國立臺灣大學電機資訊學院資訊工程學研究所

碩士論文

Graduate Institute of Computer Science and Information Engineering

College of Electrical Engineering and Computer Science

National Taiwan University

Master Thesis

Application of Cultural Differences in Expressing Apologies, Requests and Thanks in Multi-Party Dialogue and Everyday Japanese Conversation Datasets for Machine Translation

中文多方對話語料集和日語日常會話語料集在表達歉意、請求和感謝上的文化差異及其在機器翻譯上的應用

中華民國 110 年 9 月

September, 2021

doi:DOI Number

摘要

隨著機器翻譯技術的快速發展，我們使用機器翻譯與來自不同文化背景的人進行交流變得越來越普遍。在跨文化交際中，說話人的語言能力越高，在特定情況下當說話人的語言使用與聽話人的對話文化的語用規則相悖時，說話人就越有可能被認為性格不好或缺乏社交禮儀。這種跨文化對人際關係的負麵影響問題在未來可能會變得更加嚴重，更加頻繁，因為連目前的機器翻譯都擁有與外語學習者相同或更好的語言能力。因此，機器譯員有必要進行文化意識的翻譯，即在聽眾的對話文化中適當使用語言的翻譯。

在這項研究中，我們首先確認，語言使用在不同的文化中是不同的，這取決於情景和人際關係。然後，為了通過文化意識的機器翻譯減少不同文化帶來的問題，我們創建了一個文化意識的中日對話平行語料集，包括人際關係標簽和三種情況標簽：道歉、請求和感謝。我們對我們的創建的語料集進行了統計分析。此外，我們使用了一個 sequence-to-sequence 模型來對我們的語料集進行情況分類和文化感知的機器翻譯。

在本研究中，我們驗證了人際關係對日文的情景分類和中翻日的文化感知機器翻譯的貢獻。此外，對於中文的情景分類和日翻中的文化感知機器翻譯來說，這一點的結果也與一個觀點相吻合。中文在不同背景資訊中語言使用會大變化的語用規則比較模糊Chen et al. [2013]。這些結果很重要，因為該模型可以透過包括

doi:DOI Number

人際關係以及背景資訊在對話文化中實現文化感知。

關鍵字：語料集、機器翻譯、文化差異、跨文化溝通、語用

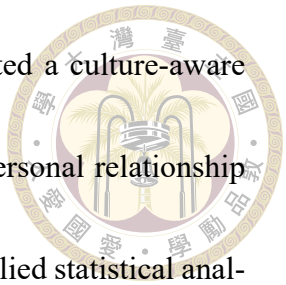
doi:DOI Number

Abstract

With the rapid development of machine translation technology, it is becoming increasingly common for us to communicate with people from different cultures using machine translation. In cross-cultural communication, the higher the speaker's language skill such as grammar skills, vocabulary, fluency, and so on, the more likely it is that the speaker will be perceived as having a bad personality or lacking social etiquette when the speaker's language use is contrary to the pragmatic rules of the listener's dialogue culture in a certain situation. This problem of cross-cultural negative impact on human relations is likely to become more serious and more frequent in the future since even current machine translators have the same or better language skills than foreign language learners. Therefore, machine translators must perform culture-aware translation which is a translation with the appropriate language use in the listener's dialogue culture.

In this study, we first confirm that language use varies from culture to culture depending on the situation and interpersonal relationship. And then to reduce the problems caused

doi:DOI Number



by different cultures by culture-aware machine translation, we created a culture-aware parallel dataset for Chinese and Japanese dialogues including interpersonal relationship labels and three situation labels: apology, request, and thanks. We applied statistical analysis to our dataset. Besides, we used a sequence to sequence model to perform situation classification and culture-aware machine translation on our dataset.

We verified that the interpersonal relationship contributes to the situation classification for Japanese and the culture-aware machine translation for Chinese to Japanese translation on our dataset in this study. Besides, for Japanese to Chinese translation, the result is consistent with a a view of Chen et al. [2013] that Chinese have less clear rules of pragmatic for using different language uses depending on the interpersonal relation. These results are important in that the model can be culture-aware in dialogue culture with contextual information including the interpersonal relationship.

Keywords: dataset, machine translation, culture-aware, cross-cultural communication, cultural difference, pragmatic

doi:DOI Number

Contents

doi:DOI Number

doi:DOI Number

doi:DOI Number

doi:DOI Number

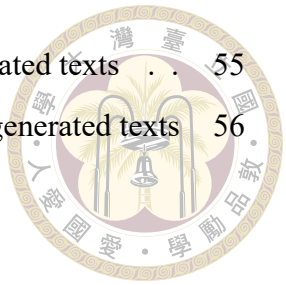
List of Figures

doi:DOI Number

doi:DOI Number

List of Tables

doi:DOI Number



doi:DOI Number

1 Introduction

We have organized this chapter as follows: 1.1 backgrounds on machine translation with respect to cross-culture, 1.2 our research motivations from the backgrounds, 1.3 goals of this research to achieve the motivation, 1.4 contributions of this research.

1.1 Background

With globalization, people have more opportunities to communicate with others who have different cultures and languages. There is a need to communicate more efficiently and comfortably with people from different countries, cultures, and languages because cultural differences often lead to communication breakdowns in cross-cultural communication.

Many empirical studies have addressed the issue of cultural differences and have shown that cultural differences exist in terms of differences in expression and strategies in specific contexts. Enomoto and Marriott [1994] state that the higher the speaker's proficiency in a foreign language, the more likely he or she is to suffer the negative effects of cultural differences.

With the development of deep learning techniques, machine translators have been making remarkable progress. Transformer-based neural networks such as BERT (Devlin

doi:DOI Number

In source language (Chinese):	请你看在我俩同学的份上吧！ 你有什么要求可以向我提提吧！ Please, for the sake of our classmates! If you have any requests, you can mention them to me
Machine Translation:	同級生のためにもお願いします。 何かリクエストがあれば、ぜひ聞いてみてくださいね。 Please do for me as your classmates. If you have any requests, please ask.
Human Translation:	お願い。 なんでも言ってくれていいから。 Please. You can tell me anything you want.

Table 1.1: Comparison between machine translation and human translation

et al. [2018]) and T5 (Raffel et al. [2020], **?**), exploiting the attention mechanism for handling complex interaction among words, perform well in a variety of challenging natural language processing tasks, including machine translation. High performance machine translation are available to the general public. In addition to available machine translators on websites, some services are equipped in browsers and applications and replace the translation results with a single click.

When it comes to conversation translators, their use in the real world is gradually increasing. They have started to be used in taxies, airports, and other situations where it is necessary to serve people in different languages. In the future, the use of translators will no doubt increase not only in service situations but also in person-to-person communication, such as between international students at universities and colleagues at work.

However, current machine translators do not perform culture-aware translation. The following is an example of the data used in this study. An example is utterances in a situation in which a couple is discussing whether to get a divorce or not. The source language is Chinese, and the target language is Japanese.

The notable difference between the machine-translated text and the human-translated

doi:DOI Number

text is that in the human translation, the part of the original sentence that refers to *for the sake of our classmates* has been removed. In the original text, this part serves as a negotiating element to make it easier for the listener to accept the request, and in the machine-translated text, there are some translations of this part. While this works in Chinese dialog culture, it does not work in the context of Japanese dialog culture. Rather, it is a statement that may bring discomfort to the listener.

Not only in the examples given above, but in various contexts, different cultures cause such cross-cultural problems. These problems can continue to occur in cross-cultural communication as long as both parties are not considerate of other cultures even using machine translators. When we assume that machine translators are used in everyday cross-cultural communication situations, it is not hard to imagine that communication breakdowns occur due to cultural differences will occur. If a speaker does not follow the rule of language use that is common sense or customary in the listener's conversation culture and inputs the speaker's utterance into the machine translator. The output from the machine translator would be an utterance that is not based on the listener's culture and does not fit the listener's culture. This can have a negative impact on their cross-cultural communication. Furthermore, given that machine translators are roughly equivalent in performance to a person who has acquired high vocabulary and grammar skills, the possibility that the communication with machine-translator brings about such negative effects is higher than that of the communication in which one of the humans uses a foreign language. Therefore, this is a critical issue in this day and age when machine translators are universally used in society.

doi:DOI Number

1.2 Motivation

We introduced the backgrounds and problems in cross-cultural communications and machine translations. There is no doubt that such problems exist now and will increase rapidly in the future. These cross-cultural problems are what we should solve.

The key to solving the problem lies in **contexts**. Many empirical studies in cultural communication have mentioned that different cultures have different expressions and strategies in specific contexts such as situations and interpersonal relationships between parties in conversations. In other words, one of these differences can manifest themselves as differences in language use.

A number of empirical studies have been attempting to find tendencies of language use in each culture by analyzing the differences in language use for specific situations or specific relations with some frameworks such as segmentation of utterances by sentence role, classification of strategies in each sentence, classification of semantic markers on words and phrases, and so on. The tendencies of differences in language use between Chinese and Japanese also have been found for specific situations (Song and Lee [1994], Lee and Song [1994]), Yu [1999], Cheng [2005], Ichihara [2016]) and for specific interpersonal relationships (Chen et al. [2013], Hill et al. [1986], Xing [2021], Huang [2015]).

The existence of tendencies is that machine learning allows us to learn from given data, discover rules, and make inferences about new data. In other words, the problems caused by cultural differences in cross-cultural communication could also be solved by machine learning.

However, there is still no dataset or task for solving the problems caused by cultural

doi:DOI Number

differences in the field of computer science. Therefore, we attempt to create a dataset and to demonstrate a sequence-to-sequence deep learning model could be aware of language uses in Chinese and Japanese.

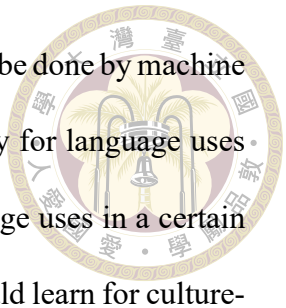
1.3 Research Goal

As we mentioned above, different cultures have different language uses in certain contexts. And there are tendencies for these differences. Therefore, in order for a computer to solve the problems caused by cultural differences in dialogue, we need to make a model learn and use the characteristics of each culture as culture-aware machine translation. We set three goals to be achieved in this study.

The first goal is to create a culture-aware parallel dialogue dataset. In order to perform culture-aware machine translation, it is essential to have a suitable dataset. In order to make the model culture-aware, it is also necessary to have the context information on the dataset. Therefore, We create a parallel dataset with parallel texts in Chinese and Japanese, situation labels, and interpersonal relationship labels based on A MultiParty Dialogue Dataset for Analysis of Emotions and Interpersonal Relationships (MPDD) (Chen et al. [2020]) and Corpus of Everyday Japanese Conversation monitor public version (CEJC) (Koiso et al. [2020]), 小磯 et al. [2019]).

The second goal is to demonstrate that the model can identify the language use that is plausible for certain situations in Chinese and Japanese culture respectively by situation classification task on our dataset. In a given culture, there is a tendency for the language uses in a certain situation. Therefore, the scores of classification would be higher, when the model is aware of appropriate language uses of a given culture in certain situations.

doi:DOI Number



The third goal is to demonstrate that culture-aware translation can be done by machine translation task on our dataset. In a given culture, there is a tendency for language uses in certain contexts. Also, in different cultures, the appropriate language uses in a certain context could be different. Thus, the language use which a model should learn for culture-aware translation depends on cultures and contexts. Therefore, a model is possible to generate a translation with more appropriate language use by providing the contexts when the model is aware of the appropriate language uses of a given culture in certain contexts.

1.4 Contribution

At first, we reduced the potential issues of cultural differences in cross-cultural communication with machine translation to the issue in natural language processing. Secondly, we created a cultural-aware parallel dialogue dataset with situation labels and interpersonal relationship label in Chinese and Japanese. Thirdly, we demonstrated that a model can be aware of the appropriate language uses of each culture in certain situations by situation classification task. Finally, we demonstrated that contextual information is important to make the model generate more culture-aware translation for each dialogue culture in Chinese and Japanese by culture-aware machine translation task.

doi:DOI Number

2 Literature Review

We organize the related work chapter into subsections focusing on different aspects, which includes the studies about the cultural difference from empirical studies and the related tasks and data in natural language processing.

