An Approach to Refine Translation Candidates for Emotion Estimation in Japanese-English Language

Kazuyuki Matsumoto, Minoru Yoshida, Kenji Kita and Fuji Ren
Department of Faculty of Engineering, Tokushima University,
Minamijosanjima cho 2-1, Tokushima, Japan

Keywords: Emotion Corpus, Emotion Estimation, Bilingual Tagged Corpus.

Abstract: Researches on emotion estimation from text mostly use machine learning method. Because machine learning requires a large amount of example corpora, how to acquire high quality training data has been discussed as one of its major problems. The existing language resources include emotion corpora; however, they are not available if the language is different. Constructing bilingual corpus manually is also financially difficult. We propose a method to convert a training data into different language using an existing Japanese-English parallel emotion corpus. With a bilingual dictionary, the translation candidates are extracted against every word of each sentence included in the corpus. Then the extracted translation candidates are narrowed down into a set of words that highly contribute to emotion estimation and we used the set of words as training data. As the result of the evaluation experiment using the training data created by our proposed method, the accuracy of emotion estimation increased up to 66.7% in Naive Bayes.

1 INTRODUCTION

Recently, there have been many researches on emotion estimation from text in the field of sentiment analysis or opinion mining (Ren, 2009), (Ren and Quan, 2015), (Ren and Wu, 2013), (Quan and Ren, 2010), (Quan and Ren, 2014), (Ren and Matsumoto, 2015) and many of them adopted machine learning methods that used words as a feature. When the type of the target sentence for emotion estimation and the type of the sentence prepared as training data are different, as in the case of terminology in the problem of domain adaptation for document classification, the appearance tendency of the emotion words differs. This causes a problem in fluctuation of accuracy. On the other hand, when a word is used as a feature for emotion estimation, the sentence structure does not have to be considered. As a result, it is easy to apply the method to other languages. Only if we prepare a large number of corpora with annotation of emotion tags on each sentence, emotion would be easily estimated by using the machine learning method. In the machine learning method, because manual definition of a rule is not necessary, we can reduce costs to apply the method to other languages.

However, just like the problem in the domain, depending on the types of the languages, sometimes it is difficult to prepare a sufficient amount of tagged corpus. For example, in comparison to English or Chinese, there are not enough tagged corpora in Japanese or Korean, as the people who use such languages are relatively small.

To solve the shortage problem of Chinese emotion corpus, Wan (Wan, 2009) used English emotion corpus as training data that was openly available for free, and attempted to classify Chinese emotions. He proposed a co-training model that combined a method translating the training data and a method translating the test data. Inui et al. (Inui and Yamamoto, 2011) mechanically translated Japanese sentences into English sentences and used them as test data or training data to classify review articles into positive or negative by using SVM. They checked whether or not the sentences included evaluation expressions. Then, based on the results, they selected the sentences by judging if the sentence should be added in the training data or in the test data. The experiment obtained approximately 80% accuracy. This accuracy was higher than the accuracy obtained when the untranslated training data was used.

Their method summarizes a document by exclusively using the sentences in a review article that have evaluation expressions. Because the method does not confirm the reliability of the translation results

Matsumoto, K., Yoshida, M., Kita, K. and Ren, F..
An Approach to Refine Translation Candidates for Emotion Estimation in Japanese-English Language.
In *Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015) - Volume 2: KEOD*, pages 74-83
ISBN: 978-989-758-158-8

to summarize, it is difficult to deal with the problem of estimation failure caused by low translation accuracy. Because our study does not aim at emotion estimation in document increments, their proposed document summarization technique cannot be applied to our study. We refined the translation candidates of each word in a sentence by narrowing them down under certain condition.

In this paper, we attempted emotion estimation by machine learning. In the sentences of Japanese-English parallel emotion corpus the translation candidates for each word were obtained in reference to the bilingual dictionary. We used them as training data for machine learning and conducted an emotion estimation experiment. If bilingual dictionaries are used to obtain translation candidates, erroneous translations might be caused as often as or more often than when machine translation is conducted.

For that reason, we proposed a refining method that narrowed down the translation candidates according to whether the kind of the sentence's emotion and the word's emotion matched or not. By removing the words that were not likely to contribute to sentence emotion, the translation candidates were refined. The aim of this method is to minimize the effects by translation error.

Section 2 describes the related works about emotion estimation based on word feature and emotion estimation based on different languages.

To remove noise feature, we propose a method for refining translation candidates extracted from bilingual dictionaries in section 3 and conduct an evaluation experiment in section 4. Then, we examine the results of the evaluation experiment and discuss the effectiveness of our method that does not use machine translation in section 5. Finally, we summarize this study in section 6.

2 RELATED WORKS

The researches on emotion estimation often adopted machine learning method that used words as a feature (Matsumoto and Ren, 2011), (Quan and Ren, 2011), (Wu and Matsumoto, 2014). Many of these methods do not consider the meanings of the words. Actually, in the task of judging a word's or a phrase's emotion polarity (positive/negative), a certain level of accuracy can be obtained without considering the word's meaning (Takamura and Okumura, 2005), (Takamura and Okumura, 2006).

