electronics

MDPI

Article

Enhance the Language Ability of Humanoid Robot NAO through Deep Learning to Interact with Autistic Children

Tianhao She * and Fuji Ren *

Faculty of Engineering, Tokushima University, Tokushima 2-1, Minamijosanjima-cho, Tokushima 770-8506, Japan
* Correspondence: zthzth2080@yahoo.co.jp (T.S.); ren@is.tokushima-u.ac.jp (F.R.)

Abstract: Autism spectrum disorder (ASD) is a life-long neurological disability, and a cure has not yet been found. ASD begins early in childhood and lasts throughout a person's life. Through early intervention, many actions can be taken to improve the quality of life of children. Robots are one of the best choices for accompanying children with autism. However, for most robots, the dialogue system uses traditional techniques to produce responses. Robots cannot produce meaningful answers when the conversations have not been recorded in a database. The main contribution of our work is the incorporation of a conversation model into an actual robot system for supporting children with autism. We present the use a neural network model as the generative conversational agent, which aimed at generating meaningful and coherent dialogue responses given the dialogue history. The proposed model shares an embedding layer between the encoding and decoding processes through adoption. The model is different from the canonical Seq2Seq model in which the encoder output is used only to set-up the initial state of the decoder to avoid favoring short and unconditional responses with high prior probability. In order to improve the sensitivity to context, we changed the input method of the model to better adapt to the utterances of children with autism. We adopted transfer learning to make the proposed model learn the characteristics of dialogue with autistic children and to solve the problem of the insufficient corpus of dialogue. Experiments showed that the proposed method was superior to the canonical Seq2sSeq model and the GAN-based dialogue model in both automatic evaluation indicators and human evaluation, including pushing the BLEU precision to 0.23, the greedy matching score to 0.69, the embedding average score to 0.82, the vector extrema score to 0.55, the skip-thought score to 0.65, the KL divergence score to 5.73, and the EMD score to 12.21.

Keywords: autism spectrum disorder; robot; dialogue generation; deep learning

Citation: She, T.; Ren, F. Enhance the Language Ability of Humanoid Robot NAO through Deep Learning to Interact with Autistic Children. *Electronics* **2021**, *10*, 2393. https://doi.org/10.3390/electronics10192393

Academic Editors: Juan M. Corchado, Stefanos Kollias and Javid Taheri

Received: 16 September 2021
Accepted: 28 September 2021
Published: 30 September 2021

1. Introduction

Autism spectrum disorder (ASD) is a lifelong neurological disability that is characterized by significant social communication and behavioral deficits. The severity of this disorder can vary greatly from one individual to another. Generally, autism is a lifelong illness, and no cure has yet been found, and autism begins early in childhood and lasts throughout a person's life [1].

Children with ASD have a unique set of characteristics but most would have difficulty socializing with others, communicating verbally or non-verbally, and behaving appropriately in a variety of settings. Left untreated, an individual with ASD may not develop effective or appropriate social skills. If a child is not making friends, sustaining a conversation, able to play in an imaginative way, inflexible with routines, or overly preoccupied with certain objects, it is important to learn the cause of these behaviors and obtain support and services to help. Through early intervention, many things can be done to improve the quality of life of children. Robots are one of the best choices for accompanying children with autism [2].

The expeditious developments in the fields of AI, deep learning technology, intelligent robots, and human–computer interactions have achieved substantial progress in recent years. Currently, intelligent robots possess increasingly human-like intelligence and abilities such as listening, speaking, reading, writing, vision, feeling, and consciousness [3]. Ren et al. analyzed large quantities of statistical data based on the latest results in neurology and psychology to derive a mental state transition network, aimed at developing emotional measurement models and computer emotional simulation models for speakers [4]. CG-MVQA [5] uses a pre-trained ResNet152 to extract medical image features and establishes a mature medical visual question and answering system to assist in diagnosis. Researchers have also explored emotional information accumulated from people's daily writings (i.e., blogs) for the detection and prevention of suicide [6].

