Towards Japanese Dialect-aware Chatbot: Adapting NLP Models for Japanese Dialect Variation

Kinga Lasek[1,*], Michal Ptaszynski[1] and Fumito Masui[1]

[1] *Text Information Processing Laboratory, Faculty of Engineering, Kitami Institute of Technology, 165 Koen-cho, Kitami-shi, Hokkaido, 090-8507 Japan*

Abstract

Chatbots are said to soon become the most frequently used consumer channel service in the world. Users' expectations are growing, especially regarding the usage of the natural language. In this study, we present our attempts of training NLP models for Japanese dialect variation with the aim of creating Japanese dialect-aware chatbot in the future. Firstly, we describe the current situation on chat-bot market. Secondly, we focus on describing attempts that have already been made to create chatbots sensitive to dialects and low-resourced languages. Furthermore, the process of collecting data needed for our research is presented. Finally, evaluation results after model training are shown and described. The results reveal key insights into the models' strengths and limitations in understanding Japanese dialects. We discuss noted occurrences and finally outline directions for our future work.

Keywords

Natural language processing (NLP), Japanese dialects, dialect-specific models, chatbot

1. Introduction

Chatbots, conversational interfaces that are able to communicate with users in anticipated natural-sounding language, greatly influenced different fields, such as healthcare [1], finance [2] as well as the broadly understood consumer realm [3]. We also know them as smart bots, chatterbots, interactive agents or digital assistants. Microsoft's Cortana and Apple's Siri serve as perfect examples of well-known chatbots. Nowadays, the main goal is to create chatbots that will use the same language as their interlocutor as it is believed that users‘ language choices are determined by their individual style, genre as well as dialect [4]. According to Marcellus Amadeus et al.:

> *Speaking like the users may not only create a good relationship between users and AI agents - consequently, the brand, the person, the company, or else that uses it as its voice - but also can make the message clearer since it is in a language variety the user understands the most.* [5]

Andre Martin and Khalia Jenkins [6] presented a framework for AI dialect introduction on the basis of research proving that dialect usage might positively influence users‘ trust and satisfaction. Moreover, it should be noted that if Artificial Intelligence (AI) systems are trained on datasets that are limited or biased, they might prioritize some dialects over others,which in consequence may unequal representation of linguistic diversity [7]. Taking those into consideration, we decided to make efforts to create Japanese dialect-aware chatbot. To our knowledge, at the moment, there is no such conversational interface system for the Japanese language and its regional variations. We are motivated by the fact that the Japanese language is rich in dialects [8], many of which are now disappearing[1]. Additionally, in 2022 in Japan goo AI x Design conducted a survey regarding chatbot usage[2]. Their report shows that neraly 80% of respondents consider chatbots as useful, fast-responding applications. At the same time,

The 9th Linguistic and Cognitive Approaches to Dialog Agents Workshop

[*] Corresponding author.

d3247180026@std.kitami-it.ac.jp (K. Lasek); michal@mail.kitami-it.ac.jp (M. Ptaszynski); f-masui@mail.kitami-it.ac.jp (F. Masui)

0009-0009-3149-9637 (K. Lasek); 0000-0002-1910-9183 (M. Ptaszynski); 0000-0001-9979-8734 (F. Masui)

[1] https://kikigengo.ninjal.ac.jp/en/data/ [access: 2024/08/13].

[2] https://aixdesign.goo.ne.jp/material/01_report.html

the correctness of the answers given and the understanding of the input text were identified as the biggest problems. It led company carrying out the survey to conclusion that technical aspects directly related to response content, such as natural language processing and chatbot response accuracy need to be improved.

The outline of the paper is as follows. Firstly, we discuss the current state of chatbot technology and actions taken so far to build chatbots sensitive to dialects. Then, we shortly describe collecting data process and how we utilized it. What is more, we explore one of the fundamental steps essential for creating Japanese dialect-aware chatbot, meaning training models assumed to handle Japanese dialects. Results of experiments are presented and discussion points are highlighted.

2. Background

2.1. Chatbots today

Around 2027, chatbots are supposed to become the dominant customer service channel for approximately 25% of organizations [3]. Moreover, the global chatbot market size was calculated at USD 5,132.8 million in 2022 and is predicted to achieve a compound annual growth rate (CAGR) of 23.3% from 2023 to 2030[4]. Such a scale indicates the improvements taking place in the world of Artificial Intelligence (AI) as well as Natural Language Processing (NLP)'s realm. Technologies that enhance the knowledge of chatbots encompass, among others, deep learning, machine learning and NLP [9]. The mutual understanding of the system and humans makes it possible to reach customer satisfaction. The better and error-free the responses of the chatbot, the greater the customer's trust. Additionally, it is supposed that users' expectations will grow as the chatbots upgrade their communication abilities or even so-called social skills [10]. For a long time, research on chatbots and their development focused mainly on appropriate grammar structures and quick responses. Still, the design of chatbots faces many problems, including biases, (lack of) emotional involvement, ethical issues or keeping privacy [9]. Posed dilemmas involve the need for new improvements, so research into the capabilities of chatbots is also being expanded. In consequence, chatbots' system design and language design are in the spotlight. Therefore, psycholinguistic, sociolinguistic, dialectometry approaches and so forth have been taken. Previous research has proved that variation within a language often comes from factors such as individual author/speaker style, dialect or genre [10], so scholars took steps to bring machine language closer to human-like style and make them more personalized and attractive to potential users. Elsholz et al. [11] conducted an experiment using two chatbots to sell theater tickets for a Shakespeare play: one communicating in modern English and one in a Shakespearean-style dialect. The results were thought-provoking for researchers as modern chatbot reached a higher usability score, whereas scores regarding interest in play were similar for both chatbots. This highlights the necessity of developing chatbot's responses that will be as close to humans as possible. Marcellus Amadeus et al. [5] also pointed out that: *integrating language variation into conversational AI will build near-real language inventories and boost user engagement.* Users might use more or less formal language, its vernacular variations, dialect or slang. Realization of language variety leads to expanding research. Scholars as well as big companies and organizations started paying more attention to the conversational style of chatbots [12] and exploring their abilities of participating in discussions conducted in diverse language aspects. Among slowly developing enhancements are chatbots/voicebots able to understand and provide an answer in low-resourced languages and a wide range of dialects. Thanks to this improvement, users will not feel obligated to use the standard form of their language, but will input information expressed in phrases most similar to those they use every day. Therefore, among papers about English and Chinese chatbots, one can find also research on conversational agents and derivatives focused on, for example, Vietnamese [13], Irish [14] or Indian languages [15]. Big steps regarding chatbot technology were also taken in the Arabic world [16].

[3] https://www.gartner.com/en/newsroom/press-releases/2022-07-27-gartner-predicts-chatbots-will-become-a-primary-customer-service-channel-within-five-years [access: 2024/04/15].

[4] https://www.grandviewresearch.com/industry-analysis/chatbot-market [access: 2024/04/15].

2.2. Existing dialect-aware and low-resource language-aware chatbots: overview

Researchers have been working on developing chatbots that are dialect-aware and capable of handling under-resourced languages to improve their performance and usability across diverse linguistic communities. We cannot yet speak of a significant number of them, however, on the basis of those available we can observe certain trends and draw conclusions. Chatbots created to understand language varieties in linguistically rich countries or communities are coming to the fore. Among 21 countries in the Arab world, each has its own dialect. In the case of Arabic, nine dialectal categories are distinguished and each of them has more varieties depending on the particular city or town [17]. Probably the most known Arabic dialect chatbots are Nabiha [18] and Botta [19]. Nabiha, developed by Dana Al-Ghadhban and Nora Al-Twairesh, is able to understand and respond in the Saudi Arabic dialect. In the early stages, authors collected, preprocessed and classified data into several text files in order to build a dialogue corpus. Authors admitted that despite good first evaluation, Nabiha should be based on a bigger dataset. Botta, another example of Arabic chatbots, was created earlier, in 2016. It speaks in Egyptian Arabic (Cairene) dialect. Botta‘s persona – female chatbot – was supposed to simulate friendly conversation and connect with as many Arab users as possible. Both Botta and Nabiha were built using artificial intelligence markup language (AIML) and launched on the Pandorabots platform. However, Botta also possesses some unique sets, being able to: detect dialectal bad language that can be offending, separate sets of months based on each dialect it identifies or words that indicate the dialect of the users.