There are also researches that judge emotion categories of emotional words in a sentence (Kang et al., 2010). In the machine learning, the quality or kind of source data used for training data is one of the most important factors that affect the classification accuracy.

To judge the emotion polarity of a sentence belonging to a different domain from the training data, Saiki et al. (Saiki and Okumura, 2008) adapted each domain by using the weighted maximum entropy model to add weight to case. Minato et al. (Minato and Kuroiwa, 2008) estimated sentence emotion by using appearance frequency weight of word for each emotion category according to Japanese-English parallel emotion corpus. The evaluation result showed that emotion estimation accuracies varied due to small size of the corpus and bias of the number of the sentences in each emotion category.

Balahur et al. (Balahur and Turchi, 2012) treated the problem of sentiment detection in several different languages such as French, German and Spanish. They translated each language resources into English by using the existing machine translation techniques and classified sentence emotion by training the n-gram feature of the translated resources based on Support Vector Machines Sequential Minimal Optimization (SVM SMO).

From the experimental result for the multilingual resources, they concluded that the statistical machine translation (SMT) was mature enough as preprocessing for sentiment classification. However, it is considered that the languages used in their study were easier to translate into English compare to translate Japanese into English. In Japanese language, with only difference of notation or intonation of the word, sense of the word sometimes changes. On the other hand, sentence structure is more complex than English or the other western languages.

Moreover, even if the machine translation system can translate Japanese into English successfully, with a little difference of the translation candidate, the nuance becomes different from the original meaning and the emotion to be conveyed might be changed.

To confirm this, it is necessary to conduct an experiment of emotion estimation by using the translation results based on Japanese-English emotion tagged corpus. Preprocessing was conducted by converting Japanese or English emotion tagged corpora into other language data by machine translation or parallel dictionary. We confirmed whether emotion estimation accuracy could be improved by refining translation candidates or not by the evaluation experiment.

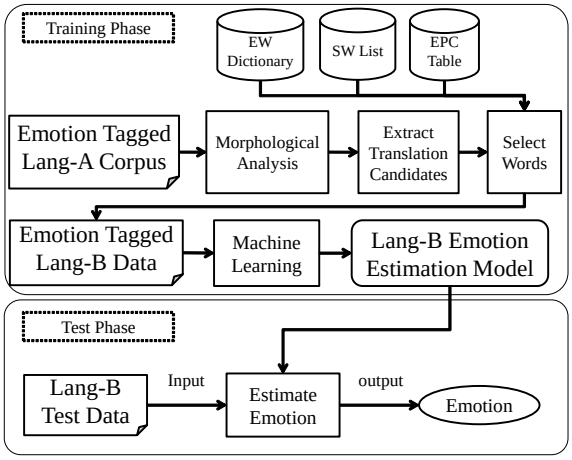

Figure 1: Emotion estimation in different languages.

3 PROPOSED METHOD

The basic flow of our proposed method is described in Figure 1. EW Dictionary means a dictionary that stored the words expressing emotion included in the Japanese-English parallel emotion corpus (Minato and Kuroiwa, 2007). SW List is a dictionary of stop word that includes nonsense words (unnecessary words). EPC Table indicates Emotion Polarity Correspondence Table constructed by Takamura (Takamura and Okumura, 2005), (Takamura and Okumura, 2006).

In the Training Phase, each sentence in the emotion tagged corpus of language A is morphologically analyzed to obtain the basic forms of the words. Then, for the basic forms of the words the translation candidates are obtained in language B with reference to the bilingual dictionary. The candidate words are refined by our proposed method and the data in language B is created that has annotation of emotion tags in sentence basis. Based on the created data in language B, the emotion estimation model for language B is made by machine learning.

3.1 Japanese-English Parallel Emotion Corpus

This section describes the Japanese-English Parallel Emotion Corpus used in this research. The Japanese-English Parallel Emotion Corpus constructed by Minato et al. (Minato and Kuroiwa, 2007) is a tagged corpus based on a Japanese-English Emotion Expression Dictionary (Hiejima, 1999).

In the corpus emotion tags are annotated on the Japanese-English bilingual conversation sentences to indicate the speaker's emotion and the emotional expression. In total, 1,190 pairs of the sentences are registered in the corpus.

There are two kinds of tags: sentence emotion tags annotated to sentences and word emotion tags annotated to words or idioms. There are nine kinds of emotion categories: (joy, anger, sorrow, surprise, hate, love, respect, anxiety and neutral).

3.2 Japanese-English Bilingual Dictionary

In this research we used the dictionaries of Eijiro Ver.2.0[1], Edict and Edict2[2] as the Japanese-English bilingual dictionaries. We selected these three dictionaries because these dictionaries included relatively newer words and a larger number of words compared to other dictionaries such as the EDR dictionary.

