Japanese Sentiment Classification Using Bidirectional Long Short-Term Memory Recurrent Neural Network

Lasguido Nio Koji Murakami
Rakuten Institute of Technology, Rakuten Inc.
{lasguido.nio, koji.murakami}@rakuten.com

1 Abstract

Conventional sentiment classification techniques often require polarity dictionaries to train the classification model. Those approaches however, require high labor cost for the dictionary creation process. Given the current neural network learning advancements, we try to generalize and simplify these labors. We propose a sentiment classification model based on the Bidirectional Long Short-Term Memory (BiLSTM) network over the distributed word representation. Investigating the effectiveness of feature and dictionaries, we also append the network hidden layer with the Part of Speech tag (POStag) feature and Japanese polarity dictionary information. During our preliminary experiments, we found interesting relationship between those given features. These findings lead our model to achieve the state-of-the-art performance in Japanese sentiment classification task.

2 Introduction

Sentiment classification is the process of identifying opinion that is implicitly or explicitly embedded in a text. Given the tremendous growth in web-content nowadays, demand in machine learning algorithm to analyze those data is increasing. In e-commerce particularly, one can monitor user behavior by understanding the positive or negative spectrum of user review data. From the user's perspective, this helps customers make a product comparison for a purchasing decision. Therefore, automatic sentiment classification of review data has attracted the interest of many researchers.

Conventional approaches on sentiment classification task rely heavily on the concept of bag-of-words or bag-of-n-grams [12, 17, 18, 28]. In these approaches, a text is viewed as terms or a short combination of terms, which omit the grammar rules or word order. Handling these syntactic factors, Nakagawa et al. [16] devised a sentiment model that incorporates a dependency tree enriched with polarity weight. However this approach is considered tedious because specialized knowledge and designing complicated feature templates are necessary [30]. More advanced techniques which are lexicon-based are also investigated [4, 26]. This approach measures sentence polarity by giving a sentimental spectrum score for each word in the sentence.

Here we point out two common challenges: 1) the complexity of building the feature and dictionary resources, and 2) highly expressive model that handles both semantic and syntactic representation of text. In this research, we aim at designing a system that is highly expressive, robust, and able to achieve a good performance result. A robust approach that works not only for an in-domain task, but also for an out-of-domain sentiment classification task.

To address the first problem we utilize a distributed word representation. The distributed word representation is an unsupervised approach that enables us to represent words into a feature vector. With it, we can avoid the usage of handcrafted features and dictionaries.

For the second problem, we employ the Bidirectional Long Short-Term Memory (BiLSTM) deep neural network. This architecture allows us to capture information regarding long-term dependency structure of sentences. In addition, to represent both semantic and syntactic aspects of the text, we train our networks not only from word information but also from POStag information.

More related works are discussed in the next section. We then explain our BiLSTM architecture in Sect. 4. The experiment is briefly explained in Sect. 5. Then, detailed evaluation results of our approaches are presented in Sect. 6. Finally, conclusions are drawn in Sect. 7.

3 Related Work

The task of sentiment classification can be seen as a subset of the text classification problem. Here we are interested in classifying text into a binary polarity state (positive or negative). This task is not new, and prior to our work there has been a lot of text classifying research. Here, we present some works that specifically focus on Japanese language only, and generally relate to our deep learning approach.

Previous approaches to Japanese sentiment analysis rely on either hand-crafted dictionaries or complex feature models. Nakagawa et al. [16] utilize Tree-

CRF which employs spares and binary feature representation. Even though it is promising, this approach is knowledge extensive and requires complex feature models. On the other hand the Stack Denoising Autoencoder approach [30] utilizes dense features through distributed word representation. While this approach performs slightly better than Tree-CRF, this method also gains more advantage because it doesn't rely on complex feature models and can be easily adapted to other domains.

Research on sentiment classification of micro-blogs have been done by Tang et al. [27] and Severyn et al. [21]. The former task utilize distributed word representation as polarity, and the later task utilizes convolutional neural network (CNN) architecture similar to Kim et al. [9] in order to train the word vector. The above-mentioned model was done with single layer CNN, whereas our model utilizes distributed word representation on recursive deep neural network structure (BiLSTM). Socher et al. [23, 24] employ recursive autoencoders over the syntactic tree to embed syntactic information into the learned model.

