Estimation Method of L2 Learners' Second Language Ability by using Features in Conversation

Xinnan Chen
s1821642@s.tsukuba.ac.jp
Tsukuba University
Ibaraki, Tsukuba, Japan

Muhammad Haseeb UR Rehman Khan
s1826098@s.tsukuba.ac.jp
Tsukuba University
Ibaraki, Tsukuba, Japan

Kei Wakabayashi
kwakaba@slis.tsukuba.ac.jp
Tsukuba University
Ibaraki, Tsukuba, Japan

ABSTRACT

We are conducting a research to train second language(L2) learners's second language ability by utilizing chat system. The main problem of existing chat systems is that it is not possible to chat with learners to adapt their second language level. In this research, in order to add a function to an existing chat system we need to measure the learner's second language level. So, to extract learners' second language capability, we propose a method to predict the language examination score of learners from chat context. This research investigates, first whether the number of utterances, number of sentences, word tokens and word types per utterance of chat context are correlated with second language examination score. Second, we build a predicting model to see the relationship between the chat context and second language examination score. As feature values of regression model for predicting the language examination score, we use variables chat time, sentence time, word token and word type. Also the unnatural sentence structure as a variable. For evaluation we use the root mean square error to check the results of prediction model, we use this model with Japanese and English chat and compare the results. We show how this chat context data is affecting the second language examination score and discuss strategies for future enhancements.

CCS CONCEPTS

• **Computing methodologies → Discourse, dialogue and pragmatics**.

iiWAS2019, December 2–4, 2019, Munich, Germany

ACM ISBN 978-1-4503-7179-7/19/12…$15.00
https://doi.org/10.1145/3366030.3366037

KEYWORDS

machine learning, neural network, chat system, second language level

ACM Reference Format:
Xinnan Chen, Muhammad Haseeb UR Rehman Khan, and Kei Wakabayashi. 2019. Estimation Method of L2 Learners' Second Language Ability by using Features in Conversation. In *The 21st International Conference on Information Integration and Web-based Applications & Services (iiWAS2019), December 2–4, 2019, Munich, Germany*. ACM, New York, NY, USA, 9 pages.
https://doi.org/10.1145/3366030.3366037

1 INTRODUCTION

In recent years, the number of second language learners studying or working abroad has increased. Two decades ago English was the de facto lingua franca for the commerce around the globe. It had been estimated that about 750M people use English as a second language, as opposed to 375M native English speakers [5]. And numbers are been growing since then, not just in commerce but in all the field all around the globe. Not only English, the amount of second language learners is growing in the recent years. For example, according to a survey [1] conducted by the Japan Student Services Organization(JASSO) in May 2017, result of an annual survey of international students in japan shown the number of foreign students are 267,042 persons. Increased by 27,755 persons (11.6 percent) compared with the result of last year.

When learners studying abroad or need to communicate with foreigners in their life, one of the problems they face is that they can not use second language to communicate well because of their hesitation. Communicating with others in the second language is very effective to be more fluent in that language, but in addition to being very nervous to talk to people in second language for the first time, it is likely to be afraid of the cultural differences of others because sometimes you can't express yourself properly, so you can not advance the conversation.

[1] https://www.jasso.go.jp/en/about/statistics/intl_student/

Therefore, it is considered effective to use a chat system as a training tool for learners in conversation skills. In recent years, research on chat systems has been actively conducted. A chat system is a dialogue system that aims to establish natural conversation with human beings without focusing on specific topics and themes, etc. It is expected to be used for the purpose of study, consulting, entertainment etc. Fryer et al. [9] chat experiment of students with two chatbots concluded that chatbots are more comfortable, enjoyable and relaxed chat partners as compared to classical teaching partners such as teachers and students and it helped the authors to claim that "chat-bots could provide a means of language practice for students anytime and virtually anywhere".

However, in these chat systems, there is a problem that, in order to be used directly by users, response sentences could be difficult for foreigners to understand because of the inability of chat systems to understand L2 learner's second language ability. As an example, english chat system sending a message like "do you fancy a drink " could be difficult to understand for students and they can not continue the conversation. For this reason, it is necessary to have a chat system that responds to the second language ability of learners so that chat system do not use sentences that are difficult for learners to understand. In order to realize this, it is important to add a function to the existing chat system that can estimate the second language ability of learners from conversations. So the question is "Can we estimate a person's language ability automatically by examining a conversation of the person?". Our focus of research is second language ability estimation from chat conversation.

