Distinguishing Japanese Non-standard Usages from Standard Ones

Tatsuya Aoki[†] **Ryohei Sasano**[‡] **Hiroya Takamura**[†] **Manabu Okumura**[†]
[†]Department of Information and Communications Engineering, Tokyo Institute of Technology
[‡]Graduate School of Informatics, Nagoya University
{aoki,takamura,oku}@lr.pi.titech.ac.jp
sasano@i.nagoya-u.ac.jp

Abstract

We focus on non-standard usages of common words on social media. In the context of social media, words sometimes have other usages that are totally different from their original. In this study, we attempt to distinguish non-standard usages on social media from standard ones in an unsupervised manner. Our basic idea is that non-standardness can be measured by the inconsistency between the expected meaning of the target word and the given context. For this purpose, we use context embeddings derived from word embeddings. Our experimental results show that the model leveraging the context embedding outperforms other methods and provide us with findings, for example, on how to construct context embeddings and which corpus to use.

1 Introduction

On social media such as Twitter, we often find posts that are difficult to interpret without prior knowledge on non-standard usage of words. For example, consider the following Japanese sentence[1]:

鯖の	負担が	増える
mackerel-POSS	load-NOM	increase-PRS

"*The load on a mackerel increases*",

which does not make sense given the standard usages for the words in the sentence. But here, *mackerel* is a non-standard usage that means *computer server*. The entire sentence should be interpreted as *"The load on the computer server increases"*. The Japanese word "鯖 (*saba*)" (i.e., mackerel) is used to mean computer server by Japanese computer geeks because *saba* happens to have a pronunciation that is similar to *sābā* (i.e., computer server). When a word is used in a meaning that is different from its dictionary meaning, we call such a usage *non-standard*.[2]

Non-standard usages can be found in many languages (Sboev, 2016). For example, the word "catfish" means a ray-finned fish as in a standard dictionary, but on social media, it can mean a person who pretends to be someone else in order to create a fake identity. Such non-standard usages would be an obstacle to a variety of language processings including machine translation; Google Translate cannot correctly interpret examples such as this. Humans, however, would be able to notice non-standard usages from the inconsistency between the expected word meaning and the context.

The purpose of this work is to develop a method for distinguishing non-standard usages of Japanese words from standard ones. Since it is impractical to construct a large labeled data set for each word, we focus on unsupervised approaches. The main idea in our method is that the difference between the target word's embedding learned from a general corpus and the embedding predicted from the given context would be a good indicator of the degree of non-standardness.

2 Data

We created a dataset for evaluating our method. First, we selected 40 words that have non-standard usages, including computer terms, company/service names, and other Internet slang. Ten

[1] This is interlinear-gloss text representation. POSS, NOM, PRS respectively represent the possessive case, nominative case, and present tense. The third line is the standard translation of the Japanese sentence.

[2] Although some non-standard usages are metaphoric, such as *sunshine* in "*you are my sunshine*", our definition of non-standard usages covers a wider variety of usages, as in the example of "mackerel".

Category	Usage	
	standard	non-standard
Computer terms	416	234
Company/Service names	440	252
Other Internet slang	817	814
Total	1,673	1,300

Table 1: Statistics of the dataset.

of the 40 words were computer terms, another 10 were company/service names, and the remaining 20 were other Internet slang. For each of the 40 target words, we found 100 tweets that contained the target word. Here, we used Twitter as the source for examples since there are many non-standard usages on it. To segment tweets into words, we used the Japanese morphological analyzer MeCab[3] with the standard IPA standard dictionary.[4]

Next, we asked two human annotators to judge whether the usage of the target word in each tweet is standard, non-standard, a named entity, or undecidable. We excluded tweets which at least one annotator judged as undecidable (96 tweets).[5] Cohen's kappa of the annotations for the remaining 3,904 tweets was 0.808. We further excluded tweets which at least one annotator judged as containing a named entity (772 tweets) in order to focus the dataset on our main purpose.[6]

Finally, to create a final dataset, we selected from the remaining 3,132 tweets the 2,973 tweets that are judged as standard by both annotators or as non-standard by them. The selected 2,973 tweets are equivalent to 94.9% of the entire set of tweets, which suggests that human can reliably distinguish non-standard usages from standard ones. The statistics of the final dataset are shown in Table 2.