2.1 Cultural Difference in Empirical Studies

One of the major issues in cross-cultural communication researches on applied pragmatics has been differences in language use in certain contexts from a culture to another culture or from a language to another language (Blum-Kulka and Olshtain [1984]). In previous studies, a part of the context is fixed and the differences in language use under that context were identified. It has already been shown that situation and interpersonal relation, as contexts, have an influence on language use.

2.1.1 Situation

There are situations in dialogue that can occur commonly in many cultures, such as apology, request, and thanks. The concept of these situations is universal, but the factors that cause these situations and the style of utterances vary from culture to culture. Previous studies have shown how language use tends to differ across cultures in certain situations.

doi:DOI Number

For instance, about apology, Guan et al. [2009] found that Chinese people are less willing to apologize than Americans, and speculated that this is because Chinese people are less likely than Americans to perceive the invasion of personal space as an act for which they should apologize. Barnlund and Yoshioka [1990] shows that Japanese prefer apologies more directly and tend to apologize more frequently than Americans. A preferred strategy used by the Japanese in apology situations is compensating, and that by the Americans is explaining the situation.

In the situation of request, Song and Lee [1994] states Chinese speakers most prefer to use *qing*, which expresses politeness and is used in imperative sentences. Lee and Song [1994] concluded that Chinese speakers prefer any direct expression and dislike euphemisms. Yu [1999] found that Chinese people prefer more direct expression in request situation than did American and expect the utterance in request situation to be elaborate, which make the length of utterance longer. On the other hand, Pizziconi [2003] Japanese speakers prefer to use euphemisms or any other indirect expression and apologies in order to show respect for the listener and to prevent the listener from feeling uncomfortable.

About thanking, Cheng [2005] states that Chinese speakers prefer invitations to dinner in situations of gratitude and are more likely to express gratitude to someone they are not close to than to someone they are close to when they receive a big favor from someone of equal status. Ichihara [2016] concluded that Japanese speakers tend to re-appreciate at a later date, which is one of the appreciation strategies, while Chinese speakers tend not to express appreciation.

As mentioned above, the previous studies have shown that there are tendencies of language use in Chinese and Japanese in various situations. There are also many previous

doi:DOI Number

studies on other situations, such as invitation and compliment, as well as on differences in other language and cultural areas. In this study, we create and analyze our dataset about apology, request, and thanks situations because we found more utterances of these situations in the corpora that we choose to use.

2.1.2 Interpersonal Relation

In different cultures, language use in the same interpersonal relation would be different because people from different cultures have different perceptions about a relationship. Cultural differences also can be seen in the way relationships are perceived and handled.

Generally, it is known that Western countries are individualistic and based on equal relationships between individuals, while Eastern countries are collectivistic and their words and actions are based on the relationship between themselves and others in the group. Also, many studies show that there are cultural differences in language use at smaller scales, such as country, language, or regions.

It has been shown that even Chinese and Japanese native speakers have different ways of perceiving relationships, and this affects their words and actions. Chen et al. [2013] analyzed the differences between Chinese, Japanese, and English in the way to extend the study of Hill et al. [1986]. His study investigated politeness using a variety of expressions to ask for a pen and the relationship between the person making the request and the person made the request and found that Chinese speakers tend to less clear distinction of the request language use of varying degrees of politeness depending on the interpersonal relationships. In other words, Chinese have less clear rules of pragmatic for using different language uses depending on the interpersonal relation, while Japanese is restricted by

doi:DOI Number

more stringent rules of pragmatic for using language uses with appropriate degrees of politeness depending on the interpersonal relationships. Xing [2021] found that when a person to whom to apologize is in a close relationship, Chinese speakers tend to avoid making a sincere apology for preventing them from an awkward situation or for showing to a person who to whom to apologize that they are close. While Japanese speakers choose an language use from various apologies depending on the degree of apology to be made and the situation, not on the relationship with the listener, and use facial expressions and attitudes to show sincerity in their apologies. Huang [2015] states that Japanese speakers used expressions of gratitude more frequently for relationship maintenance than Chinese speakers do, in the situation where the people are invited and requested.

As shown in the previous study above, the tendency of language use in Chinese and Japanese differs depending on the relationship even in the same situation. In this study, therefore, in order to pay attention to the interpersonal relationships of speakers in dialogue, we conduct experiments with interpersonal relation labels that have been given or assigned to the dataset.

2.1.3 Negative Pragmatic Transfer

Pragmatic transfer is one of the concepts in second language acquisition. There are two types of pragmatic transfer: positive transfer and negative transfer. Positive transfer occurs due to the similarity between the learner's native language and the target language and facilitates the understanding and acquisition of the target language. Negative transfer, on the other hand, occurs when learners inappropriately apply the pragmatic rules of their native language to learning the target language.

doi:DOI Number

A:	日本語から英語に訳した資料が A4 で 20 枚くらいあるんだけど、訳をチェックしてもらえない？明日の朝、出さないといけないから I have about 20 sheets of A4 paper that I've translated from Japanese to English. Could you check the translation for me? I have to turn them in tomorrow morning.
You:	

Table 2.1: An example of situation settings in the study of Meng[2008]

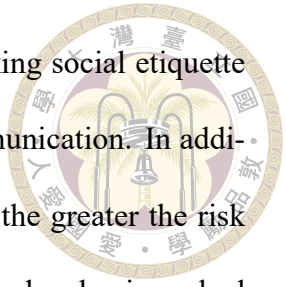
Negative pragmatic transfer often causes communication breakdown and misunderstanding in cross-cultural communication. Many studies have empirically demonstrated that second-language speakers, even if they have excellent grammatical and lexical skills in the target language, make pragmatic mistakes that prevent them from communicating effectively, resulting in misunderstandings and communication breakdowns. (Blum-Kulka [1980], Wolfson [1981], House and Kasper [2011], Thomas [1983])

Furthermore, Enomoto and Marriott [1994] mentioned that negative pragmatic transfer in low proficiency learners is more acceptable to the listener, while in high proficiency learners with high grammatical ability, vocabulary, and fluency, is less acceptable and more critical. The listener seems to think that the learner's negative pragmatic transfer is attributed to a personality issue, not a language use issue.

Meng [2008] investigated the language use of refusal in situations to be requested for native Japanese speakers native and native Chinese speakers. An example of situation settings in her study shows in Table 2.1. She states *you* react differently depending on cultures to which *you* belong. According to the results of the study, native Japanese speakers apologized at a high rate regardless of their relationship with the person who requests them, while native Chinese speakers did not apologize as much to people with whom they have a close relationship.

Based on the above studies, there is a risk that the listener could misunderstand the

doi:DOI Number



speaker and consider him/her as having personality problems or lacking social etiquette when the pragmatic transfer like above occurs in cross-cultural communication. In addition to it, the higher the speaker's ability to use the target language, the greater the risk that the speaker will be misunderstood by the listener and be perceived as having a bad character or lacking social etiquette.

In this study, in order to reduce this risk which we also would face when people from different cultures talk to each other through machine translators, we experiment whether neural network models are aware of the culture of the target language in certain contexts.

2.2 Related Tasks and Data in Natural Language Processing

Natural language processing is being applied more and more in society and is rapidly becoming more and more important in recent years. Researches in natural language processing have been conducted on a variety of tasks. Recently, many tasks have been done based on deep learning models such as BERT (Devlin et al. [2018]) and T5 (Raffel et al. [2020]). In this section, we introduce some of the tasks that are relevant to our study.

Text Classification: Automatic text classification has always been an important application and research topic. It builds a classifier by learning the characteristics of categories from the train data and then automatically classify unseen data into each category with the classifier. Typically, the data which used in research are datasets with the texts and the categories labeled on each text. There are various datasets with different categories such as news (20n, Lang [1995], Reu [2017]), topics (Lehmann et al. [2015]), reviews (Pang et al. [2002], yel) and so on.

doi:DOI Number

Machine Translation: Machine translation technology has been a very active area of research since the application of neural networks. A machine translation task is the task of automatically translating one language into another language. Various methods have been proposed in the past with the aim of improving the quality of the generated translation. Commonly, in training a model for machine translation, parallel corpora are generally used. There are various domains of the parallel corpora such as news (new ,Tiedemann [2012]), religion-related articles (Wang et al. [2020], Tiedemann [2012]), subtitles (Lison and Tiedemann [2016]), lectures (Abdelali et al. [2014], Tiedemann [2012], Reimers and Gurevych [2020]), conversation in business scene (cha, Rikters et al. [2019])and so on.

Style Transfer: Text style has recently attracted the attention of not only linguists but also computer science optimists. A text style transfer task is a task that aims to generate a sequence of text with modified style characteristics while preserving the content. To train a supervised learning model, parallel-style data are used. Jhamtani et al. [2017] and Carlson et al. [2018] made the transition between classic and modern styles with multiple styles of parallel data of Shakespeare's works and bibles respectively. Hwang et al. [2015] and Pryzant et al. [2020] created the parallel data from Wikipedia for text simplification and neutralizing respectively. Rao and Tetreault [2018] constructed a corpus including the pairs of informal and formal sentences. To train an unsupervised model, not only the parallel data but also the data with category labels as we introduced above in *Text Classification* can be used for transferring styles.

However, there is not a culture-aware dataset. And there are still no classification tasks, machine translation tasks, or style transfer tasks that focus on cultural differences, i.e., the naturalness or unnaturalness of certain language use in a given culture. In this study, we created a culture-aware dataset with the aim of reducing the possible negative

doi:DOI Number

effects of cultural differences.

doi:DOI Number

3 Dataset

In this chapter, we introduce the corpora we use, annotations on the corpora, the creation of parallel texts, and findings in statistical analysis on our datasets.

3.1 Base Corpora

In this study, we attempt to verify that a sequence-to-sequence translation model is aware of language use in Chinese culture and Japanese culture in dialogue. In order to do it, we need an appropriate dataset which is human-human interaction dialogue in both Chinese and Japanese. Therefore, we adopt two corpora in Chinese and Japanese, respectively.

- A Multi-Party Dialogue Dataset for Analysis of Emotions and Interpersonal Relationships (MPDD) (Chen et al. [2020]).
- Corpus of Everyday Japanese Conversation monitor public version (CEJC) (Koiso et al. [2020], 小磯 et al. [2019]).

MPDD is a corpus of human-human multi-party interaction dialogue from a TV script in Chinese. This corpus consists of 25,548 utterances from 4,142 dialogues and various labels such as interpersonal relation label, interpersonal seniority label, emotion label, and so on.

doi:DOI Number

CEJC is a corpus of human-human multi-party interaction daily conversation in various situations in Japanese. The corpus contains video/audio data, short unit information, transcribed texts, and meta-information about speakers and conversations. The transcribed texts in the corpus are transcribed from participants' voices. The texts consist of 126 dialogues for a total of 50 hours. The meta-information about speakers and conversation includes the list of the interpersonal relationship between an informant and other participants in each conversation.

In this study, we use the common part of the two corpora: the texts and the relational information. However, The text formats of CEJC and MPDD are different. We preprocessed the transcribed text of CEJC to be in the same format as MPDD. The transcribed text originally had various tags inserted, such as those related to pronunciation changes and stagnation, which were inconvenient for the model's train, so we removed them. In addition, each transcribed text is one sentence and stored, even when a speaker speaks multiple sentences consecutively at one time. For the convenience of later annotation, translation, and experimentation, when the same speaker's sentences are consecutive, the sentences were concatenated as a single utterance. Through this preprocessing, CEJC's transcribed texts are composed of 90,960 utterances.

