Constructing a Japanese Business Email Corpus Based on Social Situations

Muxuan Liu[1,2] **Tatsuya Ishigaki**[2]
Yusuke Miyao[3,2] **Hiroya Takamura**[2] **Ichiro Kobayashi**[1,2]
[1] Ochanomizu University
[2] National Institute of Advanced Industrial Science and Technology
[3] The University of Tokyo
{liu.muxuan, koba}@is.ocha.ac.jp
{ishigaki.tatsuya, takamura.hiroya}@aist.go.jp
yusuke@is.s.u-tokyo.ac.jp

Abstract

The use of Japanese is influenced by many social situations such as differences in social status and familiarity between speakers, so it is necessary to consider these social situations when constructing models to process Japanese text. In this paper, we attempt to build a corpus containing information about social situations for training machine learning models using the description method of social situations in Systemic Functional Linguistics, which views language as a social semiotic system. We also verify its applicability through some machine-learning models.

1 Introduction

In social situations, various factors such as age, gender, social status, and intimacy between speakers and listeners, as well as the purpose, content, flow, and setting of the conversation, influence the language used (Lee, 2016). Japanese language use is considered to strongly reflect these social factors (Matsumura and Chinami, 1998). Factors such as whether the conversation takes place face-to-face or over the phone, the type of conversation (small talk or discussion), and the tone of the conversation (friendly or confrontational, relaxed or tense) can also affect the language used. Regional differences may also play a role.

For instance, when considering a Japanese error correction task focused on emails, it is not only important to correct grammatical mistakes, but also to address situations where a writer might mistakenly use an expression such as “明日は休みます．ありがとう．” (I will take the day off tomorrow. Thank you.) when requesting a vacation from their superior. In this case, the issue at hand goes beyond mere grammatical errors and involves the unique social context of the Japanese language. It is common in Japanese communication to use an expression that seeks the approval of the receiver, such as “明日は休みたいですが、よろしいでしょうか” (I would like to take tomorrow off, would that be alright?), as a way of showing respect towards the superior. In this example, we see that different levels of politeness can be conveyed through grammatical differences in expressions that carry the same meaning. The use of interrogative sentence forms can create a more polite impression, a frequent practice in Japanese to demonstrate respect and consideration towards the other party. Such grammatical nuances are closely intertwined with the social context. This highlights the importance of considering sociolinguistic factors when approaching Japanese language tasks (Fujiwara et al., 2009). To properly capture and utilize individual social contexts, a corpus containing attribute information about these social contexts is necessary for machine learning models.

In this paper, we chose to focus on the genre of email (of the Cultural Affairs Council, 2018), which falls under document communication and is influenced by the social context of both the sender and receiver. Our objective was to create a Japanese corpus that encompasses a deeper understanding of the social context, incorporating more comprehensive “analysis information.” This “analysis information” pertains to intricate details related to the social context, as governed by the principles of systemic functional linguistics and their selectional restrictions(Halliday, 1978). Furthermore, we aimed to confirm whether the annotated social situation information in the created Japanese business email corpus can be inferred from text information, and to verify whether the social situations in the corpus are accurately reflected.

Specifically, we utilized seven pre-training models of Japanese BERT to construct a multi-label classification model for the labels representing the social relationships between the senders and receivers of business emails. This model classifies the social relationships between the email senders and receivers and assigns probability values corresponding to each relationship label. The classification experiment aims to verify whether the corpus can be sufficiently distinguished and interpreted, i.e., whether useful features can be extracted for classification. We conducted the corpus training using the pre-training models of Japanese BERT to verify whether we can accurately identify and classify the various social relationships annotated in the business emails. Furthermore, we evaluated the classification accuracy of the model for the social relationships in the email corpus and conducted a verification to confirm whether the annotations are effective in identifying and specifying social situations and relationships.

2 Related Work

Numerous studies have been conducted to analyze the grammatical and semantic characteristics of corpora and to examine how machine learning should be applied to them. For example, a recent study by O'Connor et al. (O'Connor and Andreas, 2021) investigated how the context of the text in the training data contributes to the accuracy of predictions in Transformer-based language models. Their findings suggest that using longer context is more important than finer grammatical and semantic details of words in the context.

The previous literature has reported that BERT-based classifiers can maintain a high accuracy of 75% to 90% even when the input word order is randomly shuffled (Pham et al., 2020). This indicates that BERT models are robust to variations in word order. Additionally, masked language models (MLMs) have shown excellent performance due to their ability to model higher-order word co-occurrence statistics. Sinha et al. (Sinha et al., 2021) demonstrated that pre-training MLMs on randomly shuffled word order texts achieves high accuracy in various downstream tasks, highlighting the importance of considering word order in natural language understanding.