The emergence of social robots dedicated to autism can be traced back to Emanuel's pioneering research in which computer-controlled electrical equipment, such as a turtle-like robot moving via wheels on the floor, was used as a remedial tool for autistic children [7]. In recent years, studies on the healthcare of autistic children and the elderly using intelligent robots have increased, and the robotic therapy developments in these studies are remarkable. For example, commercial animal robots, such as the AIBO developed by SONY and the NeCoRo developed by OMRON, are used with hospitalized child patients and the elderly in facilities. Through the use of volunteers, the influence of the interaction between subjects and intelligent robots has been observed, and questionnaire surveys have been conducted [8,9]. Dautenhahn et al. applied autonomous mobile robots and remotely operated robots for the treatment of autistic children, and they quantitatively analyzed their interactions [10]. Currently, nearly 30 robots have been tested as remedial tools for ASD [11].

However, in most studies, intelligent robots only act as an intermediary between autistic children and therapists. For most robots, the dialogue system uses traditional techniques to generate corresponding responses such as realizing preset responses. This requires a limited response based on large dialogue libraries. When the dialogue sentences are not recorded in the dialogue library, the robot cannot provide a meaningful response. We believe that robots must have the ability to communicate with autistic children autonomously without the intervention of therapists. In our previous research, the end-to-end deep neural network conversation model was introduced into an LEO robot, for the first time, to interact with autistic children [12]. Experiments have shown that autistic children are more focused on interacting with robots without intervention by therapists.

In this paper, we aimed at solving the problem of having an insufficient dialogue corpus for deep neural network learning and put forward a brand new dialogue model and robot strategy selection model. In the experiment, we adopted a NAO robot as the platform for the proposed model and proved the effectiveness of the proposed method through large automatic evaluation metrics and human evaluation. The main contributions of this study are summarized below:

- We designed a dialogue system based on a sequences-to-sequences model and changed its input mode to improve the sensitivity of the model to context;
- We introduced a method of transfer learning, so that the transformation model could learn the basic dialogue of children from the dialogue corpus of healthy children, and then we finetuned the model to learn the discourse characteristics of autistic children;
- We coordinated the consistency of the robot dialogue and action through a strategy selection model, which was installed in a NAO robot. A series of experiments proved the effectiveness of the language model.

2. Related Work

2.1. Autism and Communication

Autism spectrum disorder is a complicated developmental disorder that greatly disturbs and affects a wide range of functions such as human communication, cognition, social skills, and behavior. In recent years, early health examinations of 1.5 year old babies

point out these possibilities. At present, the etiology has not been determined, which seems to be caused by abnormal brain function. Usually, autism is diagnosed before the patient is 3 years old [13]. People with autism spectrum disorder may experience various obstacles in their daily lives if they are placed in an unsuitable environment. Different patients will have different diseases based on different environments. The following are the three typical diseases:

- **Qualitative disorders of social development**: Children with autism sometimes find it difficult to establish interpersonal relationships, perceive another's feelings, empathize with people around them, and to take collective action and make friends [14]. Autistic children have several associated characteristics such as not indicating and sharing their interests in items by pointing with their fingers, not making eye contact, etc. They seem to be trapped in their own world, and it is difficult for them to play together while obeying the rules [15].
- **Quality barriers to communication**: Autistic children tend to develop slowly in intelligence, which makes it difficult to detect each other's feelings, establish a sympathetic relationship with people around them, take collective action, and sometimes integrate themselves into interpersonal communication [14]. They express a series of characteristics such as not being able to organize language correctly, using special words or sentences, unilaterally saying things that interest you, and suddenly saying things that have nothing to do with the present context. In addition, even if children with autism learn to speak, sometimes they cannot understand each other's words, lack an understanding of the tone contained in the language, cannot communicate with a proper voice, and feel unable to understand implications in order to respond correctly [16].
- **Prejudice towards interests and activities**: Autistic children's interests and concerns are very limited and patterned, showing strong attachment and stubbornness to certain things or always adopting fixed action patterns [16]. Among children with autism, some physiological functions, such as memory, are excellent. For example, some autistic children play a prominent role in music and painting, handicrafts, neurasthenia games, and jigsaw puzzles. However, learning and activities that require special abilities, such as "listening", "speaking", "reading", "writing", and "calculating reasoning", would become extremely difficult [15].