3.2 Annotation

As can be seen from the related studies mentioned above, the pragmatical rules of language use in a given culture are context-dependent. Therefore, the data used in this study also needs to have contextual information. In this study, we annotate the labels of relation and situation and use them as contextual information.

doi:DOI Number

3.2.1 Relation label

MPDD is originally annotated with interpersonal relation labels on each utterance (Chen et al. [2020]). For the convenience of processing and comparison, the same type of interpersonal relation labels is used for MPDD and CEJC. Since CEJC originally has a list of relationships between informants and other dialogue participants, we assign the same relation labels as MPDD based on the list.

The number of relational label types is 25 types such as parent, spouse, child, teacher, classmate, student, and so on. The label represents the relationship of the speaker to the listener. When we assign the relation labels on each utterance in CEJC, the listener is the person who speaks next to the speaker of the utterance on which we are going to assign a label.

In Table 3.1, the ratio of each relation type in the total number of utterances is shown. In MPDD and CEJC, the relationship type with the highest rate is *friend* at 0.254 and 0.423, respectively, and the second highest one is *spouse* at 0.097 and 0.104, respectively.

3.2.2 Situation label

Many previous studies have focused on the use of words in situations such as apology, request, thanks, refusal, praise and, so on. In this study, we focus on the three situations of apology, request, and thanks, because these situations are relatively often seen in both corpora. We annotated these situation labels on each utterance when the utterance contains a meaning that corresponds to these situations.

In table 3.2, we show the number of utterances with each situation label. In MPDD

doi:DOI Number

	Rate	
	MPDD	CEJC
parent	0.074	0.082
parent-in-law	0.006	0.019
grandparent	0.004	0.005
other superior	0.011	0.002
spouse	0.097	0.104
brothers and sisters	0.058	0.039
other peer	0.023	0.016
child	0.073	0.084
son/daughter-in-law	0.006	0.017
grandchild	0.004	0.005
other inferior	0.011	0.002
teacher	0.003	0.018
classmate	0.008	0.016
student	0.003	0.018
boss	0.058	0.002
colleague	0.071	0.046
partner	0.012	0.039
subordinate	0.055	0.002
couple	0.065	0.006
friend	0.254	0.423
enemy	0.031	0.000
consignor	0.021	0.021
consignee	0.021	0.021
stranger	0.032	0.000
unknown	0.001	0.003

Table 3.1: The rate of each relation type

doi:DOI Number

Corpus	Situation	#
MPDD	Apology	134
	Request	434
	Thanks	220
CEJC	Apology	419
	Request	87
	Thanks	354

Table 3.2: # of utterances with each situation label

and CEJC, the most common situation is request and apology, respectively. The most uncommon situation is apology and request, respectively.

3.3 Creation of Parallel Texts

To make it easier to verify whether models are aware of the culture through comparison, we create two types of parallel translated texts. One is machine-translated text. It is used as NOT culture-aware translated text. The other one is human-translated text which is used as culture-aware text.

3.3.1 Machine-translated Text

We translate the texts into the target language by machine translators: the texts in MPDD are translated from Chinese to Japanese and the texts in CEJC are translated from Japanese to Chinese. Since both corpora are human interaction dialogue and the style of texts is based on spoken language, we use DeepL of free version[1]. DeepL was evaluated by professional translators without mentioning the names of the translators, DeepL consistently received the highest ratings (dee). Also, DeepL is strong in colloquial language and certain dialects (led [2020]). Therefore, it is suitable for translation of the texts of the corpora

[1] https://www.deepl.com/translator

doi:DOI Number

which we use.

3.3.2 Human-translated Text

We translate the texts into the target language by humans to create culture-aware parallel translated texts. In order to avoid communication breakdown, misunderstandings, or discomfort when interacting with someone who is from different culture, appropriate language use in the relevant context of the listener's culture is necessary. Therefore, When creating parallel human-translated texts, we must take care that the translated texts follow language use that fits the target language culture. We created parallel translated texts by paying close attention to the selection of translators and instructions to translators to ensure that the translated texts are culture-aware.

3.3.2.1 Translator

In order to create culture-aware parallel translated text, the translation must be done by people who are familiar with the culture of the target language because the translator must be aware of the naturalness of language use in the conversational culture of the mother tongue and must ensure that the translated texts are culture-aware.

The following two people translated the utterances of each corpus to their native language: one native Chinese speaker who grew up in Taiwan and who can use Japanese fluently, and one native standard Japanese speaker who grew up in Japan and who can use Chinese fluently. They both have a background in linguistics, extensive international exchange experience, and are sensitive to the discomfort of language use in native languages.

doi:DOI Number

3.3.2.2 Texts to be Human-Translated

Previous researches have shown that there are cultural differences in the utterances and the responses to each situation (張穎 [2004], Yi-bo [2015], Shi [2020]). In other words, we can see the characteristics of language use in a given culture. We had translators translate the utterances with the situation labels (Query) and the response to each of these (Response). In this study, a response is the next utterance after the query.

3.3.2.3 Information Given to the Translators

In order to do culture-aware translation, the translator needs to know the contexts including previous utterances, the people participating in the dialogue, and their relationship to each other as context. Therefore, the information provided to the translator when performing these translations is as follows: Speaker ID and interpersonal relation labels of each utterance, and both pre-translated text and machine-translated texts that contain queries, responses, and their corresponding utterances up to previous 30 turns as the context.

The translators need to take care that the translated text is culture-aware, not a verbatim translation. To encourage translators to keep this awareness in mind, we prepared two tasks. The first is a task that the translators judge whether the text to be translated is *natural*, *not-so-natural* or *unnatural* based on language use in the translator's native dialog culture. *Natural* means that the object utterance is one of the utterances which most of the people who grow up in your mother tongue culture could say in the given context. The second is a task that the translators translate the texts into translators' mother tongue to fit their mother tongue culture for each *not-so-natural* and *unnatural* utterance. On this task, the following three notes of translation are given to the translators: (1) to make each

doi:DOI Number

utterance natural in the context, (2) to keep the consistency of the meanings between the pre-translated utterance and the translated utterance, and as far as possible, (3) to prioritize pragmatically natural over the consistency of meanings between the pre-translated utterance and the translated utterance.

With the above translation, we obtained human-translated queries and responses that take into account the cultural differences. The utterances judged as *not-so-natural* and *unnatural* are translated by the translators. When an utterance is judged as *natural*, the machine-translated utterance is used as the human-translated utterance.

3.4 Findings on Our Dataset

The statistical method led to a discovery in our data set. There are two main findings: the difference between machine and human-translated utterances and the tendency of the cultural difference of language use.

3.4.1 Difference Between Machine and Human-translated Utterances

To see how many utterances are judged as having inappropriate language use, we show the number of each naturalness type judged by the translator for each machine-translated utterance with situation labels in Table 3.3. The majority of the machine-translated utterances in both the MPDD and CEJC corpora are judged to be *unnatural* or *not-so-natural*, and only a relatively small number are judged to be *natural*. Especially, in MPDD, most all utterances are judged as *unnatural* Among them, a relatively large number of utterances in the thanks situation in the CEJC are judged as *natural*. In other words, many of the machine-translated utterances are inappropriate language use in the relevant context,

doi:DOI Number

Corous	Situation	Utterance type	Natural	Not-so-natural	Unnatural	Total
MPDD	Apology	query	4	0	130	134
		response	1	0	133	134
	Request	query	0	0	436	434
		response	2	0	434	434
	Thanks	query	10	0	210	220
		response	6	0	214	220
CEJC	Apology	query	29	52	338	419
		response	30	64	325	419
	Request	query	0	5	82	87
		response	6	5	76	87
	Thanks	query	91	54	209	354
		response	26	45	283	354

Table 3.3: Number of naturalness labels on each situational utterance translated by machine.

especially in MPDD. Among them, in CEJC, when machine translation from Japanese to Chinese, relatively many of the language uses in the thanks situation were recognized as appropriate.

So, how could inappropriate language use in machine translation be changed to appropriate language use by human translation? A type of change between machine-translated utterances and human-translated utterances can be seen in the length of the utterances. Table 3.4 shows the mean length of the machine and human-translated utterances in each corpus, situation, and utterance type. In MPDD, the mean length of human-translated utterances is shorter in all situations and utterance types. The largest change in mean length is query in the request situation, which is about 15.5 points. On the other hand, in CEJC, the change in mean length between machine translation and human translation is small in all situations and all utterance types. In the MPDD, when translated from Chinese into Japanese by humans, many expressions were recognized as inappropriate and were removed.

To determine how much each machine-translated utterance and human-translated ut-

doi:DOI Number

Corpus	Situation	Utterance type	Mean of length	
			Machine Translation	Human Translation
MPDD	Apology	query	48.552	31.687
		response	39.381	28.993
	Request	query	54.166	38.666
		response	46.290	30.970
	Thanks	query	47.923	33.795
		response	31.859	21.336
CEJC	Apology	query	18.516	17.396
		response	10.489	11.000
	Request	query	27.241	29.264
		response	12.954	11.598
	Thanks	query	13.025	11.220
		response	9.644	10.630

Table 3.4: Comparison of mean length between machine and human-translated utterances

terance changed, we calculated the mean of the longest common subsequence (LCS) (Table 3.5). The higher the LSC, the more identical subsequences are contained in the two utterances. In other words, the higher the LCS, the more similar the two utterances are; similarly, the lower the LCS, the more different they are. Overall, the mean LCS of CEJC is lower and its standard deviation of CEJC is higher than that of MPDD. Also, in both MPDD and CEJC, the mean LCS of the query in the situation of Thanks is extremely high compared to the others. Therefore, in CEJC, Japanese to Chinese translation, has more subsequence changed between machine translation and human translation than in MPDD, Chinese to Japanese translation. It can also be seen that CEJC has a larger variation in the degree of changes between the machine and human-translated utterances depending on the utterance. This is consistent with the fact that the number of *natural* words is large, as can be seen from the Table 3.3. Also, commonly in MPDD and CEJC, the queries in Thanks situation are the smaller difference between machine and human translation.

To summarize the above, the following five points are recognized as differences between machine and human-translated utterances:

doi:DOI Number

Corpus	Situation	Utterance type	mean	std
MPDD	Apology	query	0.402	0.192
		response	0.373	0.175
	Request	query	0.394	0.163
		response	0.345	0.179
	Thanks	query	0.535	0.222
		response	0.373	0.227
CEJC	Apology	query	0.326	0.240
		response	0.333	0.263
	Request	query	0.302	0.132
		response	0.323	0.262
	Thanks	query	0.555	0.307
		response	0.333	0.274

Table 3.5: The ratio of the longest common subsequence between MT and HT

1. In common, most of the machine-translated utterances have inappropriate language use.

2. In the translation from Chinese to Japanese (MPDD), inappropriate language use has been removed by human translation.

3. More machine-translated utterances from Japanese to Chinese (CEJC) were judged that the language uses are appropriate than those from Chinese to Japanese (MPDD).

4. Considering 2. and 3. above, many of the utterances judged as inappropriate language use in the Japanese to Chinese machine translation (CEJC) were rephrased to a greater extent by human translation than in the Japanese to Chinese translation (MPDD).

5. In common, the *thanks* situation has the least inappropriate language use compared to the other situations.

doi:DOI Number

3.4.2 Tendency of Cultural Difference of Language Use

To find out what those cultural differences are, we do word alignment on each pre- and post-translation pair of utterances, extract the meaningful words that are unaligned and perform analysis on the unaligned words. From the above, it is clear that there are differences in language use in each dialogue culture between machine-translated utterances which are not culture-aware, and human-translated utterances which are culture-aware. As can be seen from the 1.1, the machine translation produces an almost verbatim translation of the pre-translated texts. That's why there would be the almost same difference in language use between pre-translated utterances and human-translated utterances as between machine-translated utterances and human-translated utterances. Then, the unaligned word in the pair of pre-translated utterances and human-translated utterances are objects that are worth analyzing for their differences.

