tilde{\boldsymbol{v}}_{w_t}$, $\boldsymbol{v}_c$. Conversely, when the context words are

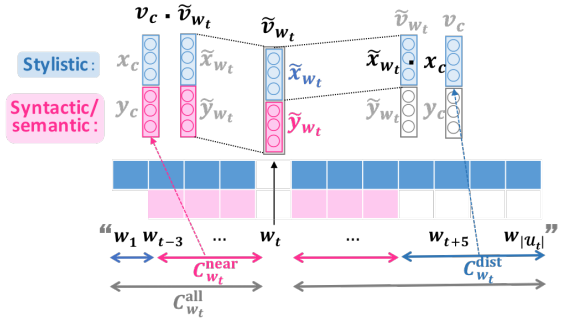

Figure 2: The architecture of CBOW-SEP-CTX.

far from the target word ($\mathcal{C}_{w_t}^{\text{dist}}$), we only update the style-sensitive vectors ($\tilde{\boldsymbol{x}}_{w_t}, \boldsymbol{x}_c$). Formally, the prediction probability is calculated as follows:

$$P_1(w_t|\mathcal{C}_{w_t}^{\text{near}}) \propto \exp\left(\tilde{\boldsymbol{v}}_{w_t} \cdot \frac{1}{|\mathcal{C}_{w_t}^{\text{near}}|} \sum_{c \in \mathcal{C}_{w_t}^{\text{near}}} \boldsymbol{v}_c\right), \quad (6)$$

$$P_2(w_t|\mathcal{C}_{w_t}^{\text{dist}}) \propto \exp\left(\tilde{\boldsymbol{x}}_{w_t} \cdot \frac{1}{|\mathcal{C}_{w_t}^{\text{dist}}|} \sum_{c \in \mathcal{C}_{w_t}^{\text{dist}}} \boldsymbol{x}_c\right). \quad (7)$$

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

3 Experiments

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

3.1 Settings

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

Hyperparameters We chose the dimensions of both the style-sensitive and the syntactic/semantic-sensitive vectors to be 300, and the dimensions of the baseline CBOWs were 300. The learning rate was adjusted individually for each part in $\{\boldsymbol{x}_w, \boldsymbol{y}_w, \tilde{\boldsymbol{x}}_w, \tilde{\boldsymbol{y}}_w\}$ such that "the product of the learning rate and the expectation of the number of updates" was a fixed constant. We ran the optimizer with its default settings from the implementation of Mikolov et al. (2013a). The training stopped after 10 epochs. We fixed the nearby window width to $\delta = 5$.

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

3.2 Stylistic Similarity Evaluation

3.2.1 Data Construction

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

In the task of selecting style-sensitive words, the pairwise inter-annotator agreement was moderate (Cohen's kappa κ is 0.51). In the rating task, the pairwise inter-annotator agreement for two classes ($\{-2, -1\}$ or $\{+1, +2\}$) was fair (Cohen's kappa κ is 0.23). These statistics suggest that, at least

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

Table 1: Results of the quantitative evaluations.

in Japanese, native speakers share a sense of style-sensitivity of words and stylistic similarity between style-sensitive words.

3.2.2 Stylistic Sensitivity

We used this evaluation dataset to compute the Spearman rank correlation (ρ_{style}) between the cosine similarity scores between the learned word vectors $\cos(\boldsymbol{v}_w, \boldsymbol{v}_{w'})$ and the human judgements. Table 1 shows the results on its left side. First, our proposed model, CBOW-ALL-CTX outperformed the baseline CBOW-NEAR-CTX. Furthermore, the $\boldsymbol{x}$ of CBOW-DIST-CTX and CBOW-SEP-CTX demonstrated better correlations for stylistic similarity judgments ($\rho_{style} = 56.1$ and 51.3, respectively). Even though the $\boldsymbol{x}$ of CBOW-SEP-CTX was trained with the same context window as CBOW-ALL-CTX, the style-sensitivity was boosted by introducing joint training with the near context. CBOW-DIST-CTX, which uses only the distant context, slightly outperforms CBOW-SEP-CTX. These results indicate the effectiveness of training using a wider context window.