Rather similar approaches to ours are proposed by Mousa et al. [15] and Yu et al. [29] which utilize a BiLSTM structure for the sentiment analysis task and extract high-level speech features. In our work however, we consider Japanese sentiment classification. Later in our experiment, we also show that adding POStag, WordNet, and Japanese polarity dictionary features can increase the classification performance.

4 Sentiment Classification

Long Short-Term Memory (LSTM) neural networks have an impressive ability to capture both semantic and syntactic information of sentences [8, 25, 1]. Bidirectional Long Short-Term Memory (BiLSTM) architecture is composed of two LSTM which process the text in both directions, this way we are able to capture both previous and subsequent sentence context [29]. In the next following section, we explain about how we utilize this BiLSTM as a classifier for Japanese text.

4.1 Long Short-Term Memory

LSTM takes words from an input sentence in distributed word representation format. Distributed word representation is a n-dimensional vector of continuous values used to represent a word in the vocabulary [3, 13, 2]. Each word in dictionary ($w \in W$) is embedded into n-dimensional space ($L \in \mathbb{R}^{n \times |W|}$). Finally, a word vector can be seen as a single vector in the column of L.

The LSTM architecture consists of a set of recurrently connected memory blocks. Each block contains one memory cell c and three multiplicative units (gates). These gates help the LSTM memory cell to perform write, read, and reset operation. They enable the LSTM memory cell to store and access information over a period of time. These gates are so-called input gate i, forget gate f, and output gate o. Mathematically one block of LSTM can be viewed as:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i), \quad (1)$$

$$f_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_f), \quad (2)$$

$$c_t = f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c), \quad (3)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o), \quad (4)$$

$$h_t = o_t \tanh(c_t), \quad (5)$$

where x_t is a single distributed vector L of word w, an input to the LSTM, σ is a logistic sigmoid function, and h is a hidden vector. The weight W and b subscript represent the edge connection matrix and bias vector.

4.2 Bidirectional Long Short-Term Memory

One drawback of LSTM architecture is that they are only considering the previous context. In order to make LSTM aware of both previous and subsequent context, it needs to process data in both directions with two separate hidden layers [19]. Later these two hidden layers are combined to the same output layer. This architecture is so called BiLSTM.

BiLSTM computes forward $\overrightarrow{h_t}$ and backward $\overleftarrow{h_t}$ hidden sequence from the LSTM output result. Then it produces the output z by iterating from $t = 1$ to T for the forward layer, and from $t = T$ to 1 for the backward layer:

$$z = W_{\overrightarrow{h}z}\overrightarrow{h_t} + W_{\overleftarrow{h}z}\overleftarrow{h_t} + b_z. \quad (6)$$

By computing both a forward and backward layer, BiLSTM allows us to exploit future and history context together at once. Recently, BiLSTM has been used intensively for real-world applications, ranging from signal processing tasks [6, 5] to text processing tasks [14, 11].

4.3 Learning from Features

Here in our experiment, we are trying to reduce the number of features used in the model. However, we employ some generic features that are widely available and relatively easy to obtain. For observation purposes, aside from the distributed word representation that is embedded by our BiLSTM model z^{word}, we also employ 3 more features: part-of-speech tag (POS tag), Japanese SentiWordnet[1][2] feature, and Japanese polar dictionary [10, 7].

The POStag feature shares the same characteristics as the word feature. It has time series information and needs to be processed over time per-token. Thus we embed it to BiLSTM network z^{pos}.

As for the the Japanese SentiWordnet and polarity dictionary, we create a feature vector that is composed by 4 different elements $v = [p_w, n_w, p_p, n_p]$. The two first elements (p_w and n_w) represent the number of the

[1] http://sentiwordnet.isti.cnr.it

[2] http://compling.hss.ntu.edu.sg/wnja/

positive and negative words in the input sentence according to SentiWordnet. It is followed by other two elements (p_p and n_p) that represent the number of the positive and negative word in the input sentence according to polarity dictionary. Later, we put feature vector v through feedforward network

$$z^{feat} = \sigma(W_{vz}v + b_v). \quad (7)$$

At the end of our network, we concatenate everything $z^{all} = [z^{word}, z^{pos}, z^{feat}]$ and put it through a Sigmoid layer, producing the label probability output

$$y = \sigma(W_{zy}z^{all} + b_{zy}). \quad (8)$$

5 Experiment

5.1 Data

In this research, sentiment classification is conducted on Japanese corpora. We use two kinds of corpora: the Tsukuba Corpus[3] and Rakuten Merchant Review Corpus [22].