3.4.2.1 Method

For analysis of the unaligned words, we create a label set based on the pragmatical analysis frameworks of Blum-Kulka and Olshtain [1984] and Cheng [2005]. The label set consists of 7 types of labels (Table 3.6). What the label represents is how the unaligned words of the pair of pre-translated and human-translated text have changed in language use for the human-translated text. We label them on each set of unaligned words, pre-translated and human-translated utterances which include the unaligned word. The labeling procedure is as follows: (1) label one in miss or segment type label on each pair, (2) label one in different language use (DLU) type label on each pair when any segment type label on the pair. (3) label more than a label in directness type label, intensity type label and perspective

doi:DOI Number

Type	Label
Miss	situation label miss, giza miss
Segment	headact, adjunct, other, response
Different language use(DLU)	DLU, no DLU, no apology, no request, no thanks
Directness	more direct, less direct
Intensity	more intense, less intense
Perspective	speaker orientated, listener orientated, speaker&listener orientated, impersonal orientated
Intense strategy	more upgrader, less upgrader, more downgrader, less downgrader, more specific, less specific, more respectful, less respectful, more humble, less humble, add expect something in return, remove expect something in return, add irony or rhetorical question, remove irony or rhetorical question

Table 3.6: Label types and labels for unaligned word analysis

type label on each pair when DLU type label on the pair, (4) label more than a label of intense strategy type label below when intensity type label on the pair.

3.4.2.2 Data Procedure

Unaligned word: Pre-translated, machine-translated and human-translated texts are tokenized by Jieba ([2]) for Chinese texts and Juman++ (Morita et al. [2015]) for Japanese text. We conduct word alignment for each pair of pre- and post-translated utterances by GIZA++ (Och and Ney [2003]). The outputs from GIZA++ are symmetrized by a so-called grow-diag-final-and (GDFA) procedure and aligned token pairs were obtained. Unaligned tokens are extracted by finding out the tokens with are not paired in the above procedure. There are 2 types of unaligned tokens. One is the deleted unaligned tokens which are in the pre-translated text but do not have tokens to be paired in corresponding post-translated text. The other one is the added unaligned tokens which are in the post-translated text but

[2]https://github.com/fxsjy/jieba

doi:DOI Number

Pre-translated text	ごめん。ぐちゃぐちゃになっちゃった
Index	0 1 2 3 4
Post-translated text	抱歉，我现在脑子有点混乱。
Index	0 1 2 3 4 5 6 7
Aligned tokens pair	0-0, 2-3, 2-4, 2-6, 3-5, 3-6, 4-6
Deleted token index	1
Added tokens index	1, 2, 7

Table 3.7: Demonstration of deleted token and added token

do not have tokens to be paired in corresponding pre-translated text. Table 3.7 demonstrates what are deleted token and added token. And then, We filter out punctuation marks and other symbols in both language and postpositional particles and undefined words in Japanese from unaligned tokens and get unaligned words. Japanese unaligned words are converted to their primitive forms.

Meaningful unaligned word: In order to extract meaningful words from unaligned words, we use Chinese linguistic inquiry and word count dictionary (CLIWC) (Huang et al. [2012], 林 et al. [2020]) and Japanese linguistic inquiry and word count dictionary (JIWC) [3] for Chinese and Japanese texts, respectively. Both CLIWC and JIWC are LIWC derived for each language. LIWC is a dictionary for text analysis using counts of psychologically meaningful words belonging to certain categories. We count the meaningful words in all unaligned words using CLIWC and JIWC, and then significance test on the differences of the meaningful unaligned words counts on categories by a sentence between the unaligned words extracted from the pair of pre-translated utterances and machine-translated utterances and the pair of pre-translated and human-translated utterances in order to ensure that the difference of the counts of the meaningful unaligned words on the category just coincidence, but is regular. We used Wilcoxon signed-rank test for each distribution for the count of unaligned words in each emotion category of CLIWC and LIWC.

[3] https://github.com/sociocom/JIWC-Dictionary

doi:DOI Number

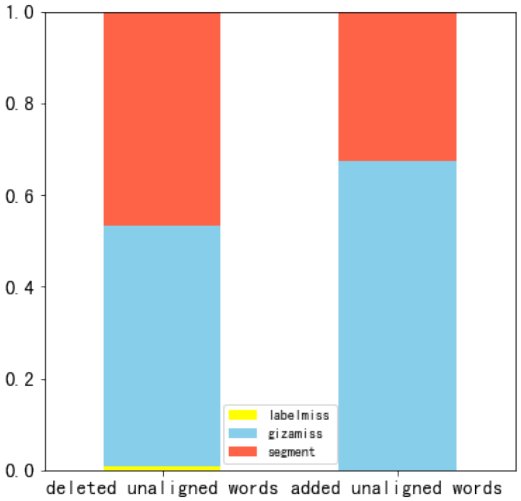


Figure 3.1: The percentage of miss and segments labels on Chinese meaningful unaligned words

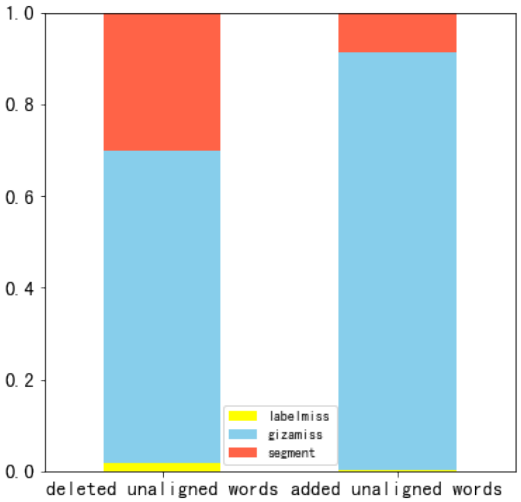


Figure 3.2: The percentage of miss and segments labels on Japanese meaningful unaligned words

Among the meaningful unaligned words obtained by the above procedure, the meaningful unaligned words that are included in the human-translated utterance but not in the corresponding machine-translated utterance are target to be labeled, since the meaningful unaligned words represent the components of language use which translators added or deleted when they do culture-aware translations.

doi:DOI Number

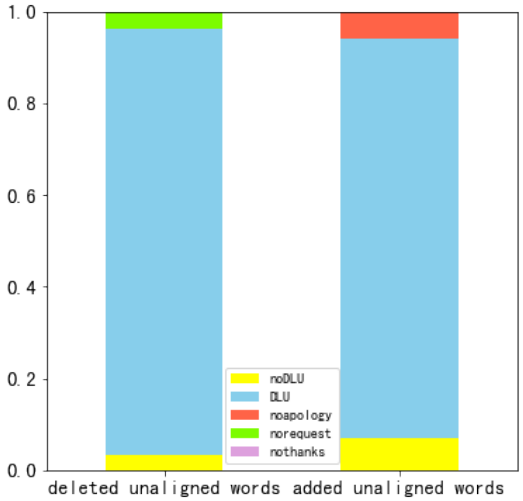


Figure 3.3: The percentage of DLU type label on Chinese meaningful unaligned words

3.4.2.3 Analysis for the labeled data

Figure 3.1 and 3.2 show the mean percentages of two miss labels and segment labels given to the target set of all meaningful unaligned words in all emotion categories, pre-translated utterances, and human-translated utterances. All have in common that there are more than a majority of miss type labels including *giza miss* and *label miss* labels and that there are more miss type labels above the added unaligned words than the deleted one. In particular, the added unaligned words are words that were added by people during culture-aware translation, and could not be aligned properly by GIZA++, although there are actually words to be aligned.

Figure 3.3 and 3.4 shows the mean percentages of all DLU type labels in all emotion category. Common to all is that DLU accounts for the highest percentage of the total. Characteristically, the ratio of *no apology* and *no request* is high for deleted unlabeled words in Japanese. In other words, the deleted and added unaligned words change the language use and constitute a translation utterance that takes cultural differences into account. In addition, about the high percentage of *no apology* and *no request* in the deleted

doi:DOI Number

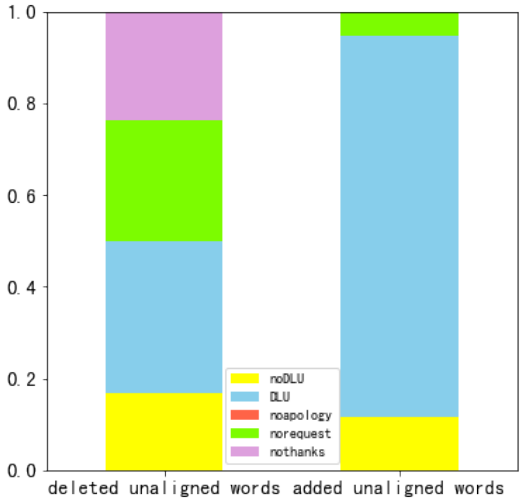


Figure 3.4: The percentage of DLU type label on Japanese meaningful unaligned words

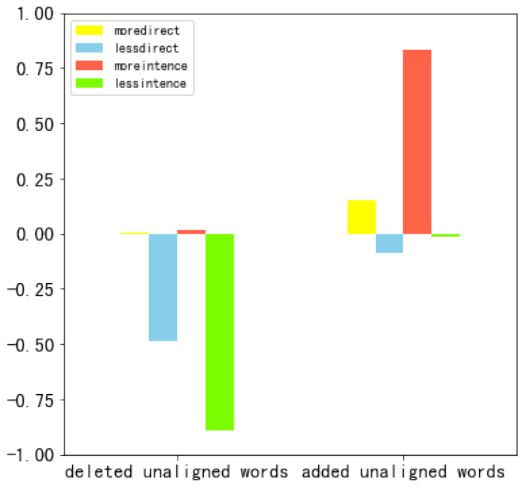


Figure 3.5: The percentage of directness and intensity type labels in DLU typelabel on Chinese meaningful unaligned words

unaligned words in Japanese, the words removed by the translator in Japanese to Chinese indicates that the unaligned words which represent each situation are inappropriate for Chinese dialogue culture in the contexts where the corresponding utterances were uttered.

Figure 3.5 and 3.6 shows that the mean percentages of directness type label and intensity type label in all DLU type labels in all emotion category. For readability of the figures, the value of *less direct* and *less intense* labels has been multiplied by -1. The percentages of *less direct* and *less intense* label of deleted unaligned words in Figure 3.5 and of added unaligned words in Figure 3.6 is higher. On the other hand, the percentages of

doi:DOI Number

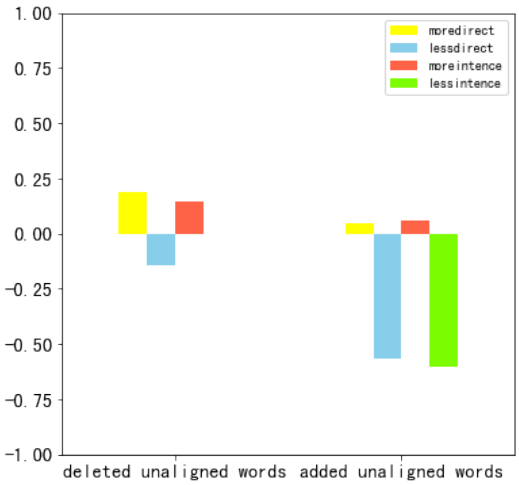


Figure 3.6: The percentage of directness and intensity type labels in DLU type label on Japanese meaningful unaligned words

more direct and more intense label of deleted unaligned words in Figure 3.6 and of added unaligned words in Figure 3.5 is higher. The former represents how meaningful unaligned words in Chinese to Japanese translation change the result of human translation, while the latter represents it in Japanese to Chinese translation. Therefore, we can mention that generally in apology, request, and thanks situations, when translating Japanese into Chinese, to use more direct or more intense expressions would be to adapting the expression to the Chinese dialogue culture, and vise versa.

doi:DOI Number

4 Situation Classification

As mentioned above, awareness of context is important for conducting dialogue with appropriate language use in a given culture. In order to do so, it is necessary to be able to correctly recognize situations as one of the context-important components in the relevant dialogue culture. In this chapter, we attempt to ensure that a sequence-to-sequence model can recognize relevant dialogue cultures by classifying situation categories.