Despite these findings, there is a lack of research on BERT models using long corpora that incorporate knowledge of social situations and systemic functional linguistics (SFL). To address this gap, our study aims to conduct a comparative experiment on various BERT models using a specially designed corpus that includes a deep understanding of social contexts and SFL principles. This evaluation will help to assess the models' performance in handling complex linguistic phenomena and contribute to advancing natural language processing techniques.

3 Expression of Social Situations

3.1 Systemic Functional Linguistics (SFL)

In this section, we introduce Systemic Functional Linguistics (SFL) as it plays a crucial role in understanding the linguistic aspects of social situations, which is the focus of our research. SFL, established by M.A.K. Halliday, views linguistic systems as social semiotic systems, emphasizing the interplay between language and social contexts.

SFL categorizes the linguistic system into three hierarchically interconnected semiotic systems: the semantic stratum, the lexicogrammar stratum, and the expression stratum, all of which are contextually conditioned. Together, they form a network of linguistic options for social communication, referred to as the "system network" (detailed in section 3.2). Importantly, SFL highlights the relationship between the selection of a situation, the meaning expressed, and the linguistic features chosen, such as vocabulary and grammar. This coordination of different symbol systems enables the effective expression and interpretation of meaning in various social contexts.

In our study, we explore how SFL principles can contribute to the analysis of social situations in the context of email communication. By understanding the role of SFL in uncovering deeper linguistic knowledge and the relationships between language and social activities, we aim to shed light on the nuances of language use in email communication. Additionally, SFL's framework allows us to investigate how BERT models, with their ability to capture contextual information, perform when applied to long corpora with social situational knowledge. The incorporation of SFL in our research offers valuable insights into the development of language models and their applicability to real-world communication scenarios. With this connection established, we proceed to present the details of SFL's linguistic system and its significance in understanding social situations in the context of our research on email communication. The selection of a situation constrains the selection of meaning, and meaning constrains the selection of vocabulary and grammar. The system of selection is thus formed through the coordination of

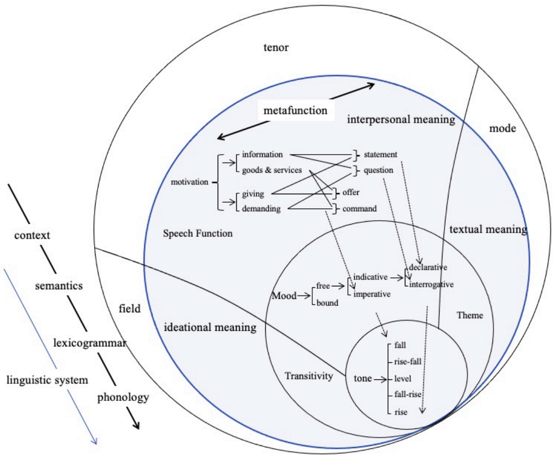

Figure 1: Language systems by systemic functional linguistics（adapted from (Halliday and Matthiessen, 2006)）

different symbol systems, including meaning, vocabulary and grammar, and expression. Figure 1 provides an overview of the linguistic system according to SFL. In Halliday's framework, the context of the situation in which dialogue occurs is explained through three frameworks: "what is happening (Field)," "who is taking part (Tenor)," and "how language is being used (Mode)" (Halliday, 1978).

In this paper, the communication discussed is limited to e-mail. The "Field" becomes the social activity of "communication through email," and the communication content can be seen as various language use domains. The tenor becomes the social role of the participants who exchange emails. The mode becomes the channel of communication that specifies the form of the text, which becomes an "electronic document."

3.2 System Network

One of the most important features of Systemic Functional Linguistics (SFL) is the representation of language resources as a network of choices, known as the "system network". The system network includes different symbolic language resources at various layers, such as the context layer, the semantic layer, the lexico-grammar layer, and the expression layer, which collectively comprise the language system. The system network operates by selecting language resources from higher layers to lower layers based on the resources selected from the context layer. For example, in a medical situation, events such as "examination" and "treatment" exist, and corresponding lexicogrammatical resources such as "surgery" and "take this medicine and monitor for a while" are selected. The system network represents the process of realizing texts, and it describes the relationships between different resources (features) and how they are selected. In terms of "choice," the system network is described using square brackets ('[') for selecting one feature and curly braces (' ') for selecting multiple features simultaneously. This way, the system network provides a framework for understanding how language resources are selected in the realization of texts.

4 Construction of a Corpus Based on SFL

In this study, we construct a corpus of emails (especially business emails) that takes into consideration social situations captured by SFL. The process of construction is as follows:

a) Constructing System Network A system network of selection for social situations targeting emails is constructed based on SFL.

b) Setting and Collection of Scenes Using cloud sourcing, diverse scenes reflecting the options of the system network constructed in 1 are collected, thereby setting various situations. The process of setting scenes by selecting options from the system network of selection corresponds to annotating social situations in emails collected in 3.

c) Collection of Mails Emails are collected using the scenes collected in 2 and cloud sourcing.

d) Annotation Based on SFL Annotation based on SFL is performed for the mails collected in 3.