The Tsukuba Corpus consists of 4,309 review sentences from the travel domain. Each sentence is annotated by two annotators. For each sentence, we took annotation intersection from two annotators. In the end, we obtained 711 negative sentences and 1,811 positive sentences.

The Rakuten Merchant Review Corpus consists of 5,277 review sentences. Each sentence has been annotated for various aspects, such as delivery, packaging, etc. We normalize this corpus by creating a general label that portrays the general sentence polarity. This works three ways: 1) if all annotated aspect is positive we assign a positive label, 2) same rule is applied for the negative label, and 3) when all annotated aspects contain both positive and negative labels, we will assign a neutral label. At last, the number of positive, negative, and neutral sentences was 1,725, 2,468, and 1,084, respectively.

5.2 Experimental Set-up

In this study, we train and test our BiLSTM classifier with Rakuten Merchant Review data. As for the baseline method, we use a stack denoising autoencoder approach[30] (sda), trained by Zhang et al. over the NTCIR-6 OPINION corpus[20][4]. In the end, we evaluate both proposed and baseline systems with the Tsukuba corpus.

6 Results and Discussion

We report two kinds of evaluations, Table 1 and Table 2 respectively show the evaluation results on the Rakuten Merchant Review Corpus and Tsukuba Corpus. In both tables we present 4 evaluation metrics: accuracy (*Acc.*), precision (*P*), recall (*R*), and F1 score (*F1*). Our BiLSTM approach is portrayed as *rnn*. We try to utilize various feature combinations such as POStag *p*, SentiWordnet *w*, and Japanese polar feature *j*. For example *rnn-p-w* means a BiLSTM approach with POStag and SentiWordnet features (BiLSTM + POStag + SentiWordnet).

Table 1: Evaluation results on merchant-review data

	Acc.	P	R	F1
sda	67.89%	64.73%	**95.22%**	77.07%
rnn	92.46%	93.45%	89.92%	91.65%
rnn-p	92.46%	90.35%	94.35%	92.31%
rnn-w[1]	**93.62%**	92.22%	94.80%	**93.49%**
rnn-j	92.26%	92.15%	91.39%	91.77%
rnn-p-w[2]	93.42%	**94.49%**	91.39%	92.92%
rnn-p-j	92.07%	94.09%	88.09%	90.99%
rnn-w-j	91.49%	90.95%	90.95%	90.95%
rnn-p-w-j[3]	92.65%	91.60%	93.09%	92.34%

Table 2: Evaluation results on Tsukuba Corpus

	Acc.	P	R	F1
sda	73.35%	76.98%	**89.73%**	82.87%
rnn	79.58%	93.20%	77.19%	84.45%
rnn-p[1]	**84.62%**	90.18%	88.18%	**89.17%**
rnn-w[2]	83.54%	91.50%	84.98%	88.12%
rnn-j[3]	83.70%	92.68%	83.93%	88.09%
rnn-p-w	82.04%	93.03%	81.06%	86.63%
rnn-p-j	80.69%	**95.16%**	77.03%	85.14%
rnn-w-j	81.52%	92.22%	81.12%	86.31%
rnn-p-w-j	81.56%	92.01%	81.39%	86.38%

As can be seen in both tables, our approach is superior on both corpora in terms of F1 score. For the Rakuten Merchant Review Corpus (Table 1), we obtained the best F1 score with BiLSTM + SentiWordnet feature *rnn-w*. And for the Tsukuba Corpus (Table 2), we obtained the best F1 score with BiLSTM + POStag feature *rnn-p*.

