Developing English Learning Support Robot Service Using Multimodal Teachers' Pedagogy

Hidenao ABE

Department of Information Systems,
Faculty of Information and Communications,
Bunkyo University（文教大学）

hidenao@bunkyo.ac.jp

Research Background & Summary

- Continuing Development of pattern recognition and machine learning algorithm environments through cloud services
  - Performance of image and speech recognition has improved to be on par with human recognition
- Clarification of individual action rule application knowledge (meta-knowledge) through the development of robot services based on task knowledge
- Development of a robot service and class support system based on a multimodal rule base aimed at supporting English classes

Practical Examples of Existing English Learning Support Robots and Services

- Musio (Toda City, Saitama Prefecture, etc.)
  - English learning robot for elementary education
  - Speech practice through voice recognition
  - Teaching materials and pedagogy set and created by the company

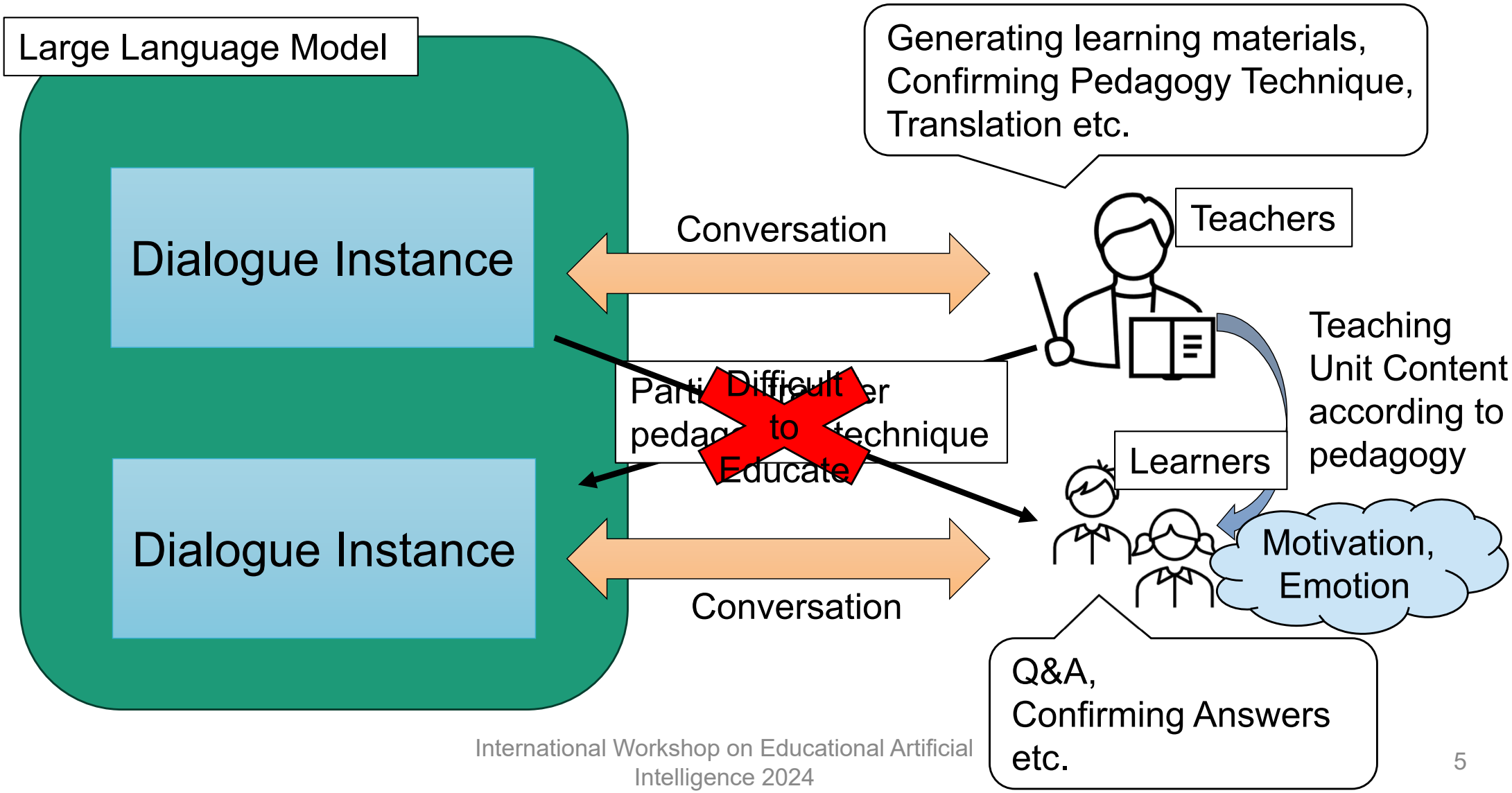
- Torepa (Private high schools in Kanagawa Prefecture, etc.)
  - English learning service for secondary education
  - Speech practice and problem exercises through voice recognition -> Creation of progress statistics
  - Teaching materials can be created by teachers, pedagogy is handled by

Teachers cannot set theirown pedagogy (questions and follow-ups for learners) or it is left to the teachers