These categories are not solely expressed in one individual but are sometimes plural and mixed to varying degrees. These characteristics affect life to some extent. The nature of communication also includes the state of being in need. Because of the meaning of conveying meaning and emotional information, people read their desires and feelings from others' words and actions, predict the next action, and adjust their actions according to the predicted information [17]. However, it is difficult for children with autism to share and communicate their psychological state with others. They struggle to understand other people's intentions and feelings according to linguistic and non-linguistic meanings and information. Their development of language lags behind and is biased. Children with autism have no problem communicating with objects. Although their interests and actions toward specific things are often restricted, they are basically adept at understanding and operating objects as a system, and their information processing is independent of human information processing to a certain extent [2]. Therefore, interactions between robots and autistic children produce good results [18–20].

2.2. Neural Conversational Model

Conversational modeling is an important task in natural language understanding and machine intelligence. Advances in end-to-end training of neural networks have led to remarkable progress in many domains such as speech recognition, computer vision, and language processing. In the expression of the conversation system, talking with users is primarily conducted through language information, but non-language information, such as facial expressions and gestures, is also very important.

Our approach was based on recent work which proposed the use of neural networks to map sequences-to-sequences [21–23]. This framework has been applied to neural machine translation and has been improved in English–French and English–German translation tasks in the WMT14 data set [24,25]. It has also been used for other tasks such as parsing [26] and image subtitles [27]. The sequences-to-sequences model is a deep learning method based on a recurrent neural network (RNN). In general, feedforward neural networks are unidirectional network structures formed from the input layer to the output layer, while RNNs are closed networks with its own output as its own input is formed. The canonical sequence-to-sequence model is an encoder–decoder structure. The encoder and decoder have the same architecture; the encoder takes the input sequence and maps it on to an encoded representation of the sequence. The RNN of the encoder compresses the input sequence into a vector and then transmits the vector to the decoder to generate an output sequence. If the data are text based, a network containing the previous words are created by inputting the words into the article one by one.

It is well known that vanilla RNNs suffer from vanishing gradients, and most researchers use variants of long short-term memory (LSTM) recurrent neural networks [28] and the gated recurrent unit (GRU) [29]. Usually, we used embedding before we entered data into the model. The embedding layer is a type of word embedding that is learned jointly with the neural network model in specific natural language processing tasks. We must first compile a "vocabulary" list containing all the words that our model should be able to use or read. The model's input must be a tensor containing the identification of words in the sequence.

On the whole, the sequences-to-sequences model has a very wide range of application scenarios, and the effect is very strong. Because the sequences-to-sequences model is an end-to-end model, it reduces many manual processing and rule-making steps. Based on the encoder–decoder structure, an attention mechanism and other technologies were introduced, which made the depth of the methods more prominent in every task.

2.3. Robots for Autism Research

Robot and artificial intelligence research are closely related. What we particularly think of as "robots" is not a computer that does not move with only intelligence but a presence that is equipped with a body and moves through the environment. It was not so long ago that it was recognized that physicality is necessary for artificial intelligence. Humans can unconsciously use various cognitive frameworks, but robots cannot separate information that is not related to the ordered behavior. At first glance, "education for children with disabilities" and the "development of artificial intelligence robots" seem to have nothing to do with each other. However, the problems associated with artificial intelligence robot development and education for children with autism, which has stalled in the past, have something in common that cannot be overlooked. Song et al. [30] showed that people have a strong tendency to look for facial features in social robots; in the process of facial recognition, the fusiform area of a face plays an important role in systematically detecting and processing facial information. As the human–computer relationship shifts towards friends and partners, people have shown increasing interest in making social judgments about such anthropomorphic objects [31]. The ability to infer other children's emotions from their facial expressions is critical for many aspects of social communication, and deficits in expression recognition are a plausible candidate marker for ASD [32]. Robots play several important roles in the interaction of children with autism and bring many benefits. Robots that interact with autistic children are designed to play different roles. By playing games and participating in activities, robots can interact with children, develop their skills, trigger specific and satisfactory behaviors, and obtain positive feedback after successfully completing tasks [2].