We can also see that the accuracy score is on par with the F1 score. The best accuracy score is achieved by the best F1 score approaches. The best system are BiLSTM + SentiWordnet feature *rnn-w* and BiLSTM + POStag feature *rnn-p* consecutively for Rakuten Merchant Review Corpus and Tsukuba Corpus.

The best precision score is achieved by BiLSTM + POStag + SentiWordnet features *rnn-p-w* on Rakuten Merchant Review Corpus, and BiLSTM + POStag + Japanese polar features *rnn-p-j* on Tsukuba Corpus. These findings indicate that the POStag feature p plays an important role in the precision score for both corpora. It also shows that SentiWordnet feature w plays a more important role for precision score in the Rakuten Merchant Review Corpus, and so does Japanese polar dictionary j for the Tsukuba Corpus.

[3] `http://nlp.mibel.cs.tsukuba.ac.jp/~inui/SA/corpus/`

[4] This might be not a fair comparison because we train and test our model within the same domain corpus. However, in the next evaluation, we evaluate both the proposed and baseline system with the Tsukuba corpus, which is considered out of domain for both systems.

For both evaluation schemes, the baseline approach *sda* manages to get the high recall score. But it has lower performance in terms of precision and accuracy, which make the overall F1 score drop.

Interestingly, we found that if we add more features to the classifier, the performance will slightly decrease. But it is still better than the baseline approach. Overall we can conclude that POStag p and SentiWordnet w features play a more important role in our BiLSTM model. It is indicated by the top three results for each corpus, which is dominated by these features.

7 Conclusion

In this work, we presented our preliminary works on high-performance sentiment classifier for Japanese language using BiLSTM. The proposed method can run without any dictionaries or features. However, adding some features that can be easily obtained resulted in more robust performance.

There are still many things can be done on this topic. Future works might investigate the semantic relations between feature and corpus, and add an attention model to the BiLSTM architecture.

References

[1] Auli, M., Galley, M., Quirk, C., and Zweig, G. Joint language and translation modeling with recurrent neural networks. In *Proc. of the 2013 Conference on Empirical Methods in Natural Language Processing* (October 2013), EMNLP'13, ACL, pp. 1044–1054.

[2] Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. A neural probabilistic language model. *The Journal of Machine Learning Research 3* (mar 2003), 1137–1155.

[3] Collobert, R., and Weston, J. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proc. of the 25th International Conference on Machine Learning* (2008), ICML '08, ACM, pp. 160–167.

[4] Ding, X., Liu, B., and Yu, P. S. A holistic lexicon-based approach to opinion mining. In *Proc. of the 2008 International Conference on Web Search and Data Mining* (2008), WSDM '08, ACM, pp. 231–240.

[5] Fan, Y., Qian, Y., Xie, F.-L., and Soong, F. K. TTS synthesis with bidirectional LSTM based recurrent neural networks. In *The 15th Annual Conference of the International Speech Communication Association* (2014), INTERSPEECH'14, pp. 1964–1968.

[6] Graves, A., Jaitly, N., and Mohamed, A.-r. Hybrid speech recognition with deep bidirectional LSTM. In *Proc. of the 2013 Workshop on Automatic Speech Recognition and Understanding* (2013), ASRU'13.

[7] Higashiyama, M., Inui, K., and Matsumoto, Y. Learning sentiment of nouns from selectional preferences of verbs and adjectives. In *Proc. of the 14th Annual Meeting of the Association for Natural Language Processing* (2008), pp. 584–587.

[8] Hochreiter, S., and Schmidhuber, J. Long short-term memory. *Neural Computation 9*, 8 (Nov. 1997), 1735–1780.

[9] Kim, Y. Convolutional neural networks for sentence classification. In *Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing* (2014), ACL, pp. 1746–1751.

[10] Kobayashi, N., Inui, K., Matsumoto, Y., Tateishi, K., and Fukushima, T. Collecting evaluative expressions for opinion extraction. In *Proc. of the First International Joint Conference on Natural Language Processing* (2005), IJCNLP'04, Springer-Verlag, pp. 596–605.

[11] Lin, B. Y., Xu, F., Luo, Z., and Zhu, K. Multi-channel bilstm-crf model for emerging named entity recognition in social media. In *Proc. of the 3rd Workshop on Noisy User-generated Text* (2017), ACL, pp. 160–165.