Conversational AI (Large Language Models) vs. "Good" Teachers (Educators)

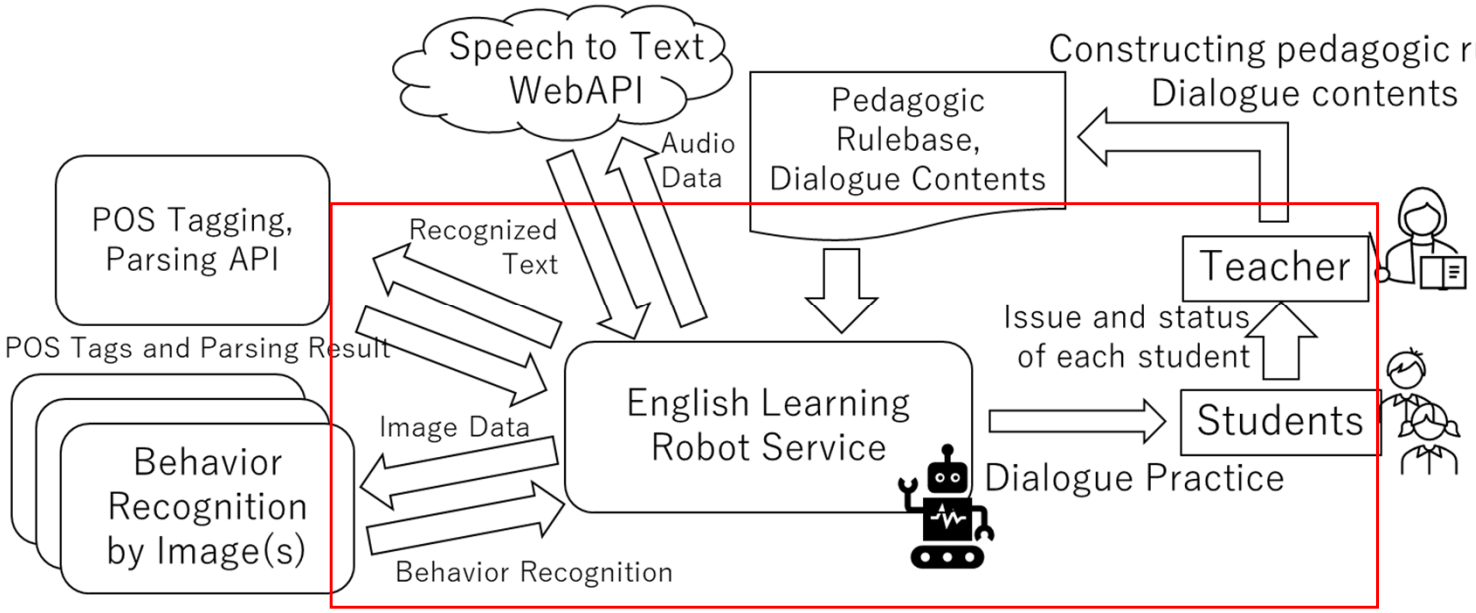
- Conversational AI based on large language resources
  - Multitask dialogue possible with models like Chat GPT based on GPT models
  - Generates responses that may include mistakes or nonsense
  - Humans need to detect and correct mistakes (teach the correct answers)
  - Good questions lead to good answers (dialogue AI and learners are equal)
- "Good" Teachers (Educators)
  - Requires appropriate and flexible responses based on correct subject knowledge and pedagogy in classes
  - Enhances learners' abilities through appropriate responses (detects and corrects mistakes) and motivates them
  - Not "Instruction" but "Education"
  - Appropriate and flexible responses based on pedagogy are acquired through research lessons and mock lessons
- Teachers (educators) need skills accompanied by meta-knowledge (from the learners' perspective)

Relationship Between Large Language Models in Dialogue Practice and Teachers & Learners



Research Purpose

- Development of a hierarchical multimodal teacher's business rule base description format and input interface
  → Description of user model
- Development of an educational support robot service using a multimodal teacher's business rule base



Related Work

- Evaluation of learning effects through question-and-answer format with robots [Muramoto 2023]
  - Mid-term retention improved through question-and-answer format learning compared to repeating learning (e.g., Audio-Lingual Method)
- Skills required for ICT utilization in dialogue practice (English) [Compton 2009]
  - Skills to utilize ICT itself
  - Skills to handle pedagogy
  - Skills to evaluate the retention of subject content

[Compton2009]Compton, L. K. L. (2009). Preparing language teachers to teach language online: A look at skills, roles, and responsibilities. Computer Assisted Language Learning, 22, 1, 73–99. doi:10.1080/09588220802613831

Previous Work

[Akimoto2018] Momoko Akimoto, Hidenao Abe, Yuko Ikuta, Takashi Morita, and Takahira Yamaguchi: Implementation and Evaluation of Pronunciation Practice in English by Using Interactive Robot and Pedagogic Process Rule Analysis, Proceedings of the Information Education Symposium, No. 26, pp. 185-188 (2018) (In Japanese).
[Akimoto2019] 秋本桃子, 阿部秀尚, 森田武史, 山口高平：対話型ロボットサービスにおける教師業務ルール実装のための基本動作認識システムの開発, 人工知能学会 第117回知識ベースシステム研究会, 2019.