The following section will describe each step in detail.

4.1 Constructing System Network

Tenor(Role Relationship) "Tenor" refers to the relationship between the speaker and the listener in the exchange of language expressions, or between the sender and receiver in email communication. To consider the social standing of the participants in a typical business email conversation, we constructed a selectional system for the tenor relationship as shown in Figure 2. The attributes of "internal" and "external" represent the internal and exter-

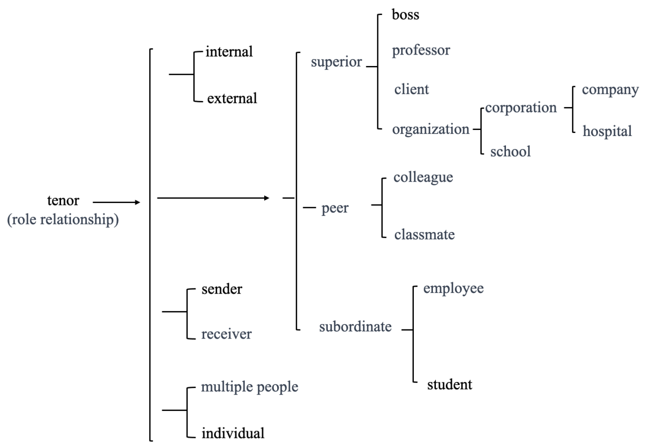

Figure 2: System Network of "Tenor"

Situation You are under the care of department A of your client. Please write a year-end greeting email to all members of department A at your client.	
Text	
Subject: Greetings for the End of the Year To all members of department A at XX Corporation, I am writing to express my gratitude for your continuous support throughout the year. My name is XX from XX Corporation. As the year-end approaches, there is only a little time left in this year. I would like to express my sincere appreciation for your significant cooperation during this fiscal year. We will continue to do our best in our business as much as possible in the coming years, so we would appreciate your continued support. Finally, I would like to express my best wishes for your further prosperity. I hope you have a wonderful new year. From XX at XX Corporation	
Labels (Participants)	
Superiority relationship (receiver)	Superior
Superiority relationship (sender)	Subordinate
Sender's role	Employee
receiver's role	All members of a department in a client company
Internal/External	External
Number of senders	Individual
Number of receivers	Multiple
Labels (Speech function)	
Sender's action	Assertion
Sender's detailed action	Greeting
Exchange role	Giving
Exchange item	Information

Table 1: Example corpus: Email text and its labels for an employee greeting all members of a department in a client company

nal positional relationships of the conversation participants. Generally, "internal" refers to "family, colleagues, or members of the same group," while "external" refers to "unfamiliar people, outsiders, people from other companies, or people from other groups" (Hirabayashi and Hamada, 1988). Additionally, to represent the sender's position, we divided the characters and organizations commonly used in business emails into three attributes: superior, peer, and subordinate, from the sender's perspective.

Speech Function In SFL, "Tenor" affects the "Speech Function" in the semantic layer. Regarding speech functions, Teruya (Teruya et al., 2022) analyzed human relationships and interpersonal meanings using SFL and summarized the interpersonal roles in speech functions (see Appendix 7 for details).

Based on the interpersonal relationships and speech functions, this paper constructed a selectional system for common speech motives in business emails, as shown in Figure 3.

4.2 Setting and Collection of situations

In order to represent the social situation of communication through business emails, we set the attributes of the corpus based on the options of the selection system network shown in the previous section. In the domain of "communication through business emails," we gathered a corpus of language usage in diverse contexts through crowdsourcing, where communication scenarios were created. Specifically, we established 20 pairs of typical sender-receiver relationships in business emails based on the role relationships depicted in Figure 2 (refer to Appendix: Table 9). Furthermore, we set eight common purposes for senders, including greetings, expressions of gratitude, apologies, rejections, inquiries, requests, notifications, and reminders, based on the "sender's actions" shown in Figure 3.

We collected a total of 1,040 communication situations by hiring 52 Japanese native speaker workers through crowdsourcing, with each worker creating approximately 20 situations. In order to improve the quality of the crowdsourced data, we provided a large number of examples, as shown in Appendix Table 10, when commissioning the work. Additionally, we requested that workers refer to the receiver with designations such as "Mr./Ms. A" or "President A" in order to facilitate receiver identification in future data analyses.