3 Methodology

Our basic idea for distinguishing word usages is that if a word is used in a non-standard manner, the context words around it will tend to differ from standard context words. To implement this idea, we employed word embeddings. Below, we review the Skip-gram model used for obtaining the word embeddings in Section 3.1 and present our method in Section 3.2.

3.1 Skip-gram

Skip-gram (Mikolov et al., 2013) is widely used for obtaining word embeddings. Given a sequence of words w_1, w_2, ..., w_T as training data, Skip-gram maximizes the likelihood $\frac{1}{T}\sum_{t=1}^{T}\sum_{-m\leq i\leq m, i\neq 0}\log p(w_{t+i}|w_t)$, where m is the window size. w_{t+i} is a context word nearby w_t. $p(w_k|w_t)$ is given by

$$p(w_k|w_t) = \frac{\exp(v_{w_t}^{IN}\cdot v_{w_k}^{OUT})}{\sum_{w=1}^{W}\exp(v_{w_t}^{IN}\cdot v_{w}^{OUT})},$$

where *W* is the vocabulary size of the training data. Skip-gram learns a model predicting context words using word embeddings v^{IN} and v^{OUT}, which are called *input* embedding and *output* embedding respectively.

The embeddings are learned in such a way that $v_{w_t}^{IN}\cdot v_{w_k}^{OUT} - \log\sum_{w=1}^{W}\exp(v_{w_t}^{IN}\cdot v_{w}^{OUT})$ increases if word w_t occurs near w_k in the training corpus. As a result, $v_{w_t}^{IN}\cdot v_{w_k}^{OUT}$ tends to be large for such words and small for word pairs that do not co-occur in the training corpus. We exploited this tendency for recognizing non-standard usages; if the dot-product between the embeddings of the target word and the context words is small, it should indicate a non-standard usage, on the condition that the embeddings have been learned on a general balanced corpus where words correspond to their standard meanings in most cases.

v^{IN} is widely used as a *word embedding* in many studies, while v^{OUT} has not been in the limelight; only a few researchers have examined the effectiveness of v^{OUT} (Mitra et al., 2016; Press and Wolf, 2017). In recent studies, embeddings v^{IN} are usually used for measuring the similarity between words. However, given the characteristics described in the previous paragraph and SGNS's equivalence with shifted positive pointwise mutual information (Levy and Goldberg, 2014), if we want to measure to what extent word w_t tends to co-occur with w_k in the training data, then we should use the similarity of $v_{w_t}^{IN}\cdot v_{w_k}^{OUT}$, instead of $v_{w_t}^{IN}\cdot v_{w_k}^{IN}$.

In this study, we show the importance of using v^{OUT} in a task where we need to see if a word matches its context.

[3] http://taku910.github.io/mecab/

[4] https://ja.osdn.net/projects/ipadic/

[5] The undecidable tweets are meaningless expressions such as the emoticon "(* ´ 茸 ` *)", which includes the word "茸".

[6] Most of the discarded target words are in a named entity, such as the word "尻" in "利尻島". These expressions are different from our definition of non-standard usage.