In recent years, many robots have been developed to interact with autistic children. KASPAR [18] is a child-sized humanoid robot aimed at improving the lifestyle of autistic children. By interacting and behaving in a childlike way, KASPAR helps autistic children

to better interact and communicate with adults and other children. It uses a series of simplified facial and body expressions, gestures, and sounds to interact with children and uses sensors on the cheeks, arms, body, hands, and feet to automatically respond to touch, to help children understand socially acceptable tactile interactions [33]. Keepon is a yellow robot that aims to study social development by interacting with children. It has been observed that Keepon's minimalist design can attract children with different social abilities to actively participate [34]. To date, nearly 30 robots have been used to study interactions with autistic children as a medium [11], e.g., Probo [35], Maria [36], Sphero [37], CARO [38], KiliRo [39], MINA [40], QTrobot [41], Milo [42], Leo [12], Daisy [43], SAM [44], SPRITE [45], Actroid-F [46]. These studies show that robots can overcome or even outperform some sensory or linguistic obstacles encountered by autistic children when interacting with human partners. They can be used as "catalysts" to promote autistic children's perceptions of things, e.g., eye or head orientation, physical contact, and pointing to objects of shared interest [47].

3. Materials and Methods

3.1. Dialogue System

Human–robot conversational agents can be divided into two categories: retrieval-based agents and generation-based agents. Instead of generating new text, the retrieval-based model accesses a repository of predefined responses and selects the appropriate response based on the input. This method is usually used in the dialogue systems of traditional robots for autism treatment. The use of this method needs to be based on a large session database, which requires a large amount of guidance and help from a therapist in the process of interaction. Different from healthy people, when conversations are not recorded in the database, autistic children often do not have enough patience to interact. We advocate for the building of a dialogue system that can generate more meaningful responses without the intervention of a therapist.

In this section, we adopted the method of generation-based agents that uses recurrent neural networks to create effective models based on the sequences-to-sequences model, with the aim of generating meaningful and coherent dialogue responses given the dialogue history.

3.1.1. Model

Figure 1 shows the framework of the dialogue model. A shared embedding layer provides input to the encoder–decoder structure.

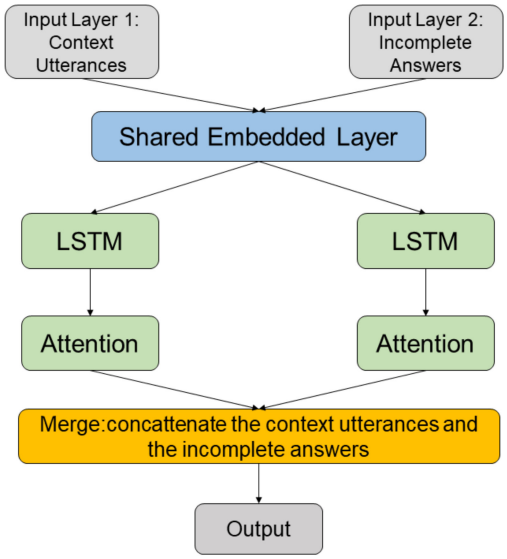


Figure 1. The framework of the proposed model.

Our work is closely related to the proposed generative conversational agents (GCAs) model of Ludwig et al. [48], which shares an embedding layer between the encoding and decoding processes through the adoption of the model. The proposed end-to-end model adopts pre-trained GloVe [49] as a shared embedding to generate dialogue vectors. The dialogue context utterances are arranged as a vector, x, that contains a sequence of token indexes filled with zeros and having the dimension s_s, which is an arbitrary value for sentence length. The elements $x_i,\ i \in \{1 \dots s_s\}$ of x are encoded into one-hot vector representation, $\overline{x}_i$. The same happens with the elements $y_i,\ i \in \{1 \dots s_s\}$ from the dialogue of incomplete answers, y. These vectors are arranged to comprise the matrices $X = [\overline{x}_1\ \overline{x}_2 \dots\ \overline{x}_{s_s}]$ and $Y = \left[\overline{y}_1\ \overline{y}_2 \dots\ \overline{y}_{s_s}\right]$.