[12] Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. Learning word vectors for sentiment analysis. In *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1* (2011), HLT '11, ACL, pp. 142–150.

[13] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. Distributed representations of words and phrases and their compositionality. In *Proc. of the 26th International Conference on Neural Information Processing Systems* (2013), NIPS'13, Curran Associates Inc., pp. 3111–3119.

[14] Misawa, S., Taniguchi, M., Miura, Y., and Ohkuma, T. Character-based bidirectional lstm-crf with words and characters for japanese named entity recognition. In *Proc. of the First Workshop on Subword and Character Level Models in NLP* (2017), ACL, pp. 97–102.

[15] Mousa, A. E.-D., and Schuller, B. W. Contextual bidirectional long short-term memory recurrent neural network language models: A generative approach to sentiment analysis. In *Proc. of the 2017 European Chapter of the Association for Computational Linguistics* (2017), EACL '17, ACL.

[16] Nakagawa, T., Inui, K., and Kurohashi, S. Dependency tree-based sentiment classification using crfs with hidden variables. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (2010), HLT '10, ACL, pp. 786–794.

[17] Pang, B., and Lee, L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proc. of the 42Nd Annual Meeting on Association for Computational Linguistics* (2004), ACL '04, ACL.

[18] Pang, B., Lee, L., and Vaithyanathan, S. Thumbs up? sentiment classification using machine learning techniques. In *Proc. of the 2002 Conference on Empirical Methods in Natural Language Processing* (July 2002), ACL, pp. 79–86.

[19] Schuster, M., and Paliwal, K. K. Bidirectional recurrent neural networks. *Trans. Sig. Proc. 45*, 11 (nov 1997), 2673–2681.

[20] Seki, Y., Evans, D. K., Ku, L.-W., Chen, H.-H., and Kando, N. *Overview of opinion analysis pilot task at NTCIR-6*. 2007, pp. 265–278.

[21] Severyn, A., and Moschitti, A. Twitter sentiment analysis with deep convolutional neural networks. In *Proc. of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2015), SIGIR '15, ACM, pp. 959–962.

[22] Shinzato, K., and Oyamada, Y. What do people write in reviews for sellers?- investigation and development of an automatic classification system. *Journal of Natural Language Processing 25*, 1 (to appear).

[23] Socher, R., Pennington, J., Huang, E. H., Ng, A. Y., and Manning, C. D. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proc. of the 2011 Conference on Empirical Methods in Natural Language Processing* (2011), EMNLP '11, ACL, pp. 151–161.

[24] Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proc. of the 2013 Conference on Empirical Methods in Natural Language Processing* (2013), EMNLP '13, ACL, pp. 1631–1642.

[25] Sutskever, I., Vinyals, O., and Le, Q. V. Sequence to sequence learning with neural networks. In *Proc. of the 27th International Conference on Neural Information Processing Systems - Volume 2* (2014), NIPS'14, MIT Press, pp. 3104–3112.

[26] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. Lexicon-based methods for sentiment analysis. *Computational Linguistics 37*, 2 (June 2011), 267–307.

[27] Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., and Qin, B. Learning sentiment-specific word embedding for twitter sentiment classification. In *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics* (June 2014), ACL, pp. 1555–1565.

[28] Tieleman, T. Training restricted boltzmann machines using approximations to the likelihood gradient. In *Proc. of the 25th International Conference on Machine Learning* (2008), ICML '08, ACM, pp. 1064–1071.

[29] Yu, Z., Ramanarayanan, V., Suendermann-Oeft, D., Wang, X., Zechner, K., Chen, L., Tao, J., Ivanou, A., and Qian, Y. Using bidirectional lstm recurrent neural networks to learn high-level abstractions of sequential features for automated scoring of non-native spontaneous speech. In *Proc. of the 2015 Workshop on Automatic Speech Recognition and Understanding* (2015), ASRU'15.

[30] Zhang, P., and Komachi, M. Japanese sentiment classification with stacked denoising auto-encoder using distributed word representation. In *Proc. of the 2015 Pacific Asia Conference on Language, Information and Computation* (2015).