The contributions of the paper are summarized as follows:

- This problem hasn't been discussed earlier and has no such dataset available which can be used for this research so we collect text conversations data from students.
- Extract second language ability insights from the results of statistical analysis of linguistic features and error features based on collected data.
- Propose a prediction model which can estimate the second language score using features obtained from statistical analysis.
- From the result, simple conversations without any guideline showed no strong relationship between linguistic features and language examination score. But even these simple conversations, using error features in predict model showed a little better result than baseline. The future studies should focus on conversations which can extract language proficiency, and also other language features.

2 RELATED WORK

As related work in this research, we have research on fields of computer assist language learning, utterance analysis and automatic grading.

In the computer assist language learning filed, humans can teach or learn anytime and anywhere on web-based education which is the most beneficial feature of this field [14]. This field has evolved a lot recently and now people can communicate easily with native speakers using different web-based tools, such as Youtube, Facebook, and Udemy etc. These tools can help people in the second language acquisition process.

Virous et al. [19] described a web based voice tutor which is intelligent enough to provide feedback and instructions to individual students based on ITS (intelligent tutoring system) and AH (adaptive hypermedia) techniques. Moreover with the help of long term student model Web PVT can perform diagnostic error and ambiguity resolution. But these language learning computer system have some limited conditions, like special topics Q&A, facilitator dependency, access control, and need more time to learn student's profile. Our goal is to make second language learning accessible in a convenient way. So, we think of chatbot is great tools.

The use of chatbots has notably increased recently in many fields especially education and artificial intelligence make it more promising. The goal of Unriza and Carolina's master thesis [18] was to make a chatbot with user having a natural conversation which can assess the user's level into the CEFR framework. They collected the training data of particularly and short answers. In their research they proposed method couldn't able to classify the conversation sentence data into different CEFR levels, and they didn't get the original language level data from participants for testing. So we want to use different method to detect the learners' language ability from conversation. And find what kind of feature is effect on learner's language ability, and these features whether have a correlation with second language level.

According to extract features from chat, we search the utterance analysis field, and to see what kind of features used in past research. A 25 years plan by William [20] mentioned as "quality" factors use for automatic grading essay's features. such like Fluency, Spelling, Diction. Mihai's [4] research based on William's [20] paper, proposed few same "quality" factors along with "utterance" for chat system assessment.

Because text in chat is short than essay, so we focus on feature about numbers of unnatural places in the conversation, and create Unnatural Place Miodel based on Language models in which the features of sentences are learned using neural networks is called neural network language model

(NNLM). One of the statistical language models is Bengio Model [1] and our unnatural place model is based on it.

So, we want to extract some features from conversation and language examination score to build a predict model.

3 DATA COLLECTION

In this research, as data for verifying whether it is possible to predict the score of the Japanese and English language examination from the contents of conversation, two surveys were conducted where students were asked to be part of an actual conversation for a limited amount of time, and conversations in the form of texts were collected. Each student also had to provide their scores of the Japanese and English language examination.

For English conversation, the participants were students studying English as second language. Using the chat tools Facebook or LINE, we asked 23 experimental participants to interact with a character called "Jam" who's a new foreign student came to japan for the first time. The conversation was in English. Of the 23 participants, 2 were Japanese, and the remaining 21 were Chinese international students. The subject of the dialogue was not set in particular, and they talked freely for 30 minutes. In order to continue the conversation, 10 topics were selected for "Jam" in advance. "Jam" was allowed to change the topic if a participant wasn't seem interested or hadn't anything to say about the current topic. The topics were nationality, food, travel, future plans, job, hobby, study, vacation, abroad life and some advice.

A summary of the collected data is shown in **Table 1**. "Utterances" column in this Table indicates the number of utterances from each speaker in this approximately 30 minutes dialogue. And number of utterances means, if conversation partner speaks with multiple speech bubbles before and after Jam speaking, the symbol representing the boundary of the speech balloon is described in the sentence and treated as one utterance. "Time" represents the time taken for the entire conversation. "Test Name" and "Test Year" respectively indicate the test type and year when the English language proficiency test was last taken.

For Japanese conversation, the participants were students studying Japanese as second language. Using the chat tool LINE or WeChat, we asked 17 participants to interact with one of the researchers. The conversation was in Japanese. All of the participants were Chinese students who knows Japanese. The subject of the dialogue was not set in particular, and they talked freely for 30 minutes. In order to continue the conversation, 10 topics were selected in advance. Researcher was allowed to change the topic if a participant wasn't seem interested or hadn't anything to say about the current topic. The topics were same as for English conversation.