3.3 Refining Translation Candidates

When training data and test data are constructed by extracting translation candidates from the dictionary, the problem is that the translation candidates include unnecessary words that cause estimation error. We considered the following perspectives to refine the translation candidates.

- The stop words included in translations should be removed from training data.
- When the sentence's (writer's) emotion and the emotion of the word included in the sentence are different, the word should be removed from training data.
- The more the word contributes to the emotion expressed by the sentence, the more the weight is added to the word and the word should be included in training data.

We made the training data as shown in Table 1. In Table 1 "Dictionary for Refining" indicates the kind of the dictionaries used for refining the translation candidates. "EW" represents Emotion Word Dictionary, "SW" represents Stop Word List, and "EPC" represents Emotion Polarity Correspondence Table. In these columns, J, E indicates whether the dictionary is Japanese or English. The column of Code represents the abbreviation of each model. In the paper, $J \rightarrow E$ represents extracting English translation candidates from Japanese sentences and $E \rightarrow J$ represents extracting Japanese translation candidates

[1]http://www.alc.co.jp/support/eijiro2/

[2]http://www.csse.monash.edu.au/ jwb/edict_doc.html

from English sentences. For $M_{E\to J}, M_{J\to E}$ in Baseline Model and Machine Translated Model, translation candidates were not refined. The following subsections elaborate the training data used in this paper.

Table 1: Definition of training data.

Model Name	Code	Target Lang.	Dictionary for Refining		
			EW	SW	EPC
Baseline Model	$M_{E\to J}$	J	-	-	-
	$M_{J\to E}$	E	-	-	-
	$M'_{E\to J}$	J	J	-	-
	$M'_{J\to E}$	E	E	-	-
Machine Translated Model	$G_{E\to J}$	J	-	-	-
	$G_{J\to E}$	E	-	-	-
Stop Word Model	$R_{E\to J}$	J	-	J	-
	$R_{J\to E}$	E	-	E	-
Polarity Model	$P_{E\to J}$	J	-	-	J
	$P_{J\to E}$	E	-	-	E
	$P'_{E\to J}$	J	-	-	J
	$P'_{J\to E}$	E	-	-	E

3.3.1 Baseline Method

The Japanese translation candidates corresponding to each word in the English sentences were extracted from the three kinds of bilingual dictionaries. Then, these candidates were used as the feature for creating training data $M_{E\to J}$.

In the same way, $M_{J\to E}$ was created by extracting English translation candidates of each word in Japanese sentence and using these candidates as the feature. Training data $M'_{E\to J}$ and $M'_{J\to E}$ was also created by extracting specific translation candidates of the words that had been marked as emotion expression in the corpus from a bilingual dictionary.

3.3.2 Machine Translated Model

To compare the proposed method and the method using the result of machine translation as training data, the words in Japanese-English sentences were translated by Google translation[3] and used them as feature to create the training data $G_{E\to J}, G_{J\to E}$. Google translation is based on a statistical machine translation method. However, its translation accuracy is not very high. To judge the quality of translation various measures have been proposed.

One of the famous measures is BLEU Papineni and Zhu (2002) that uses the word's N-gram precision. IMPACT (Echizen-ya and Araki, 2007) has a relatively high correlation with evaluation by human in adequacy and fluency. METEOR Banerjee (2005), (Lavie and Agarwal, 2007) was proposed as an evaluation method without using a word's N-gram. The details of METEOR are described in the paper written by Banerjee et al. (Banerjee, 2005). To investigate the accuracy of Google translation in corpus translation, the average score of IMPACT, METEOR[4] and BLEU was calculated. The obtained scores were shown in Table 2.

Table 2: Evaluation of machine translation (Google MT).

Evaluation Method	$J\to E$	$E\to J$
IMPACT	0.48	0.35
METEOR	0.28	0.34
BLEU	0.19	0.15

The quality of web translation system such as Google translation is highly controversial. However, the average scores obtained were not especially low.

3.3.3 Stop Word Model

Besides the problem of a simple translation error, there is another problem that decreases the accuracy of emotion estimation. The problem is caused by the words that do not contribute to the speaker's (writer's) emotion.

These words should be removed from the training data by refining according to the rule. Therefore, we focused on a method based on stop words. SMART (Buckley and Singhal, 1995) is a famous list of unnecessary words. However, it is an English word list and we cannot apply it to Japanese language. Therefore, we attempted to refine the words by parts of speech for Japanese language. If the part of speech of the word annotated by morphological analysis was not included in Table 3, the word was regarded as an unnecessary word.

If the translation candidates of the word w included the unnecessary word sw, the word sw was removed from the training data. In this way, the training data were created and defined as $R_{E\to J}, R_{J\to E}$.

3.3.4 Polarity Model

To refine the translation, we also focused on the emotion polarity of words. If the candidates of the word w include ew, whose emotion polarity is different from the sentence's emotion polarity, the word ew should be removed from the training data. The correspondence between emotion polarity and emotion category is shown in Table 4.