4.3 Collection of Emails

We selected 770 valid situations from the previous step and obtained five emails for each situation through crowdsourcing. To ensure the qual-

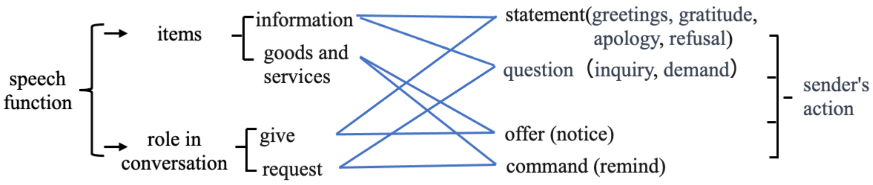

Figure 3: System Network of "Speech Function"

Sender's Action	Number of situations	situations (%)	Number of Emails	Average Sentence Length	Total Number of Words	Number of Unique Words
Decline	70	0.09	350	16.69	23406	1086
Request	100	0.13	500	17.60	35961	1561
Apology	100	0.13	500	17.14	42326	1576
Urging	100	0.13	500	20.57	42758	1130
Gratitude	100	0.13	500	15.82	37232	1499
Greeting	100	0.13	500	15.62	38370	1641
Notice	100	0.13	500	18.31	44822	1903
Inquiry	100	0.13	500	19.07	40734	1614
Total	770	1	3850	17.67	302521	3869

Table 2: Statistics showing the characteristics of the corpus

ity of the data, we provided examples, as shown in Appendix Table 8, at the time of commissioning and requested that they be created while considering interpersonal relationships and social hierarchies within reasonable common sense, such as using polite language with superiors and casual language with friends. It is worth noting that we did not set specific criteria for collecting scenarios; instead, we asked workers to provide scenarios within the bounds of common sense. We think that this approach would result in a more diverse and contextually rich collection of scenarios.

Additionally, to make it easier to use for tasks such as morphological analysis, we requested that the subject and addressee fields be filled in and proper nouns such as location and the participant's name be replaced with "XX". For example, "My friend at AAA University BBB Faculty" would be replaced with "My friend at XX University XX Faculty."

4.4 SFL-Based Annotation

An example of the overall corpus is shown in Table 1. To facilitate use in machine learning, we simplified the structure of the selection system network listed in Section 4.1 when using the options as annotation names in the corpus. As shown in the example email text illustrated in Table 1, it is a notification email from an employee to a customer. The hierarchical relationship of "superior", "peer", and "subordinate" in the participant selection system network shown in Figure 2 is represented by the annotation "subordinate (sender)" and "superior (receiver)". In addition, the inner and outer relations between the sender and receiver's specific identities and their belongingness are expressed by the annotation "internal-external relationship".

Regarding the annotation that represents speech functions, we set it based on the selection system network shown in Figure 3. Since we collected the email texts from the sender's perspective, we excluded "receiver's actions" when setting the annotations. "Sender's actions" have four items, "statement", "question", "offer", and "command", and the details of each are expressed by the annotation "sender's actions (details)". The "roles in the interaction" are selected based on whether the sender "gives" or "requests" what they want to communicate, and the interaction is either about "information" or "goods and services"(Teruya et al., 2022). In the case of the example email text, the employee is giving information to the customer. In the case of the exchange of "goods and services", such as a conversation that starts with a request to "return a book", since the purpose is achieved by returning the book to the sender, the role played by language in this interaction is different from that of an exchange of "information"(Teruya et al., 2022).

5 Computational Analysis of the Corpus

The statistical data of scenes and emails in the corpus are presented in Table 2. The total number of words and the number of unique words (including symbols) were calculated using the National Institute for Japanese Language and Linguistics' morphological analysis tool, “Web ChaMame”[1].

In terms of the act of “refusal” in the sender's actions, according to a study on the human relationship between the “requester” and the “refuser” by Cai (Cai, 2005), it is generally considered to be an act that occurs between an individual sender and an individual receiver. As it is rare for one sender to refuse an entire group of receivers (for example, a student refusing all members of a club), such one-to-many interpersonal pairs of “refusal” were excluded during the corpus construction. Therefore, there were only 70 scenes of “refusal”, compared to other items.

5.1 Experimental Setup and Evaluation Methods

In this study, we consider an experiment to test our corpus to determine whether our corpus can help the model learn the social context better. Specifically, we experiment on a multi-label classification task that predicts 11 different labels (Receiver's social position, Sender's social position, Sender's identity, Receiver's identity, Relationship, Number of senders, Number of receivers, Sender's action, Sender's action (details), Role in conversation, Items). However, it is important to note that actual emails were difficult to collect due to privacy concerns. Therefore, for this study, we conducted experiments solely using the corpus we created.

We evaluate the multi-label classification models using macro-F1 score and visualize the classification results using a confusion matrix. The confusion matrix enables us to analyze the performance of the models by visualizing the predicted and actual results for each label.