ID	Test Name	Test Year	Utterances	Time(min)
1	TOEFL iBT	2016	24	30
2	IELTS	2017	29	34
3	TOEFL iBT	2017	18	30
4	TOEIC	2015	22	32
5	TOEIC	2015	25	34
6	TOEFL iBT	2017	30	35
7	TOEIC	2016	46	34
8	CET-6	2010	30	31
9	TOEIC	2016	20	30
10	TOEIC	2017	25	31
11	TOEIC	2018	30	30
12	TOEIC	2016	12	31
13	IELTS	2017	13	32
14	TOEIC	2015	15	30
15	TOEFL iBT	2017	28	31
16	TOEIC	2014	20	33
17	TOIEC	2016	37	32
18	TOEIC	2013	17	30
19	TOEIC	2016	20	31
20	IELTS	2014	16	30
21	TOFEL iBT	2016	25	31
22	TOEFL iBT	2017	37	31
23	TOEIC	2017	25	30

Table 1: English conversation data

A summary of the collected data is shown in **Table 2**. "Level" and "Text Year" respectively indicate the level and year when the Japanese language examination was taken. And the other two columns are as same with **Table 1**.

All IELTS scores are between 0 and 9, TOEIC score is between 0 to 990, TOEFL iBT score is between 0 to 120, CET-6 is from 0 to 710. These are English Proficiency tests taken by participants. For Japanese Proficiency test JLPT is the only test taken by participants but it has different level. e.g. N1, N2, N3, N4 and N5. And each level has different grading policy. So, it's a difficult to compare the scores of examinations with each other. That's why in this study, the statistical analysis was conducted using only the data of 13 participants at JLPT N1 level, and the data of 13 participants at TOEIC because these were the highest common tests taken by different participants.

4 STATISTICAL ANALYSIS

Linguistic features role has been actively studying by researchers since 4 decades ago in the second language proficiency writing. The difference in writing proficiency levels have traditionally depended on surface analysis features

ID	Level	Test Year	Utterances	Time(min)
1	N2	2016	26	33
2	N1	2015	56	35
3	N1	2012	52	32
4	N2	2016	40	35
5	N2	2016	32	30
6	N1	2011	49	30
7	N1	2012	38	35
8	N1	2016	28	31
9	N1	2016	19	30
10	N1	2014	41	30
11	N1	2016	49	31
12	N1	2013	50	31
13	N1	2017	11	30
14	N2	2014	15	30
15	N1	2014	42	31
16	N1	2011	20	33
17	N1	2015	40	37

Table 2: Japanese conversation data

such as word type, length of text, word frequency and repetition. (e.g. [2], [7], [8], [6], [10], [11], [15], [16], [17]) Because of in our research chat ability is discourse ability and writing proficiency. Therefore we will extract some linguistic features mentioned in these research.

Linguistic Features

In our research, we extract the few features based on the traditional linguistic features (surface), combined with basic discourse features. These are "number of utterances", "number of sentences per utterance" ,"number of tokens per utterance" and "number of types per utterance". We extract these features from the chat. A scatter plot for each language to see the correlation with features and language examination score, is shown as follow in **Figure 1** and **Figure 2**.

From the graphs it can been seen there is no clear relation found between linguistic features and language examination score. Error production along with other features outside of the linguistic features might play an important role in language skills [3]. so additional error features are added for language proficiency level prediction model.

Error Features

Sentence structure is an important aspect of language proficiency and tells a lot about the second language skill level. So finding the unnatural sentence structure is helpful in the language skills estimation for which we have developed a neural network model. And unnatural words places are considered as error features for the prediction model.

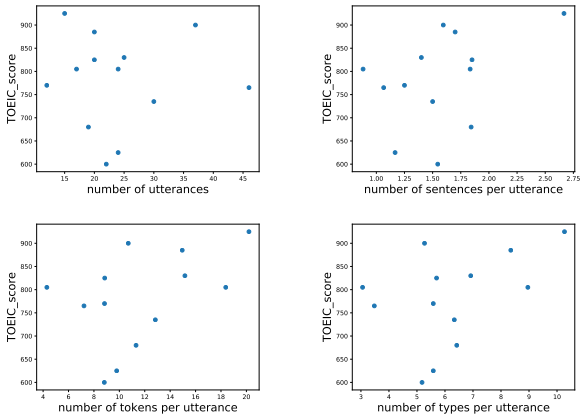

Figure 1: TOEIC tester linguistic features

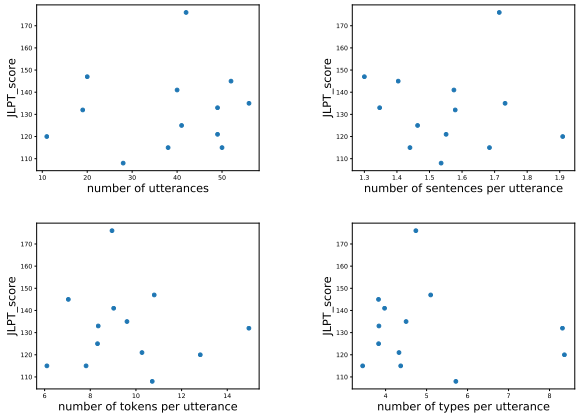

Figure 2: JLPT tester linguistic features

Unnatural Place Model. To find the unnatural sentence structures we tried to use a language model. Language models in which the features of sentences are learned using neural networks is called neural network language model (NNLM). In NNLM, when the number of $n-1$ word strings were given, estimate the probability that the word w will appear next position n. One of the statistical language models is Bengio Model [1] and our unnatural place model is based on it.

