International Journal of Advanced Intelligence
Volume 9, Number 4, pp.541-555, December, 2017.

AIA International Advanced Information Institute

Review Score Estimation Based on Transfer Learning of Different Media Review Data

Kazuyuki Matsumoto, Fuji Ren, Minoru Yoshida and Kenji Kita
Tokushima University, Minamijosanjima-cho 2-1, Tokushima
Tokushima, 770-8506, Japan
matumoto@is.tokushima-u.ac.jp

Received (15 Sep. 2017)
Revised (28 Dec. 2017)

We propose a model to classify reviews based on review data from different media sources. Recently, research has been actively conducted on transfer learning between different domains with various kinds of big data as the target. The fact that evaluation expressions often vary in different domains presents a barrier to reputation analysis. Users commonly use various linguistic expressions to refer to creative works, depending on the specific media form. For example, the terms or expressions used in *anime* to describe creative works within that medium are different from the expressions used in *comics*, or *games* or *movies*. These differences can be considered as features of each individual medium. We should expect, then, that there would be differences in evaluation expressions among the various media, as well. We analyze the effects of such differences on classification accuracy by conducting transfer learning between review data from different media and demonstrate compatibility between the original (pre-transfer) and target (post-transfer) media by constructing a review classification model. As a result of our evaluation experiments, we are able to more accurately estimate review scores without using SO-Scores for training review fragments based on Long Short-Term Memory (LSTM) rather than using a method based on SO-Scores.

Keywords: review classification; transfer learning; Long Short-Term Memory; different media.

1. Introduction

Recently, it has become relatively easy to analyze the thoughts and feelings of people in their daily lives through a qualitative analysis of social big data. In particular, given the spread of Internet shopping among the general population, sellers are now able to use much more extensive data for consumer analysis. As a result, techniques for opinion/reputation analysis have progressed dramatically.

As important input into a potential decision to buy creative works, word-of-mouth repetition from customers who have already bought or consumed one of these works can be extremely helpful. However, the appeal of creative works such as movies or music, which include multiple artistic elements, is strongly dependent on personal sensibilities; if the artistic elements that comprise these works do not match the tastes of those who have seen or listened to them, an unreasonable,

irrelevant or unfair evaluation may be given. Furthermore, the fact that reviewers tend to provide higher evaluations exclusively on the works they like can make it difficult to obtain truly useful, objective information.

Today, a number of word-of-mouth websites offer reviews to potential consumers. However, if unfair reviews are included, informed choice is impeded. As a consequence, we sought to devise a method to classify reviews obtained from large word-of-mouth data sets that would support human judgment.

With the development of consumer-generated media (CGM), it is also possible to collect opinions from SNS. Such opinions are even more unedited than those from the word-of-mouth websites. However, existing studies have not yet succeeded in constructing an evaluation expressions dictionary with full consideration of the detailed categories of the creative works targeted for review.

Importantly, there is no evaluation expression dictionary specifically intended for review analysis of creative works, and it is not clear that existing dictionaries are applicable. In this study, we focus on the media of creative works, to which little previous attention has been paid. We propose a method to create a review classification model from review sentences relating to different media. In so doing, we can create rather extensive training data for the various media with a relatively few review sentences. We also evaluate the effectiveness of the proposed method by comparing it with existing review classification methods.

2. Related Works

2.1. *Review Classification*

A number of researchers have studied review classification [1],[2]. Manually constructed evaluation expression dictionaries have proven to be very helpful and have been used in numerous studies on evaluation analysis.

There have been a variety of studies on binary classification of review sentences, identifying reviews as either positive or negative. There have also been three-value classification studies, which classify review sentences as positive, negative, or neutral, as well as five-level scoring schemes.

Algorithms such as SVR (Support Vector Regression) and Metric Labeling [3],[4] are suitable for performing these sorts of tasks. However, review sentences often include both positive and negative elements. Therefore, before using review sentences and their evaluation scores as training data, each sentence of a given review should be judged in terms of both its positive and negative points.

There are existing studies dealing with how to classify reviews related to reputation into binary categories such as subjective or objective [5]. However, most of the studies validate the effectiveness of their methods in a particular domain, such as cars, music, etc. Because there are distinctive characteristics in the review sentences for each domain, it would seem natural to create classifiers by dividing review sentences in each domain category. However, this means that it is necessary to create applicable dictionaries or corpora each time the target domain category changes.

In recent years, it has become common to release creative works in several different media. For example, animation or comics have also been released as a TV game product. As a result, one may have to analyze reviews using the same dictionary or corpus even though the target media for the various reviews may be different.