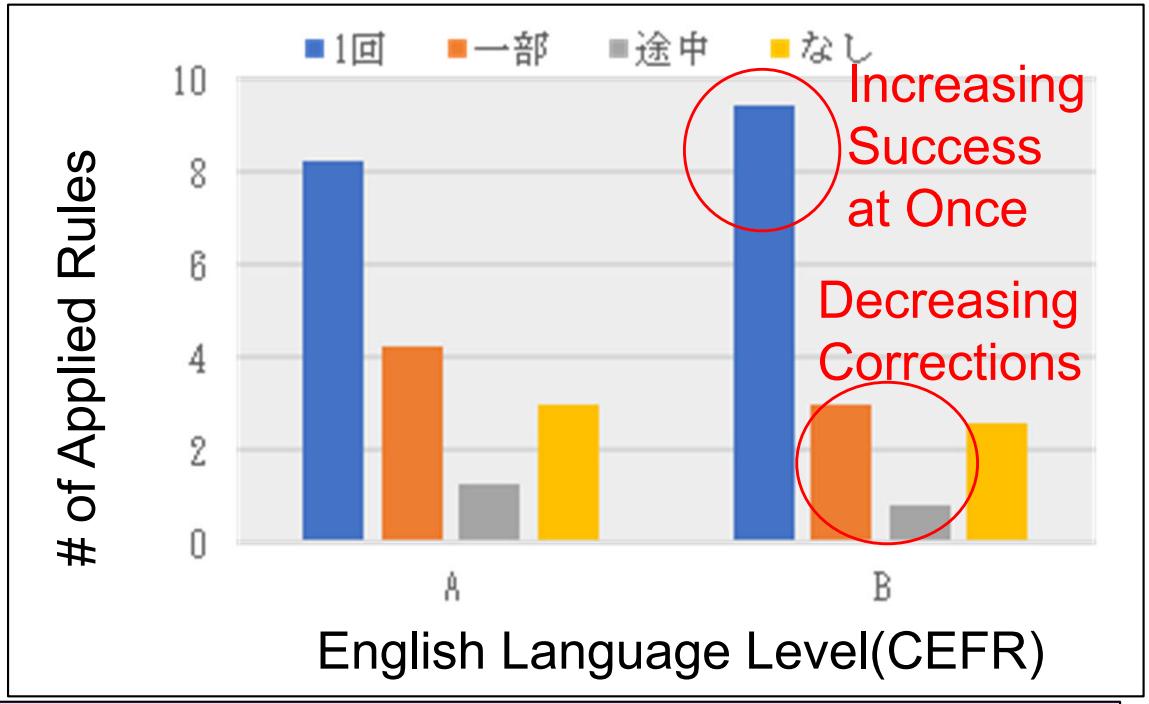
- Extraction of teacher task rules (behavior patterns) in language activities in English classes and evaluation of rules by experts (university teachers in charge of English teacher training courses)
  - Implementation of a robot service for pronunciation correction [Akimoto 2018]
    - Confirmed differences in learner proficiency and the number of applied rules
- Evaluation of basic behavior recognition performance in class environments using speech and image recognition [Akimoto 2019]
  - Speech recognition using cloud services, basic behavior recognition with over 96% accuracy using various traditional machine learning algorithms

Our Previous Work: Difference of Rule Application according to Learners' Language Levels[Akimoto2018]

Experiment Details:
- Pronunciation practice using the Audio-Lingual Method [Pedagogy]
- Subjects: 9 university students (4 at CEFR A2 level, 5 at B1/B2 level)
- Three rules for evaluation and correction (see table below)
- 12 random sentences at the level of Japanese junior high school students