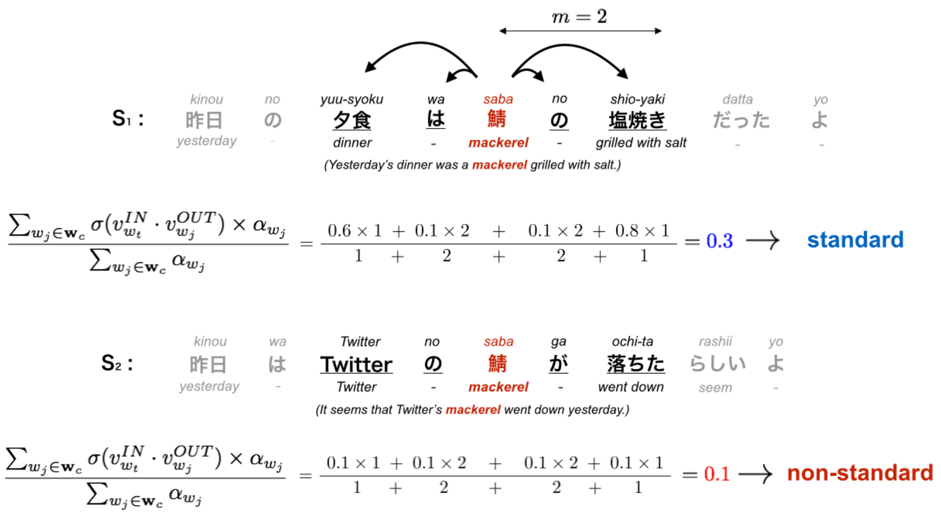
m = 2
S₁ :
kinou 昨日 yesterday
no の -
yuu-syoku 夕食 dinner
wa は -
saba 鯖 mackerel
no の -
shio-yaki 塩焼き grilled with salt
datta だった -
yo よ -
(Yesterday's dinner was a mackerel grilled with salt.)
0.6 × 1 + 0.1 × 2 + 0.1 × 2 + 0.8 × 1 / 1 + 2 + 2 + 1 = 0.3 → standard
S₂ :
kinou 昨日 yesterday
wa は -
Twitter Twitter Twitter
no の -
saba 鯖 mackerel
ga が -
ochi-ta 落ちた went down
rashii らしい seem
yo よ -
(It seems that Twitter's mackerel went down yesterday.)
0.1 × 1 + 0.1 × 2 + 0.1 × 2 + 0.1 × 1 / 1 + 2 + 2 + 1 = 0.1 → non-standard

Figure 1: Overview of our method. Our model exploits context embedding and weighting.

3.2 Distinguishing Non-standard Usages from Standard Ones

Following the idea described in Section 3.1, we propose a method for distinguishing non-standard usages from standard ones by leveraging word embeddings. An overview of our method is shown in Figure 1. We use Skip-gram with Negative Sampling (SGNS) (Mikolov et al., 2013) for obtaining the word embeddings.

Given a target word w_t and its context $\mathbf{w}_c$ as input, we calculate the following weighted average of scaled dot-products as a measure of standardness:

$$\frac{\sum_{w_j \in \mathbf{w}_c} \sigma(v_{w_t}^{IN} \cdot v_{w_j}^{OUT}) \times \alpha_{w_j}}{\sum_{w_j \in \mathbf{w}_c} \alpha_{w_j}}, \quad (1)$$

where $v_{w_t}^{IN}$ is the input embedding for the target word w_t and $v_{w_j}^{OUT}$ is the output embedding for the context word w_j. α_{w_j} is a non-negative weight for the word w_j, and σ is the sigmoid function used for scaling dot-products into a range from 0 to 1. Although the values of α_{w_j} are arbitrary, we will use the values given by the training algorithm used in `word2vec`[7] and `gensim` (Řehůřek and Sojka, 2010), popular tools for obtaining word embeddings. In their training of word embeddings, context words that are closer to the target word are weighted higher.[8] We therefore set α_{w_j} to be $m+1-d_{w_j}$, where m is the window size and d_{w_j} is an integer that represents the distance between w_j and the target word. Hence, this is a decaying weighting. In contrast, with uniform weights, we set α_{w_j} to be 1 for all w_j in the context. We call the score of Equation (1) *standardness*. If the standardness is low, our method regards the instance as non-standard; otherwise, our method regards it as standard. We should note again that, in our method, word embeddings should be learned on a general balanced corpus that is different from the domain of the target instances.

4 Experiment

4.1 Methods for Comparative Evaluation

Our model has three characteristics: (input and output) word embeddings, decaying weights, and a general balanced corpus. We evaluated each of these characteristics in a task distinguishing non-standard usages from standard ones.

First, we verified the effectiveness of the input and output embeddings. We tested a method in which only input embeddings are used to calculate the similarity: the cosine similarity between $v_{w_t}^{IN}$ and $v_{w_j}^{IN}$ instead of $\sigma(v_{w_t}^{IN} \cdot v_{w_j}^{OUT})$, which is a similar framework to that of previous work (Neelakantan et al., 2014; Gharbieh et al., 2016). We then tested a method based on the positive pointwise mutual information (PPMI) (Levy et al., 2015; Hamilton et al., 2016). Here, suppose that M is a matrix in which each element is a PPMI of words w_i and w_j. $v_{w_t}^{IN} \cdot v_{w_j}^{OUT}$ in Equation (1) is replaced with the (t, j)-element of the low-rank approximation of M obtained through singular value decomposition (SVD). We refer to this model as SVD.