4.1 Method

To ensure that the sequence-to-sequence model recognizes appropriate language uses in the relevant dialogue cultures, we classify situations using our dataset and compare each classification score.

4.1.1 Model

The model which we use is Multilingual-T5 (mT5) (Xue et al. [2021]) which is a variance of T5 (Raffel et al. [2020]), supports 101 languages (including Chinese and Japanese) and the pre-trained models are published [1]. The pre-trained model is mt5-base. The high performance of mT5 is supported by a very large dataset called ”mC4”, which is collected

[1] https://github.com/google-research/multilingual-t5

doi:DOI Number

by web scraping and is cleaned by removing duplicates, incomplete sentences, extreme content, and noise. In addition, mT5 is a model that can perform many different kinds of tasks simultaneously. When a specific task is to be trained, it is specified using a prefix text that tells the model what to do with the input Considering this prefix as a condition, the model generates a sequence as an output. In other words, this prefix can be interpreted to act as a condition for some kind of conditional generation.

Since mT5 is capable of powerful transfer learning, supports Chinese and Japanese, and considers the prefix as a kind of condition, it was selected as the model for our dataset which is comparatively. For the implementation, we use Simple Transformers [2].

4.1.2 Data

The texts data in our dataset is composed of six parts: texts in two corpora, machine-translated and human-translated texts from MPDD and CEJC. These six parts can be divided into two categories by language. Chinese texts are pre-translated texts in MDPP and two types of translated texts in CEJC by machine and human. Japanese texts are the rests. In addition, as we mentioned in Section 3.2, the situation labels and relation labels are annotated on each instance of text data.

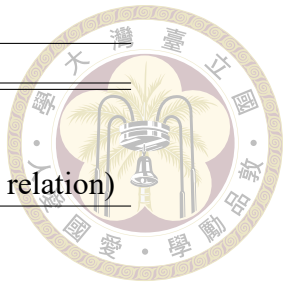
For this situation classification, the queries, responses, situation labels and relation labels in our dataset are used. The queries and responses with each situation label are used as positive labeled data for the relevant situation category, and the others are used as negative labeled data. The relation labels are used as contextual information in the way to be given to the model as a prefix text which we explain in Section 4.1.3.

[2]https://github.com/ThilinaRajapakse/simpletransformers

doi:DOI Number

Type	Prefix format (variable)
None:	
Base:	(corpus name) (utterance type)
Base + Rel:	(corpus name) (utterance type) (interpersonal relation)

Table 4.1: Prefix format



4.1.3 Prefix Option

To ensure whether the context information improves the model's perception of the appropriate language uses in each culture, we prepare different sets of contexts for the input prefix text of mt5 (Table 4.1). Since mT5 considers the prefix as a kind of condition, it can be expected that the knowledge of the condition specified by the prefix would be used for prediction.

Corpus name in Table 4.1 could have *MPDD* or *CEJC*. The domains of the corpora would be adapted for the model by the contextual information passed as a prefix. The two corpora used in our dataset are both dialogues, but they differ in the content of the dialogues. For example, in MPDD, there is a lot of talk about the Chinese Communist Party, but in CEJC, there is little talk about politics. Also, *utterance type* in Table 4.1 could have *query* or *response*. The role of the utterance type would be adapted for the model by this prefix. Since a query is an utterance that corresponds to the relevant situation, and the response is the utterance that follows the query and is often a reply to the query, the role of the utterances of both query and response is different. In addition, *interpersonal relation* would have the name of relation label on the corresponding utterance. The relationship knowledge which the model learned would be used as one of the conditions to determine if the language use is appropriate for the category to be classified.

doi:DOI Number

4.1.4 Measurement

For the measurement of the classifications, we use the mean of the macro F1-scores of the binary classifications in each situation by language. For the classification results, the model predicts 0 (negative) or 1 (positive) for each relevant situation category. The macro F1-scores are calculated for the results of each situation category, and then the mean of them are calculated by language where each situation category s_j is a member of the set S and N denotes the number of instances that were predicted to be 0 or 1.

$$\text{Macro F1-score}^{s_j} = \frac{1}{N} \sum_{i \in 0,1}^{N} \text{F1-score}_i^{s_j} \tag{4.1}$$

$$\text{Score}^{s_j} = \frac{1}{n(S)} \sum_{j \in \text{S}}^{n(S)} \text{Macro F1-score}^{s_j} \tag{4.2}$$

4.1.5 Comparison of Scores

The pre-translated texts are the data that native speakers had conversations. The human-translated texts are the culture-aware translated data that the conversations which different language speakers had. The machine-translated texts are the NOT culture-aware translated data that different language speakers had conversations. Each text in a language has a different degree of culturally appropriate language use and in descending order of the degree: pre-translated texts, human-translated texts, machine-translated texts. Therefore, We can build a hypothesis that the descending order of the scores is the score for pre-translated texts, human-translated texts, and machine-translated texts. In Figure 4.1, we illustrate the marks of the order of the scores in our hypothesis and the data used as train

doi:DOI Number

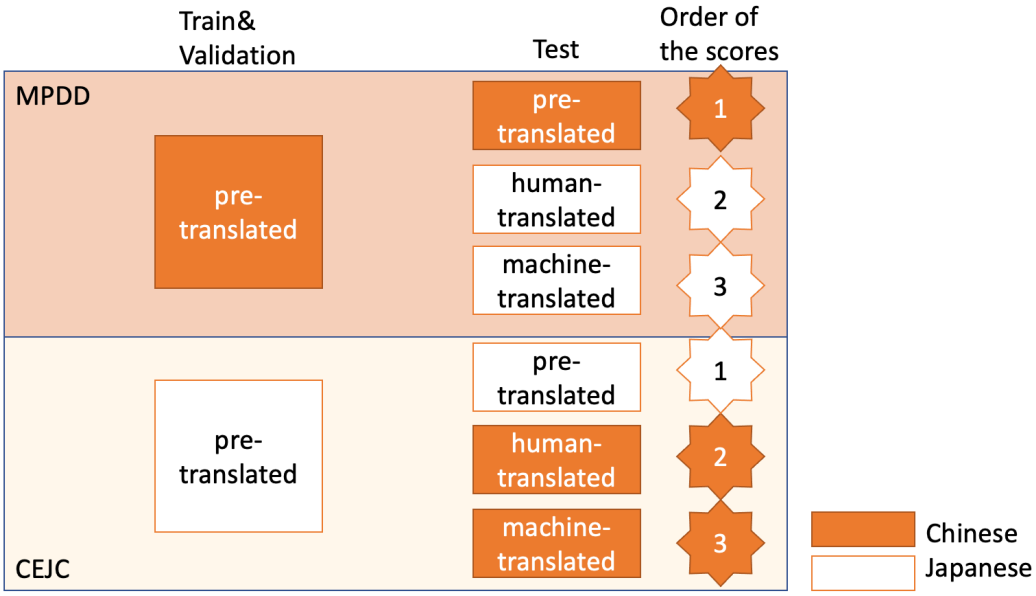


Figure 4.1: Data used in our experiment and the order of the scores in hypothesis and test data, in Chinese and Japanese respectively.

4.2 Experiment Setup

In this experiment, the parameters setting is as following: optimizer, learning rate, max sequence length, and rate of dropout are AdamW, 1e-5, 128, and 0.1, respectively. We set early stopping at the point when the validation loss did not drop more than 3 epochs. The best model used in the test was the one that had the lowest validation loss in every epoch of training.

For training, we use pre-translated data. The queries and responses with each situation label in pre-translated data are used as positive labeled data for the relevant situation category and the others are used as negative data. Both positive and negative data are divided into 8:1:1 as train, validation, and test set of data. Also, we sample randomly from utterances other than queries and responses, and use as negative label data. The number of samples is the same as the number of instances of queries and responses. For

doi:DOI Number

the test set, we prepare three types of test sets which are pre-translated, machine-, and human-translated data. The test sets of the machine and human-translated data contain the corresponding instances of the test set of the pre-translated data.

For an input text of every instance, we give a set of a prefix and an utterance (a query or a response) to the model. The model classifies the utterance into each relevant category or not. We train models for each prefix type. And then, we test all models of each prefix type for the three types of test sets which are pre-translated, machine-, and human-translated data.

4.3 Results

Table 4.2 shows that the scores of the situation classification by each language, each prefix set type, and each test set of data. The scores are the overall scores which are the mean of the scores for each situation. Our hypothesis is that the descending order of the scores is the score of pre-translated, machine-translated, and human-translated texts when the model is culture-aware.

In Chinese, the result does not support our hypothesis. For all prefix types in Chinese, the descending order of the scores for each test data type is pre-translated, machine-translated, and human-translated. This result infers that the language use for each situation in machine-translated texts is similar to that in pre-translated text than to that of human-translated text. However, For most of the scores in Chinese, there is an upward trend as more prefix information is added, which means that the contextual information contributes to the classification of situations.

In Japanese, our hypothesis was partially supported. For *Base* and *Base+Rel* in

doi:DOI Number

Language	Prefix type	Test data type	f1_macro
Chinese	None	pre-translated	0.510
		human-translated	0.343
		machine-translated	0.425
	Base	pre-translated	0.477
		human-translated	0.403
		machine-translated	0.404
	Base+Rel	pre-translated	0.481
		human-translated	0.419
		machine-translated	0.430
Japanse	None	pre-translated	0.455
		human-translated	0.396
		machine-translated	0.434
	Base	pre-translated	0.436
		human-translated	0.504
		machine-translated	0.490
	Base+Rel	pre-translated	0.433
		human-translated	0.466
		machine-translated	0.462

Table 4.2: The scores of the situation classification by each setting

Japanese, the scores of human-translated texts are higher than that of machine-translated texts. It means that the language uses in human-translated texts with the contexts are similar to the language uses in pre-translated texts which are used as train data than machine-translated data. However, for *Base* and *Base+Rel*, the scores of pre-translated are the lowest among all test data types, although for *None* it is the highest score than other test data types.

As for the improvement of situation classification performance with the increase of contextual information, we can see an improvement in the classification performance of scores for Chinese human-translated data, but this is not inevitable considering the fact that there is no upward trend in scores for pre-translated data.

doi:DOI Number

4.4 Discussion

The above results show that the results of situation classification just partially support our hypothesis (Table 4.2). In this section, we discuss the causes of some parts of the result which we did not expect.

4.4.1 The Model in Confusion

Regarding the fact that the score of pre-translated texts in Japanese for *Base* and *Base+Rel* was lower than that of other test data types. It is weird that a test set in a different corpus has a higher score than a test set in the same corpus as the train set of data. The reason may be that the data was too complex because these scores in Table 4.2 were trained using both the data of queries and responses when the model was trained. Therefore, we classified only queries in the model trained with only queries and classified responses in the same way.

The results shows in Table 4.3 and Table 4.4, respectively. As can be seen in these results, our hypothesis is supported. The scores are pre-translated, human-translated, and machine-translated in order of decreasing. Therefore, we can mention that at least one of the reasons why the scores for pre-translated texts in Table 4.2 are lower is that the data mixed with queries and responses confused the model.

4.4.2 Difference between Culture in Mainland of China and Taiwan

The scores for human-translated texts in Chinese are lower than that for machine-translated texts, as shown in Table 4.2. There are two possible reasons.

doi:DOI Number

Language	Prefix type	Test data type	f1_macro
Japanese	Base	pre-translated	0.532
		human-translated	0.508
		machine-translated	0.493
	Base+Rel	pre-translated	0.541
		human-translated	0.465
		machine-translated	0.447

Table 4.3: The scores of the situation classification on queries in Japanese

Language	Prefix type	Test data type	f1_macro
Japanese	Base	pre-translated	0.455
		human-translated	0.337
		machine-translated	0.340
	Base+Rel	pre-translated	0.537
		human-translated	0.297
		machine-translated	0.292

Table 4.4: The scores of the situation classification on responses in Japanese

One is that adding the relation label as part of the prefix to the input is confusing the model. This reasoning is consistent with a view of Chen et al. [2013] that Chinese have less clear rules of pragmatic for using different language uses depending on the interpersonal relation, while Japanese is restricted by more stringent rules of pragmatic for using language uses with appropriate degrees of politeness depending on the interpersonal relationships. Therefore, this result could be plausible.