The dialogue context utterances and the dialogue of incomplete answers are processed into two dense matrices, E_c and E_a, by a pre-trained GloVe embedding layer. The model has two LSTM layers with the same architecture, one to process E_c and another to process E_a. The processes of extracting the embedded vectors of context utterances and the dialogue of incomplete answers are expressed as follows:

$$e_c = LSTM(E_c; \omega_c) \tag{1}$$

$$e_a = LSTM(E_a; \omega_a) \tag{2}$$

ω_c and ω_a are sets of parameters of two LSTM layers. We adopted an attention mechanism to focus on the core content of the dialogue. Attention mechanisms calculate the scores of hidden vectors of the decoder at time t, and the hidden vectors of the encoder at each time to decide the weight of highlighting those words. This score can be used to calculate the weighted average of the hidden layer vector of the encoder and then calculate the hidden layer vector at time t. The outputs of the two attention layers were concatenated and provided to two dense layers that utilized the ReLU activation function and softmax activation function, respectively.

$$u_c = Attention(e_c) \tag{3}$$

$$u_a = Attention(e_a) \tag{4}$$

$$k = Relu(W_1[u_c\ u_a] + b_1) \tag{5}$$

$$p = Softmax(W_2 k + b_2) \tag{6}$$

The proposed model adopts the greedy decoding approach method, feeding the value of the larger output of the model back into the input layer on the model. This process continues until the token representing the end of the sentence is predicted.

3.1.2. Input Method

In the proposed conversation model, the input part is different from the canonical sequence-to-sequence model, and it is a structure with two input layers. In the dialogue corpus, the entire conversation was generally divided into QA pairs (query–answer) and provided to the neural network model for training. Such a trained conversation model can only realize single-round dialogue and poor understanding of the context of the conversation. In the data preprocessing stage, all dialogues in the corpus were processed into $\{Q_1, Q_2, \dots, Q_n\}$ and $\{A_1, A_2, \dots, A_n\}$ like normal question and answer tasks. Table 1 shows the two inputs corresponding to each step.

Table 1. The input order of the proposed model.

Step	Input Layer 1	Input Layer 2
1	Q_1	A_1
2	Q_1A_1	Q_2
3	A_1Q_2	A_2
4	Q_2A_2	Q_3
5	A_2Q_3	A_3
6	Q_3A_3	Q_4
7	A_3Q_4	A_4
...	...	...
$2_n - 2$	$Q_{n-1}A_{n-1}$	Q_n
$2_n - 1$	$A_{n-1}Q_n$	A_n

3.2. Implementation of Robot Systems

NAO is a small humanoid robot developed by Aldebaran Robotics. Each part has joints with 26 degrees of freedom, which are balanced by an inertial unit, and the surrounding environment is detected by multiple touch sensors and acoustic sensors on the head, hands, and feet. The dialogue is realized by 4 directional microphones and speakers, and the surrounding cameras capture the environment with high-resolution images. With such a rich expressive force, NAO can create various verbal and non-verbal interactions with people and provide communication support in important fields such as medical care, nursing, and education. This humanoid robot was purposely designed to look approachable and portray emotions like a toddler. Figure 2 shows the appearance of the humanoid robot NAO.

Figure 2. The humanoid robot NAO created by Aldebaran Robotics in France.

The accompanying software, Choregrphe [50], was used to generate NAO operations. Choregrphe was developed and designed by SoftBank Robotics Europe, and it is used for visual editing instructions for NAO. It has a graphical user interface (GUI), and uses software developed as an interface for humanoid robots such as Pepper, NAO, and Romeo. As mentioned in the related research, we put forward some basic movement strategies, such as standing, sitting, touching, and joint attention, as methods of interacting with autistic children.