5.1.1 Pre-trained Language Models

As shown in Table 3, we use seven BERT models and their variants in the classification task and compare their performance. As each model has different architecture and parameter settings, their performance and characteristics in processing Japanese text differ. By comparing the performance of these models in the classification task, we can better understand the strengths and weaknesses of each model in Japanese text classification, which can serve as a reference for future research.

6 Results and Discussion

6.1 Multi-class classification

Table 3 shows the experimental settings and accuracy results of each language model. In this study, we conducted experiments by dividing the corpus into two subsets: training and validation, in a ratio of 8:2. For the batch size setting, BERT$_{large}$, RoBERTa $_{base}$, DeBERTa $_{base}$, and DeBERTa $_{large}$ were set to a batch size of 16, while ALBERT, BERT$_{base}$, and BERT$_{base-wwm}$ were set to a batch size of 90.

As shown in the experimental results of the language models presented in the table, their accuracies range from 67.4% to 83.5%.

Language Model	Multi-label Classification Accuracy (%)
ALBERT [2]	67.4
BERT$_{base}$ [3]	82.3
BERT$_{base-wwm}$[4]	80.7
BERT$_{large}$[5]	**83.5**
RoBERTa$_{base}$ [6]	68.8
DeBERTa$_{base}$ [7]	78.6
DeBERTa$_{large}$ [8]	79.1

Table 3: Experimental results for each model

Among these models, BERT$_{large}$ achieved the highest accuracy of 83.5%, making it the top-performing model in this experiment. The results demonstrate the varying performance of different language models in the multi-label classification task.

Overall, the results suggest that the BERT series of models tend to have relatively high accuracies. Moreover, in most cases, increasing the training batch size led to better model learning, indicating that the training batch size has a significant impact on model performance.

Table 4 presents the macro-average F1 scores for each label of the language models. Specifi-

[1] https://chamame.ninjal.ac.jp

[2] https://huggingface.co/ALINEAR/albert-japanese-v2

[3] https://huggingface.co/cl-tohoku/bert-base-japanese

[4] https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking

[5] https://huggingface.co/cl-tohoku/bert-large-japanese

[6] https://huggingface.co/rinna/japanese-roberta-base

[7] https://huggingface.co/ku-nlp/deberta-v2-base-japanese

[8] https://huggingface.co/ku-nlp/deberta-v2-large-japanese

Language model	Receiver's social position	Sender's social position	Sender's identity	Receiver's identity	Relationship	Number of senders	Number of receivers	Sender's action	Sender's action (details)	Role in conver- sation	Items
ALBERT	0.846	0.846	0.879	0.806	0.941	1.000	0.933	0.867	0.823	0.919	0.918
BERT$_{base}$	0.891	0.891	0.928	**0.848**	**0.956**	1.000	0.930	0.896	0.852	0.917	0.932
BERT$_{base\text{-}wwm}$	**0.897**	**0.897**	**0.929**	0.840	0.938	1.000	0.925	**0.901**	0.852	0.938	**0.941**
BERT$_{large}$	0.894	0.894	0.926	0.836	0.953	1.000	**0.945**	0.885	**0.855**	**0.945**	0.922
RoBERTa$_{base}$	0.874	0.874	0.891	0.831	0.960	1.000	0.936	0.892	0.846	0.940	0.927
DeBERTa$_{base}$	0.847	0.847	0.859	0.777	0.951	1.000	0.929	0.856	0.818	0.921	0.902
DeBERTa$_{large}$	0.874	0.874	0.890	0.824	0.938	1.000	0.935	0.884	0.834	0.935	0.935

Table 4: Comparison of macro-F1 scores by each model (by label)

Language model	Receiver's social position	Sender's social position	Sender's identity	Receiver's identity	Relationship	Number of senders	Number of receivers	Sender's action	Sender's action (details)	Role in conver- sation	Items
ALBERT	87.5	87.5	90.0	78.3	96.2	100.0	94.6	88.8	82.4	92.3	93.8
BERT$_{base}$	91.1	91.1	**94.2**	83.1	96.8	100.0	94.3	90.5	85.2	92.3	94.7
BERT$_{base\text{-}wwm}$	90.8	90.8	93.4	82.5	95.9	100.0	94.2	**91.3**	85.4	94.0	**95.5**
BERT$_{large}$	**91.6**	**91.6**	93.2	**84.0**	**96.9**	100.0	**95.7**	90.2	**85.5**	**94.7**	93.6
RoBERTa$_{base}$	89.3	89.3	91.1	81.2	97.3	100.0	96.1	90.7	84.6	94.3	94.3
DeBERTa$_{base}$	86.6	86.6	87.9	76.7	96.8	100.0	94.2	87.8	81.8	92.6	92.3
DeBERTa$_{large}$	89.4	89.4	90.8	80.6	95.9	100.0	94.7	90.2	83.6	93.9	94.9

Table 5: Comparison of accuracy in each model (by label)

cally, for each language model, we compared the performance for labels such as "Receiver's social position", "Sender's social position", "Sender's identity", "Receiver's identity", "Relationship", "number of senders", "number of receivers", "Sender's action", "Sender's action (details)", "Roles in communication" and "Items".