[3] https://translate.google.com

[4] METEOR used parameters: $\alpha = 0.9; \beta = 3.0; \gamma = 0.5$

Table 3: Part of speech regarded as necessary word.

Part of Speech	Sub category
Prefix	adjective connection, noun connection
Conjunction	-
Adnominal adjective	-
Noun	general, "Sa"-connection, "Nai"-adjective connection, Adverb-possible, Adjective verb stem
Verb	independent
Adjective	independent
Adverb	general, auxiliary word connection
Interjection	-

Table 4: Emotion polarity and emotion category.

Positive (+)	joy, love, respect
Negative (-)	anger, sorrow, hate, anxiety
Neutral (0)	surprise, neutral

We used the emotion polarity correspondence table created by Takamura et al. (Takamura and Okumura, 2005) to judge the word's emotion polarity. Because many of the words with extremely small emotion polarity value actually do not express emotion, the threshold value was set for the emotion polarity value. In this paper, the threshold value was set as $th = 0.5$. If the absolute emotion polarity value of a word is smaller than this threshold value, the emotion polarity of the word was set as 0 (neutral). The created training data were $P_{E\to J}, P_{J\to E}$.

However, above method cannot extract features if there are no words that match to the emotion polarity of the sentence. For that reason, we also considered another method to judge whether the word to be included as feature or not according to the degree of contribution to the emotion expression of the sentence.

In this method we used the words that did not match to the emotion polarity of the sentence as feature and we did not to set the threshold for emotion polarity. In this research, the words whose degree of contribution is ranked in the top rc are used as features. How to calculate the contribution degree of the word w_j to the sentence S_i is shown in Equation 1, 2, 3, 4.

$$\mathrm{EM}(w_j, S_i) = \begin{cases} 1 & if \quad \mathrm{EMP}(w_j) = \mathrm{EMP}(S_i) \\ 0 & otherwise \end{cases} \quad (1)$$

$$\mathrm{EMS}(w_j, S_i) = \begin{cases} 1 & if \quad \mathrm{EMP}(w_j) = \mathrm{EMP}(S_i) \\ 0.5 & otherwise \end{cases} \quad (2)$$

$$\alpha = \frac{|S_i|}{\sum_{w_k \in S_i} \mathrm{EM}(w_k, S_i)} \quad (3)$$

$$\mathrm{CScore}(w_j \in S_i) = \begin{cases} \alpha \cdot \mathrm{EMS}(w_j, S_i) & if \quad \mathrm{EM}(w_j, S_i) = 1 \\ \mathrm{EMS}(w_j, S_i) & otherwise \end{cases} \quad (4)$$

$\mathrm{EMP}(w_j)$, $\mathrm{EMP}(S_i)$ respectively indicate a sign of emotion polarity of the word w_j and sentence S_i. EM shown in Equation 1 is a function that returns 1 if the signs of emotion polarity of the word and the sentence are the same, otherwise returns 0. EMS shown in Equation 2 is a function that returns 1 if the signs of emotion polarity of the word and the sentence are the same, otherwise returns 0.5. The Equation 4 calculates CScore that represents the contribution value of the word to the sentence by using the functions EM and EMS.

$|S_i|$ in the Equation 3 indicates the total number of the words in the sentence S_i. By multiplying x, the more the weight of the words whose sign of the emotion polarity matched with that of the sentence increases relatively, the less there are the words whose sign of the emotion polarity matched with that of the sentence.

CScore is calculated for each word in each sentence and the words are sorted by descending order of the CScore. By using only the top r_c words as features, training data $P'_{E\to J}, P'_{J\to E}$ are made. We conduct several experiments to calculate the value of r_c and used the value with the highest accuracy as the value of r_c.

With these methods, we thought that we could prevent acquiring unnecessary translation candidates for constructing training data in another language, thereby could create a highly accurate compact estimation model. Moreover, the calculation for training could be reduced.

3.3.5 Classification Method

In the research of emotion estimation from text, the Support Vector Machine (SVM) and the Naive Bayes classifier (NB) are often used for machine learning. In this paper, Naive Bayes classifier was used because it has a simple algorithm and can be easily applied to multiclass classification.

4 EXPERIMENT

The proposed method was evaluated by experiment. The target data was 1,190 pairs of sentences in Japanese-English parallel emotion corpus and 4,652 pairs of sentences in open Japanese-English parallel corpus with annotation of emotion. The information of the morphemes included in the sentences was used as feature.