The Bengio model estimates the probability of occurrence of a word by n-gram model by using a neural network. This model learns from the word sequences probability and distributional representation of words. Sentence that never appeared gets a high probability if it is constructed by the similar words appearing in an already existing sentence.

To detect unnatural word in a sentence, we adopt an idea in the vector Log-bilinear Language Model (vLBL) [13] that is proposed for obtaining better word embeddings. In vLBL model, the appearance probability of each word $\mathbf{w}_j$ in the

conversation is estimated by using the surrounding words $\mathbf{w}_{j-2}, \mathbf{w}_{j-1}, \mathbf{w}_{j+1}, \mathbf{w}_{j+2}$ as the feature. The unnatural place model we use in this work is obtained by modifing the Bengio model to use the surrounding words as input. **Figure 3** shows the architecture of the unnatural place model.

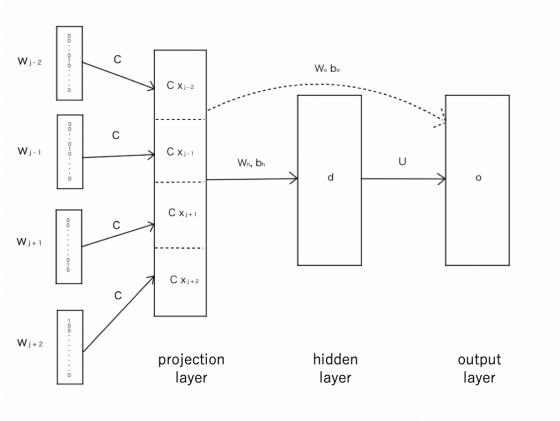

Figure 3: Architecture of Unnatural Place Model

First, each word is represented by a vector called one-hot expression, in which only the elements of the index assigned to that word are 1 and the other is 0. By one-hot expression, we express sentences as word sequence $\mathbf{S} = w_1, w_2, ..., w_{n_S}$. Here, w_j is a N -dimensional vector. When predicting the probability of any j -th word w_j of $\mathbf{S}$, the surrounding words $\mathbf{w}_{j-2}, \mathbf{w}_{j-1}, \mathbf{w}_{j+1}, \mathbf{w}_{j+2}$ be model inputs. Each input vector is transformed by using $N \times P$ projection matrix C that represents the word embedding vectors of dimension P. The transformed vectors are concatenated as $\mathbf{c} = (C\mathbf{w}_{j-2}, C\mathbf{w}_{j-1}, C\mathbf{w}_{j+1}, C\mathbf{w}_{j+2})$. As a hidden layer of dimension H, we apply a linear transformation to c with $P \times H$ weight matrix W_h and H dimensional bias vector $\mathbf{b}_h$, i.e., $\mathbf{d}' = W_h\mathbf{c} + \mathbf{b}_h$. Nonlinear transformation is applied to $\mathbf{d}'$ by using an activation function of tanh to obtain $\mathbf{d}$, i.e., $d_k = \tanh d'_k$.

The output layer $\mathbf{o}$ of dimension N is composed as a sum of two factors: (i) embedded vectors $\mathbf{c}$ that is linearly transformed with $4P \times N$ weight matrix W_o and N dimension bias vector $\mathbf{b}_o$ and (ii) hidden vectors $\mathbf{d}$ that is linearly transformed with the weight matrix U of $H \times N$, formally, $\mathbf{o}' = \mathbf{W_o c} + \mathbf{b_o} + \mathbf{Ud}$. Softmax function is applied to obtain the probabilistic distribution of the target word as $o_k = e^{o'_k} / \sum_{l=1}^{N} e^{o'_l}$. When we denote the index corresponding to the word $\mathbf{w}_j$ to be predicted by v, the value o_v is the probability of the word $\mathbf{w}_j$ is in this context. We can simply express this probability in the output as follows.

$$P(\mathbf{w}_i \mid \mathbf{w}_{i-2}, \mathbf{w}_{i-1}, \mathbf{w}_{i-1}, \mathbf{w}_{i-2}) \tag{1}$$