As shown in Table 4, model BERT$_{base}$ demonstrated the highest scores for "Receiver's identity" and Relationship" whereas model BERT$_{base-wwm}$ achieved the highest macro-F1 scores for "Receiver's social position".

Table 5 presents a comparison of the accuracies for each label in the corpus used in this study. The results showed that the BERT$_{base}$ and BERT$_{base-wwm}$ models demonstrated similar performance, but the BERT$_{base}$ model showed slightly better performance in identifying the sender's identity. The BERT$_{large}$ model showed superior performance to all other models in identifying factors such as the "Receiver's social position", "Sender's action" and "Roles in communication". The DeBERTa$_{large}$ model exhibited excellent performance in identifying the "Number of receivers," "Sender's action (details)" and "Roles in communication". These results indicate that the performance of language models in identifying various social factors in communication may vary depending on the specific context.

	Sender	Receiver
Pair 1	Subordinate	Superior
Pair 2	Peer	Peer
Pair 3	Superior	Subordinate

Table 6: Pairings of Social Positions

In addition, for other labels, since each email has only one sender, the macro-F1 value for the "Number of senders" label would always be 1, and the accuracy would also be 100%. For the "Receiver's social position" and "Sender's social position" labels, there are, as shown in Table 6, only three pairs. Therefore, when one social position is fixed, the other social position is also fixed, resulting in similar macro-F1 values and accuracy for these two labels.

6.2 Confusion Matrix

A confusion matrix is a method of comparing the predicted labels of a classifier with the true labels, providing more detailed information to evaluate the performance of the classifier.

By using a confusion matrix, we can identify the error patterns of the model for specific labels and gain a deeper understanding of the model's performance through further analysis.

As shown in Table 5, the accuracy for "Receiver Social Position" and "Sender's Action (Details)" is lower to 90%, so we mainly focused on the confusion matrix of these two labels and created the visualizations in Figure 4 and Figure 5.

Contrast with other labels (number of classes is from 2 to 4), Receiver Social Position" and "Sender's Action (Details)" has more classes (19 and 8 classes), which makes it difficult to predict.

When analyzing the content of "Sender's Movement (Details)", it was found there are more mispredicting between "greetings" and "gratitude". As shown in Table 1, one of the reasons why BERT's classification results are incorrect is thought to be the presence of expressions of gratitude such as "Thank you" and "Thank you for your help" in the "greetings" email.

In various languages, greeting expressions are commonly used to establish a rapport with others, and it is a universal practice for topics to follow greetings. In the Japanese context, you often encounter highly standardized expressions like 'Good morning (こんにちは)' and 'Good evening (こんばんは),' which serve as courteous ways to initiate interactions. Additionally, there are 'quasi-greetings' that convey a stronger intention on the part of the speaker. These 'quasi-greetings,' including phrases like 'Hello' and 'Nice to meet you,' are often followed by discussions related to 'gratitude' and 'apology' (Xiao, 2019). In this experiment, it was found that the classification accuracy of emails containing typical greeting expressions was high, while the accuracy of quasi-greeting emails containing many expressions of gratitude was slightly low. Overall, it can be said that the model was able to learn the social situation based on the SFL for all label accuracies. The labeling of social situations for the corpus was confirmed to be accurate.

7 Conclusion

Based on systemic functional linguistics (SFL), this study created a Japanese corpus annotated with information on social role relationships embodied in business emails. The labels used for annotation were adopted from the choices within "Tenor" and "Speech Function" in the system network. Specifically, this corpus was constructed with a focus on social roles, especially those with clear social hierarchies, in business emails. It should be noted that not all the selection system network options of SFL were used as labels for annotation, as the emphasis was placed on social role relationships.

In addition, we constructed classification models using seven Japanese pre-trained BERT models based on our corpus and conducted comparative experiments. As a result of the experiments, classification models with high accuracy of about 90% in many cases were achieved. This suggests that our corpus is useful for machine learning models to learn interpersonal features of Japanese.

We have plans to release this corpus in the near future. One of the future prospects of this study is to apply the created classifier to an extended model designed for controlled language generation. To be more specific, we aim to enhance natural language generation by utilizing the classifier developed in this study for conditioning or controlling the generation model. This approach is anticipated to elevate the degree of control and accuracy in a range of natural language generation tasks, including dialogue systems and automatic translation models.