Japanese sentences were morphologically analyzed by ChaSen[5]. English sentences were morphologically analyzed by Brill's Tagger (Brill, 1994) then basic forms of the parts of stems were obtained by using the Porter stemming algorithm[6]. We evaluated the results by calculating the accuracy with the Equation 5.

$$match_i = \begin{cases} 1 & if \quad |T_{o,i} \cap T_{c,i}| \geq 1 \\ 0 & otherwise \end{cases} \tag{5}$$

$$Accuracy(\%) = \frac{\sum_i^{|S|} match_i}{|S|} \times 100 \tag{6}$$

In the equation, S indicates a set of sentences targeted for the evaluation. $T_{o,i}$, and $T_{c,i}$ respectively indicate a set of tag sets outputted by classifier and a set of correct tags for the sentence $s_i \in S$. $|T_{o,i} \in T_{c,i}|$ indicates the number of the matched tags between the tags outputted by the classifier and the correct tags. $match_i$ means the score of the correct answers for the sentence s_i. Arithmetic average of these scores is calculated as Accuracy. Although the classifier outputs the probability for each emotion category, the category with the highest probability is extracted and evaluated in the experiment.

4.1 Experiment-1

First, to evaluate emotion estimation by using Japanese-English Parallel Emotion Corpus as training data and test data, we created the training data whose feature is the translation candidates of each word in the bilingual sentences.

Then, we conducted Experiment-1 using the test data whose feature is the words in the source language's sentence. Experiment was conducted by using 10-fold cross validation.

4.2 Experiment-2

Experiment-1 used the basic Japanese-English parallel emotion corpus to evaluate our proposed method. In Experiment-2, the proposed method was evaluated in whether it was valid to the open data or not. As open data, we used four kinds of bilingual corpora: The Tanaka Corpus[7], Natsume Corpus extracted from Japanese English Emotion Expression Dictionary (Enozawa and Guruensutain, 1999), Gogakuru Corpus extracted from a Web site for language education called "Gogakuru,"[8] and Web Eikaiwa Corpus collected from several educational Web sites for English conversation. Emotion tags were annotated to these four kinds of corpora to create test corpora. Table 5 shows the numbers of the registered sentence pairs in the constructed corpora.

Table 5: Size of open J-E translation corpus.

Corpus	# of Sentence Pairs
Tanaka Corpus	1,055
Gogakuru Corpus	1,555
Web Eikaiwa Corpus	935
Natsume Corpus	1,107

We used these corpora as test data. From the training data created in Experiment-1, $M_{J\to E}, M_{E\to J}, R_{J\to E}, R_{E\to J}, P_{J\to E}, P_{E\to J}, P'_{J\to E}, P'_{E\to J}$ were used in Experiment-2. Because the training data scale was much smaller than the test data, the translation candidate data was gradually added 100, 200, 300, and 400 sentences for each emotion in each training data from the open data; the remaining data in the open data was used as test data.

$G_{J\to E}, G_{E\to J}$ were not used because the aim of Experiment-2 was not to compare the results with the Web translation result. $M'_{J\to E}, M'_{E\to J}$ were not used in Experiment-2 because they were not able to be applied to unknown emotion expressions in the open data.

5 RESULTS AND DISCUSSIONS

The result of Experiment-1 is shown in Table 6. $P'_{E\to J}, P'_{J\to E}$ were made by setting the threshold rc of contribution value as 2^9.

The result of the training data $M'_{E\to J}, M'_{J\to E}$, which used only emotion expression's translation candidates as the feature, had higher accuracy than using all words' translation candidates as the feature.

[5]http://chasen-legacy.osdn.jp/

[6]http://tartarus.org/ martin/PorterStemmer/

[7]http://www.edrdg.org/wiki/index.php/Tanaka_Corpus

[8]http://gogakuru.com/

[9]We evaluated the results by setting the threshold values: 1, 2, 3, 4, 5. When the threshold was $r_c = 2$, the accuracy was the highest.

The result also showed that the conditions of unnecessary words for refining the translation candidates did not greatly increase the accuracy.

However, in the experiment based on Japanese test data, using $R_{E\to J}, P_{E\to J}$ as the training data decreased accuracy only about 2%. Therefore, these refining methods will become more effective when the size of the data increases because the calculation amount decreases.

In the experiments with $P'_{E\to J}, P'_{J\to E}$, over 50% accuracies were obtained. When $P'_{E\to J}$ was used as training data and the edict was used as a bilingual dictionary, the best accuracy of 66.7% was obtained.

These results suggested that our method solved the weak points in the Polarity Model, which is considering solely the concordance of emotion polarities refines too much the translation candidates and provides all words the same weights.

- By adding higher weight to the words that contribute to the emotion expression in the sentence, effective features can be emphasized.
- It is able to extract feature even though none of the words in the sentence contribute to the emotion expression in the sentence.

Table 6: Result of experiment-1.