2.2. *Transfer Learning*

The studies referred to above have problems in domain adaptation, corpus construction with domain adaptation, or general versatility. Some domain categories lack labelled data. Consequently, a versatile method is desirable. Recently, a number of studies have focused on methods to "transfer learning." These methods adapt the corpora or dictionary of one domain to another domain [6].

Transfer learning here essentially means adapting data from one domain to a target domain when a new classifier is to be devised for the target domain but the target domain lacks sufficient training data. Methods to perform this sort of adaption vary. Some are superficial and replace semantically similar words. Others are more complex and judge words that change emotional polarities depending on the domain.

The "Fine Tuning" method effectively trains a model for other tasks by using models and weights learned in other tasks. The models and weights are trained from huge image data sets and are frequently used in the image recognition task of Fine Tuning.

3. Proposed Method

In this study, we conduct transfer learning for the review classification of creative works. Specifically, we classify the broad review target of "creative works" into media or genre categories such as *comics*, *animations*, *dramas*, *games*, *novels*, *sfx*, *Japanese films*. We then attempt to find the best combination of media for transfer learning suitable for review classification by conducting transfer learning between the different media in each media category.

From review posting websites, we collect target review sentences with an evaluation score of *good*, *bad* or *ordinary*, using the sentences explicitly identified as either *good* or *bad* evaluations as training data.

Other sentences are used as evaluation data. The construction flow of our review score estimation model is shown in Fig.1. Below, we explain the proposed method, which enables the transfer learning of a review score estimation model from different media.

3.1. *Extraction of Review Fragment*

Reviews consist of multiple sentences. These can be roughly divided into two classifications: the *body*, which involves direct evaluation, and *others*, which are not directly evaluative, such as a synopsis.

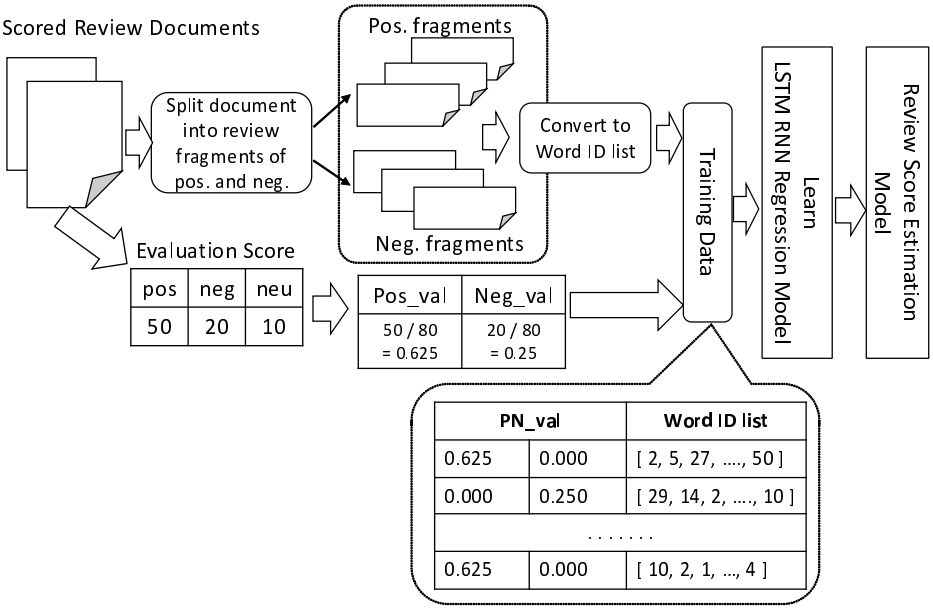


Fig. 1. Construction flow of review score estimation model.

Many of the reviews posted on our target websites clearly indicate that they are the bodies of reviews, as shown in Fig.2. We extract these by pattern matching and use them as labelled training data.

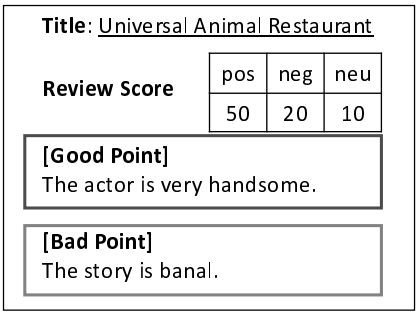


Fig. 2. Example of review data.

Specifically, we judge review fragments as indicating a *good point*, a *bad point* or *neither*. Both positive and negative examples are needed to train the review evaluation scores. As the length of the target review increases, the learning cost will rise, which can make training more difficult. Consequently, we divide reviews into fragments of positive and negative, then train them with the

evaluation scores. For example, if a review has evaluation scores of $positive = 20, negative = 10, and neutral = 5$, we determine the good point score for the fragment as $Score_{pos} = 20/(20 + 10 + 5) = 0.571$ and the bad point score as as $Score_{neg} = 10/(20 + 10 + 5) = 0.286$ and use these for training.