Boulesnane et al. [20] created a communication platform for patients in need of consultation. They named it DZchatbot and developed it with a focus on the Algerian Arabic dialect. Differently from Botta and Nabiha, this chatbot was hinged on the sequence-to-sequence model (seq2seq) with RNN encoder and decoder. Larger, more detailed healthcare chatbot, MedicalBot [21], was proposed through application of three deep learning techniques, but any particular dialect was itemized. Scholars mentioned, though, the importance of the multitude of dialects. Among Arabic dialect-aware chatbots one can also mention a medical chatbot for Tunisian dialect [22] or chatbot system focused on dialogue acts, which was built in order to cope with Levantine Arabic dialect [23].

In India, despite the common perception that English is mainly spoken there, other so-called 21 modern Indian languages are used and English is spoken by only ten percent of India population [15]. Among Indian languages one can also observe a quite wide scope of chatbots successfully communicating in its variations. For instance, Golpo, which can converse in Bengali, was developed with a focus on storytelling and engaging users with narrative-based interactions [15]. On the contrary, Doly [24] was supposed to be a support for Bengali-speaking users. The development process involved training Doly on a diverse dataset of Bengali conversations to enhance its linguistic capabilities and dialect recognition. It can provide the appropriate response thanks to the list using Naive Bayesian. However, probably the most known Hindi chatbot is AskDISHA, which nowadays is available as AskDISHA2.0 (Digital Interaction To Seek Help Anytime)[5] as it has been upgraded. Currently this chatbot, created thanks to AI and machine learning, can answer its interlocutor both through text message and voice message in Hindi, English and so-called Hinglish (macaronic hybrid use of Indian English and the Hindustani language). Its main management is coping with railway reservations. Kaleem et al. [25] faced the challenge of designing an Urdu conversational agent, which covers novel features such as the Word Order Wizard (WOW) algorithm and scripting language in its architecture. They named it Umair and developed it with the aim of applying it as customer service representative for Pakistan’s National Database and Registration Authority (NADRA). To improve the relevance and coherence of its responses, UMAIR may utilize context-aware techniques to understand the ongoing conversation and maintain context over multiple turns. Some years later, Shabbir et al. [26] scrutinized whether and how it is possible to automate the process of user‘s intent generation by using AI and deep learning techniques so that human endeavor could be reduced. They made their experiment on grounds of Roman Urdu and RASA Framework. They took into consideration two major factors of this particular framework: RASA NLU, as it can perform intent classification and entity extraction from the training dataset, and Dialog Management Model (DMM) that prepares the specific response according to intent.

[5] https://corover.ai/askdisha/ [access: 2024/04/25].

A knowledge graph with RASA Framework has also been embedded to preserve the dialog history *for semantic based natural language mechanism for chatbot communication* [26]. Brixey and Traum [27]'s chatbot, Masheli, represents a significant advancement in exploiting technology to support language revitalization efforts and cultural preservation. They used ChoCo, a Choctaw language corpus to create the chatbot's replies and later formed questions to place them in QA corpus. In order to build Covid-19 chatbot sensitive to African dialects Aymen Ben, Mabrouk et al. [28] gathered needed data and divided it into two main categories: Frequently Asked Questions (FAQ) and chitchat. Chatbot possesses an ability of providing answers in English, French, Arabic, Tunisian as well as spoken in Nigeria Igbo, Yoru'ba and Hausa. Sarma and Pathak [29] presented "Shiksha Mitra", an artificial intelligence chatbot able to reply to user queries in Assamese. Just as for chatbots using the best-known languages, the Artificial Intelligence Markup Language (AIML) method is used for under-resourced language chatbots [30], [31]. The AIML conversational agent operates on the principle of pattern matching, where responses are generated through the mapping of keywords within each request to their corresponding patterns. Utilizing the AIML Interpreter facilitates the pattern matching process between queries and responses. Sandhini, Binu et al. [31] introduced Malayalam (Dravidian language spoken in the Indian state of Kerala and assigned as classical language of India) chatbot. The comparison between AIML based version of this chatbot and machine learning type of this chatbot was made. The performance of machine learning based chatbot was better. Compared to AIML, machine learning requires substantial amounts of high-quality training data to perform effectively.

As the interest in language varieties is constantly growing [32], one can assume that deployment of chatbots related to dialects and low-resource languages will also take on meaning. One of the most serious challenges is scarcity of data. That is why among scholars, a corpus-driven approach can be observed. Researchers collect large datasets of conversations in different dialects, annotate them with linguistic features, and use them to train dialect-aware chatbots. Usually, due to a small amount of data, creators of a chatbot must collect data themselves. This is what Al-Ghadhban and Al-Twairesh [18] did to build Nabiha. It has been trained on a diverse corpus of Arabic conversations, encompassing various dialectal variations, to learn from real-world linguistic data and adapt to different linguistic contexts. Brixey and Traum [27] also chose this path to bridge the gap between Choctaw language (native American language) and English. They used ChoCo, a Choctaw language corpus to create the chatbot's replies and later formed questions to place them in QA corpus. The methodology of Orosoo et al.[33] was also to collect diverse linguistic dataset from different languages in order to propose system that improves NLP in multilingual chatbots. Authors gave prominence to significant obstacles, such as difficulties with low-resource languages, biases in training data and dynamic between language and culture. The attempts of researchers mentioned above prove the complexity of studies on chatbots.

3. Data Collection and Data Preprocessing

3.1. Used corpuses and datasets

In this study, a comprehensive data collection approach was employed, utilizing several key datasets, including:

- **Japanese Dialect Conversations Data**[6]: includes transcripts of short dialogues in different Japanese dialects in diverse formality of language.

- **Corpus of Japanese Dialects: COJADS**[7]: a parallel corpus consisting of standard Japanese, (*hyōjungo*) text and dialect texts as well as audio. It has been sorted according to 47 Japanese prefectures. Among those corpuses and datasets that we decided to incorporate in our research, COJADS has the the widest variety of dialectal data.

[6] *Hōgen rōru purei kaiwa deetabeesu* [Japanese Dialect Conversations Data] http://hougen-db.sakuraweb.com/[access: 2024/06/27].

[7] Corpus of Japanese Dialects (COJADS). NINJAL. We used version published in March 2023. https://www2.ninjal.ac.jp/cojads/index.html [access: 2024/06/27].

- **The Corpus of Kansai Vernacular Japanese** [34]: was divided according to the place (Osaka-Kobe area, Kyoto area, towns of Takacho and Nishiwaki) consisted of four files. Among them, one was created on the basis of interviews with international students living in the Kansai area. Aiming for natural language used by Japanese people, we did not include students speech based data in our research.

- **Crowdsourced Parallel Speech Corpus of Japanese Dialects** [35]: consists of parallel text and speech data of 21 Japanese dialects. Dialogues for each dialect were short and included 250 sentences/turns.

- **JMD: Japanese multi-dialect corpus**[8]: that was made of audio and text versions of Osaka and Kumamoto dialects. It was constructed so that each sentence was used as a separate entity with a separate recording. Files for both dialects consist of 1300 lines of text and present the same content.