We denote a conversation by D. Each conversation is represented by a set of utterances $D = \{D^{(1)}, D^{(2)}, ..., D^{(M)}\}$ where M is the number of utterances. Each utterance is a series of words $D^{(m)} = \mathbf{w}_1^{(m)}, \mathbf{w}_2^{(m)}, ..., \mathbf{w}_n^{(m)}$. The probability of each word $\mathbf{w}_j^{(m)}$ is calculated using the unnatural place model using the surrounding words as the context $p(\mathbf{w}_j^{(m)} \mid \mathbf{w}_{j-2}^{(m)}, \mathbf{w}_{j-1}^{(m)}, \mathbf{w}_{j+1}^{(m)}, \mathbf{w}_{j+2}^{(m)})$. The number of unnatural words e is calculated as follows.

$$e = \sum_{m,j:\mathbf{w}_j^{(m)} \in C} \delta\left(p(\mathbf{w}_j^{(m)} \mid \mathbf{w}_{j-2}^{(m)}, \mathbf{w}_{j-1}^{(m)}, \mathbf{w}_{j+1}^{(m)}, \mathbf{w}_{j+2}^{(m)}) \geq \theta\right)$$

where $\delta(x)$ is Dirac's delta function that returns 1 when the predicate x is true and 0 when it is false. When we use e as the feature for prediction of the language examination score, we normalize the number of unnatural words by the number of utterances, i.e., $\tilde{e} = \frac{e}{M}$.

The probability of each word is determined using the unnatural place model for the conversation text of the participants, and the number of words whose probability exceeds the threshold θ is used as the feature quantity of the number of unnatural parts. However, the unnatural place model simply assigns low probability even to natural words for low frequency vocabulary. For this reason, we limit the set of detected vocabularies C to frequent vocabularies. In order to decide C, the top T vocabularies are considered as elements of C, arranged in order of the frequency of vocabulary in Corpus called T, the number of vocabulary to be detected.

Corpus. From this model it is possible to find words that are considered unnatural to the words before and after. To train our model we need a big corpus of sentences. We used 3 corpus separately for the training of this model.

For Japanese chat experiment, training of this model was done using a natural Japanese corpus, In this research, a natural Japanese corpus, from Yahoo! Wisdom Bag [2] from April 1, 2004 to April 7, 2009 of 3,000,000 sentences out of the questions were collected.

For English chat experiment, we trained our model using two different English corpus, the first is NLTK Brown corpus of 57,340 sentences and other one is the Twitter from January 21, 2009 to February 8, 2009 of 240,000 tweets.

The first million-word electronic corpus of English, called Brown Corpus was created at Brown University in 1961. 500 sources text data is part of this corpus and sources are also categorized by genre, such as news and editorial etc. We used this corpus using NLTK python package. This package covers the symbolic and statistical natural language processing and have program modules, datasets, exercises and tutorials related to NLP [12].

[2] https://chiebukuro.yahoo.co.jp/

5 EXPERIMENT

Prediction Model

We extracted linguistic features and error features from chat dataset now we want to predict participants language examination score based on these features so we used two machine learning algorithms with different combinations of these features as input and predicted score as output. The used algorithm in this research as mentioned below.

Linear Regression. To check the linearity of features with language examination scores we used linear regression algorithm as part of prediction model. Linear Regression is a supervised learning algorithm which predicts the straight line relationship between features and labels by calculating the weights for each feature.

k-nearest neighbor Regression. We also checked this as a classification problem and used k-nearest neighbor regression algorithm in prediction model. This algorithm attempt to estimate the mapping function from the features to language examination score.

First we used only linguistic features, which are number of sentences per utterance, number of tokens per utterance, number of types per utterance and this combination is called **S(3)**. Next combination of features included error feature along with **S(3)** and it is denoted as **S + E(4)** in this paper. And third combination of features is only the error feature and mentioned as **E(1)** in evaluation results.

So one output of prediction model means one machine learning algorithm with one combination of features. A example of the model algorithms and combinations are shown in **Figure 4**.

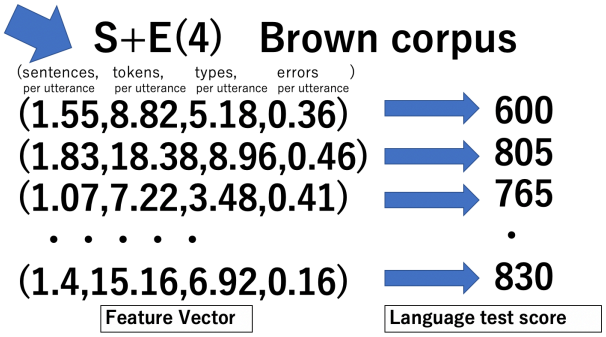

Figure 4: Predict Model

Evaluation Method

We perform feature extraction on the data of 13 people (JLPT N1 tester) and 13 people (TOEIC tester), and divided it into training data and testing data based on cross validation. In each test set, the language examination score of the test data predicted by the regression model learned using training data is evaluated by the root mean square error (RMSE). RMSE is the method to determine the accuracy of our model in predicting the target values. The smaller the value of the RMSE, the better is the predictive of the model. RMSE is defined as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad (2)$$

N is the total number of predictions, y_i is the actual value, and $\hat{y}_i$ is the predicted value. In this experiment, the total number of predicted targets is 13 for Japanese chat experiment (JLPT N1 tester), and 13 in English chat case (TOEIC tester).