3.2. *Transferring*

To process transfer learning, we use distributed representations of words [7]. A word distributed representation expresses a word in a corpus as a real-valued fixed dimension vector based on peripheral information. It is possible to express a sentence's distributed representation by using the sum of the word distributed representations; however, the position of the word in a sentence is often important.

We quantize word distributed representations by unsupervised clustering and estimate the positive/negative score of a sentence by using Long Short-Term Memory, which is a kind of Recurrent Neural Network mainly used for sequential data learning. We then attempt to fine-tune the LSTM recurrent neural networks by using weights learned for six emotion estimations as initial weights for the model. Fine tuning enables more efficient construction of a score estimation model than using "from scratch" model training with a huge amount of data. The pre-training in our method is shown in Fig.3.

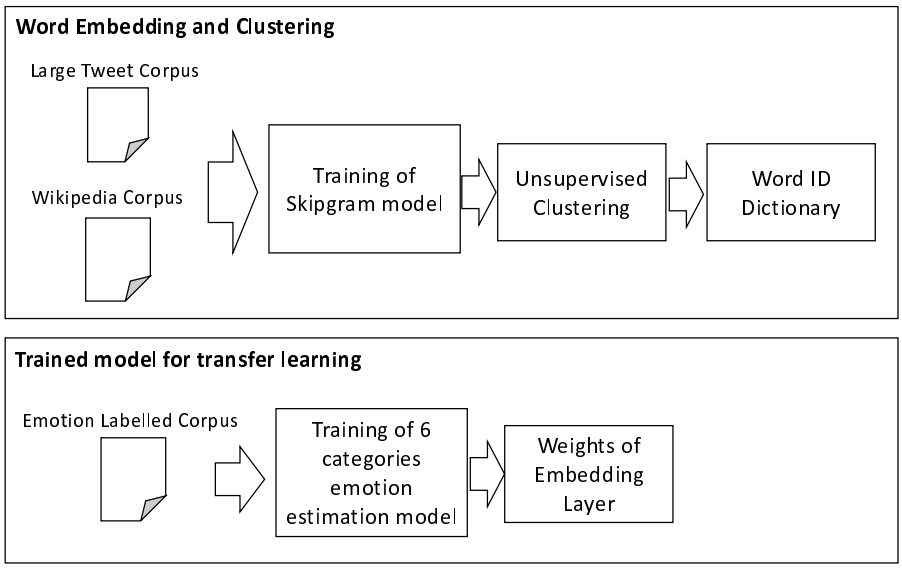


Fig. 3. Pretraining for LSTM training

We use a repeated bisection method[9] as our word clustering algorithm and use "bayon"[10] as a clustering tool. The number of clusters was empirically set at 20,000.

Because we aimed to estimate an evaluation score as a sequential value, we train the LSTM regression prediction model instead of using a simple binary classification model.

3.3. *Extended Long Short-Term Memory*

Extended Long Short-Term Memory Recurrent Neural Networks (Extended LSTM, hereafter, LSTM)[11] is a recurrent neural network that can memorize past information by using sequential information as input and can forget information as time advances. The schematic diagram of the LSTM block is shown in Fig.4.