Other one could be the difference between the language uses in the conversation culture of Mainland of China and Taiwan. The text of MPDD is from TV script and the language use of it is similar to the language use of Mainland of China, while the translator who translated CEJC from Japanese to Chinese have been grown up in Taiwan. Therefore, it is possible that there is a cultural difference in language use between pre- and post-translated texts.

doi:DOI Number

4.4.3 Utterance Length

Comparing to the scores for human-translated texts in Japanese, all of the scores for it in Chinese are much lower. One of the causes could be that the utterance lengths are too long. It is too confusing for the model to classify them into grand-truth situation categories because each utterance has too much information. It is possible that a single utterance can contain both positive and negative data features. It is also possible that each instance is too confusing to learn the features for classification. On the contrary, compared to the scores for human-translated texts in Chinese, all of the scores for it in Japanese are higher. The model was trained by CEJC corpus whose utterance lengths are comparatively short (Table 3.2).

doi:DOI Number

5 Culture-Aware Machine Translation

As mentioned above, awareness of contexts is important for having conversation with appropriate language use in a given culture. In order to do so, it is necessary to be able to generate appropriate translations for the language use of the target language's culture in a certain context. In this chapter, we attempt to ensure that contexts contribute to generating culture-aware translation.

5.1 Method

To demonstrate that our dataset is capable of culture-aware translation, we attempt to perform culture-aware machine translation using a sequence-to-sequence model. Also, to ensure that contexts are helpful to generate culture-aware translation, we change the amount of contextual information given to the model and compare the results.

5.1.1 Model

For this experiment, the model which we use is Multilingual-T5 (mT5) (Xue et al. [2021]) and the pre-trained model is mt5-base. They are the same as the ones used in Chapter 4, so please refer to Section 4.1.1 for the model.

doi:DOI Number

5.1.2 Data

For this machine translation, the queries, responses, situation labels, and relation labels in our dataset are used. Also, we only use the pre-translated and human-translated texts. The pre-translated texts and labels are used as input data into the model. The human-translated data are used as grand truths. Also, the situation labels and relation labels are used as the contextual information to be given to the model as a prefix text which we explain in Section 5.1.3.

The pre-translated texts are the texts in each corpus originally and the human-translated texts are the culture-aware translation of the pre=translated texts (Section 3.3.2). Therefore, the model should learn the features of appropriate language uses of each culture for each relevant situation.

Furthermore, to ensure whether the contexts improve the model's perception of appropriate language uses of each culture in certain situations, we prepare different sets of contextual information. We use the previous utterances as an additional input sequence as contexts. The three different turns of previous utterances are prepared: 0, 1, and 2 turn(s).

5.1.3 Prefix Option

To ensure whether the contexts improve the model's perception of cultural differences in the appropriate language uses of each culture in certain situations, we prepare different sets of contextual information as input prefix of mt5 (Table 5.1). we train models using each set of contexts and compare them. Since mT5 considers the prefix as a kind of condition, it can be expected that the knowledge of the condition specified by the prefix

doi:DOI Number

Type	Prefix format (variable)
None:	
Base:	(corpus name) (situation type) (utterance type)
Base+Rel:	(corpus name) (situation type) (utterance type) (interpersonal relation)

Table 5.1: Prefix format

would be used for generation.

Corpus name could have *MPDD* or *CEJC*. The domains of the corpora would be adapted for the model by the contextual information passed as a prefix. The two corpora used in our dataset are both dialogues, but they differ in the content of the dialogues. For example, in MPDD, there is a lot of talk about the Chinese Communist Party, but in CEJC, there is little talk about politics. Also, *situation type* would have *apology*, *request* or *thanks*. As we mentioned in Section 2.1.1, each culture has its own particular language use tendencies in certain situations of dialogue. *situation type* in the prefix would serve as a condition for the generation of appropriate language use in the culture of the relevant language. In addition, *utterance type* could have *query* or *response*. The role of the utterance type would be adapted for the model by this prefix. Since a query is an utterance that corresponds to the relevant situation, and the response is the utterance that follows the query and is often a reply to the query, the role of the utterances of both query and response is different. Furthermore, *interpersonal relation* would have the name of relation label on the corresponding utterance. The relationship knowledge which the model learned would be used as one of the conditions to generate appropriate language use in the culture of the relevant language.

doi:DOI Number

5.1.4 Measurement

The metric we use is F1-score of BERTScore (Zhang et al. [2019]). BERTScore, which is similar to a human evaluation in the sense that it determines how close the meaning of a sentence is, makes an evaluation based on the similarity between the candidate sentence and the reference sentence using the vector representation obtained from the pre-trained BERT.

In the sense of measuring whether the generated text fits the dialogue culture in the target language, measuring the similarity between the generated text and the culture-aware grand truth is suitable since the meaning of the text could change when the pre-translated text is made culture-aware translations into the target language. We attempt to ensure that contexts contribute to generating culture-aware translations by measuring and comparing the similarity of the human-translated utterances, which are grand-truth, and translations generated by each model which was trained on different context settings.

5.2 Experiment Setup

In this experiment, the parameters setting is as following: optimizer, learning rate, max sequence length, length penalty, and rate of dropout are AdamW, 1e-5, 128, 20, and 0.1, respectively. We set early stopping at the point when the validation loss did not drop more than 3 epochs. The best model used in the test was the one that had the lowest validation loss in every epoch of training.

The texts we use are the pairs of utterances in pre-translated data and human-translated data. The pre-translated texts are used as input texts and human-translated data are used

doi:DOI Number

Input text
query: 好啦！现在请大家安静下来，我要接着说下去。context: 门外站的是我父母和外公。她们怕打扰你，影响你的工作！

Table 5.2: An example of input text

as grand-truth. The utterances include queries and responses in each situation. *n* turn(s) of previous utterance(s) are also used as contextual information which is given to the model as the additional input texts. An example of the format of input texts shows in Table 5.2. The target utterance to be machine-translated are put just after ”query: ” and the previous utterances are put just after ”context: ”. Furthermore, situation and relation labels on each utterance are used as contextual information which is given to the model as the input prefix options. The data for each situation are divided into 8:1:1 as train, validation, and test set of data, respectively.

For each contextual information setting including in each prefix type and each *n* turn(s) of previous utterance(s), we train models and test them.

5.3 Results

The results for each context-setting including contextual information in prefix types and previous utterances shows in Table 5.3 that target language is Chinese and 5.4 that target language is Japanese. The highest scores in each prefix setting are underlined, The highest scores in each table are used in bold font.

For Japanese to Chinese translation, the highest score is obtained when the *Base* prefix and previous utterances up to 1 turn before were given as context. For *Base+Rel* prefix, the highest score was obtained when previous utterances up to 2 turns ago were given as context, and for the *None* prefix when no previous utterance was given. What we can see

doi:DOI Number

from this is that there is only a 0.001 point difference between when the most and least contextual information is given. Figure 5.1 and 5.2 shows that the comparison of BERTScores for each contextual information setting. The average and maximum BERTScore for each prefix type, and the BERTScores for the previous utterance(s) of each turn in each prefix type are not right-ascending graphs. By providing contextual information to the model, the BERTScore could get higher or also get lower. In other words, for Japanese-Chinese translation, the increase in contextual information does not simply lead to culture-aware translation. This result does not support our hypothesis that contextual information contributes to the generation of translated sentences with appropriate language use in the target language. About this result, we analyzed it in the discussion section.

For Chinese to Japanese translation, as can be seen in Table 5.4, we verified that contextual information contribute to generating culture-aware translation. For both *Base* and *Base+Rel* prefix types, the highest scores are ones when 2 turns of precious utterances were given to the models. Also, the highest score among every contextual information settings is one when *Base+Rel* prefix type and 2 turns of precious utterances were given to the models, where the models are given the most contextual information. For convenience of understanding, we shows that that the comparison of BERTScores for each contextual information setting in Figure 5.3 and 5.4. The average and maximum BERTScore for each prefix type are right-ascending graphs (5.3). It can be seen that as contextual information increases from *None*, *Base*, to *Base+Rel*, the BERTScore also increases. Also, the BERTScores in both *Base* and *Base+Rel* prefix types for the previous utterance(s) of each turn are right-ascending graphs (5.4). In each prefix type setting, as previous utterances increase, the score also tends to increase.

doi:DOI Number

Prefix type	# of previous utterance turns	BERTScore
None	0	0.775
	1	0.771
	2	0.769
Base	0	0.764
	1	**0.780**
	2	0.774
Base+Rel	0	0.774
	1	0.767
	2	0.776

Table 5.3: BERTScores for each context setting in Japanese to Chinese translation

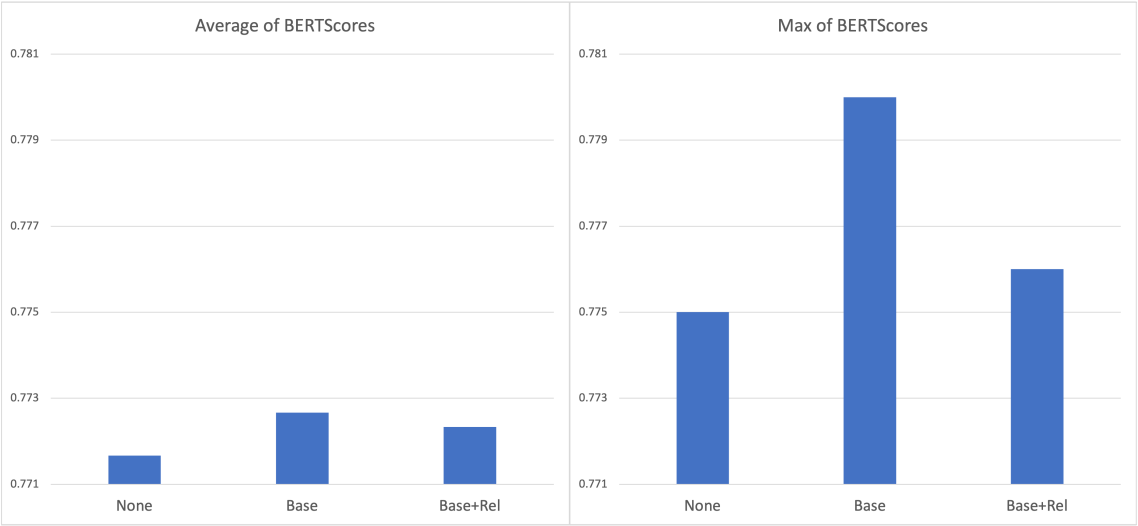


Figure 5.1: Comparison of BERTScores for each prefix type in Japanese to Chinese translation

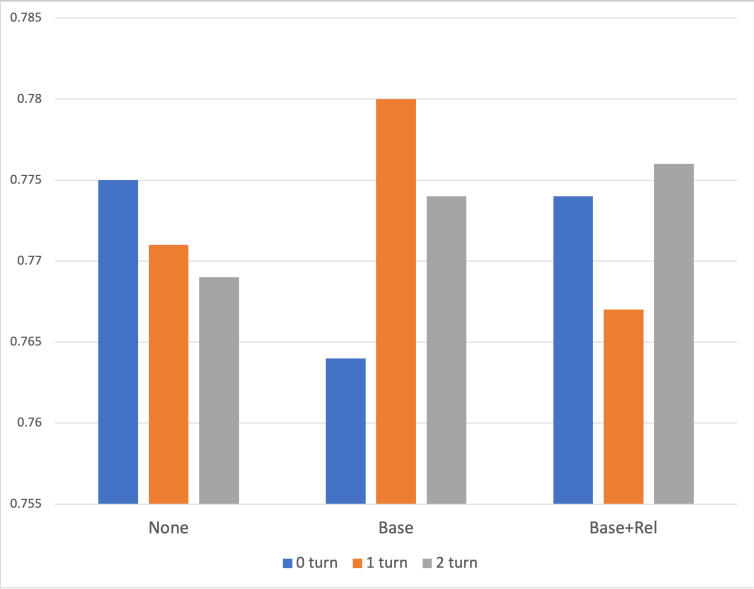


Figure 5.2: Comparison of BERTScores for each turn of previous utterances in each prefix type in Japanese to Chinese translation

doi:DOI Number

Prefix type	# of previous utterance turns	BERTScore
None	0	0.764
	1	0.769
	2	0.744
Base	0	0.760
	1	0.769
	2	0.770
Base+Rel	0	0.764
	1	0.764
	2	**0.773**