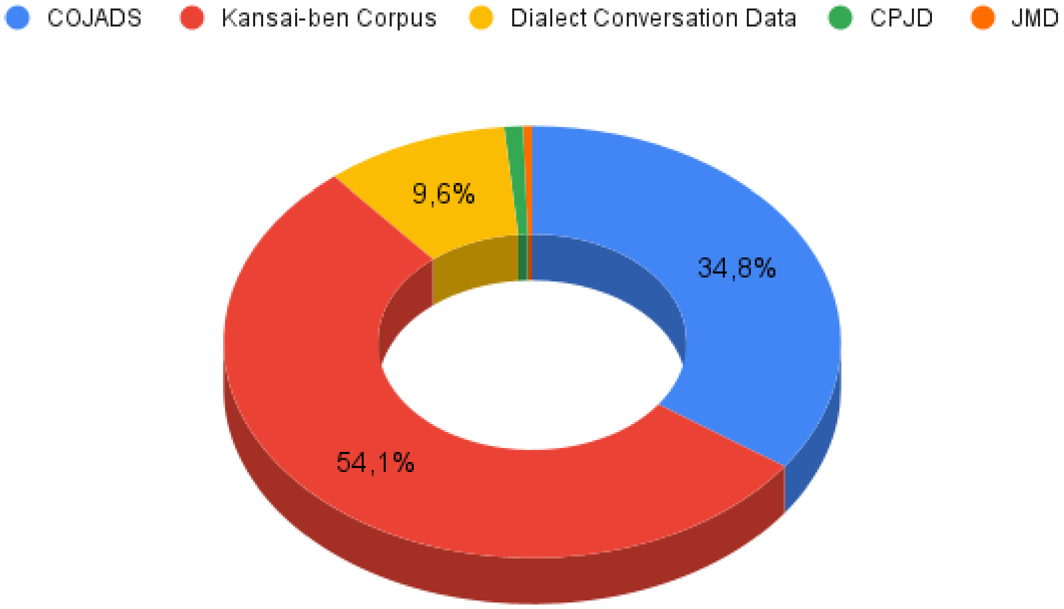

Figure 1: Used corpuses and their size

3.2. Data preprocessing

By incorporating those, it was necessary to pay heed to the Japanese writing system. Japanese is written as a combination of three types of symbols as well as roman letters and Arabic numerals. The oldest symbols, *kanji*, came to Japan from China and were adapted to the Japanese language. These logographic characters are used for conceptual words and names. Two syllabaries, *hiragana* and *katakana*, were developed later. *Hiragana*‘s main function is recording words and grammatical endings that are not written in *kanji*, while *katakana* serves as a method of writing words of foreign origin [36]. In COJADS, mixed *kanji* and *hiragana* script marks text in standard form of Japanese. On the contrary, dialect texts are written in *katakana*.

[8] Shinnosuke Takamichi, "JMD: Japanese multi-dialect corpus," https://sites.google.com/site/shinnosuketakamichi/publication/research-topics/jmd_corpus [access: 2024/08/08].

We decided to convert dialectal texts from the other four corpuses into *katakana*. Before that, we adjusted data for our needs: Firstly, we decided not to include tagged parts of each sentence. For instance, we cleaned the text from non-verbal sounds, like cough or laugh, that were originally indicated by curly brackets as we focus on the text, not the audio version. The same was done to the *hyōjungo* version. Consequently, we did the same with other corpus that treated non-verbal sounds in the same way in the text, namely Japanese Dialect Conversations Data. For the same reason, we did not take into consideration symbol /// that was supposed to highlight unknown meanings of words. Moreover, in Japanese Dialect Conversations Data, one could find the arrows that indicated up or down intonation. Those were also deleted, together with signs // which meant that the immediately following utterance started at the same time as the next speaker's utterance [37]. Some of the talks were held in standard Japanese, however, it needs to be emphasized that each dialogue differs and we do not possess parallel data. When it comes to the Corpus of Kansai Vernacular Japanese, we used a morphologically untagged version of the texts. In JMD: Japanese multi-dialect corpus for speech synthesis neither tags nor special signs were used. We only had to clear the text of some *kanji* readings (probably those considered as hard), which were taken in brackets and written in the *hiragana* syllabary, and transform the text into *katakana*.

3.3. *Katakana* Conversion

For *katakana* conversion, we used MeCab [38], a highly optimized Japanese morphological analyzer. It effectively processes mixed *kanji* and kana sentences, converting them into uniform *katakana* or *hiragana* versions. It is believed to achieve a high efficiency at breaking down Japanese text into individual morphemes, which we find essential for accurate katakana transcription. We started from converting each character into *katakana* (e.g., お父さん became オトウサン, read: *otōsan*). Then we focused on transcription that can provide a standardized phonetic nuance representation (so here お父さん became オトーサン), which is believed to be useful when dealing with various Japanese dialects. We hope that this method will benefit in capturing the subtle differences between accents and pronunciation among dialects.