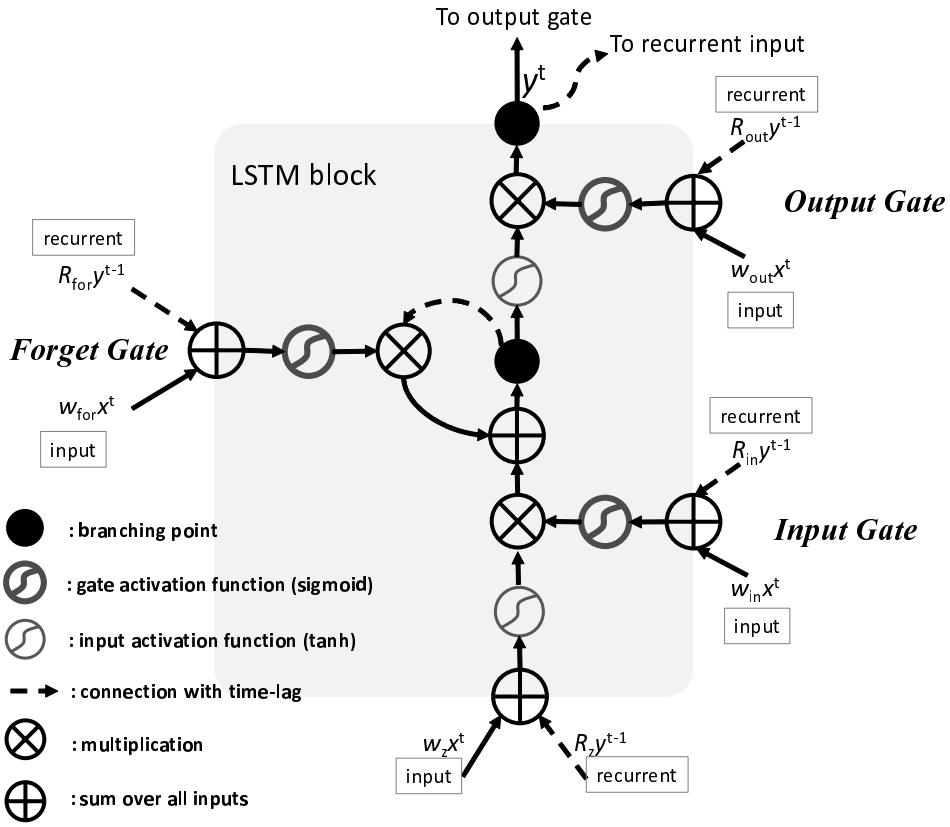


Fig. 4. LSTM block (Forget gate).

Normal Recurrent Neural Networks can train by keeping an internal state when a hidden layer of past time is used for current time as input; however, the gradient fades away during training by backpropagation. LSTMs have a storage cell called the Constant Error Carousel (CEC) and can avoid gradient disappearance by saving

errors in this storage cell.

As a result of analyzing the target review data, we found that the lengths of the review fragments used for our study were concentrated between 10 and 20 words. Therefore, we decided to set a maximum of 20 valid words. At that time, we encountered problems stemming from the fact that unknown words were included in the pre-training corpus, and some words with different polarity were included in the same cluster. Therefore, we defined words that co-occurred with evaluation expressions in the training review data as SO-words; we separated them from the clustering result and allocated different IDs. With this strategy, we are able to use important words in the training data as valid words and can avoid the problem of having words with different polarity assigned to the same cluster ID.

Bidirectional LSTM is another type of neural network that extends an input to a hidden layer of the LSTM bi-directionally. Normal LSTM uses the t-1 state as input; however, bi-LSTM also uses the t+1 state as input. Because it is known that bi-LSTM achieves high accuracy depending on the task, we decided to use bi-LSTM for training the review score estimation model.

Table1 shows parameters for each layer of the LSTM and bi-LSTM. *cid* indicates the number of word ID types. The dropout rate between each layer was set as 0.1. ReLU was used as an activation function.

Table 1. Parameters for each layer of LSTM/bi-LSTM

type	epochs	layer-1(unit)	layer-2(unit)	layer-3(unit)	minibatch size
LSTM	10	Embedding(*cid*)	LSTM(128)	Dense(2)	32
bi-LSTM	10	Embedding(*cid*)	bi-LSTM(128)	Dense(2)	32

4. Experiment

We evaluate the results of the experiment by using Accuracy (Eq.1) and Mean Squared Error (MSE: Eq.2) as the basis for the evaluation. In Eq.1, SgnMax() is a function that returns polarity (1 or -1) based on a comparison of positive and negative evaluation score values. If the values are the same, the equation returns 0.

Match() is a function that returns 1 when polarities match and 0 when they do not match. N indicates the number of evaluation data. $Sc_p^a, Sc_n^a, Sc_p^o, Sc_n^o$ give, respectively, the positive/negative scores of correct evaluations and the positive/negative scores of estimated evaluations.

$$Accuracy = \frac{1}{N} \sum Match(SgnMax(Sc_p^a, Sc_n^a), SgnMax(Sc_p^o, Sc_n^o)) \tag{1}$$

$$MSE = \frac{1}{N} \sum \left((Sc_p^o - Sc_p^a)^2 + (Sc_n^o - Sc_n^a)^2 \right) \tag{2}$$

Table 2. Number of Review Flagment for Training

Test \ Train	comic	anime	drama	novel	sfx	game	jmovie
comic	46724	114606	56136	56227	51640	80043	56179
anime	114606	67882	77294	77385	72798	101201	77337
drama	56136	77294	9412	18915	14328	42731	18867
novel	56227	77385	18915	9503	14419	42822	18958
sfx	51640	72798	14328	14419	4916	38235	14371
game	80043	101201	42731	42822	38235	33319	42774
jmovie	56179	77337	18867	18958	14371	42774	9455

Table 3. Number of Test Data

Media Type	Number of Review Documents
comic	18321
anime	23673
drama	5016
novel	8109
sfx	1219
game	11579
jmovie	5851

As a comparative method, we use a score estimation method based on a Support Vector Machine (SVM). We use Primal Estimated sub-GrAdient SOlver (Pegasos) for our SVM. This is an L2-regularization and L1-loss SVM using a loss function. We use word cluster as the dimension of the feature, and use the sum of the affiliation degree to cluster as the value of the feature. We also conduct evaluation experiments with the following two conditions:

- SO-Score words as other dimensions of the feature
- averaged word distributed representation vector as the feature

4.1. *Dataset*

The number of training review fragments for each combination of media is shown in Table2, and the number of reviews for each medium is shown in Table3. We use all data for the experiment although the numbers of data values vary in each medium. The review documents were collected from the review site *sakuhin database*[a].