NAO has two joints in its head, two joints in its waist, six joints in its arms, and five joints in its feet. Considering the combination of each joint angle of the robot as a parameter, a pose of the robot with a degree of freedom n can be regarded as a point in space. Human attitude knowledge is regarded as fuzzy without a clear definition, for example, "attitude is the attitude when the joint is X degree". The action is generated by adjusting the angle of each branch of the NAO as shown in Figure 3.

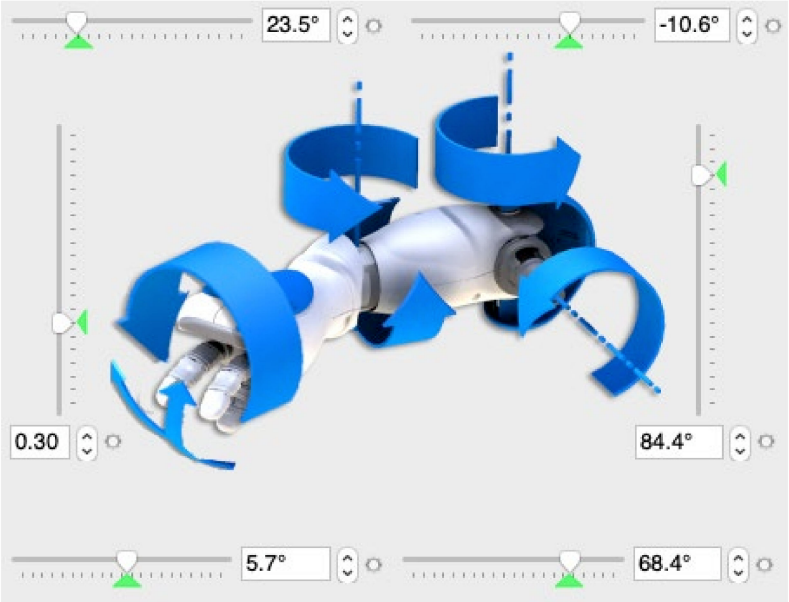


Figure 3. Adjustment of the NAO robot's joint degree of freedom.

The robot's movements can be subdivided into fixed postures. It can be considered that one action is a process of continuously generating multiple postures. Each posture has a wide range and is considered as using ambiguous knowledge of posture. We postulate that behavior is a transitional path composed of fuzzy postures, $P = [p_1, p_2, \ldots, p_n]$, which can generate actions that interpolate among these postures. For example, the action of "waving" requires a combination of several poses. As shown in Figure 4, the arm is first raised, then it swings diagonally back and forth, and finally the arm is lowered. Figure 5 shows two basic movements of the NAO robot: "standing posture" and "sitting posture".

When interacting with autistic children in a physical contact, it is necessary to be able to recognize the physical and psychological states of the touched object and predict the effect after contact. Each contact event is an important decision. Robots must have a clear understanding of their strength. When autistic children feel uncomfortable, they must cease contact. However, with today's cognitive ability of robots, it is difficult to identify and predict each other's state in the process of contact. Similarly, under such conditions, it is difficult for the NAO robot to have a clear understanding of its own state. Instead of allowing robots to try to initiate contact with autistic children, it would be a better strategy for autistic children to actively contact robots.