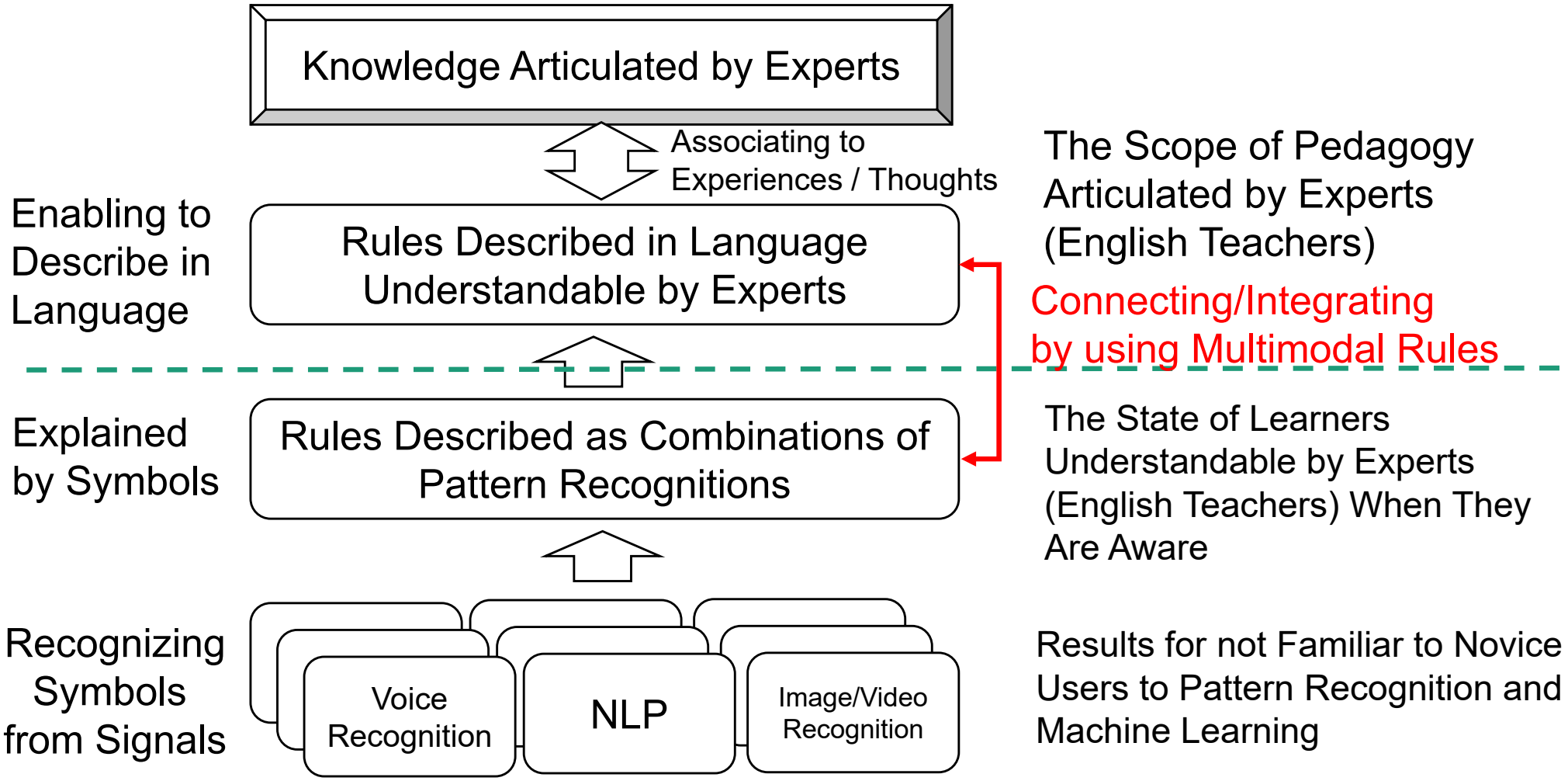
音声認識結果	処理	判定
文頭から8割以上一致	ほめる：Excellent!	一回
文頭から5割以上一致	もう一度発話指示する：One more time.	途中
2割以上一致(誤3単以内)	英文を発話する：Repeat after me.	一部

Experiment Results(Right Figure):
A correlation is suggested between learners' proficiency and the applied the business rules.



This aligns with the fact that on-site teachers appropriately apply evaluation and correction actions (rules) according to the learners.