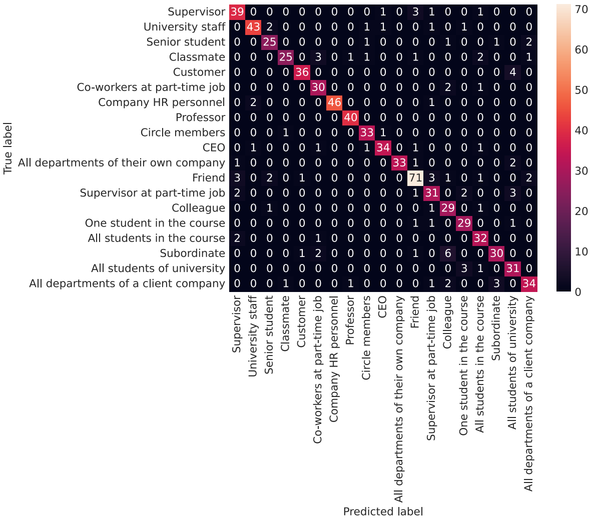

Figure 4: Visualization chart with confusion matrix of "Receiver's identity"

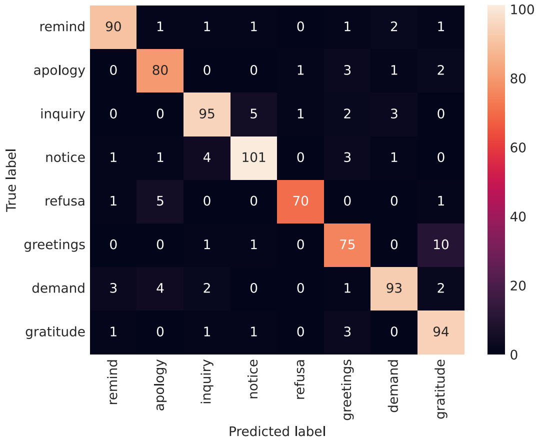

Figure 5: Visualization chart with confusion matrix of "Sender's Action (detail)"

Acknowledgements

This paper is based on results obtained from a project JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO) and the Avanade Research Grant. For experiments, computational resource of AI Bridging Cloud Infrastructure (ABCI) provided by National Institute of Advanced Industrial Science and Technology (AIST) was used.

References

Yinzhu Cai. 2005. Declination in emails by native japanese speakers: A perspective on politeness communication(nihongo bogo washa noi- me-ru niokeru kotowari taigū komyunike-shon no kanten kara, in japanese). *Waseda University Japanese Language Education Research*.

Asa Fujiwara, Hitomi Abe, yuko Ooi, Hiroko Tsubahara, and Noriko Yoshida. 2009. Teaching expressions related to consideration in japanese language education(nihongo kyoiku niokeru hairyo ni kakawaru hyogen no shido). *Minutes of Hokkaido University Graduate School of Education*.

M.A.K. Halliday. 1978. *Language as Social Semiotic: The Social Interpretation of Language and Meaning*. Open University set book. Edward Arnold.

M.A.K. Halliday and Christian Matthiessen. 2006. *Construing Experience Through Meaning: A Language-Based Approach to Cognition*. Continuum. Illustrated edition, 672 pages.

Shusuke Hirabayashi and Yumiko Hamada. 1988. *Series of Japanese Example Sentences and Problems for Foreigners 10 Honorifics(gaikokujin no tame no nihongo reibun mondai shiri-zu 10 keigo ,in japanese)*. Aratake.

Soon-hyeong Lee. 2016. Discourse analysis of learners and native speakers of japanese an empirical study of listener verbal behavior of learners of japanese and native speakers（danwa bunseki kara mita nihongo gakushusha to bogo washa no kikite gengo kodo no jissho teki kenkyu, in japanese）. *D. Thesis, Department of Japanese Language Education, Graduate School of Humanities and Sciences, Tokyo Metropolitan University*.

Yoshiko Matsumura and Kyoko Chinami. 1998. The actual situation and effects of style alternation in japanese discourse(nihongo danwa niokeru sutairu kotai no jittai to sono koka,in japanese). *Linguistic Science(gengo kagaku,in japanese)*, (33).

Joe O'Connor and Jacob Andreas. 2021. What context features can transformer language models use? In *Annual Meeting of the Association for Computational Linguistics*.

National Language Sub-Committee of the Cultural Affairs Council. 2018. Verbal communication for mutual understanding (report)(wakariau tame no gengo komyunike-shon hokoku,in japanese). *Kokugo Subsection Meeting*.