Test	Training	Dictionary			
		-	edict	edict2	eijiro
J	$M_{E\to J}$		50.1	50.3	49.8
	$M'_{E\to J}$		52.3	52.0	51.3
	$G_{E\to J}$	37.5			
	$R_{E\to J}$		48.6	42.4	49.4
	$P_{E\to J}$		47.5	48.2	48.5
	$P'_{E\to J}$		**66.7**	58.1	64.1
E	$M_{J\to E}$		39.4	39.7	35.0
	$M'_{J\to E}$		51.1	51.3	50.0
	$G_{J\to E}$	46.6			
	$R_{J\to E}$		47.5	48.6	49.2
	$P_{J\to E}$		37.5	38.2	48.8
	$P'_{J\to E}$		51.1	50.1	**53.9**

The result of Experiment-2 is shown in Figure 2, 3, 4, 5. The result showed that the accuracy in Japanese increased by using a refining process, although the accuracy in English sometimes decreased. Comparing the accuracies in M that refined the translation candidates and those in P, P' and R that did not refine the translation candidates, the percentage of the accuracy increased approx.15% at a maximum by adding training data. From this result, it was found that the larger the training data became, the more effectively the refining of the translation candidates worked.

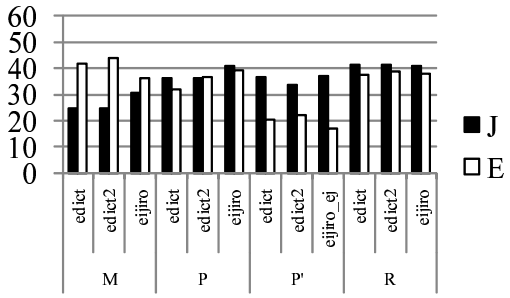

Figure 2: Result of Experiment-2: add 100 sentence each emotion.

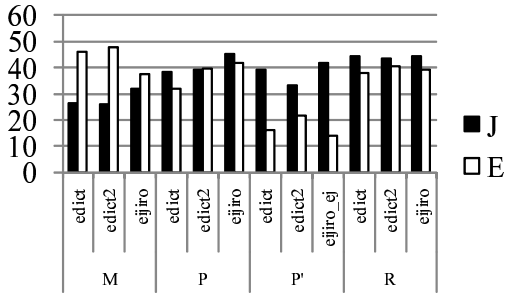

Figure 3: Result of Experiment-2: add 200 sentence each emotion.

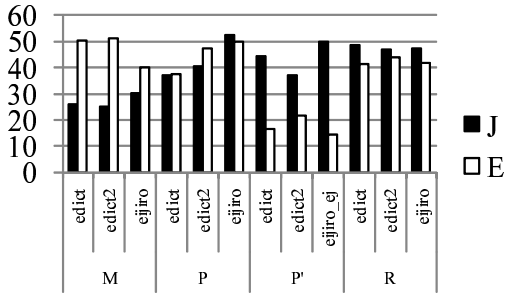

Figure 4: Result of Experiment-2: add 300 sentence each emotion.

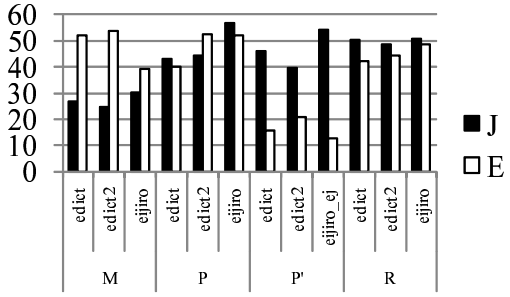

Figure 5: Result of Experiment-2: add 400 sentence each emotion.

However, it is not necessarily appropriate to suggest that accuracy will improve by increasing the training data because the test data decreased instead.

It is also considered that the sentences added randomly from the open data included the words that were not included in the original training data. These words might have been effective for emotion estimation.

The results of the experiment used $P'_{J\to E}$ were significantly lower in accuracy. In the case of English, even though the words are included in EPC, many of them have more varieties of meanings and higher ambiguity than Japanese. For that reason, many words in the open data might be used in different meaning and have contributed to the emotion expression. Therefore, it is considered that over-training might have been caused by weighting contribution value. It will be necessary to improve the calculation of contribution by considering the number of the translation candidates.

Figure 6, 7, 8, 9, 10, 11, 12, 13 shows the accuracy for each emotion category. From this figure, we found the accuracies of neutral are higher than those of the other emotions, especially neutral. Because many of those sentences with neutral annotations did not include emotional expressions, it might have been easy to be judged as neutral. On the other hand, because there were large numbers of sentences with neutral tag annotations, it seemed that estimating emotion as neutral became easier and easier as the amount of the training data increased.

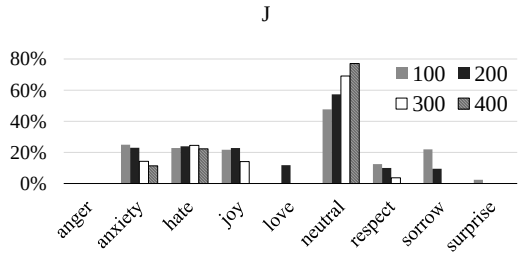

Figure 6: Result of Experiment-2: $P_{E\to J}$, each emotion category, use eijiro.