Because there is bias in the number of labels (positive/negative), we use an over sampling method to complete the numbers. For over sampling, we extend the

[a]https://sakuhindb.com/

example data. We randomly select sentences that have annotated labels with a smaller number of examples and randomly dropout words in the sentences with probability of 0.1.

By regarding the dropout words as unknown words, the newly generated sentences become similar to, but not identical to, the original sentences. By adding the generated sentences to the training data, we produce the data values for each label. For the pre-training word distributed representation vector, 1) approximately five million tweets were collected randomly from Twitter, together with 2) articles from Wikipedia Japanese (2017.Jan.1). We removed noise from both and tokenized them by morphological analyzer MeCab[12]. As the training model for the Twitter data, the skipgram model on word2vec[7] was used. Vector dimension was set as 500 and the context window size was set at 5. As the training model for the Wikipedia data, the skipgram model on fastText[8] was used. The number of subword characters was set at 3 to 6, the vector dimension was set at 300, and the context window size was set at 5.

A Japanese evaluation polarity dictionary[13],[14] was used for SO-Score calculation for the baseline method. The calculation equations for SO-Score and SO-Vec are shown in Eq.8.

N_{pos} indicates the number of positive expressions, while N_{neg} indicates the number of negative expressions. $h(w, pos)$ indicates the frequency with which w co-occurs with positive expressions, and $h(w, neg)$ indicates the frequency with which w co-occurs with negative expressions. $pmi(w, pos)$ and $pmi(w, neg)$ indicate the pointwise mutual information (PMI) of word w for positive and negative. A higher PMI means that there is a stronger relationship to either positive or negative. SO-Vec is a vector having values obtained by calculating the ratio of $pmi(w, pos)$ and $pmi(w, neg)$ as vector elements; we use it as the estimated evaluation score.

$$pmi(w, pos) = log(\frac{h(w, pos)(N_{pos} + N_{neg})}{h(w, pos) + h(w, neg))N_{pos}}) \tag{3}$$

$$pmi(w, neg) = log(\frac{h(w, neg)(N_{pos} + N_{neg})}{h(w, pos) + h(w, neg))N_{neg}}) \tag{4}$$

$$SO - Score = pmi(w, pos) - pmi(w, neg) \tag{5}$$

$$score_{w,pos} = \frac{pmi(w, pos)}{(pmi(w, pos) + pmi(w, neg))} \tag{6}$$

$$score_{w,neg} = \frac{pmi(w, neg)}{(pmi(w, pos) + pmi(w, neg))} \tag{7}$$

$$SO - Vec = \left(\sum_{w \in T} score_{w,pos}, \sum_{w \in T} score_{w,neg} \right) \tag{8}$$

4.2. *Results*

Fig.5 shows score estimation results in each experiment. The horizontal axis indicates the media types added to the training data; the vertical axis indicates the types of evaluation media.

Fig.6 shows accuracies using a calculation based on a comparison of positive/negative scores.

5. Discussions

It is apparent that better estimation results were obtained when larger amounts of training data were used. This might be because we did not adjust the amount of data for the various media.

We did not observe a significant difference in MSE for the different distributed representation models. However, the method based on the regression estimation model using LSTM or bi-LSTM produced smaller estimation errors than the baseline method using SO-Score or SVM. Thus, it was found that LSTM is effective in estimating the evaluation scores of reviews.

On the other hand, when LSTM and bi-LSTM are compared, LSTM produced better results, on average. We observed a tendency for bi-LSTM results to improve when the review data for *anime*, which included many examples, were used as the training data.