3.3 Syntactic and Semantic Evaluation

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

3.3.1 Syntactic Sensitivity

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

$$\frac{1}{|\mathcal{V}|N} \sum_{w \in \mathcal{V}} \sum_{w' \in \mathcal{N}(w)} \mathbb{I}[\mathrm{POS}(w) = \mathrm{POS}(w')], \quad (8)$$

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

Table 1 shows SyntaxAcc@N with $N = 5$ and 10. For both N, the $\boldsymbol{y}$ (the syntactic/semantic part) of CBOW-NEAR-CTX, CBOW-ALL-CTX and CBOW-SEP-CTX achieved similarly good. Interestingly, even though the $\boldsymbol{x}$ of CBOW-SEP-CTX used the same context as that of CBOW-ALL-CTX, the syntactic sensitivity of $\boldsymbol{x}$ was suppressed. We speculate that the syntactic sensitivity was distilled off by the other part of the CBOW-SEP-CTX vector, i.e., $\boldsymbol{y}$ learned using only the *near* context, which captured more syntactic information. In the next section, we analyze CBOW-SEP-CTX for the different characteristics of $\boldsymbol{x}$ and $\boldsymbol{y}$.

3.3.2 Semantic and Topical Sensitivities

To test the model's ability to capture the semantic similarity, we also measured correlations with the Japanese Word Similarity Dataset (JWSD) (Sakaizawa and Komachi, 2018), which consists of 4,000 Japanese word pairs annotated with semantic similarity scores by human workers. For each model, we calculate and show the Spearman rank correlation score (ρ_{sem}) between the cosine similarity score $\cos(\boldsymbol{v}_w, \boldsymbol{v}_{w'})$ and the human judgements on JWSD in Table 1[2]. CBOW-DIST-CTX has the lowest score ($\rho_{sem} = 15.9$); however, surprisingly, the stylistic vector $\boldsymbol{x}_{w_t}$ has the highest score ($\rho_{sem} = 28.9$), while both vectors have a high ρ_{style}. This result indicates that the proposed stylistic vector $\boldsymbol{x}_{w_t}$ captures not only the stylistic similarity but also the captures semantic similarity, contrary to our expectations (ideally, we want the stylistic vector to capture only the stylistic similarity). We speculate that this is because not only the *style* but also the *topic* is often consistent in single utterances. For example, "サンタ (Santa Clause)" and "トナカイ (reindeer)" are topically relevant words and these words tend to appear in a single utterance. Therefore, stylistic vectors $\{\boldsymbol{x}_w\}$ using all the context words in an utterance also capture the topic relatedness. In addition, JWSD contains topic-related word pairs and synonym pairs; therefore the word vectors that capture the topic similarity have higher ρ_{sem}. We will discuss this point in

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

	Word w	The top similar words $\{w'\}$ to w w.r.t. cosine similarity	
		$\cos(\boldsymbol{x}_w, \boldsymbol{x}_{w'})$ (stylistic half)	$\cos(\boldsymbol{y}_w, \boldsymbol{y}_{w'})$ (syntactic/semantic half)
Japanese	俺 (I; male, colloquial)	おまえ (you; colloquial, rough), あいつ (he/she; colloquial, rough), ねーよ (not; colloquial, rough, male)	僕 (I; male, colloquial, childish), あたし (I; female, childish), 私 (I; formal)
	拙者 (I; classical*) *e.g., samurai, ninja	でござる(be; classical), ござる(be; classical), ござるよ(be; classical)	僕 (I; male, childish), 俺 (I; male, colloquial), 私 (I; formal)
	かしら (wonder; female)	わね (QUESTION; female), ないわね (not; female), わ (SENTENCE-FINAL; female)	かな (wonder; childish), でしょうか (wonder; fomal), かしらね (wonder; female)
	サンタ (Santa Clause; shortened)	サンタクロース (Santa Clause; -), トナカイ (reindeer; -), クリスマス (Christmas; -)	お客 (customer; little polite), プロデューサー (producer; -), メイド (maid; shortened)
English	shit	fuckin, fuck, goddamn	shitty, crappy, sucky
	hi	hello, bye, hiya, meet	goodbye, goodnight, good-bye
	guys	stuff, guy, bunch	boys, humans, girls
	ninja	shinobi, genin, konoha	shinobi, pirate, soldier

Table 2: The top similar words for the style-sensitive and syntactic/semantic vectors learned with proposed model, CBOW-SEP-CTX. Japanese words are translated into English by the authors. Legend: (translation; impression).

the next section.