Overview of Hierarchical Multimodal Task Rule Base Construction



Description Format of Hierarchical Multimodal Rule Base

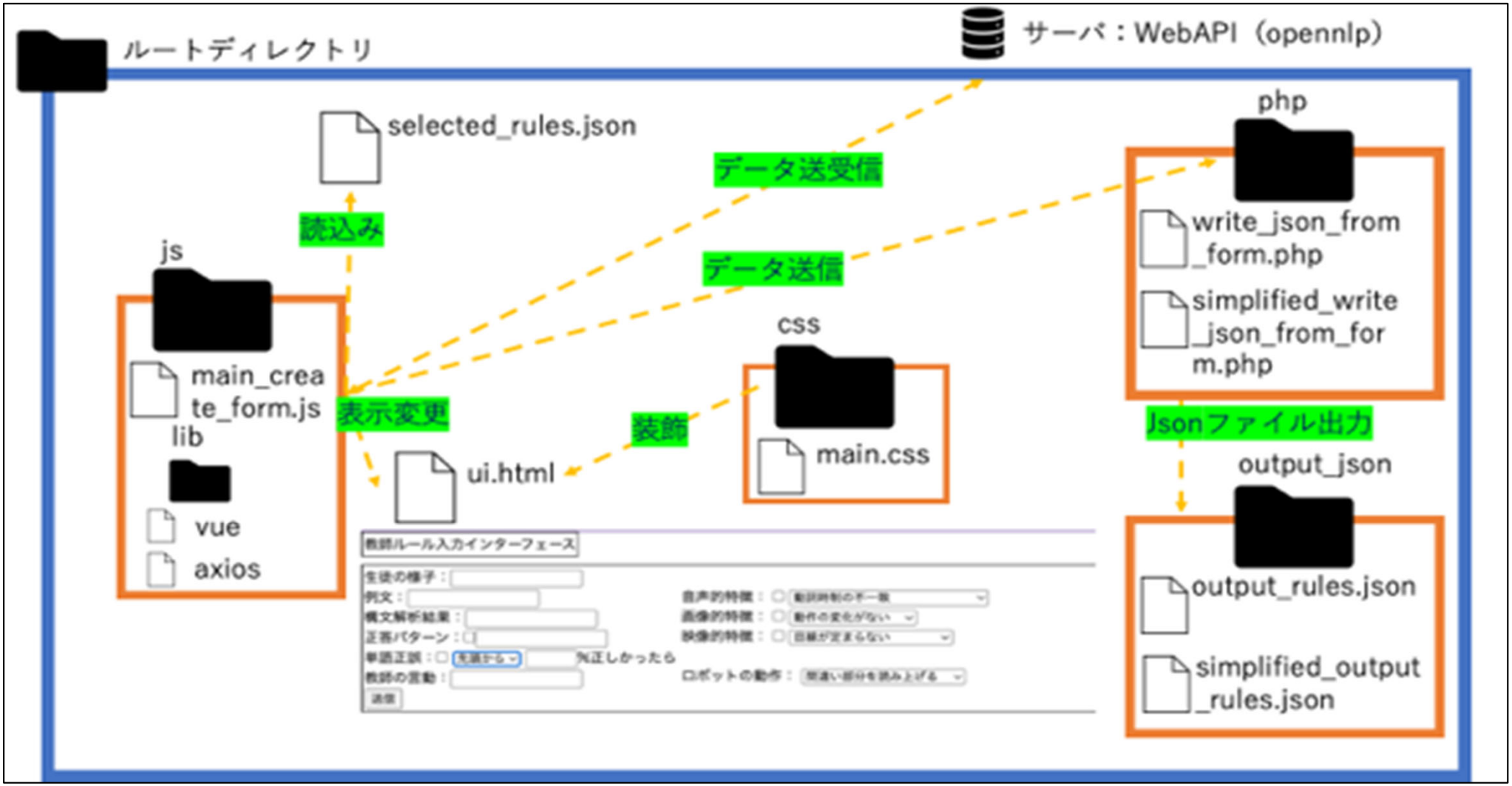
- Created a schema in JSON format based on Drools decision tables and rule description format
  - Condition part: when, Conclusion part: then
- Added items corresponding to the language used by users to describe the state of learners, questions, and

```
“rule”: {} //Each rule
 “basic_information”: {} // Basic information for this rule
  “student_situation”: “Text for explaining student situation”
  “example_sentence”: “Example sentence expecting student reaction”
  “parsing_result”: “parsing result of above example sentence”
 “when”: {} //Conditions of this rule
  “word_accuracy”: {} // Accuracy of words by speech recognition
  “voice_recognition”: [] // Conditions for results of speech recognition
  “image_recognition”: [] // Conditions for results of behavior classification by image
  “video_recognition”: [] // Conditions for results of behavior classification by images
 “then” : {} //Reaction for the student answer and behavior
  “teacher_behavior”: {} //Teacher’s original behavior
  “action”: {} //Robot actions
```

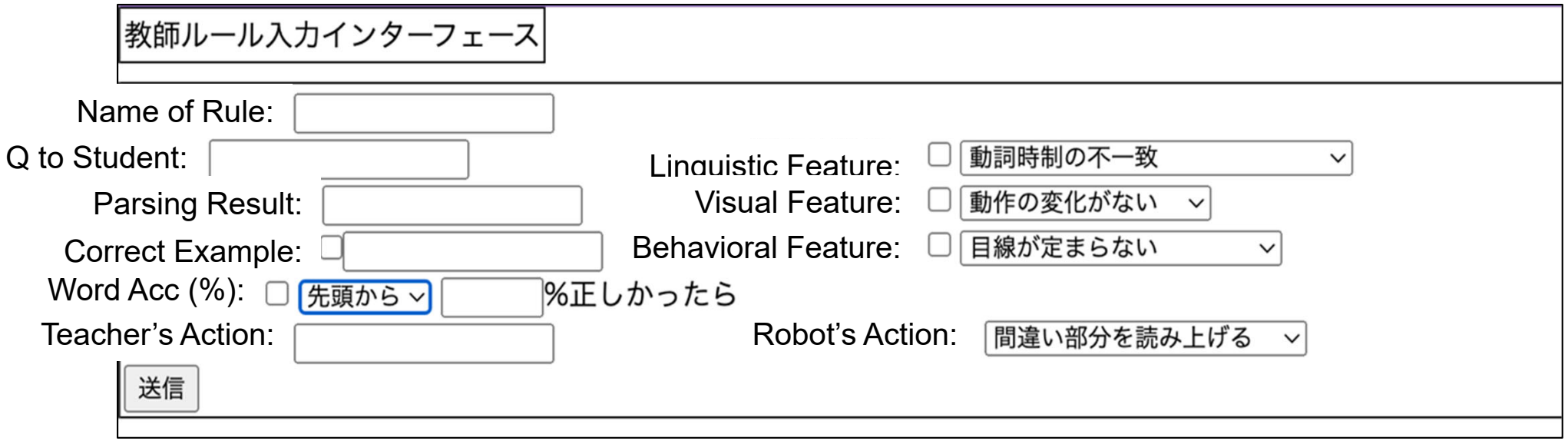
Input Interface for Hierarchical Multimodal Rule Base [Abe2023]