[7] `code.google.com/archive/p/word2vec/`

[8] This weighting scheme is mentioned in (Levy et al., 2015).

Corpus	Description	#word	#token
BCCWJ[9]	Japanese Balanced Corpus	131,913	1.1b
Web[10]	Sentences randomly picked from the Web	336,048	6.0b
Wikipedia[11]	Japanese Wikipedia	1,081,154	8.9b
Newspaper[12]	Japanese Newspapers	1,204,914	15.0b

Table 2: Description of corpora.

Next, we replaced the decaying weights α with uniform weights to examine the impact of decaying weights.

Finally, we conducted experiments with different training corpora to examine the impact of the balanced corpus. We used four corpora as training data for obtaining word embeddings. These corpora are described in Table 2.

4.2 Experimental Settings

In the training of the word embeddings, we set the window size to 5, and the dimensions of the word embeddings to 300. We regarded the words with frequency counts of 5 or less in the training data as unknown words and replaced those words with "<unk>". We used `gensim` (Řehůřek and Sojka, 2010) as an implementation of SGNS, where we set the number of negative samples to 10. We used the code provided by Levy et al. (2015) as the SVD implementation. For the evaluations, we ranked test instances in ascending order of standardness score and evaluated the ranking in terms of the area under the ROC curve (AUC) (Davis and Goadrich, 2006).

4.3 Results

Table 3 shows the AUC for each model.[13] First, we examined the impact of the choice of training corpus for obtaining word embeddings. The models with BCCWJ are constantly better than those with other corpora, although BCCWJ is smaller than the others (Table 2). This result suggests that use of a balanced corpus is crucial in our method for this task.

corpus	SGNS IN-OUT		SGNS IN-IN		SVD	
	decay	uni	decay	uni	decay	uni
BCCWJ	**.875**	**.870**	.846	.837	.821	.813
Web	.846	.842	.817	.807	.771	.765
Wikipedia	.827	.821	.824	.805	.739	.732
Newspaper	.844	.839	.825	.810	.770	.764

Table 3: Area under the ROC curve (AUC) in usage classification task for each model.

Next, we examined the impact of context embeddings. Table 3 shows that our model (SGNS IN-OUT) with BCCWJ achieved the best AUCs (.875 and .870), better than the AUCs of SGNS IN-IN with BCCWJ (.846 and .837). This result suggests that input embeddings should be used in combination with output embeddings for the task of judging whether a word matches its context or not. Table 3 also shows that SGNS-based models are better than SVD-based models.

As we discussed in Section 3.2, we used two weighting schemes for each model. Although the AUC of each decaying weight model is larger than that of the corresponding uniform weight model, the differences were not statistically significant.

5 Related Work

The previous studies focused on distinguishing non-standard usages that are multi-word expressions or idiomatic expressions (Kiela and Clark, 2013; Salehi et al., 2015; Li and Sporleder, 2010). The task of this research is similar to new sense detection (Cook et al., 2014). Our research target includes jargon, whose actual meaning is difficult to infer without specific knowledge about its usage (Huang and Riloff, 2010). Recent studies in computational linguistics have used word embeddings and other techniques to capture various semantic changes in words, such as diachronic changes, geographical variations, and sentiment changes (Mitra et al., 2014; Kulkarni et al., 2015; Fremann and Lapata, 2016; Eisenstein et al., 2010; Hamilton et al., 2016; Yang and Eisenstein, 2016).

A few researchers have exploited output embeddings for natural language applications such as document ranking (Mitra et al., 2016) and improving language models (Press and Wolf, 2017).

[9] The Balanced Corpus of Contemporary Written Japanese (Maekawa et al., 2010).

[10] Japanese sentences are collected using the method described in (Kawahara and Kurohashi, 2006).

[11] We downloaded Japanese Wikipedia articles in July 2016 from `https://dumps.wikimedia.org/jawiki/`.