Because our dataset is small, for feature combination and machine learning algorithms, so we use the leave-one-out method for cross validation. In other words, the testing data is the score value of each person, and the data of the remaining 12 people are the training data. In baseline evaluation, the ith value was calculated by averaging all the other values of language examination scores.

Evaluation Result

The Minimum Value of RMSE. First, we investigated the effect of the detection target vocabulary number T and the threshold θ to RMSE, which are the hyperparameters in the unnatural place model. We examined the effect of T and θ under **E(1)** condition that uses the number of errors per utterances alone as a feature of the prediction model. We tried different Ts and θs and check the change of RMSE. We controlled the T from 0 to 2000 (increase by 100) and the threshold θ from 0.0 to 0.8 (increase by 0.01) and found the best combination of T and θ in each configuration.

Table 3: RMSE of JLPT scores

Feature Pattern	Linear	K-Neighbors
Baseline	**18.99**	
$S(3)$	22.32	19.86
$T = 600$, $\theta = 0.27$		
$S + E_{Yahoo}(4)$	25.51	19.86
$E_{Yahoo}(1)$	20.83	**18.38**

Table 4 show the result in Japanese conversation data. In **Table 3**, E_{Yahoo} means the error feature was computed by using an unnatural place model trained with Yahoo corpus. The value of RMSE by using the baseline method is 18.99. The minimum RMSE 18.72 is obtained when **E(1)** is used as

the feature value of the k-nearest neighbor regression algorithm with $T = 600$, $\theta = 0.27$.

Table 4: RMSE of TOEIC scores

Feature Pattern	Linear	K-Neighbors
Baseline	**104.61**	
$S(3)$	116.13	118.48
$T = 200$, $\theta = 0.40$		
$S + E_{Brown}(4)$	117.99	119.4
$E_{Brown}(1)$	104.84	**91.76**
$T = 250$, $\theta = 0.10$		
$S + E_{Twitter}(4)$	109.97	119.55
$E_{Twitter}(1)$	102.51	**85.36**

Table 4 show the results in English conversation data. In **Table 4**, E_{Brown} means the error feature was computed by using an unnatural place model trained with Brown corpus, and $E_{Twitter}$ means one with Twitter corpus. Findings have shown that the value of RMSE by using the baseline method is 104.61. In Brown corpus condition, the minimum value of RMSE 91.76 is obtained when **E(1)** is used as the feature value of the k-nearest neighbor regression algorithm with $T = 200$, $\theta = 0.40$. In Twitter corpus condition, the minimum RMSE 85.36 is obtained when **E(1)** is used as the feature value of the k-nearest neighbor regression algorithm with $T = 250$, $\theta = 0.10$.

We can see all RMSE values shown in the tables are almost the same as the baseline method, which merely uses the average of the language examination score. **Figure 5** shows scatter plots to examine correlations between the number of errors per utterance and the RMSE with **E(1)** feature. From these plots, there seems to be no clear correlation between the number of errors and the RMSE. These results show that the estimation of the language examination score from the chat is a difficult task at least for the proposed method.

Effect of T ***and*** θ. We checked the sensitivity of RMSE against the configurations of hyperparameters T and θ in unnatural place model. **Figure 6** shows the changes of RMSE on Japanese conversation data with Yahoo corpus using JLPT score. The blue lines show the RMSEs with linear regression and the orange broken lines show the RMSEs with k-nearest neighbor regression. In the left side of **Figure 6**, we show the result when we set T to be 600 and vary the threshold θ from 0 to 0.8 (increase by 0.01). The plots in the right side show the result when we set θ to be 0.27 and vary the detection target vocabulary number T from 0 to 2000 (increase by 100). We examined the $E(1)$ and $S + E(4)$ feature combinations in each configuration. These plots show that the proposed method is less sensitive to settings of the hyperparameters T and θ.