- Target users
  - Japanese-speaking English teachers (middle and high school)
  - Not familiar with the processing results or contents of image recognition, speech recognition, and natural language processing
  - Can describe the state of learners in language
- Two types of interfaces
  - HTML form-based input interface
  - Chat-style input interface
- Programming languages used
  - Interface: HTML/CSS, PHP, Vue.js, Axios
  - Natural language processing: Python (Flask), OpenNLP

システム構成（1）：入力フォームベース

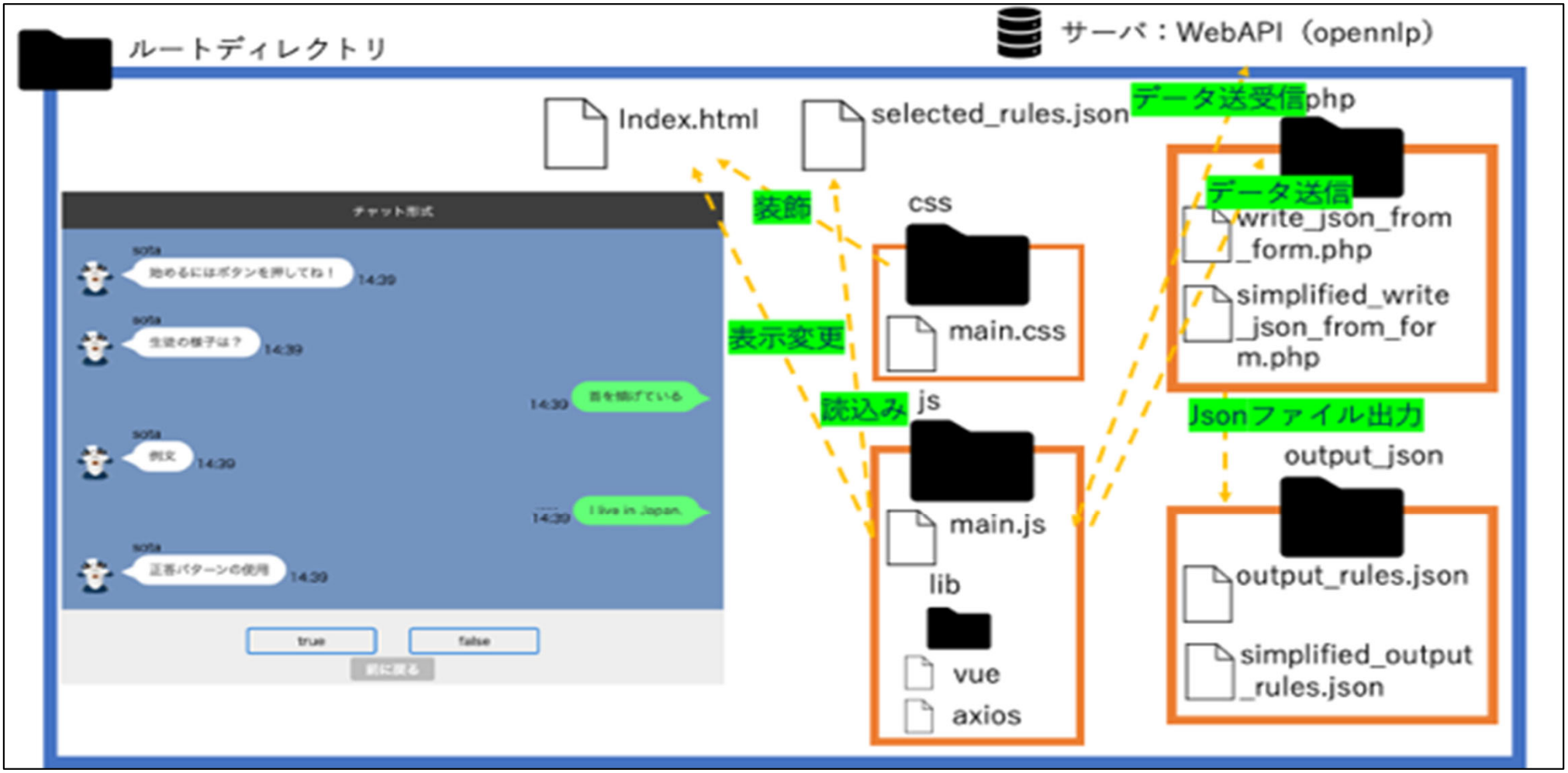


System Configuration (1): Form-Based Input



- Hierarchical multimodal rule input interface using HTML input forms
  - Text, checkboxes, and dropdown selections using HTML
  - Can grasp the input of items at once
  - Difficult for target users to intuitively understand the selection of responses and states from learners