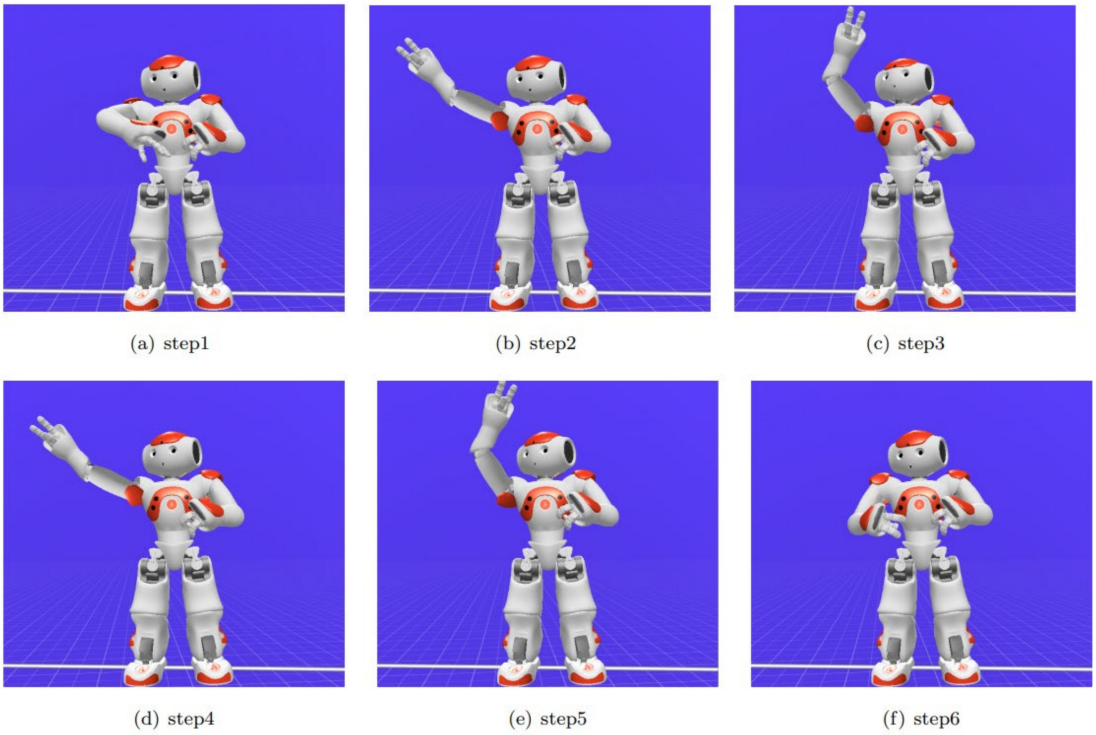
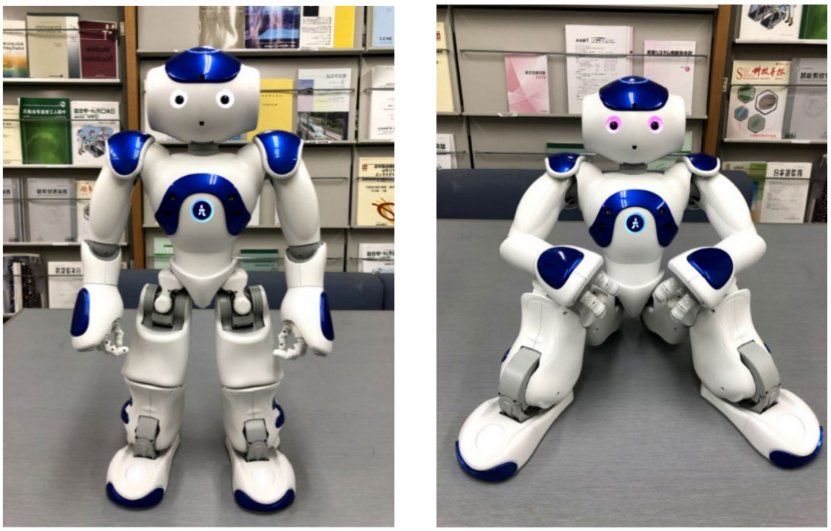


Figure 4. The waving of the robot can be seen as 6 fixed poses generated without interruption.



(**a**) (**b**)

Figure 5. Two basic movements of the humanoid robot NAO: (**a**) standing posture; (**b**) sitting posture.

First, in order to find an object correctly, face recognition must continue until the end of the program. The NAO performs several eye-catching actions and greeting voices, and the voice synthesis part adopts the ALTextToSpeech module that the NAO comes with. Incidentally, the accuracy of the NAO's own voice recognition function is not high; thus, the voice recognition API Google Speech Recognition was used in this system. In the process of communicating with children with autism, it is considered most important to maintain their attention and dialogue. When designing the dialogue, we used simple and easy-to-understand words, as much as possible, and rhetorical questions. Regarding motion generation, we also conferred top priority to motions that may interact with each other. However, it takes many man-hours to manually set the rules every time a new

dialogue mode or new action selection was added. To address this problem, we aimed to use a method to automatically acquire rules using machine learning.