Table 5.4: BERTScores for each context setting in Chinese to Japanese translation

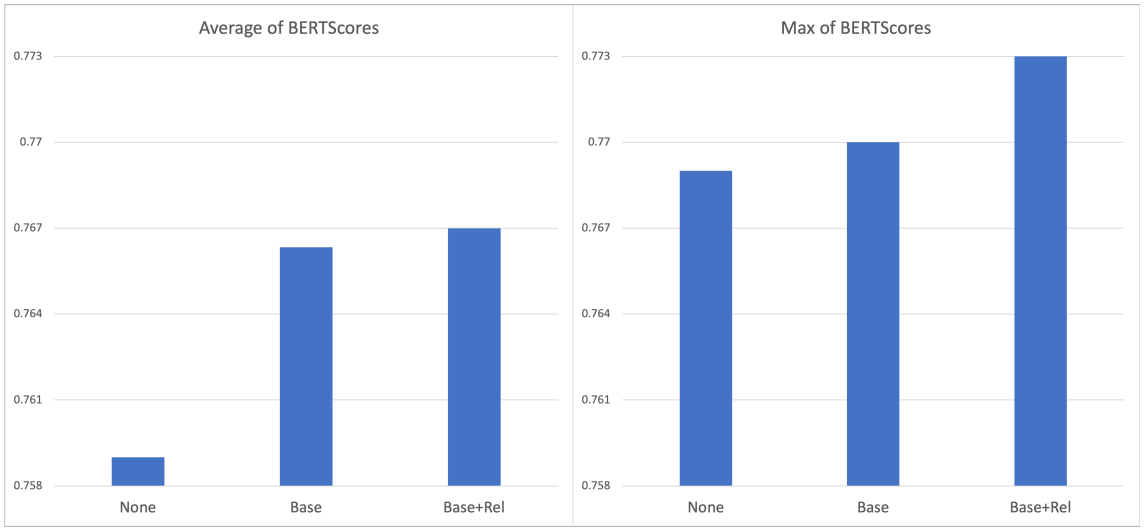


Figure 5.3: Comparison of BERTScores for each prefix type in Chinese to Japanese translation

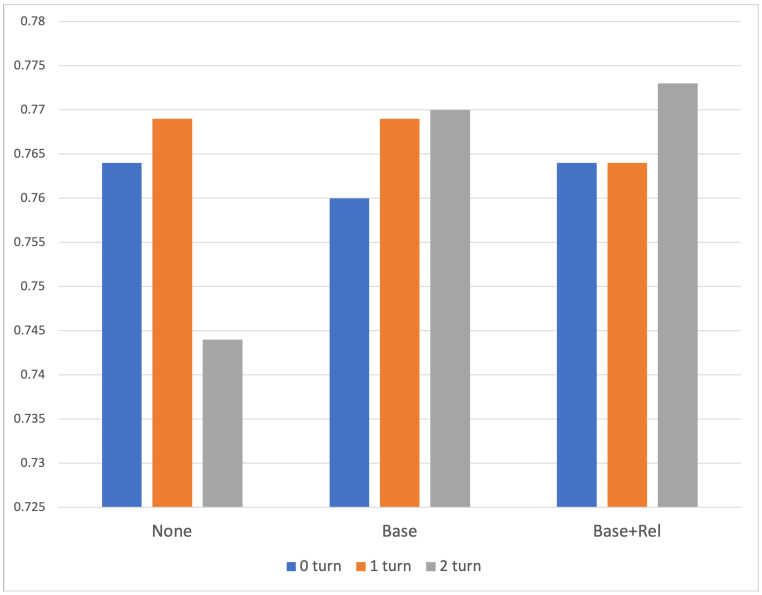


Figure 5.4: Comparison of BERTScores for each turn of previous utterances in each prefix type in Chinese to Japanese translation

doi:DOI Number

5.4 Discussion

The above result for Chinese to Japanese supports our hypothesis that contextual information contribute to generate culture-aware translation (Table 5.4) but the result for Japanese to Chinese translation does not (Table 5.3). Also, the generated sequences not only should be measured by automatic metrics but also should be evaluated by a human. Therefore We will review the generated texts and analyze what is being culture-aware translated well and what is not being translated well. In this section, we discuss the causes of results for Japanese to Chinese, and about well culture-aware translation and not well translation.

5.4.1 Confusion by Relation Label

For Japanese to Chinese translation, the highest score in *(*Base+Rel) prefix is lower than the highest score in *(*Base) prefix and it is the almost same score as of *None* prefix. This means that even though the relation information was added as a prefix and given to the model, the score did not get higher. This implies that the relation information does not contribute to the improvement of classification performance. To investigate the cause of this, we calculated the BERTScore for each utterance pair of generated text and grand-truth in the two text cases which the number of turns of previous utterances is up to two in both, and the prefix settings are *Base+Rel* or *Base* in. The top three instances which have the largest difference among the instances, and whose scores for *Base+Rel* are lower than *Base* and are shown in Table 5.5. In the first and third instances with the largest difference in scores, the generated texts of *Base* have the same words as grand-truth, while that of *Base+Rel* have different words. In the second instance, the generated text of *Base+Rel* is an apparently bad translation. Considering this, adding the relation label as part of

doi:DOI Number

the prefix to the input is confusing the model. This reasoning is consistent with a view of Chen et al. [2013] that Chinese have less clear rules of pragmatic for using different language uses depending on the interpersonal relation, while Japanese is restricted by more stringent rules of pragmatic for using language uses with appropriate degrees of politeness depending on the interpersonal relationships. Therefore, this result could be plausible.

5.4.2 Well Culture-Aware Translation

Especially for Chinese to Japanese translation, as can be seen from the scores (Table 5.4), the generated texts are more similar to grand-truth when more contextual information is given, which infer that it is more culture-aware translated. The following examples (Table 5.6 and 5.7) are a part of the generated texts which is more appropriate language use in each target language's dialogue culture.

For the example of Japanese to Chinese translation (Table 5.6), there is a gratitude and an apology in the input text but the expression of gratitude is not in the sentence of grand-truth. Similarly, the expression of gratitude is missing from the generated text. About our dataset, in both the pre-translated text and the human-translated text, there are not few sentences that contain expressions of apology and gratitude within a single sentence. In addition to it, when we consider the utterance of the same interpersonal relation, similar changes to this example utterances rarely occur between pre-translated texts and human-translated texts except for in the utterances of the consignee. Considering these, we can assume that the model never simply erased the expression of gratitude when translating it into Chinese, and generated translations with considering the previous utterances or other contextual information.

doi:DOI Number

Order	Text Type (Prefix)	Utterance
1st	input	うん? ちょっとごめんね。 Hmm? I'm sorry.
	grand-truth	嗯？抱歉啊。 Hmm? I'm sorry.
	generated (Base)	嗯? 抱歉啊。 Hmm? I'm sorry.
	generated (Base+Rel)	嗯? 不好意思诶。 Hmm? Excuse me.
2nd	input	それ上級者じゃなくてさなんか一番つらいとこってさなんか何時間コースみたいのあるじゃん。 I did not talk about the expert course. I mean that the hardest course in some place is the one that takes many hours.
	grand-truth	好不好爬倒不是什麼问题。不是有那种一次要爬几小时的那种行程吗？ It's not a matter of whether it's easy to climb or not. Isn't there the kind of trip that takes several hours to climb at a time?
	generated (Base)	诶? 你到底要怎麼跟他比较好吧? Eh? How the hell are you going to compare with him?
	generated (Base+Rel)	诶, 你说, 你说, 你说, …(repeated)…你说 Eh? you said, you said, you said …(repeated)…you said.
3rd	input	ごめん。 Sorry.
	grand-truth	不好意思，不是诶。 Excuse me, it is not.
	generated (Base)	不好意思诶。 Excuse me.
	generated (Base+Rel)	抱歉诶 Sorry,

Table 5.5: The three instances where there is the largest difference in BERTScore.

doi:DOI Number

Scene:	The speaker apologizes and thanks after a companion of the speaker said 'sorry' that they answered wrongly where the clerk should put the ordered item.
Pre-translated text: (Input text)	はい。ありがとうございます。すいません。 Yes. Thank you very much. I'm sorry.
Human-translated text: (Grand-truth)	没关系啦。那个，不好意思。 It's okay. Um, sorry.
Generated text: (# of turn: 2, prefix: Base+Rel)	诶... 不好意思诶 Ugh... Sorry.

Table 5.6: Example of well culture-aware translation from Japanese to Chinese

Scene:	The speaker responds to his friend's apology for the misunderstanding and his friend's wish for a Happy New Year.
Pre-translated text: (Input text)	小弟弟！听起来好亲切哦！大哥，我就不打扰你了！拜！ Little brother! You sound so friendly! Brother, I'll leave you alone! Bye!
Human-translated: (Grand-truth)	ありがとう、じゃあ失礼します。 Thank you. Excuse me then.
Generated text: (# of turn: 2, prefix: Base+Rel)	ごめん、ありがとう Sorry, thanks.

Table 5.7: Example of well culture-aware translation from Chinese to Japanese

For the example of Japanese to Chinese translation (Table 5.7), there are the expressions of the speaker's impression for the previous utterances and of goodbye with consideration for the listener. However, after translation, Impressions of the previous utterance are removed, and a humble expression and an expression of gratitude were generated. The generated text is not similar to direct translation of pre-translated texts but much similar to grand-truth. Furthermore, the generated text by the model without any contextual information is following: お兄さん、ありがとう。(Elder brother, thank you). Vocatives are rarely used in conversation in Japanese, and also much fewer vocatives can be seen in Japanese texts in our dataset. Considering these, when the model is given contextual information, the model succeeds in learning the language use of the Japanese conversation culture in given contexts.

doi:DOI Number

Translation	grand-truth	generated text
Ch to Ja	31.506	21.888
Ja to Ch	13.799	9.885

Table 5.8: Comparison of mean length between grand-truths and generated texts

5.4.3 Not Well Translation

The quality of a part of the generated texts is low. We review the generated texts and mention what is not well about the generated translation.

Repetition of the phrase: A few generated have a repetition of the same phrase or words like the generated text (*Base+Rel*) of the second instance with the largest difference in scores in Table 5.5. Besides that, for Chinese to Japanese translation in the condition of *Base+Rel* prefix and two turns of previous utterances, several generated texts have a repetition of the same phrase or words.

Shorter length: When the length of input text is long, the generated text tends to be much shorter such as the generated text (*Base*) of the second instance with the largest difference in scores in Table 5.5. We measured the mean length for both of grand-truths and the generated texts in the condition of *Base+Rel* prefix and two turns of previous utterances in test set. As can be seen in Table 5.8, both for Chinese to Japanese and for Japanese to Chinese translation, the mean lengths of generated texts are much shorter than that of grand-truths, though the length of grand-truth tends to be shorter than that of input, especially for Chinese to Japanese in MPDD (Table 3.4).