Thang M. Pham, Trung Bui, Long Mai, and Anh M Nguyen. 2020. Out of order: How important is the sequential order of words in a sentence in natural language understanding tasks? *ArXiv*, abs/2012.15180.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joëlle Pineau, Adina Williams, and Douwe Kiela. 2021. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. In *Conference on Empirical Methods in Natural Language Processing*.

Kazuhiro Teruya, C.M.I.M. Matthiessen, John Bateman, Byrnes Heidi, M.A.K.Halliday, Kaori Okuizumi, and Yumiko Mizusawa. 2022. *Linguistics for Better Understanding Meaning - Invitation to Systemic Functional Linguistics(imi ga wakaru you ni naru tame no gengogaku, in japanese)*. Kuroshio.

Jie Xiao. 2019. A study on the criteria and classification of greetings and greetings expressions : Based on japanese perspectives(aisatsu to aisatsu hyogen no handan kijun oyobi bunrui nikansuru kosatsu : nihongo no shiten o moto ni,in japanese). *Hokkaido University Graduate School of Humanities and Human Sciences*.

A Appendix

Table 7: Speech Functions and Responses (adapted from (Teruya et al., 2022))

Role in Conversation	Exchangeable Items	Speaker’s Move	Hearer’s Move	
		Initiation	Response	
			Expected Response	Unexpected Response
Give	Information	Statement	Agreement	Disagreement
Request		Question	Answer	Avoidance
Give	Goods and Services	Offer	Acceptance	Rejection
Request		Command	Compliance	Refusal

Table 8: Example of email collection

Situation	あなたは大学生です。今回家庭の事情で半年休学するようになったことを、お世話になっている A 教授にメールでこの件を知らせてください。	You are a college student and have decided to take a six-month leave of absence due to family circumstances. Write an email to Professor A, who you are grateful to for their support, to inform them of this matter.
Example of answer	件名：休学について A 先生 いつもお世話になっております。XX 学部 1 年生の XX です。 実は先月、私の母が交通事故に遇い、大腿骨骨折で入院することになりました足を骨折して歩けなくなりました。私は母子家庭で、家では私しか面倒を見る人がいないので、学校に通いながら、同時に面倒を見るようにしてきました。しかし、1 ヵ月間を試して、やはり介護と学業との両立が難しいと感じており、先日、半年休学を申請しました。 これまでいつも丁寧にご指導いただきありがとうございます。来年復学したら、もう一度先生の授業を履修させていただきます。 これからも何卒よろしくお願いいたします。 ーーーーーーーーーーーーーーーーーーーーーーー XX 学部 1 年生の XX	Subject: Regarding the Leave of Absence Dear Professor A, I am XX, a first-year student in the XX department. Thank you for your guidance thus far. I regret to inform you that my mother was involved in a traffic accident last month and has been hospitalized with a fractured thigh bone, making it impossible for her to walk. I come from a single-parent household and am the only one available to take care of my mother, so I have been juggling my studies and caregiving responsibilities at home. However, after a month of trying, I have found it challenging to balance both and have decided to apply for a leave of absence for the next six months. Thank you for your continuous support and guidance thus far. When I return to school next year, I would like to take your class again. Sincerely, ーーーーーーーーーーーーーーーーーーーーーーー XX, a first-year student in the XX department

Table 9: Settings for the relationship between the sender and receiver

	Sender	Receiver
Pair 1	Student	Professor
Pair 2	Student	Company HR personnel
Pair 3	Student	Circle members
Pair 4	Student	Supervisor at part-time job
Pair 5	Student	Senior student
Pair 6	Student	Friend
Pair 7	Student	Classmate
Pair 8	Student	Co-workers at part-time job
Pair 9	Student	University staff
Pair 10	Student	All students of university
Pair 11	Employee	Supervisor
Pair 12	Employee	Customer
Pair 13	Employee	All departments of their own company
Pair 14	Employee	All departments of a client company
Pair 15	Employee	Colleague
Pair 16	Employee	Subordinate
Pair 17	Employee	CEO
Pair 18	Teacher	Colleague
Pair 19	Teacher	One student in the course
Pair 20	Teacher	All students in the course

Table 10: Example of a situation setting

【従業員が『自社のある部門全員』にメールで催促する必要がある場面】の解答例	An example answer for a scenario where an employee needs to send an email to ”all members of a department in their own company” to urge them for something.
「あなたは経理部に所属しています。あなたは最近、旅費の払い戻し期限が12月10日17:00であることを、全社員に電子メールで送りました。しかし、期限を過ぎても、まだ提出していない社員がいます。全員に対してできるだけ早く提出するよう促すメールを書いてください。」	”You belong to the accounting department. Recently, you sent an email to all employees stating that the deadline for reimbursement of travel expenses is December 10th, 17:00. However, there are still some employees who have not submitted their expenses. Please write an email urging all employees to submit their expenses as soon as possible.”