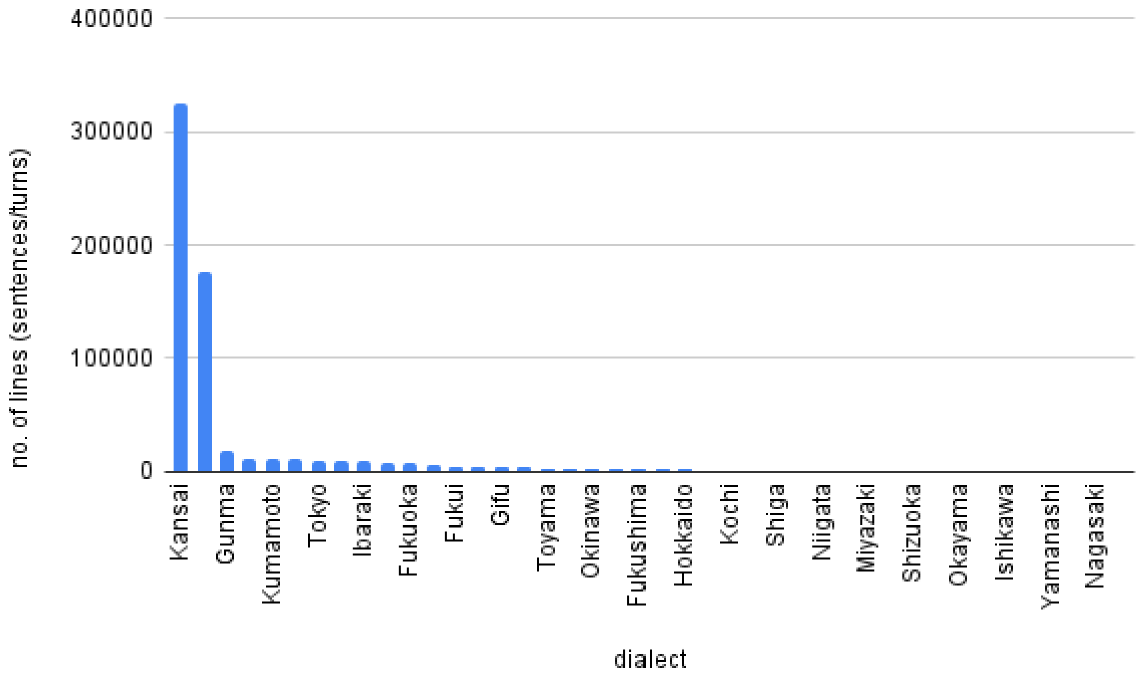

Figure 2: Number of lines (sentences/turns) per dialect in our dataset

3.4. Data distribution

In the next stages of our study, removing empty lines was needed. Possessing all data in their *katakana* version, we were able to incorporate them into relevant classes. We have adhered to the division used in COJADS to a large extent as grouping dialects by prefectures can provide a convenient framework for study. However, due to the apparent dominance of *hyōjungo* and the Kansai dialect, we decided to combine Nara, Kyoto, Osaka and Hyōgo based texts into one data, as Kansai-ben. These dialects share many features, which can make them more compatible for grouping. Due to scarcity of data, dialect imbalances in our dataset were, for the time being, unavoidable. On the other hand, dataset constraints made it impossible to capture dialects spoken in territories more extensive or narrower than prefectures and between different social groups. For instance, Tsugaru dialect (hereafter, Tsugaru-ben), spoken on the west side of Aomori prefecture [9], was not mentioned as specified group, but still covers some of the data from class called Aomori. This choice, with all of its advantages and disadvantages, enabled us to start testing. We also made a decision to include *hyōjungo* in our research. It was developed from the educated middle-upper-class Tokyo dialect for the purpose of spreading a sense of a nation unity within the country in the Meiji era (1868-1912) [39]. This denotes that *hyōjungo* is also a variation of a language. It is believed that recognizing the standard language as one of many varieties, among that each with separate rules and functions, enables scholars to perceive it within the broader context of language variations, meaning dialects and social variants [40]. We also felt an obligation to discriminate between *hyōjungo* and the rest of the data operated under the name of Tokyo. In COJADS, Tokyo was treated purely territorially, but authors provided users information about particular places (namely: district, villages, towns etc.). It could be easily observed that language used within a metropolis was standardized, while language spoken in rural areas near Tokyo (like, for example, Hinohara village) possessed noticeable differences from it. That is why we draw a line between those two.

Next stage consisted in making dialect distribution. We checked the number of lines (sentences/turns) per each dialect area. The greatest number was observed in the Kansai dialect (more than 300,000 lines),

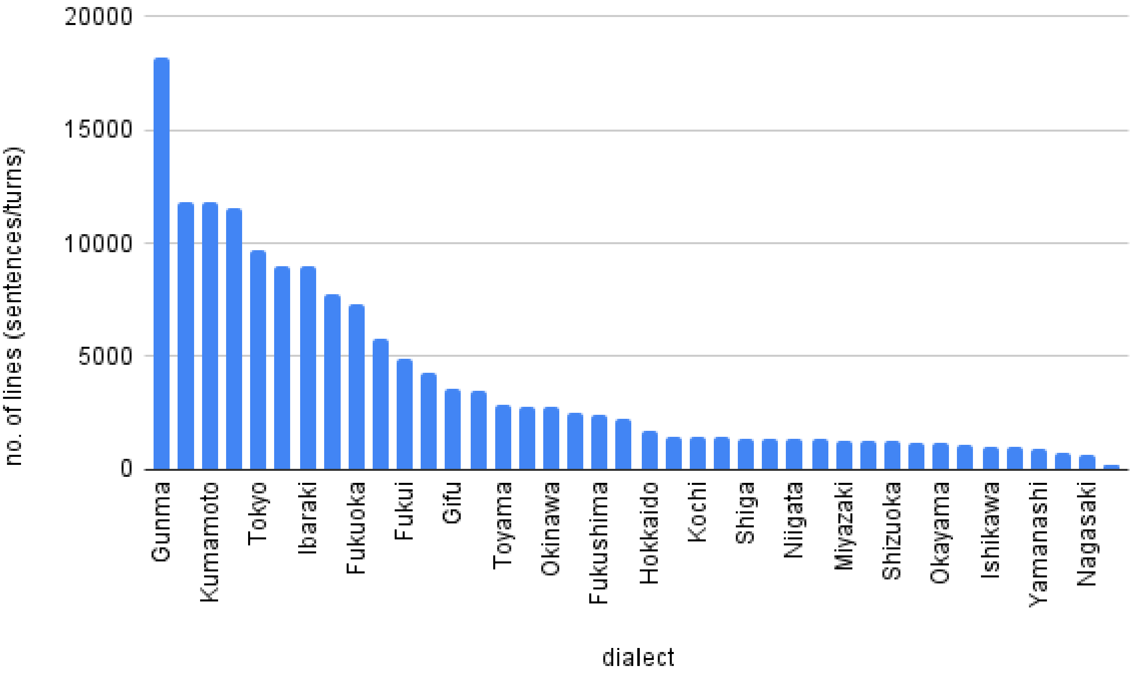

Figure 3: Number of lines (sentences/turns) per dialect in our dataset without Kansai and *hyōjungo*

[9] See more here: https://tsugaruben.net [access: 2024/10/11].

followed by *hyōjungo* that noted only 175,849 lines due to the division we have applied. There is a significant gap between those two classes and the rest of the groups. Remaining groups did not even reach 100,000 lines or at least 50,000 lines. Third place went to Gunma, with 18,165 lines. The lowest number of lines was recorded in Yamanashi, Saga, Nagasaki and Tottori, respectively. Tottori had less than 200 lines. Next step involved an investigation of the average length of each line, calculating the average length of sentence for each dialect and finally checking their standard deviation as it is known as a universal and functional measure as it shows the average distance of each score from the mean.

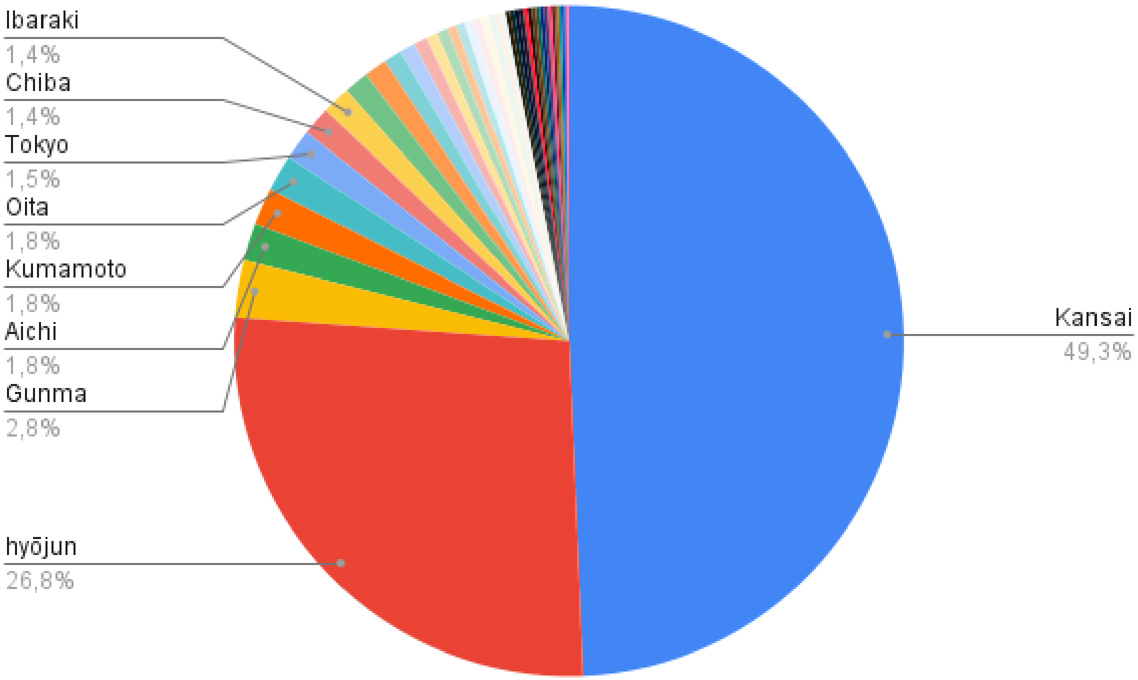

Figure 4: Dialect data distribution

4. Experimental results

First of all, in order to create a Japanese dialect-aware chatbot, we need a dialect identification model that will be able to detect dialectal text written in *katakana* quickly and successfully. Texts are categorized by dialect/area, and the neural network is trained on *katakana* representations for each dialect category. In the experiments, for training we used 525,569 out of 656,826 sentences, while the rest of them was left for testing. All four models were trained using learning rate 5e-05 and a batch size of 32. We set 3 as a number of epochs. In the first place, we trained four models with preprocessed data:

- **tohoku-nlp/bert-base-japanese**[10]: A BERT model pre-trained on a Japanese corpus and trained on Japanese Wikipedia.
- **tohoku-nlp/bert-large-japanese**[11]: A bigger BERT model that contains around 30M sentences.
- **tohoku-nlp/bert-large-japanese-char**[12]: A character-level version of the larger BERT model, trained with the whole word masking enabled for the masked language modeling (MLM) objective.