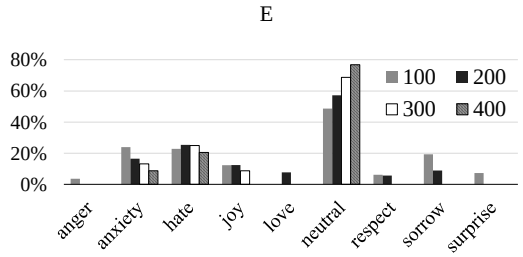

Figure 7: Result of Experiment-2: $P_{J\to E}$, each emotion category, use eijiro.

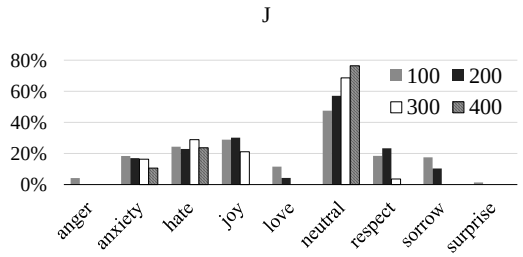

Figure 8: Result of Experiment-2: $P'_{E\to J}$, each emotion category, use eijiro.

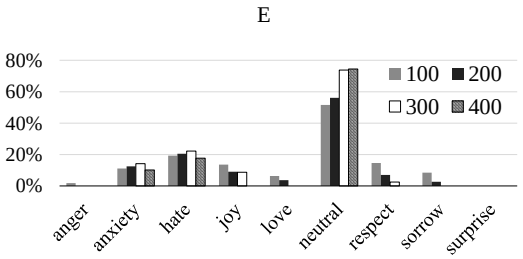

Figure 9: Result of Experiment-2: $P'_{J\to E}$, each emotion category, use eijiro.

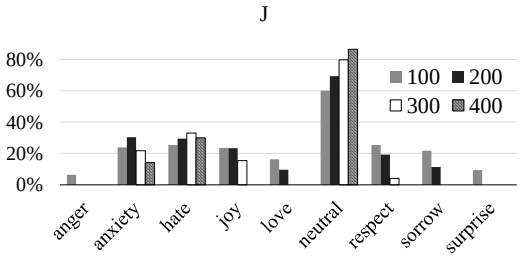

Figure 10: Result of Experiment-2: $M_{E\to J}$, each emotion category, use eijiro.

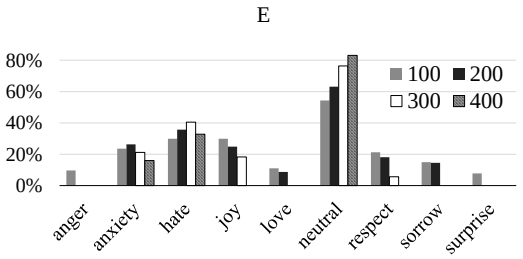

Figure 11: Result of Experiment-2: $M_{J\to E}$, each emotion category, use eijiro.

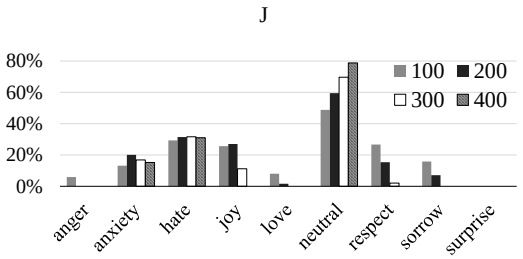

Figure 12: Result of Experiment-2: $R_{E\to J}$, each emotion category, use eijiro.

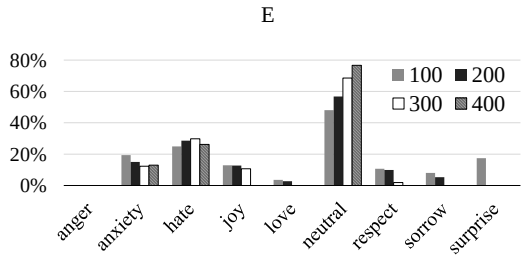

Figure 13: Result of Experiment-2: $R_{J \to E}$, each emotion category, use eijiro.

In this paper, we increased the training data randomly. However, the method to remove bias of each emotion category should be taken. Actually, there are many biases in the corpora that are available. The sentences that are annotated neutral tags do not express any emotion. Therefore, to eliminate estimation bias, we think that the two levels of classification might be necessary. After judging whether emotion is expressed or not, the emotion should be classified into eight emotion categories as for the sentences expressing emotions.

6 CONCLUSIONS

In this paper, existing bilingual dictionaries were used to convert the linguistic resources for emotion estimation into another language. To avoid including the noise feature in training data by converting the resource into other languages, a method to refine the translation candidates was proposed based on emotion polarity or stop word list.

The evaluation experiment using the basic Japanese-English Bilingual Dictionary obtained approximately 66.7% accuracy in emotion estimation when the translation candidates exclusively corresponding to the emotional expressions were included in the training data. On the other hand, from the experimental result using open data, it was found that the process of refining translation candidate worked effectively.