システム構成（2）：チャット風インタフェース



System Configuration (2): Chat-Style Interface

- Hierarchical multimodal rule input using chat-style input interface
  - Sequential input of rule contents while receiving information such as text, images, and videos
  - Familiar input interface for target users
  - Can input while illustrating the state of learners with images
- Inferior in listing the entered items

User Model Description Format Corresponding to Hierarchical Multimodal Teacher Task Rule Base

- User model description in JSON format
  - History of applied rules
  - History of the accuracy of responses heard from learners
  - History of speech recognition results
  - History of image recognition results → Used for behavior recognition judgment
  - History of behavior recognition results
  - Learner identification information

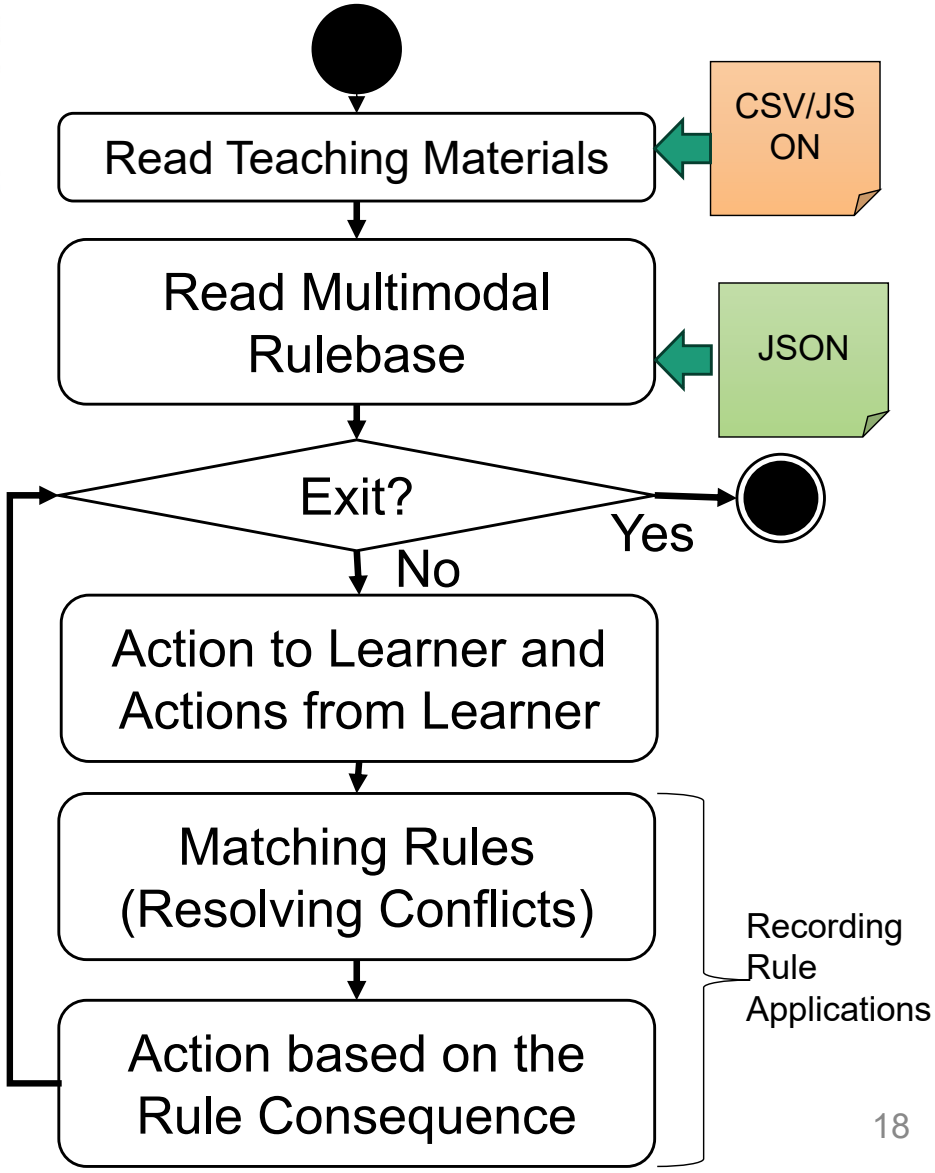
```
“user_model”: {} //Object for user model, including applied rules
  “applied_rules”: [], //Applied rule history
  “word_accuracy”: [], // Speech recognition history
  “voice_recognition”: [], //History of speech recognition results and natural language processing results
  “image_recognition”: [], //Student status (face orientation, etc.) using image recognition
  “video_recognition”: [], //Student behavior recognition results using multiple images
  “student_reaction”: [] //Student's state when applying rules (when executing action)
```

English Education Support (Dialogue Practice) Robot Service Using Multimodal Teacher Task Rule Base