[10] https://huggingface.co/tohoku-nlp/bert-base-japanese
[11] https://huggingface.co/tohoku-nlp/bert-large-japanese
[12] https://huggingface.co/tohoku-nlp/bert-large-japanese-char

- **tohoku-nlp/bert-large-japanese-v2**[13]: A word-level variant of the large BERT model, enriched by the Japanese portion of CC-100 dataset that incorporates approximately 392M sentences alone.

Our evaluation measures, namely Precision, Recall and F1 Score, are widely used for Text Classification [41]. Their formulas demonstrate the significance of retrieval of positive examples in text classification. Table 1 presents evaluation results for models mentioned above.

All four models were based on Bidirectional Encoder Representations from Transformers (BERT) [42] - language framework that deals with pretraining techniques such as Mask Language Modeling (MLM) and Next Sentence Prediction (NSP) [43]. The tohoku-nlp/bert-base-japanese model entirely outperforms others, reaching the highest overall F1 score. The tohoku-nlp/bert-large-japanese model achieved, surprisingly, slightly lower scores in comparison to the base version, despite possessing more parameters. Moreover, the tohoku-nlp/bert-large-japanese-char model performance is poor across almost all dialects. It may suggest that the character-level approach is not effective for this task. The tohoku-nlp/bert-large-japanese-v2 model showed some improvement over the character-level model. However, it still noticeably shows small effectiveness compared to the base and large models.

After getting unfavorable results, we decided to compare the performance of these with two other models:

- **google-bert_bert-base-multilingual-uncased**[14]: trained on multilingual data, does not make distinction between capital and lowercase letters. It was described in detail by Jacob Delvin et al. [42]

- **google-bert_bert-base-multilingual-cased**[15]: almost identical model, but prepared to detect differences between uppercase and lowercase letters.

Letter case might can heavily impact meaning in some languages. Texts in languages like Japanese, Chinese or Korean usually do not have any spaces between words. This fact was considered during the development process.

The results are summarized in Table 2. Both models performed very well dealing with *hyōjungo* and Kansai dialect. Strong performance could be also observed for Okinawian dialect, especially high F1 score for cased model. A score of over 0.50 was achieved by Aomori and Oita for both models. Both for tohoku-nlp models as well as google-bert models, poor performance across all metrics can be noted among Tottori, Tochigi, Shiga and Hiroshima regions. Most of the dialects can be found in the group of dialects with visible moderate performance, rating F1 score between 0.20 and 0.40.

Significant disparity between Precision and Recall might be observed within some dialects, like Fukuoka (for models 1, 2, 5 and 6). Same thing applies to Gifu. Discrepancy is highly noticeable among dialects with overall bad performance that mark very low precision and even lower recall, like Aichi, Niigata, Kanagawa or Mie.

5. Discussion

Based on these results, we can observe a significantly strong correlation between results and amount of data. Kansai dialect, with the largest amount of sentences, reaches the best results whereas low-sourced dialects note scores close to 0. It indicates the need for more extensive data collection and consequently searching for more training data in general. The google-bert_bert-base-multilingual-uncased model generally performs better than the cased model, especially in dialects with lower performance. This leads to an assumption that case insensitivity might be helpful in effectively handling diverse language forms.

[13] https://huggingface.co/tohoku-nlp/bert-large-japanese-v2
[14] https://huggingface.co/google-bert/bert-base-multilingual-uncased
[15] https://huggingface.co/google-bert/bert-base-multilingual-cased

Table 1

Evaluation results. Model 1: tohoku-nlp/bert-base-japanese. Model 2: tohoku-nlp/bert-large-japanese. Model 3: tohoku-nlp/bert-large-japanese-char-v2. Model 4: tohoku-nlp/bert-large-japanese-v2.