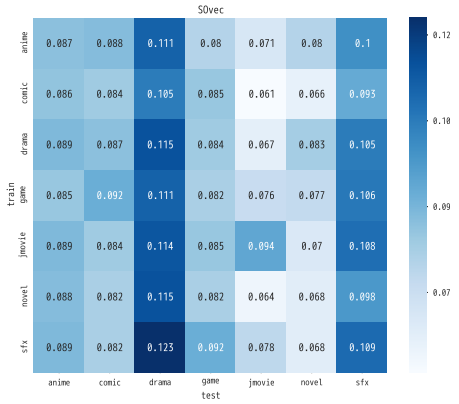
The bi-LSTM might have been over-trained versus LSTM due to a lack of sufficient training data. Although the SO-Score-based method is quite simple, it produced a lower MSE (01., on average) than the SVM-based method. This suggests that the SO-Score-based method is effective for review score estimation. However, it is still not clear how effective this method is for tweet sentences that include a large number of evaluation expressions that are not included in dictionaries. In future work, we would like to further investigate this by comparing results with review score estimation of reviews by tweets.

As a result of fine tuning by using the weights of the embedding layer from the pre-trained six emotions estimation model as initial weights, we were able to estimate review scores more accurately than when the methods were used without fine tuning.

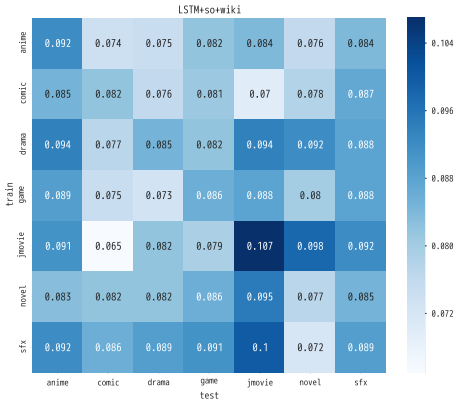
6. Conclusion

In this paper, we verified the compatibility of transfer among media by conducting transfer learning among various media and creating evaluation score estimation models. We found that there are still some problems in compatibility; significant differences were not found in score estimates when word clustering was used as pre-training.

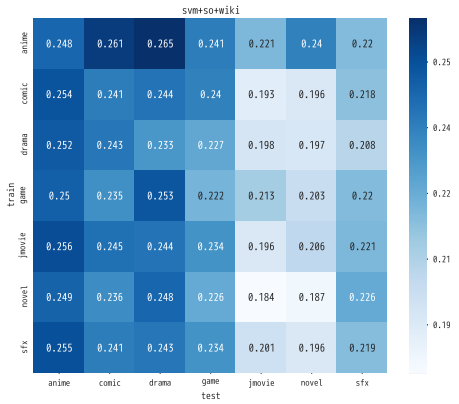
Our proposed method uses features that were not obtained by SO-Score values, but rather by clustering word distributed representation vectors to create review