3.4 Analysis of Trained Word Vectors

Finally, to further understand what types of features our CBOW-SEP-CTX model acquired, we show some words[3] with the four most similar words in Table 2. Here, for English readers, we also report a result for English[4]. The English result also shows an example of the performance of our model on another language. The left side of Table 2 (for stylistic vector $\boldsymbol{x}$) shows the results. We found that the Japanese word "拙者 (I; classical)" is similar to "ござる (be; classical)" or words containing it (the second row of Table 2). The result looks reasonable, because words such as "拙者 (I; classical)" and "ござる (be; classical)" are typically used by Japanese *Samurai* or *Ninja*. We can see that the vectors captured the similarity of these words, which are stylistically consistent across syntactic and semantic varieties. Conversely, the right side of the table (for the syntactic/semantic vector $\boldsymbol{y}$) shows that the word "拙者 (I; classical)" is similar to the personal pronoun (e.g., "僕 (I; male, childish)"). We further confirmed that 15 the top similar words are also personal pronouns (even though they are not shown due to space limitations). These results indicate that the proposed CBOW-SEP-CTX model jointly learns two different types of lexical similarities, i.e., the stylistic and syntactic/semantic similarities in the different parts of the vectors. However, our stylistic vector also captured the topic similarity, such as "サンタ (Santa Clause)" and "トナカイ (reindeer)" (the fourth row of Table 2). Therefore, there is still room for improvement in capturing the stylistic similarity.

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

[4]We trained another CBOW-SEP-CTX model on an English fan-fiction dataset that was collected from the Web (https://www.fanfiction.net/).

4 Conclusions and Future Work

This paper presented the unsupervised learning of style-sensitive word vectors, which extends CBOW by distinguishing nearby contexts and wider contexts. We created a novel dataset for style, where the stylistic similarity between word pairs was scored by human. Our experiment demonstrated that our method leads word vectors to distinguish the stylistic aspect and other semantic or syntactic aspects. In addition, we also found that our training cannot help confusing some styles and topics. A future direction will be to addressing the issue by further introducing another context such as a document or dialog-level context windows, where the topics are often consistent but the styles are not.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 15H01702. We thank our anonymous reviewers for their helpful comments and suggestions.

References

Reina Akama, Kazuaki Inada, Naoya Inoue, Sosuke Kobayashi, and Kentaro Inui. 2017. Generating stylistically consistent dialog responses with transfer learning. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*. pages 408–412.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matians, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: The concept revisited. *ACM Transactions on Information Systems* 20(1):116–131. https://doi.org/10.1145/503104.503110.

Lucie Flekova, Daniel PreoŢiuc-Pietro, and Lyle Ungar. 2016. Exploring stylistic variation with age and income on twitter. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. pages 313–319. https://doi.org/10.18653/v1/P16-2051.

Daniela Gerz, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen. 2016. Simverb-3500: A large-scale evaluation set of verb similarity. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. pages 2173–2182. https://doi.org/10.18653/v1/D16-1235.

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

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*. pages 994–1003. https://doi.org/10.18653/v1/P16-1094.

Francois Mairesse and Marilyn Walker. 2007. Personage: Personality generation for dialogue. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. pages 496–503.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at the International Conference on Learning Representations*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *The 26th Annual Conference on Neural Information Processing Systems*. pages 3111–3119.

Chiaki Miyazaki, Toru Hirano, Ryuichiro Higashinaka, Toshiro Makino, and Yoshihiro Matsuo. 2015. Automatic conversion of sentence-end expressions for utterance characterization of dialogue systems. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*. pages 307–314.

Xing Niu and Marine Carpuat. 2017. Discovering stylistic variations in distributional vector space models via lexical paraphrases. In *Proceedings of the Workshop on Stylistic Variation at the 2017 Conference on Empirical Methods in Natural Language Processing*. pages 20–27. https://doi.org/10.18653/v1/W17-4903.

Xing Niu, Marianna Martindale, and Marine Carpuat. 2017. A study of style in machine translation: Controlling the formality of machine translation output. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. pages 2804–2809. https://doi.org/10.18653/v1/D17-1299.

Ellie Pavlick and Ani Nenkova. 2015. Inducing lexical style properties for paraphrase and genre differentiation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pages 218–224. https://doi.org/10.3115/v1/N15-1023.

Ellie Pavlick and Joel Tetreault. 2016. An empirical analysis of formality in online communication. *Transactions of the Association of Computational Linguistics* 4:61–74.

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

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pages 35–40. https://doi.org/10.18653/v1/N16-1005.

Yuya Taguchi, Hideaki Tamori, Yuta Hitomi, Jiro Nishitoba, and Kou Kikuta. 2017. Learning Japanese word distributional representation considering of synonyms (in Japanese). Technical Report 17, The Asahi Shimbun Company, Retrieva Inc.

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

Di Wang, Nebojsa Jojic, Chris Brockett, and Eric Nyberg. 2017. Steering output style and topic in neural response generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. pages 2140–2150. https://doi.org/10.18653/v1/D17-1228.

Wei Xu. 2017. From shakespeare to twitter: What are language styles all about? In *Proceedings of the Workshop on Stylistic Variation at the 2017 Conference on Empirical Methods in Natural Language Processing*. pages 1–9. https://doi.org/10.18653/v1/W17-4901.