[12] We used editions of the Mainichi Shimbun, Nihon Keizai Shimbun, and Yomiuri Shimbun published from 1994 to 2004.

[13] Although we also conducted experiments with a sigmoid function for the SGNS IN-IN model and with the cosine similarity for the SVD model, their accuracies were worse than those in Table 3.

6 Conclusion

We presented a model that uses context embeddings to distinguish Japanese non-standard usages from standard ones on social media. Our experimental results show that our model is better than the other models tested. They indicate the importance of context embeddings. To sum up, to distinguish non-standard usage, (1) using a balanced corpus as training data for obtaining word embeddings is crucial, (2) exploiting context embeddings derived from *input* and *output* word embeddings of SGNS achieves the best AUC, and (3) decaying weights have little impact on performance.

We are interested in expanding our method for detecting words that have non-standard usages. We are also interested in finding the meanings of the detected non-standard usages.

References

Paul Cook, Jey Han Lau, Diana McCarthy, and Timothy Baldwin. 2014. Novel word-sense identification. In *Proceedings of COLING '14*, pages 1624–1635.

Jesse Davis and Mark Goadrich. 2006. The relationship between precision-recall and roc curves. In *Proceedings of ICML '06*, pages 233–240.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of EMNLP '10*, pages 1277–1287.

Lea Frermann and Mirella Lapata. 2016. A bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4:31–45.

Waseem Gharbieh, Bhavsar Virendra, and Paul Cook. 2016. A word embedding approach to identifying verb-noun idiomatic combinations. In *Proceedings of the 12th Workshop on Multiword Expressions*, pages 112–118.

William L. Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of EMNLP '16*, pages 595–605.

Ruihong Huang and Ellen Riloff. 2010. Inducing domain-specific semantic class taggers from (almost) nothing. In *Proceedings of ACL '10*, pages 275–285.

Daisuke Kawahara and Sadao Kurohashi. 2006. Case frame compilation from the web using highperformance computing. In *Proceedings of LREC '06*, pages 1344–1347.

Douwe Kiela and Stephen Clark. 2013. Detecting compositionality of multi-word expressions using nearest neighbours in vector space models. In *Proceedings of EMNLP '13*, pages 1427–1432.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of WWW '15*, pages 625–635.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Proceedings of NIPS '14*, pages 2177–2185.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Linlin Li and Caroline Sporleder. 2010. Linguistic cues for distinguishing literal and non-literal usages. In *Proceedings of COLING '10*, pages 683–691.

Kikuo Maekawa, Makoto Yamazaki, Takehiko Maruyama, Masaya Yamaguchi, Hideki Ogura, Wakako Kashino, Toshinobu Ogiso, Hanae Koiso, and Yasuharu Den. 2010. Design, compilation, and preliminary analyses of balanced corpus of contemporary written Japanese. In *Proceedings of LREC '10*, pages 1483–1486.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS '13*, pages 3111–3119.

Bhaskar Mitra, Eric T. Nalisnick, Nick Craswell, and Rich Caruana. 2016. A dual embedding space model for document ranking. *arXiv: 1602.01137*.

Sunny Mitra, Ritwik Mitra, Martin Riedl, Chris Biemann, Animesh Mukherjee, and Pawan Goyal. 2014. That's sick dude!: Automatic identification of word sense change across different timescales. In *Proceedings of ACL '14*, pages 1020–1029.

Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2014. Efficient non-parametric estimation of multiple embeddings per word in vector space. In *Proceedings of EMNLP '14*, pages 1059–1069.

Ofir Press and Lior Wolf. 2017. Using the output embedding to improve language models. In *Proceedings of EACL '17*, pages 157–163.

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of Workshop on NewChallenges for NLP Frameworks*, pages 45–50.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2015. A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of NAACL-HLT '15*, pages 977–983.

Aleksandr Sboev. 2016. The sources of new words and expressions in the Chinese internet language and the ways by which they enter the internet language. In *Proceedings of PACLIC '16*, pages 355–361.

Yi Yang and Jacob Eisenstein. 2016. Overcoming language variation in sentiment analysis with social attention. *arXiv:1511.06052*.