Operations

- Dialogue robot: Sota by Vstone

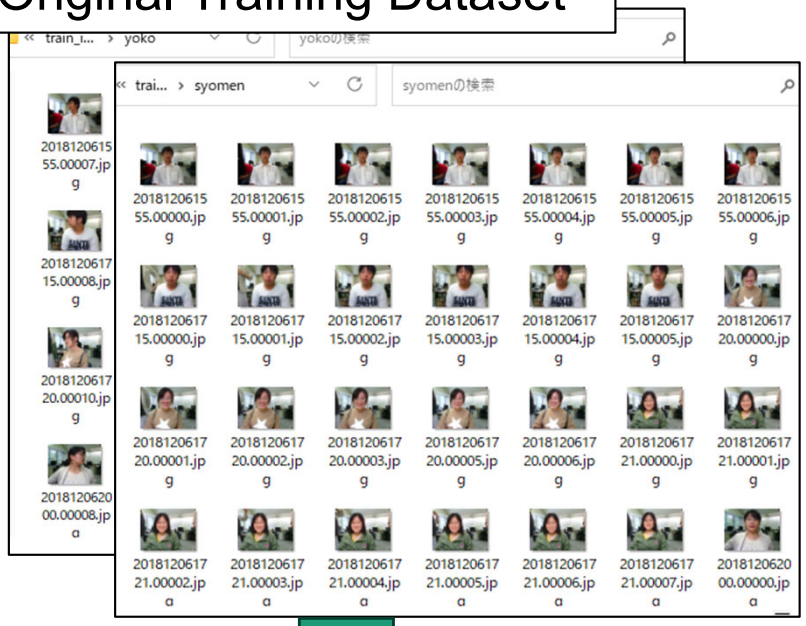
- Programming language: Java
- Performance of microphone and camera depends on Sota
- Speech recognition: IBM Cloud Speech-to-Text
  - Speech recognition accuracy and functionality depend on the quality of the STT service
- Natural language processing: Apache OpenNLP
- Image recognition: Custom CNN
  - Programming language: Python (PyTorch)



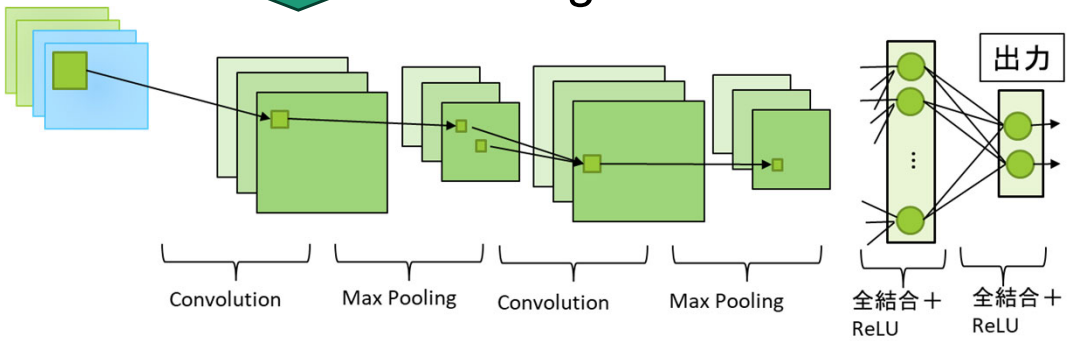
音声認識＋自然言語処理結果へのルールの適用（動作確認）

Visual Recognition and Applying Rules (with partial example)

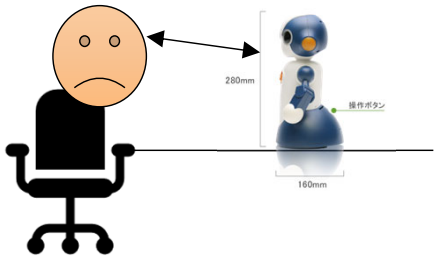
Original Training Dataset



Training



approx.70cm



Prediction

Rulebase (partial)

```
{
   "name" : "正面・かしげる以外",
   "when" : { "visual_recognition":[
                {
                    "id": 2,
                    "boolean": false,
                    "feature": "首を傾げる",
                    "type": "顔の動作"
                },
                {
                    "id": 3,
                    "boolean": false,
                    "feature": "頭が揺れる",
                    "type": "顔の動作"
                },
                {
                    "id": 4,
                    "boolean": false,
                    "feature": "常に動いている",
                    "type": "顔の動作"
                },
                {
                    "id": 101,
                    "boolean": false,
                    "feature": "正面を向いている",
                    "type": "顔の向き"
                },
                {
                    "id": 102,
                    "boolean": true,
                    "feature": "正面を向いていない",
                    "type": "顔の向き"
                }
   ] },
   "then" : { "action" : { "speak": "正面を向いていません"}}
}
```

Applying Rules

```
Input: sita, Id: 102
Input: sita,Result: 正面を向いていません
```

Conclusion

- Formulated a description format for hierarchical multimodal rule base and user model description format aimed at supporting English learning
  - Developed an English education support (learning based on dialogue practice) robot service using a hierarchical multimodal rule base
  - Work In Progress state
- Future Works
  - Implementation and evaluation of English education support (learning based on dialogue practice) robot service using a hierarchical multimodal rule base
  - Conducting user evaluations for the two types of input interfaces for multimodal rulebase
  - LLM utilization for supporting education of students who want to be teachers