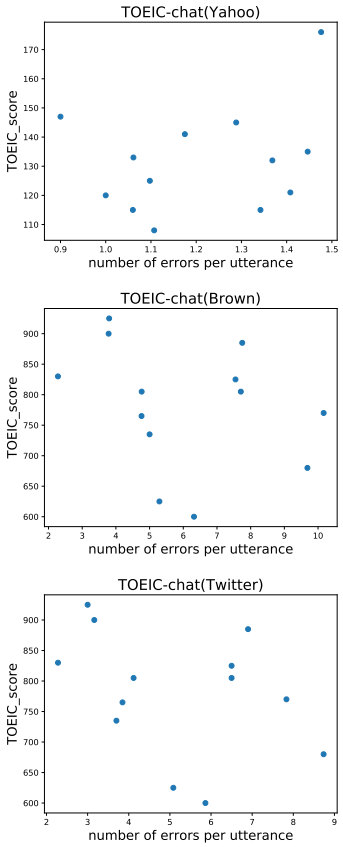

Figure 5: Error feature of each corpus

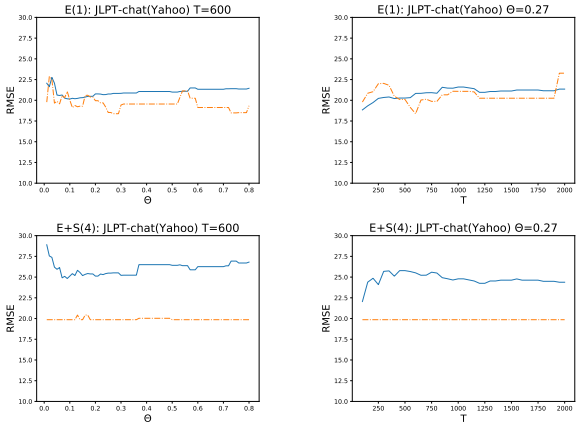

Figure 6: Effect of θ and T in unnatrual place model trained with Yahoo corpus

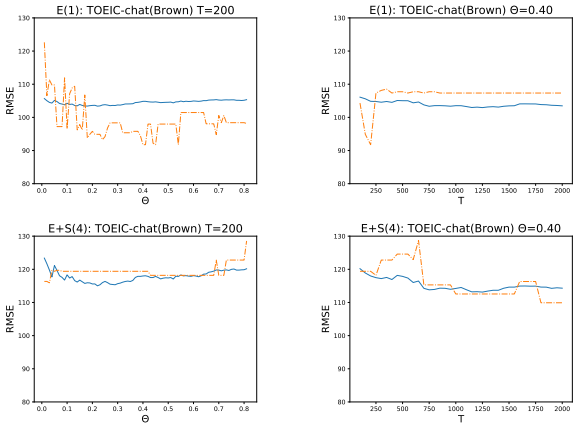

Figure 7: Effect of θ and T in unnatrual place model trained with Brown corpus

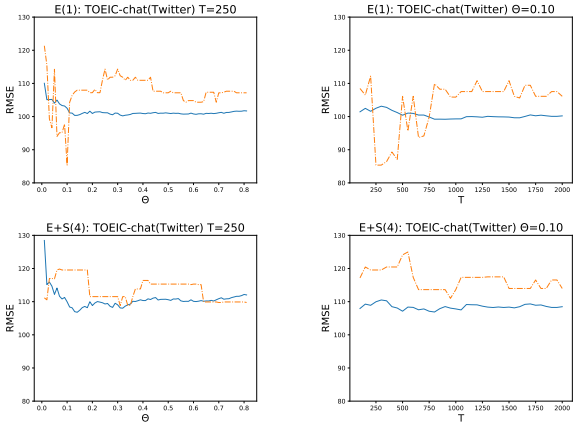

Figure 8: Effect of θ and T in unnatrual place model trained with Twitter corpus

Figure 7 shows the changes of RMSE in the condition of English conversation data with the Brown corpus using TOEIC score and **Figure 8** is one with the Twitter corpus. The plots in the left side show the result when we set T to be 200 and 250 respectively, and the plots in the right side show the result when we set θ to be 0.40 and 0.10, respectively. $E(1)$ and $S+E(4)$ feature combinations are examined in each configuration. These plots show that the proposed method is a little more sensitive to the hyperparameters T and θ for English conversation data. This result suggests that we need a hyperparameter tuning when we apply the proposed method to another language.

6 CONCLUSION

In this research, we aim at adding a function to measure the learner's second language level in existing chat systems to allow chat systems to adapt second language learners' language ability to continue conversations. To extract learners' second language capability, we propose a method to predict the language examination score of learners from chat context.