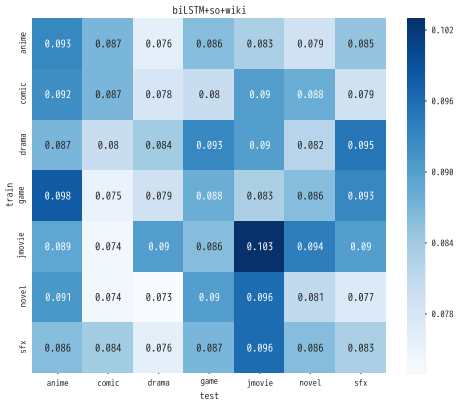
Baseline, SO-Vec



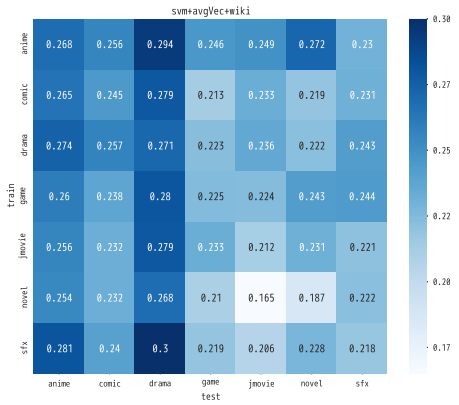
Proposed, LSTM, wikipedia



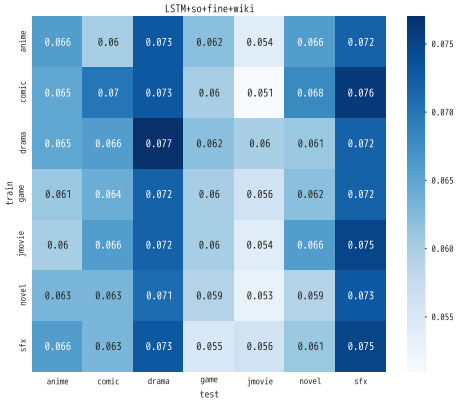
Baseline, SVM, wikipedia



Proposed, bi-LSTM, wikipedia

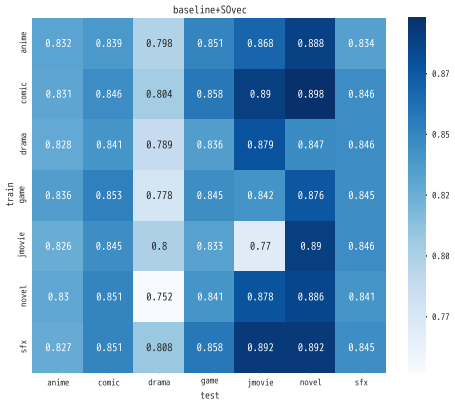


Baseline, SVM, avgVec, wikipedia

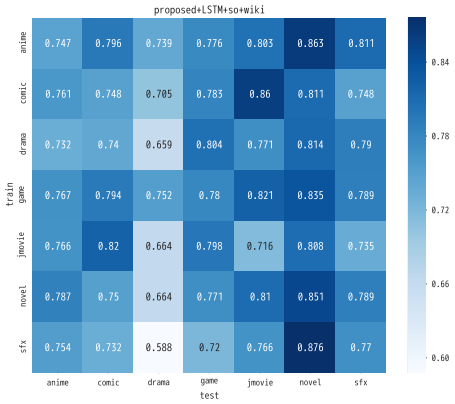


Proposed, LSTM fine-Tuning, wikipedia

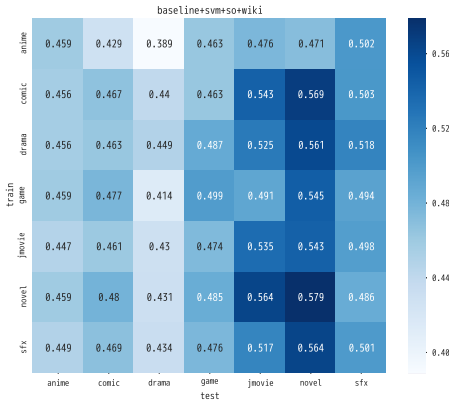
Fig. 5. Comparison of MSE between the experiments



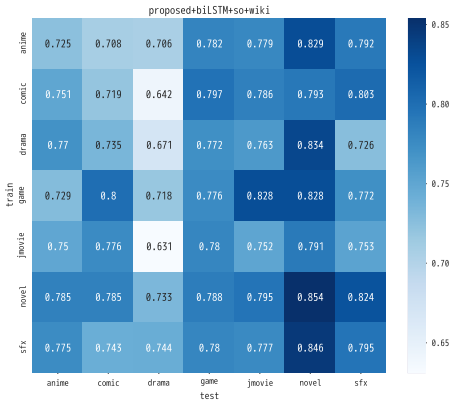
Baseline, SO-Vec



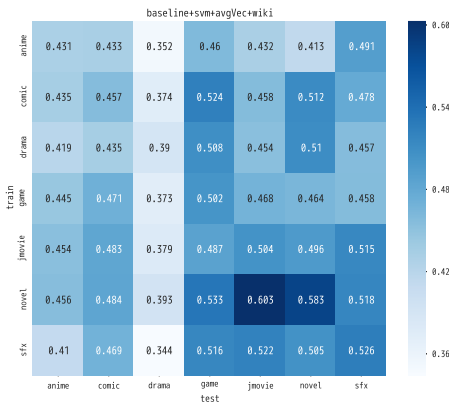
Proposed, LSTM, wikipedia



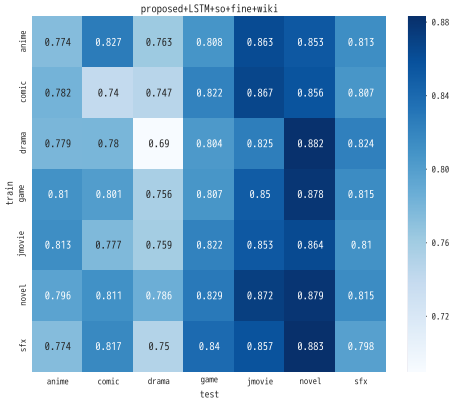
Baseline, SVM, wikipedia



Proposed, bi-LSTM, wikipedia



Baseline, SVM, avgVec, wikipedia



Proposed, LSTM fine-tuning, wikipedia

Fig. 6. Comparison of Accuracy between the experiments

score estimation models using LSTM or bi-LSTM. We were able to demonstrate the effectiveness of our proposed method. For evaluation score estimation, both LSTM and bi-LSTM-based methods showed higher performance than the baseline method based on SVM.

On the other hand, we also obtained satisfactory results by using a simple SO-Score-based method that used an evaluation polarity dictionary. From these results, we believe that better performance can be achieved by adding dictionary knowledge such as evaluation polarity.