Number of	**Class**	**Dialect/Area**	**Model 1**			**Model 2**			**Model 3**			**Model 4**		
sentences/turns			**Precision**	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**
140682	0	hyōjungo	0.72	0.90	0.80	0.72	0.90	0.80	0.27	1.00	0.42	0.68	0.83	0.75
9431	1	Aichi	0.14	0.06	0.08	0.12	0.05	0.08	0.00	0.00	0.00	0.00	0.00	0.00
1805	2	Akita	0.50	0.24	0.32	0.39	0.19	0.25	0.00	0.00	0.00	0.00	0.00	0.00
6217	3	Aomori	0.70	0.57	0.63	0.69	0.55	0.61	0.00	0.00	0.00	0.00	0.00	0.00
7187	4	Chiba	0.29	0.14	0.19	0.27	0.13	0.18	0.00	0.00	0.00	0.00	0.00	0.00
1171	5	Ehime	0.16	0.07	0.10	0.19	0.07	0.10	0.00	0.00	0.00	0.00	0.00	0.00
3906	6	Fukui	0.25	0.18	0.21	0.25	0.16	0.20	0.00	0.00	0.00	0.00	0.00	0.00
5824	7	Fukuoka	0.50	0.33	0.40	0.52	0.32	0.40	0.00	0.00	0.00	0.00	0.00	0.00
1933	8	Fukushima	0.53	0.26	0.34	0.42	0.23	0.30	0.00	0.00	0.00	0.00	0.00	0.00
2892	9	Gifu	0.46	0.30	0.36	0.42	0.28	0.34	0.00	0.00	0.00	0.00	0.00	0.00
14535	10	Gunma	0.59	0.60	0.59	0.58	0.59	0.59	0.00	0.00	0.00	0.00	0.00	0.00
2265	11	Hiroshima	0.03	0.03	0.03	0.04	0.02	0.02	0.00	0.00	0.00	0.00	0.00	0.00
1349	12	Hokkaido	0.24	0.12	0.16	0.14	0.05	0.07	0.00	0.00	0.00	0.00	0.00	0.00
7159	13	Ibaraki	0.36	0.15	0.21	0.33	0.14	0.20	0.00	0.00	0.00	0.00	0.00	0.00
813	14	Ishikawa	0.52	0.32	0.39	0.33	0.17	0.22	0.00	0.00	0.00	0.00	0.00	0.00
1094	15	Kagawa	0.51	0.21	0.30	0.42	0.16	0.23	0.00	0.00	0.00	0.00	0.00	0.00
1995	16	Kagoshima	0.11	0.06	0.08	0.15	0.09	0.11	0.00	0.00	0.00	0.00	0.00	0.00
3465	17	Kanagawa	0.42	0.30	0.35	0.37	0.26	0.30	0.00	0.00	0.00	0.00	0.00	0.00
259283	18	Kansai	0.87	0.86	0.87	0.86	0.86	0.86	0.00	0.00	0.00	0.66	0.91	0.77
1154	19	Kochi	0.15	0.05	0.07	0.10	0.03	0.05	0.00	0.00	0.00	0.00	0.00	0.00
2800	20	Mie	0.39	0.19	0.26	0.35	0.15	0.21	0.00	0.00	0.00	0.00	0.00	0.00
1054	21	Miyazaki	0.23	0.35	0.28	0.17	0.23	0.19	0.00	0.00	0.00	0.00	0.00	0.00
4603	22	Nagano	0.34	0.44	0.38	0.32	0.44	0.37	0.00	0.00	0.00	0.00	0.00	0.00
538	23	Nagasaki	0.32	0.19	0.24	0.24	0.14	0.18	0.00	0.00	0.00	0.00	0.00	0.00
1086	24	Niigata	0.36	0.15	0.21	0.35	0.16	0.22	0.00	0.00	0.00	0.00	0.00	0.00
9269	25	Oita	0.64	0.44	0.53	0.63	0.44	0.52	0.00	0.00	0.00	0.00	0.00	0.00
960	26	Okayama	0.59	0.42	0.49	0.56	0.40	0.47	0.00	0.00	0.00	0.00	0.00	0.00
2203	27	Okinawa	0.66	0.72	0.69	0.73	0.69	0.71	0.00	0.00	0.00	0.00	0.00	0.00
584	28	Saga	0.54	0.39	0.45	0.40	0.34	0.37	0.00	0.00	0.00	0.00	0.00	0.00
1079	29	Saitama	0.11	0.10	0.11	0.08	0.03	0.04	0.00	0.00	0.00	0.00	0.00	0.00
1123	30	Shiga	0.19	0.04	0.07	0.15	0.03	0.04	0.00	0.00	0.00	0.00	0.00	0.00
1030	31	Shimane	0.37	0.24	0.29	0.29	0.15	0.19	0.00	0.00	0.00	0.00	0.00	0.00
992	32	Shizuoka	0.58	0.42	0.49	0.54	0.32	0.40	0.00	0.00	0.00	0.00	0.00	0.00
867	33	Tochigi	0.08	0.02	0.03	0.11	0.02	0.03	0.00	0.00	0.00	0.00	0.00	0.00
1139	34	Tokushima	0.13	0.06	0.08	0.11	0.05	0.07	0.00	0.00	0.00	0.00	0.00	0.00
144	35	Tottori	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2278	36	Toyama	0.25	0.21	0.23	0.21	0.20	0.20	0.00	0.00	0.00	0.00	0.00	0.00
801	37	Wakayama	0.23	0.09	0.13	0.19	0.06	0.09	0.00	0.00	0.00	0.00	0.00	0.00
967	38	Yamaguchi	0.51	0.18	0.27	0.50	0.12	0.20	0.00	0.00	0.00	0.00	0.00	0.00
746	39	Yamanashi	0.39	0.06	0.10	0.17	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00
9428	40	Kumamoto	0.26	0.33	0.29	0.25	0.31	0.28	0.00	0.00	0.00	0.00	0.00	0.00
7718	41	Tokyo	0.27	0.24	0.25	0.25	0.24	0.25	0.00	0.00	0.00	0.00	0.00	0.00
OVERALL SCORE			**0.37**	**0.26**	**0.29**	**0.33**	**0.23**	**0.26**	**0.01**	**0.02**	**0.01**	**0.03**	**0.04**	**0.04**

Table 2
Evaluation results. Model 5: google-bert_bert-base-multilingual-uncased, Model 6: google-bert_bert-base-multilingual- cased.

Number of	Class	Dialect/Area	Model 5			Model 6		
sentences/turns			**Precision**	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**
140682	0	hyōjungo	0.72	0.90	0.80	0.72	0.91	0.80
9431	1	Aichi	0.11	0.05	0.07	0.11	0.05	0.07
1805	2	Akita	0.42	0.19	0.26	0.27	0.17	0.21
6217	3	Aomori	0.67	0.56	0.61	0.73	0.55	0.63
7187	4	Chiba	0.31	0.15	0.20	0.32	0.13	0.19
1171	5	Ehime	0.14	0.04	0.07	0.13	0.03	0.05
3906	6	Fukui	0.24	0.18	0.21	0.29	0.17	0.21
5824	7	Fukuoka	0.51	0.33	0.40	0.51	0.31	0.39
1933	8	Fukushima	0.40	0.17	0.24	0.35	0.09	0.14
2892	9	Gifu	0.53	0.31	0.39	0.49	0.26	0.34
14535	10	Gunma	0.56	0.61	0.58	0.56	0.63	0.60
2265	11	Hiroshima	0.04	0.02	0.03	0.05	0.01	0.01
1349	12	Hokkaido	0.19	0.06	0.09	0.10	0.03	0.05
7159	13	Ibaraki	0.34	0.18	0.23	0.36	0.16	0.22
813	14	Ishikawa	0.41	0.18	0.25	0.00	0.00	0.00
1094	15	Kagawa	0.38	0.18	0.25	0.33	0.15	0.21
1995	16	Kagoshima	0.08	0.05	0.06	0.13	0.07	0.09
3465	17	Kanagawa	0.36	0.20	0.25	0.38	0.19	0.25
259283	18	Kansai	0.86	0.86	0.86	0.86	0.87	0.86
1154	19	Kochi	0.08	0.03	0.04	0.19	0.03	0.05
2800	20	Mie	0.42	0.20	0.27	0.39	0.20	0.26
1054	21	Miyazaki	0.22	0.34	0.27	0.14	0.33	0.19
4603	22	Nagano	0.32	0.40	0.36	0.30	0.39	0.34
538	23	Nagasaki	0.25	0.12	0.16	0.18	0.02	0.03
1086	24	Niigata	0.45	0.12	0.19	0.42	0.08	0.13
9269	25	Oita	0.63	0.41	0.50	0.62	0.43	0.51
960	26	Okayama	0.53	0.38	0.44	0.35	0.12	0.18
2203	27	Okinawa	0.69	0.69	0.69	0.87	0.68	0.76
584	28	Saga	0.50	0.37	0.43	0.39	0.29	0.33
1079	29	Saitama	0.11	0.08	0.09	0.13	0.03	0.05
1123	30	Shiga	0.20	0.01	0.02	0.00	0.00	0.00
1030	31	Shimane	0.29	0.20	0.24	0.05	0.02	0.03
992	32	Shizuoka	0.59	0.33	0.42	0.18	0.11	0.13
867	33	Tochigi	0.40	0.01	0.02	0.00	0.00	0.00
1139	34	Tokushima	0.13	0.06	0.09	0.01	0.00	0.01
144	35	Tottori	0.00	0.00	0.00	0.00	0.00	0.00
2278	36	Toyama	0.25	0.19	0.22	0.19	0.15	0.17
801	37	Wakayama	0.25	0.04	0.07	0.00	0.00	0.00
967	38	Yamaguchi	0.46	0.16	0.24	0.63	0.12	0.20
746	39	Yamanashi	0.20	0.01	0.02	0.00	0.00	0.00
9428	40	Kumamoto	0.28	0.36	0.31	0.26	0.36	0.26
7718	41	Tokyo	0.26	0.22	0.24	0.28	0.26	0.27
	OVERALL SCORE		**0.35**	**0.24**	**0.27**	**0.29**	**0.20**	**0.22**

An interesting case in an Okinawa region that reaches a good performance, despite the small amount of data. The reasons for this may be found in many differences between Japanese and Ryukyuan languages. It is essential to notice that Ryukyuan languages have been downgraded to the status of Japanese dialects [44] despite being mutually unintelligible in many aspects with standard Japanese. Aleksandra Jarosz [45] as major features of Ryukyuan languages points out, among others, the raising of short mid-close vowels and a phonological appearance of the glottal stop. The question is, whether other models will achieve similar ratings for Okinawa or is it only used models specificity.