To verify our method, we did as follows. First, we collected data by using chat tools (Line, Facebook, Wechat) by a chat experiment with 40 second language learners in total, each person talks with free topics in 30 minutes. Because every examination has different level, we chose participants having the same test and the same level, and examined the relationship with the features of conversation data and their examination score. Second, we confirmed if linguistic features (the number of utterances, number of sentences, word tokens and word types per utterance of chat context) are correlated with language examination score. But it can be seen there is no clear relation found between linguistic features and language examination score. Then, we build a prediction model of language examination score by using the feature of chat context. Because error features are also important features that affect language ability, we added it to the prediction model. We used the unnatural place model to detect error word place in chat context to extract the error features. In the predicting model, we used two machine learning algorithms, linear regression and K-neighbors regression. Finally, we used RMSE to evaluated predicting model in each condition compared with the baseline method. The results show that the estimation of the language examination score from conversations is a difficult task, at least for the proposed method.

The limitation of our experiment is small data size (chat conversation data of 13 testers in each language). And linguistic features for the predicting model is not quality factors to find some correlations with conversation and language examination score. Because conversation data in a limited time is shorter than essays, the participants in the experiment cannot make serious, deep and formal discourse in the chat. Although the error feature extracted by the unnatural place model shows a sensitive result compared with the baseline method, we think error feature in short conversation is a quality factor effect L2 learners' second language ability. The result obtained from our research allows us to conclude that the achievement of this study includes the evaluation of collected data about second language leaners' chat conversation data and language examination score contribution in a chat environment. With some improvements, our research can be a standalone evaluation method to check a person's second language ability. And we will improve the method of how to extract error feature from chat in future work. We hope this function on chatbots can make an environment for practice of L2 learner's second language

ability in the future and it can be used as a guidance for chatbot developers.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Number 19K20333 and 16H02904.

REFERENCES

[1] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research* 3, Feb (2003), 1137–1155.

[2] U Connor. 1984. A study of cohesion and coherence in ESL students' writing. Papers in Linguistics. *International Journal of Human Communication* 17, 3 (1984), 301–304.

[3] Scott A Crossley and Danielle S McNamara. 2012. Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading* 35, 2 (2012), 115–135.

[4] Mihai Dascalu, Stefan Trausan-Matu, and Philippe Dessus. 2010. Utterances assessment in chat conversations. *Research in Computing Science* 46 (2010), 323–334.

[5] Crystal David. 1997. English as a global language. *UK: Cambridge University Press. Print* (1997).

[6] María José De La Fuente. 2002. Negotiation and oral acquisition of L2 vocabulary: The roles of input and output in the receptive and productive acquisition of words. *Studies in second language acquisition* 24, 1 (2002), 81–112.

[7] Cheryl A Engber. 1995. The relationship of lexical proficiency to the quality of ESL compositions. *Journal of second language writing* 4, 2 (1995), 139–155.

[8] Dana R Ferris. 1994. Lexical and syntactic features of ESL writing by students at different levels of L2 proficiency. *Tesol Quarterly* 28, 2 (1994), 414–420.

[9] Luke Fryer and Rollo Carpenter. 2006. Bots as language learning tools. *Language Learning & Technology* 10, 3 (2006), 8–14.

[10] Scott Jarvis. 2002. Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing* 19, 1 (2002), 57–84.

[11] Scott Jarvis, Leslie Grant, Dawn Bikowski, and Dana Ferris. 2003. Exploring multiple profiles of highly rated learner compositions. *Journal of Second Language Writing* 12, 4 (2003), 377–403.

[12] Edward Loper and Steven Bird. 2002. NLTK: the natural language toolkit. *arXiv preprint cs/0205028* (2002).

[13] Andriy Mnih and Koray Kavukcuoglu. 2013. Learning word embeddings efficiently with noise-contrastive estimation. In *Advances in neural information processing systems*. 2265–2273.

[14] Christoph Peylo, Wilfried Teiken, Claus-Rainer Rollinger, and Helmar Gust. 2000. An Ontology as Domain Model in a Web-Based Educational System for Prolog.. In *FLAIRS Conference*. 55–59.

[15] Joy Reid. 1986. Using the writer's workbench in composition teaching and testing. *Technology and language testing* (1986), 167–188.

[16] Joy Reid. 1990. Responding to different topic types: A quantitative analysis from a contrastive rhetoric perspective. *Second language writing: Research insights for the classroom* (1990), 191–210.

[17] Randi Reppen. 1995. Variation in elementary student language: A multi-dimensional perspective. (1995).

[18] Unriza Salamanca and Carolina Eugenia. 2019. Study and design of a chatbot as support of foreign languages learning within an e-learning platform.

[19] Maria Virvou and Victoria Tsiriga. 2001. Web passive voice tutor: an intelligent computer assisted language learning system over the WWW. In *Proceedings IEEE International Conference on Advanced Learning Technologies*. IEEE, 131–134.

[20] William Wresch. 1993. The imminence of grading Essays by computer–25 Years Later. *Computers and composition* 10, 2 (1993), 45–58.