We observed changes in accuracy depending on the training method of the word distributed representation or training sources. This suggests that accuracy in pre-training has some effects on estimation performance. In the future, we plan a more detailed analysis by increasing the combinations of transfer learning – for example, by mixing multi media. We also plan to focus on differences among review sites as well as differences among media.

Acknowledgments

This research has been partially supported by JSPS KAKENHI Grant Number 15H01712, 15K16077, 15K00309, 15K00425.

References

1. P., D., Turney: "Thumbs up or thumbs down? Senmantic orientation applied to unsupervised classification of reviews," In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 417–424, 2002.
2. B., Pang, L., Lee, and S., Vaithyanathan: "Thumbs up?: sentiment classification using machine learning techniques," In Proceedings of the International Conference on Empirical Methods in Natural Language Processing, pp. 79–86, 2002.
3. J., Kleinberg, and E., Tardos: "Approximation Algorithms for Classification Problems with Pairwise Relationships: Metric Labeling and Markov Random Fields," Journal of the ACM, Vol.49, No.5, pp.616–639, 2002.
4. H., Do, A., Kalousis, J., Wang, and A., Woznica: "A metric learning perspective of SVM: on the relation of SVM and LMNN," In Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 308–317, 2012.
5. A., Finn, N., Kushmerick, and B., Smyth: "Genre Classification and Domain Transfer for Information Filtering," In Proceedings of the European Conference on Information Retrieval, pp. 353–362, 2002.
6. S., Thurn, and L., Pratt: "Learning to Learn," Springer Science & Business Media, 2012.
7. T., Mikolov, I., Sutskever, K., Chen, G., Corrado, J., Dean: "Distributed Representations of Words and Phrases and their Compositionality," Advances in Neural Information Processing Systems 26, NIPS2013, pp.1–9, 2013.
8. P., Bojanowski, E., Grave, A., Joulin, and T., Mikolov: "Enriching Word Vectors with Subword Information," arXiv preprint arXiv:1607.04606, 2016.
9. Y., Zhao and G., Karypis: "Comparison of agglomerative and partitional doc-ument clustering algorithms," Technical report, Department of Computer Science, University of Minnesota, Minneapolis, MN 55455, 2002.
10. Bayon - a simple and fast clustering tool: `http://code.google.com/p/Bayon/`.
11. F., A., Gers, J., Schmidhuber, and F., Cummins: "Learning to Forget: Continual Prediction with LSTM," Neural Computation, Vol.212, Issue 10, pp.2451–2471, 2000.
12. MeCab: Yet Another Part-of-Speech and Morphological Analyzer, `http://taku910.github.io/mecab/`.

13. N., Kobayashi, K., Inui, Y., Matsumoto, K., Tateishi, T., Fukushima: "Collecting Evaluative Expressions for Opinion Extraction," Journal of Natural Language Processing, Vol. 12, No. 3, pp. 203–222, 2005 (in Japanese).
14. Japanese Evaluation Expression Dictionary: `http://www.syncha.org/evaluative_expressions.html`

Kazuyuki Matsumoto

He received the PhD degree in 2008 from Tokushima University. He is currently an assistant professor of Tokushima University. His research interests include Affective Computing, Emotion Recognition, Artificial Intelligence and Natural Language Processing. He is a member of IPSJ, ANLP, IEICE and IEEJ.

Fuji Ren

He received the Ph.D. degree in 1991 from Faculty of Engineering, Hokkaido University, Japan. He worked at CSK, Japan, where he was a chief researcher of NLP. From 1994 to 2000, he was an associate professor in the Faculty of Information Sciences. From 2001 he joined the faculty of engineering, the University of Tokushima as a professor. His research interests include Natural Language Processing, Artificial Intelligence, Language Understanding and Communication. He is a member of the IEICE, CAAI, IEEJ, IPSJ, JSAI, AAMT and a senior member of IEEE.

Minoru Yoshida

He is a lecturer at the Department of Information Science and Intelligent Systems, University of Tokushima. After receiving his BSc, MSc, and PhD degrees from the University of Tokyo in 1998, 2000, and 2003, respectively, he worked as an assistant professor at the Information Technology Center, University of Tokyo. His current research interests include Web Document Analysis and Text Mining for the Documents on the WWW.

Kenji Kita

He received the B.S. degree in mathematics and the PhD degree in electrical engineering, both from Waseda University, Tokyo, Japan, in 1981 and 1992, respectively. From 1983 to 1987, he worked for the Oki Electric Industry Co. Ltd., Tokyo, Japan. From 1987 to 1992, he was a researcher at ATR Interpreting Telephony Research Laboratories, Kyoto, Japan. Since 1992, he has been with Tokushima University, Tokushima, Japan, where he is currently a Professor at Faculty of Engineering. His current research interests include multimedia information retrieval, natural language processing, and speech recognition.