Striking variability between Precision and Recall results was noticeable in some classes. Recall is the proportion of Real Positive cases that are accurately Predicted Positive while Precision represents the proportion of predicted positive cases that are correctly identified as true positives [46]. Imbalance between them might suggest that while some models are fairly effective at recognizing instances they can detect, they still fail to capture a significant amount of relevant dialect data. This situation might a a consequence of a lack of sufficient feature representation. Moreover, results achieved by some classes like Aichi, Niigata or Mie indicate that models are often incorrect with their predictions. Their low recall reflects ineffectiveness in identifying the majority of actual instances.

For certain dialects, specialization might be necessary since specialized models can be fine-tuned with dialect-specific data, resulting in higher Precision, Recall, and F1 scores for dialects with poor performance. This poses a dilemma of whether to create several specialized models for various dialects or a single, highly generalized model.

6. Conclusions

In our study we monitored the performance of some models working on Japanese dialects. We presented the results and outlined conclusions. We also introduced some already existing dialect-aware chatbots and chatbots that were made with the thought of low-rescourced languages. The data collection process was also explained in detail.

Models successfully dealt with standard Japanese and Kansai dialect, but showed an inability in detecting less-resourced dialects. This contrast suggests that standard NLP models might have problems in generalizing across dialectal features without adequate data representation. This stresses also the need for more targeted training or extensive data collection. Moreover, it gave us a hint to try implementing other learning rates and scrutinize what impact they will have on the model performance.

Our study highlights the challenges in dialect detection due to data scarcity, emphasizing the complexity of dialect detection within Japanese language. Facing mentioned challenges is crucial for advancing NLP systems that can accurately reflect the rich linguistic landscape of Japan.

7. Future work

In the future, besides learning rates implementation, we plan to check the performance of some other models in order to have a comparison and choose the best performing model as a way to focus on its application, namely Japanese dialect-aware chatbot. This paper revolved around dialect recognition, but it cannot be excluded that in the future we will focus on style shifts within dialects. Style adaptation might greatly influence the chatbot performance and users‘ impression. With the increasing need of personalization [47], incorporating both dialect recognition and the ability to adjust for conversational style can provide a more customized and responsive user experience.

In the case of Japanese, formality's level of the language plays an important role in customer service and everyday relationships. Speaker‘s awareness of his position in particular situation in Japan is reflected by honorific language, *keigo* [48]. However, among dialects *keigo* expressions can also differ. On the other hand, there are also regions, like southern part of the Tohoku Region, that do not have the linguistic forms corresponding to the honorific and humble forms of the national lingua franca, and where honorific expressions are considered to be uncommon[16]. Hence, profound research on honorific forms among dialects will be needed. Taking into consideration the scarcity of dialectal data, for the time being successful distinguishing between *keigo* expressions within dialects might be impossible. Nonetheless, collecting dialectal data focused on honorific forms might be another goal to be set.

References

[1] N. Bhirud, S. Tataale, S. Randive, S. Nahar, A literature review on chatbots in healthcare domain, International journal of scientific & technology research 8 (2019) 225–231.

[2] K. Patil, M. S. Kulkarni, Artificial intelligence in financial services: Customer chatbot advisor adoption, Int. J. Innov. Technol. Explor. Eng 9 (2019) 4296–4303.

[3] A. Følstad, C. Nordheim, C. Bjørkli, What makes users trust a chatbot for customer service? an exploratory interview study, in: Internet Science, 2018, pp. 194–208. doi:10.1007/978-3-030-01437-7_16.

[16] *Keigo no shishin* (Guide to *keigo*), Bunkashingikai (Agency for Cultural Affairs), 2019: https://www.bunka.go.jp/seisaku/bunkashingikai/kokugo/hokoku/pdf/keigo_tosin.pdf, [access: 2024/10/8].

[4] A. P. Chaves, M. A. Gerosa, How should my chatbot interact? a survey on social characteristics in human–chatbot interaction design, International Journal of Human–Computer Interaction 37 (2019) 729 – 758. URL: https://api.semanticscholar.org/CorpusID:102350801.

[5] M. Amadeus, J. R. Homeli da Silva, J. V. Pessoa Rocha, Bridging the language gap: Integrating language variations into conversational AI agents for enhanced user engagement, in: N. Hosseini-Kivanani, S. Höhn, D. Anastasiou, B. Migge, A. Soltan, D. Dippold, E. Kamlovskaya, F. Philippy (Eds.), Proceedings of the 1st Worskhop on Towards Ethical and Inclusive Conversational AI: Language Attitudes, Linguistic Diversity, and Language Rights (TEICAI 2024), Association for Computational Linguistics, St Julians, Malta, 2024, pp. 16–20. URL: https://aclanthology.org/2024.teicai-1.3.

[6] A. Martin, K. Jenkins, Speaking your language: The psychological impact of dialect integration in artificial intelligence systems, Current Opinion in Psychology 58 (2024) 101840. doi:10.1016/j.copsyc.2024.101840.

[7] A. Martin, K. Jenkins, Speaking your language: The psychological impact of dialect integration in artificial intelligence systems, Current Opinion in Psychology (2024) 101840.

[8] Y. Kawaguchi, F. Inoue, Japanese dialectology in historical perspectives, Revue belge de philologie et d'histoire 80 (2002) 801–829. doi:10.3406/rbph.2002.4642.

[9] S. Aggarwal, S. Mehra, P. Mitra, Multi-purpose nlp chatbot : Design, methodology & conclusion, ArXiv abs/2310.08977 (2023). URL: https://api.semanticscholar.org/CorpusID:264127993.

[10] A. P. Chaves, J. Egbert, T. Hocking, E. Doerry, M. A. Gerosa, Chatbots language design: The influence of language variation on user experience with tourist assistant chatbots, ACM Trans. Comput.-Hum. Interact. 29 (2022). URL: https://doi.org/10.1145/3487193. doi:10.1145/3487193.

[11] E. Elsholz, J. Chamberlain, U. Kruschwitz, Exploring language style in chatbots to increase perceived product value and user engagement, in: Proceedings of the 2019 Conference on Human Information Interaction and Retrieval, CHIIR '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 301–305. URL: https://doi.org/10.1145/3295750.3298956. doi:10.1145/3295750.3298956.

[12] R. Hoegen, D. Aneja, D. McDuff, M. Czerwinski, An end-to-end conversational style matching agent, in: Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents, IVA '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 111–118. URL: https://doi.org/10.1145/3308532.3329473. doi:10.1145/3308532.3329473.

[13] O. T. K. Tran, T. C. Luong, Understanding what the users say in chatbots: A case study for the vietnamese language, Eng. Appl. Artif. Intell. 87 (2020). URL: https://api.semanticscholar.org/CorpusID:208947162.

[14] N. N. Chiaráin, A. N. Chasaide, Chatbot technology with synthetic voices in the acquisition of an endangered language: Motivation, development and evaluation of a platform for Irish, in: N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), European Language Resources Association (ELRA), Portorož, Slovenia, 2016, pp. 3429–3435. URL: https://aclanthology.org/L16-1547.

[15] A. George, G. Muralikrishnan, L. Ninan, P. Varrier, D. L K, Survey on the design and development of indian language chatbots, 2021 International Conference on Communication, Control and Information Sciences (ICCISc) (2022) 1–6. doi:10.1109/ICCISc52257.2021.9484891.

[16] M. Gammoudi, S. Yassine, Trends and challenges of arabic chatbots: Literature review, Jordanian Journal of Computers and Information Technology 9 (2023) 1. doi:10.5455/jjcit.71-1685381801.

