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 sentence1:

| 負担が増える | 異なる | 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 “鰤 (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”.

2313
Table 1: Statistics of the dataset.

<table>
<thead>
<tr>
<th>Category</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>standard</td>
</tr>
<tr>
<td>Computer terms</td>
<td>416</td>
</tr>
<tr>
<td>Company/Service names</td>
<td>440</td>
</tr>
<tr>
<td>Other Internet slang</td>
<td>817</td>
</tr>
<tr>
<td>Total</td>
<td>1,673</td>
</tr>
</tbody>
</table>

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

\[
\frac{1}{T} \sum_{t=1}^{T} \sum_{-m \leq i \leq m, i \neq 0} \log p(w_{t+i} | w_t), \quad m
\]

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 \(v^IN \cdot v^OUT - \log \sum_{w=1}^{W} \exp(v_{w_t}^IN \cdot v_{w}^OUT)\) increases if word \(w_t\) occurs near \(w_k\) in the training corpus. As a result, \(v^IN \cdot v^OUT\) 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 \(v^IN \cdot v^OUT\), instead of \(v^IN \cdot v^IN\).

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.
3.2 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 \( w_c \) as input, we calculate the following weighted average of scaled dot-products as a measure of standardness:

\[
\frac{\sum_{w_j \in w_c} \sigma(v^N_{w_t} \cdot v^{OUT}_{w_j}) \times \alpha_{w_j}}{\sum_{w_j \in w_c} \alpha_{w_j}} = \frac{0.6 \times 1 + 0.1 \times 2 + 0.1 \times 0.8 \times 1}{1 + 2 + 1} = 0.3 \rightarrow \text{standard}
\]

where \( v^N_{w_t} \) is the input embedding for the target word \( w_t \) and \( v^{OUT}_{w_j} \) is the output embedding for the context word \( w_j \), \( \alpha_{w_j} \) is a non-negative weight for the word \( w_j \), and \( \sigma \) is the sigmoid function used for scaling dot-products into a range from 0 to 1. Although the values of \( \alpha_{w_j} \) are arbitrary, we will use the values given by the training algorithm used in word2vec\(^7\) and gensim (Rehůřek 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 \( \alpha_{w_j} \) 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 \( \alpha_{w_j} \) to be 1 for all \( w_j \) in the context. We call the score of Equation (1) \textit{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 \( v^N_{w_t} \) and \( v^{OUT}_{w_j} \) instead of \( \sigma(v^N_{w_t} \cdot v^{OUT}_{w_j}) \), 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 \( w_j \). \( v^N_{w_t} \cdot v^{OUT}_{w_j} \) in Equation (1) is replaced with the \((i,j)\)-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).
Table 2: Description of corpora.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Description</th>
<th>#word</th>
<th>#token</th>
</tr>
</thead>
<tbody>
<tr>
<td>BCCWJ(^9)</td>
<td>Japanese Balanced Corpus</td>
<td>131,913</td>
<td>1.1b</td>
</tr>
<tr>
<td>Web(^{10})</td>
<td>Sentences randomly picked from the Web</td>
<td>336,048</td>
<td>6.0b</td>
</tr>
<tr>
<td>Wikipedia(^{11})</td>
<td>Japanese Wikipedia</td>
<td>1,081,154</td>
<td>8.9b</td>
</tr>
<tr>
<td>Newspaper(^{12})</td>
<td>Japanese Newspapers</td>
<td>1,204,914</td>
<td>15.0b</td>
</tr>
</tbody>
</table>

Next, we replaced the decaying weights \(\alpha\) 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 (ˇReh˚uˇrek 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.

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

<table>
<thead>
<tr>
<th>Corpus</th>
<th>SGNS IN-OUT</th>
<th>SGNS IN-IN</th>
<th>SVD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>decay</td>
<td>uni</td>
<td>decay</td>
</tr>
<tr>
<td>BCCWJ</td>
<td>0.875</td>
<td>0.870</td>
<td>0.846</td>
</tr>
<tr>
<td>Web</td>
<td>0.846</td>
<td>0.842</td>
<td>0.817</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>0.827</td>
<td>0.821</td>
<td>0.824</td>
</tr>
<tr>
<td>Newspaper</td>
<td>0.844</td>
<td>0.839</td>
<td>0.825</td>
</tr>
</tbody>
</table>

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; Frermann 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).

\(^{13}\)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