However, the bilingual dictionaries and the emotion polarity correspondent table included unnecessary words for emotion estimation. As the result, noise features could not be removed even though threshold value was set.

In future, we would like to improve the method to refine the translation candidates and propose a method to remove unnecessary word from emotion polarity correspondent table and also a method to automatically construct a bilingual dictionary suitable for emotion estimation.

ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Numbers 15H01712, 15K16077, 15K00425.

REFERENCES

Balahur, A. and Turchi, M. (2012). Multilingual sentiment analysis using machine translation? In *the 3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, pages 52–60.

Banerjee, S. (2005). Meteor : an automatic metric for mt evaluation with improved correlation with human judgments. In *ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

Brill, E. (1994). Some advances in transformation-based part of speech tagging. In *National Conference on Artificial Intelligence*, pages 722–727.

Buckley, B., S. G. A. J. and Singhal, A. A. (1995). Automatic query expansion using smart: Trec-3. In *the Third Text REtrieval Conference(TREC-3)*, pages 500–238.

Echizen-ya, H. and Araki, K. (2007). Automatic evaluation of machine translation based on recursive acquisition of an intuitive common parts continuum. In *the Eleventh Machine Translation Summit (MT SUMMIT XI)*, pages 151–158.

Enozawa, K. and Guruensutain, D. (1999). *Kaiwa de oboeru kanjouhyougen Wa-Ei jiten -English merry-go-round series- (in Japanese)*. Natsumesha.

Hiejima, I. (1999). Tokyodo Shuppan.

Inui, T. and Yamamoto, M. (2011). Usage of different language translated data on classification of evaluation document (in japanese). pages 119–122.

Kang, X., Ren, F., andWu, Y. (2010). Bottom up: exploring word emotions for chinese sentence chief sentiment classification. In IEEE International Conference on Natural Language Processing and Knowledge Enginerring, pages 422426.

Lavie, A. and Agarwal, A. (2007). Meteor : an automatic metric for mt evaluation with high levels of correlation with human judgments. In *ACL Second Workshop on Statistical Machine Translation*, pages 228–231.

Matsumoto, K. and Ren, F. (2011). Estimation of word emotions based on part of speech and positional information. *Computers in Human Behavior*, 2011(27):1553–1564.

Minato, J., M. K. R. F. and Kuroiwa, S. (2007). Corpus-based analysis of japanese-english of emotional expressions. In *IEEE International Conference on Natural Language Processing and Knowledge Enginerring*, pages 413–418.

Minato, J., M. K. R. F. T. S. and Kuroiwa, S. (2008). Evaluation of emotion estimation methods based on statistic features of emotion tagged corpus. *International Journal of Innovative Computing, Information and Control*, 4(8):1931–1941.

Papineni, K., R. S. W. T. and Zhu, W. (2002). Bleu: a method for automatic evaluation of machine translation. In *the 40th Annual Meeting on Association for Computational Linguistics(ACL'02)*, pages 311–318.

Quan, C. and Ren, F. (2010). A blog emotion corpus for emotional expression analysis in chinese. *Computer Speech and Language*, 24(1):726–749.

Quan, C. and Ren, F. (2011). Recognition of word emotion state in sentences. *IEEJ Transactions on Electrical and Electronic Engineering*, 6:34–41.

Quan, C. and Ren, F. (2014). Unsupervised product feature extraction for feature-oriented opinion determination. *Information Sciences*, 272(2014):16–28.

Ren, F. (2009). Affective information processing and recognizing human emotion. *Electronic Notes in Theoretical Computer Science*, 225:39–50.

Ren, F. and Matsumoto, K. (2015). Semi-automatic creation of youth slang corpus and its application to affective computing. *IEEE Transactions on Affective Computing*.

Ren, F., K. X. and Quan, C. (2015). Examining accumulated emotional traits in suicide blogs with an emotion topic model. *IEEE Journal of Biomedical and Health Informatics*.

Ren, F. andWu, Y. (2013). Predicting user-topic opinions in twitter with social and topical context. IEEE Transactions on Affective Computing, 4(4):412424.

Saiki, Y., T. H. and Okumura, M. (2008). Domain adaptaion in sentiment classification by instance weighting (in japanese). In *IPSJ SIG Notes*, volume 2008, pages 61–67.

Takamura, H., I. T. and Okumura, M. (2005). Extracting semantic orientations of words using spin model. In *the 43rd Annual Meeting on Association for Computational Linguistics*, pages 133–140.

Takamura, H., I. T. and Okumura, M. (2006). Latent variable models for semantic orientations of phrases (in japanese). *Transactions of Information Processing Society of Japan*, 47(11):3021–3031.

Wan, X. (2009). Co-training for cross-lingual sentiment classification.

Wu, Y., K. K. and Matsumoto, K. (2014). Three predictions are better than one: sentence multi-emotion analysis from different perspectives. *IEEJ Transactions on Electrical and Electronic Engineering (TEEE)*, 9(6):642–649.