[17] A. Aliwy, H. Taher, Z. AboAltaheen, Arabic dialects identification for all Arabic countries, in: I. Zitouni, M. Abdul-Mageed, H. Bouamor, F. Bougares, M. El-Haj, N. Tomeh, W. Zaghouani (Eds.), Proceedings of the Fifth Arabic Natural Language Processing Workshop, Association for Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 302–307. URL: https://aclanthology.org/2020.wanlp-1.32.

[18] D. Al-Ghadhban, N. Al-Twairesh, Nabiha: An arabic dialect chatbot, International Journal of

Advanced Computer Science and Applications 11 (2020). doi:10.14569/IJACSA.2020.0110357.

[19] D. Abu Ali, N. Habash, Botta: An Arabic dialect chatbot, in: H. Watanabe (Ed.), Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations, The COLING 2016 Organizing Committee, Osaka, Japan, 2016, pp. 208–212. URL: https://aclanthology.org/C16-2044.

[20] A. Boulesnane, Y. Saidi, O. Kamel, M. M. Bouhamed, R. Mennour, Dzchatbot: A medical assistant chatbot in the algerian arabic dialect using seq2seq model, in: 2022 4th international conference on pattern analysis and intelligent systems (PAIS), 2022, pp. 1–8. doi:10.1109/PAIS56586.2022.9946867.

[21] M. Abdelhay, A. Mohammed, H. Hefny, Deep learning for arabic healthcare: Medicalbot, Social Network Analysis and Mining 13 (2023) 71. doi:10.1007/s13278-023-01077-w.

[22] S. Rekik, M. Elamine, L. H. Belguith, A medical chatbot for tunisian dialect using a rule-based and machine learning approach, in: 20th ACS/IEEE International Conference on Computer Systems and Applications, AICCSA 2023, Giza, Egypt, December 4-7, 2023, IEEE, 2023, pp. 1–7. URL: https://doi.org/10.1109/AICCSA59173.2023.10479249. doi:10.1109/AICCSA59173.2023.10479249.

[23] A. Joukhadar, H. Saghergy, L. Kweider, N. Ghneim, Arabic dialogue act recognition for textual chatbot systems, in: Proceedings of The First International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2019) co-located with ICNLSP 2019-Short Papers, 2019, pp. 43–49.

[24] M. Kowsher, F. S. Tithi, M. A. Alam, M. N. Huda, M. M. Moheuddin, M. G. Rosul, Doly: Bengali chatbot for bengali education, in: 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), 2019, pp. 1–6. URL: https://api.semanticscholar.org/CorpusID:209460125.

[25] M. Kaleem, J. O'Shea, K. Crockett, Development of umair the urdu conversational agent for customer service, Lecture Notes in Engineering and Computer Science 1 (2014) 86–91.

[26] J. Shabbir, M. U. Arshad, W. Shahzad, Nubot: Embedded knowledge graph with rasa framework for generating semantic intents responses in roman urdu, ArXiv abs/2102.10410 (2021). URL: https://api.semanticscholar.org/CorpusID:231986364.

[27] J. Brixey, D. R. Traum, Masheli: A choctaw-english bilingual chatbot, in: International Workshop on Spoken Dialogue Systems Technology, 2020, pp. 1–9. URL: https://api.semanticscholar.org/CorpusID:226325503.

[28] A. B. E. Mabrouk, M. B. H. Hmida, C. Fourati, H. Haddad, A. Messaoudi, A multilingual african embedding for faq chatbots, ArXiv abs/2103.09185 (2021). URL: https://api.semanticscholar.org/CorpusID:232240431.

[29] S. Sarma, N. Pathak, Shiksha mitra: An assamese language ai chatbot using deep learning, International Journal of Scientific Research in Computer Science Engineering and Information Technology 9 (2023) 48–57. doi:10.32628/CSEIT2390572.

[30] G. Y. Hailu, S. Welay, Deep learning based amharic chatbot for faqs in universities, arXiv preprint arXiv:2402.01720 (2024).

[31] S. Sandhini, R. Binu, R. Rajeev, M. Reshma, A proposal of chatbot for malayalam, International Journal of Computer Sciences and Engineering 06 (2018) 21–25. doi:10.26438/ijcse/v6si6.2125.

[32] A. Joshi, R. Dabre, D. Kanojia, Z. Li, H. Zhan, G. Haffari, D. Dippold, Natural language processing for dialects of a language: A survey, ArXiv abs/2401.05632 (2024). URL: https://api.semanticscholar.org/CorpusID:266933497.

[33] M. Orosoo, I. Goswami, F. R. Alphonse, G. Fatma, M. Rengarajan, B. Kiran Bala, Enhancing natural language processing in multilingual chatbots for cross-cultural communication, in: 2024 5th International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), 2024, pp. 127–133. doi:10.1109/ICICV62344.2024.00027.

[34] K. Heffernan, < research note> an introduction to the corpus of kansai spoken japanese, Journal of Policy Studies (2012) 157–163.

[35] S. Takamichi, H. Saruwatari, CPJD corpus: Crowdsourced parallel speech corpus of Japanese dialects, in: N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara,

B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, T. Tokunaga (Eds.), Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan, 2018, pp. 434–437. URL: https://aclanthology.org/L18-1067.

[36] W. Hadamitzky, M. Spahn, Japanese Kanji & Kana:(JLPT All Levels) A Complete Guide to the Japanese Writing System (2,136 Kanji and 92 Kana), Tuttle Publishing, 2013.

[37] I. Fumiko, A Study of Regional and Generation Differences in Discourse Pattern, Technical Report, The National Institute for Japanese Language and Linguistics, 2014. Application/pdf, NINJAL Collaborative Research Project Reports ; 13-04.

[38] Y. Sim, A morphological analyzer for japanese nouns, verbs and adjectives, ArXiv abs/1410.0291 (2014). URL: https://api.semanticscholar.org/CorpusID:15195083.

[39] S. Saito, M. Turner, Language diversity in "monolingual" japan: Language awareness among high school teachers of english, Journal of Language, Identity & Education (2024) 1–15. URL: http://dx.doi.org/10.1080/15348458.2024.2385837. doi:10.1080/15348458.2024.2385837.

[40] R. Quirk, Language varieties and standard language, English Today 6 (1990) 3–10. doi:10.1017/S0266078400004454.

[41] M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks, Information Processing & Management 45 (2009) 427–437. doi:10.1016/j.ipm.2009.03.002.

[42] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[43] C. Zhou, Q. Li, C. Li, J. Yu, Y. Liu, G. Wang, K. Zhang, C. Ji, Q. Yan, L. He, et al., A comprehensive survey on pretrained foundation models: A history from bert to chatgpt, arXiv preprint arXiv:2302.09419 (2023).

[44] P. Heinrich, S. Miyara, M. Shimoji, Handbook of the Ryukyuan Languages: History, Structure, and Use, 2015. doi:10.1515/9781614511151.

[45] A. Jarosz, Japonic languages: an overview, Silva Iaponicarum (2017). doi:10.14746/sijp.2017.41/42.6.

[46] D. M. Powers, Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation, arXiv preprint arXiv:2010.16061 (2020).

[47] T. Ait Baha, M. El Hajji, Y. Es-Saady, H. Fadili, The power of personalization: A systematic review of personality-adaptive chatbots, SN Computer Science 4 (2023). URL: http://dx.doi.org/10.1007/s42979-023-02092-6. doi:10.1007/s42979-023-02092-6.

[48] E. T. Rahayu, Japanese honorific language in various domains, in: Proceedings of the Fourth Prasasti International Seminar on Linguistics (Prasasti 2018), Atlantis Press, 2018/08, pp. 25–34. URL: https://doi.org/10.2991/prasasti-18.2018.5. doi:10.2991/prasasti-18.2018.5.