Low diversity: The model tends to generate the phrases or words which occur frequently in the train set of data. Some characteristic phrases in each situation are generated regardless of whether the input texts contain the corresponding phrase in the target lan-

doi:DOI Number

Translation	grand-truth	generated text
Ch to Ja	37.854	44.134
Ja to Ch	20.573	40.059

Table 5.9: Comparison of mean self-BLEU between grand-truths and generated texts

guage or not. To ensure that those characteristic phrases are generated too often, we measure diversity for both grand-truths and the generated texts in the condition of *Base+Rel* prefix and two turns of previous utterances in the test set using self-BLEU. Self-BLEU score is frequently used for measuring the diversity of a set of generated texts. It calculates the BLEU scores (Papineni et al. [2002]) by selecting each sentence in the set of generated sentences and the other sentences and takes the mean of the BLEU scores of all generated sentences where s_i and s_j are members of the set of the generated texts S.

$$\text{self-BLEU} = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} (100 - \text{BLEU}(s_i, s_j)) \tag{5.1}$$

Table 5.9 shows that self-BLEU scores of generated texts both for Chinese to Japanese and for Japanese to Chinese translation are higher than that of grand-truths, which means that the diversities of the generated texts are lower. However, this fact also can be seen as a part of the result of the model's consideration for culture-aware translation.

doi:DOI Number

6 Conclusion

In this work, we stated the necessity of solving the problems caused by cultural differences in language use in dialogue and the importance of contextual information for appropriate language use in dialogue. To approach the problem, we created and analyzed our dataset. Also, we conduct experiments with our dataset. We organize this chapter with our contributions and future work as our conclusion.

6.1 Contributions

So far, to have a model perform culture-aware translation for solving the problems caused by cultural differences in dialogue, we created our dataset and conducted analysis, experiments, and discussions. We describe in this section the four contributions we have made through this study.

At first, we reduced the potential issues of cultural differences in cross-cultural communication with machine translation to the issue in natural language processing. This is the first study that attempts to solve the problems caused by cultural differences in dialogue using natural language processing techniques.

Secondly, we created a cultural-aware parallel dialogue dataset with situation labels

doi:DOI Number

and interpersonal relationship labels in Chinese and Japanese. This dataset is the first parallel dataset for culture-aware dialogue.

Thirdly, we demonstrated that the model can be aware of appropriate language uses of Japanese culture in certain situations by situation classification task. For the Chinese situation categorization, the scores were not as we expected, but the consistency with the results of Chen et al. [2013]'s study and the cultural differences in the dialogue between China and Taiwan were suggested as reasons for this.

Finally, we demonstrated that contextual information is important to make the model generate more culture-aware translation for each dialogue culture in Chinese and Japanese by culture-aware machine translation task. In both Chinese and Japanese culture-aware machine translation, we confirmed that the translations were generated in a culture-aware manner. Also, for Chinese to Japanese cultural-machine translation, we verified that the contextual information including situations and interpersonal relationships contributes to the generation of culture-aware translation. In addition, the result of culture-aware machine translation for Japanese to Chinese did not meet our expectations, but the consistency with the results of the Chen et al. [2013] study was suggested as a reason for this.

6.2 Future work

Our dataset size which we created in this study is small and includes limited three situations (Table 3.2). Therefore, we indicate four potentially fruitful directions of data expansion.

The first direction is to expand the amount of data in each situation. There are countless possible contexts in the real world, and they are so complex that no two contexts are exactly the same. Therefore, in order for the model to better learn language uses in the

doi:DOI Number

dialog culture of each language, more various data for each situation should be included in the dataset.

The second is to expand the variety of situations and to create data. In this study, we focused on only three situations including apology, request, and thanks. However, there are more situations where problems could occur due to cultural differences in language use. In the previous studies, other situations are also featured such as compliment, invitation, refusal, and so on. These situations are worth investigating.

The third is to expand data in other languages or cultures. In this study, we created data only in Chinese and in Japanese. The other languages also have distinctive dialogue culture and problems could arise due to these cultural differences between languages. Therefore, expanding data in other languages is worthy. In addition to it, In section 4.4.2, we mentioned the possibility of differences in the appropriate language uses in a certain situation between Mainland of China and Taiwan. Even if the language is the same, the language use that is considered appropriate in a certain situation is different, so it is also meaningful to extend the data by region.

The fourth one is to annotate the different language use labels which we introduced in 3.4.2 section. In this study, we annotated them only on a part of parallel texts. The different language use labels would help the model learn the appropriate language use for culture-aware translation more directly because each label means the strategies or styles of language uses of translation in a certain context.

doi:DOI Number

References

Rong Chen, Lin He, and Chunmei Hu. Chinese requests: In comparison to american and japanese requests and with reference to the “east-west divide”. Journal of Pragmatics, 55:140–161, 2013.

Sanae Enomoto and Helen Marriott. Investigating evaluative behavior in japanese tour guiding interaction. 1994.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21 (140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Mei Song and Wong Lee. Qing/please–a polite or requestive marker?: Observations from chinese. 1994.

Wong Lee and Mei Song. Imperatives in requests: Direct or impolite–observations from chinese. Pragmatics, 4(4):491–515, 1994.

doi:DOI Number

Ming-Chung Yu. Universalistic and culture-specific perspectives on variation in the acquisition of pragmatic competence in a second language. Pragmatics, 9(2):281–312, 1999.

Stephanie Weijung Cheng. An exploratory cross-sectional study of interlanguage pragmatic development of expressions of gratitude by Chinese learners of English. The University of Iowa, 2005.

Asuka Ichihara. ' re-thanking on previous indebtedness' as a thanking strategy of japanese language : Compared with chinese language discourse. Japanese language education, (51):21–29, 2016.

Beverly Hill, Sachiko Ide, Shoko Ikuta, Akiko Kawasaki, and Tsunao Ogino. Universals of linguistic politeness: Quantitative evidence from japanese and american english. Journal of pragmatics, 10(3):347–371, 1986.

Junjie Xing. Practical applications of the understanding of foreign cultures in japanese-language teachings. Journal of Contemporary Educational Research, 5(6):19–22, 2021.

Ming-Shu Huang. A comparative study of chinese and japanese linguistic behavior in the closing section in agreement situation of invitational discourse : Focusing on two situations with different burden degrees. Japanese language education, (48 • 49):22–31, 2015. ISSN 0917-4206. URL https://ci.nii.ac.jp/naid/120005770816/.

Yi-Ting Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. Mpdd: A multi-party dialogue dataset for analysis of emotions and interpersonal relationships. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 610–614, 2020.

Hanae Koiso, Haruka Amatani, Yuriko Iseki, Yasuyuki Usuda, Wakako Kashino, Yoshiko Kwabata, Yayoi Tanaka, Yasuharu Den, and Ken'ya Nishikawa. Design, evaluation,

doi:DOI Number

and preliminary analysis of the monitor version of the corpus of everyday japanese conversation. NINJAL Research Papers, (18):17–33, jan 2020.

花絵小磯, 晴香天谷, 祐一石本, 泰如居關, 友里子 nad 臼田, 和佳子柏野, 良子川端, 弥生田中, 康晴伝, and 賢哉西川. 『日本語日常会話コーパス』モニター公開版コーパスの設計と特徴. 国語研究所日常会話コーパスプロジェクト報告書 3, mar 2019.

Shoshana Blum-Kulka and Elite Olshtain. Requests and apologies: A cross-cultural study of speech act realization patterns (ccsarp). Applied linguistics, 5(3):196–213, 1984.

Xiaowen Guan, Hee Sun Park, and Hye Eun Lee. Cross-cultural differences in apology. International Journal of Intercultural Relations, 33(1):32–45, 2009.

Dean C Barnlund and Miho Yoshioka. Apologies: Japanese and american styles. International Journal of Intercultural Relations, 14(2):193–206, 1990.

Barbara Pizziconi. Re-examining politeness, face and the japanese language. Journal of pragmatics, 35(10-11):1471–1506, 2003.

Shoshana Blum-Kulka. Learning to say what you mean in a second language; a study of the speech act performance of learners of hebrew as a second language. 1980.

Nessa Wolfson. Compliments in cross-cultural perspective. TESOL quarterly, 15(2): 117–124, 1981.

Juliane House and Gabriele Kasper. Politeness markers in english and german. In Conversational routine, pages 157–186. De Gruyter Mouton, 2011.

Jenny Thomas. Cross-cultural pragmatic failure. Applied linguistics, 4(2):91–112, 1983.

doi:DOI Number

Yun Meng. Pragmatic transfer by chinese advanced learners of japanese: expressing politeness in the context of refusals to requests. Forum of international development studies, (36):241–254, 2008. ISSN 13413732.

20 newsgroups. URL http://qwone.com/~jason/20Newsgroups/.

Ken Lang. Newsweeder: Learning to filter netnews. In Proceedings of the Twelfth International Conference on Machine Learning, pages 331–339, 1995.

The reuters dataset, Jul 2017. URL https://martin-thoma.com/nlp-reuters/.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. Semantic web, 6(2):167–195, 2015.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. arXiv preprint cs/0205070, 2002.

Yelp open dataset. URL https://www.yelp.com/dataset.

News-commentary v16. URL https://opus.nlpl.eu/News-Commentary.php.

Jörg Tiedemann. Parallel data, tools and interfaces in opus. In Lrec, volume 2012, pages 2214–2218. Citeseer, 2012.

Changhan Wang, Anne Wu, and Juan Pino. Covost 2 and massively multilingual speech-to-text translation. arXiv preprint arXiv:2007.10310, 2020.

Pierre Lison and Jörg Tiedemann. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. 2016.

doi:DOI Number

Ahmed Abdelali, Francisco Guzman, Hassan Sajjad, and Stephan Vogel. The amara corpus: Building parallel language resources for the educational domain. In LREC, volume 14, pages 1044–1054, 2014.

Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 11 2020. URL https://arxiv.org/abs/2004.09813.

Chat translation task - emnlp fifth conference on machine translation. URL http://www.statmt.org/wmt20/chat-task.html.

Matīss Rikters, Ryokan Ri, Tong Li, and Toshiaki Nakazawa. Designing the business conversation corpus. In Proceedings of the 6th Workshop on Asian Translation, pages 54–61, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5204. URL https://www.aclweb.org/anthology/D19-5204.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Eric Nyberg. Shakespearizing modern language using copy-enriched sequence-to-sequence models. arXiv preprint arXiv:1707.01161, 2017.

Keith Carlson, Allen Riddell, and Daniel Rockmore. Evaluating prose style transfer with the bible. Royal Society open science, 5(10):171920, 2018.

William Hwang, Hannaneh Hajishirzi, Mari Ostendorf, and Wei Wu. Aligning sentences from standard wikipedia to simple wikipedia. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 211–217, 2015.

doi:DOI Number

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. Automatically neutralizing subjective bias in text. In Proceedings of the aaai conference on artificial intelligence, volume 34, pages 480–489, 2020.

Sudha Rao and Joel Tetreault. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. arXiv preprint arXiv:1803.06535, 2018.

Translation quality. URL https://www.deepl.com/en/quality.html.

Ai 企業が考察する google 翻訳超え機械翻訳「deepl」のスゴさ, Apr 2020. URL https://ledge.ai/deepl/.

張穎. 依頼会話の展開パタンに関する日中対照研究. 言語文化と日本語教育, 28: 8–14, 2004.

YAN Yi-bo. The comparative study on politeness between chinese and english. US-China Foreign Language, 544, 2015.

Xuqing Shi. Study on pragmatic failures in cross-culture communication. 2020.

Hajime Morita, Daisuke Kawahara, and Sadao Kurohashi. Morphological analysis for unsegmented languages using recurrent neural network language model. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2292–2297, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1276. URL https://aclanthology.org/D15-1276.

Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. Computational Linguistics, 29(1):19–51, 2003.

doi:DOI Number

Chin-Lan Huang, Cindy K Chung, Natalie Hui, Yi-Cheng Lin, Yi-Tai Seih, Ben CP Lam, Wei-Chuan Chen, Michael H Bond, and James W Pennebaker. The development of the chinese linguistic inquiry and word count dictionary. Chinese Journal of Psychology, 2012.

瑋芳林, 金蘭黃, 以正林, 嘉玲李, and James W. Pennebaker. 語言探索與字詞計算詞典 2015 中文版之修訂. 調查研究-方法與應用, 45:73–118, 2020.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 483–498, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41. URL https://aclanthology.org/2021.naacl-main.41.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675, 2019.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pages 311–318, 2002.

doi:DOI Number