3.3. Data Sets and Learning Strategy

In order to achieve autonomous communication with autistic children, we used multiple data sets to train our conversation model and compare it with the baseline Seq2Seq model. The data sets used in this experiment were all collected and integrated from the "Child Language Data Exchange System" [51]. Furthermore, our report summarizes the experimental results generated by the sessions from the text information of all benchmark data sets. Table 2 shows the specific data distribution of the two corpora. Conversation scripts are different dialogue scenes, and each dialogue scene has several conversation utterances.

Table 2. The statistics of data sets.

Data Set	#Conversation Scripts			#Utterances		
	Train	Value	Test	Train	Value	Test
Healthy Children	4398	32	28	348,256	2000	2000
Autistic Children	374	26	23	35,336	2000	2000

3.3.1. Healthy Children's Dialogue Corpus

In the research field of dialogue generation, there is no open-source database related to children's dialogue. Thus, we collected dialogue data on children's health education and linguistics research and integrated them into a large data set. This data set contained the dialogue contents of ten healthy children's dialogue datasets: "Gelman Corpus" [52], "EllisWeismer Corpus" [53], "Demetras-Working Corpus" [54], "Demetras-Trevor Corpus" [54], "Brown Corpus" [55], "Braunwald Corpus" [56], "Bohannon Corpus" [57], "Bloom73 Corpus" [58], "Bloom70 Corpus" [59], and "Bliss Corpus" [60]. These dialogues included a total of 352,256 sentences of dialogue data between mentally and physically healthy children and their parents, teachers, and friends in various scenarios such as games, eating, and work. All dialogue data were converted from audio or video to text data. Although these dialogue data were very realistic and close to daily life situations, which is suitable as samples for linguistic research, it was not the highest quality dialogue corpus for the neural network model.

3.3.2. Autistic Children's Dialogue Corpus

Collecting conversational data sets of children with autism is an arduous task. Related research is inadequate in this direction, and it is difficult to extract effective conversations from materials or videos. We used three different English dialogue corpora of children with autism to train the model. The three corpora were as follows:

Tager–Flusberg Corpus [10]: This corpus contains files from children with autism and children with Down's syndrome. It contains behavioral dialogue observations of six children between the ages of 3 and 6 years, with 8–13 visits per child over a period of 1–2 years.

Nadig ASD English Corpus [11]: This corpus contains files from videos. The overall goal of this project is to longitudinally examine word learning in children with autism (36–74 months). This corpus employs a variety of measures, including a natural language sample, during parent–child interactions. Twenty children participated at three-time points over the course of a year (between 2009 and 2012). The language sample that comprises this corpus was collected during free-play tasks with parents, children with autism, and children with Down's syndrome. It contains behavioral dialogue observations of six children between the ages of 3 and 6 years, with 8–13 visits per child over a period of 1–2 years.

Rollins Corpus [12]: This corpus consists of transcripts of video recordings of 5 male children with autism between the ages of 2 and 7 years. These children with autism met the following criteria:

1. Received an initial diagnosis of autistic symptoms by a psychologist or neurologist;
2. At least one year of a preschool program;
3. After the completion of the preschool program, children with autism had the ability to express several rich vocabularies.

In the preschool program, several videos were recorded for all autistic children who participated in the program over the entire semester. To capture each child's optimal level of on-task communicative functioning, only intervals where the child was interacting one-on-one with his clinician were transcribed and coded for analyses.

The integrated conversational corpus for autistic children contains scripts that play various roles, totaling 39,336 conversations. Like a healthy children's dialogue corpus, these corpora were used as conversational data in the field of medical language research for children with disabilities. We did not believe that these data sets were high-quality training data and test data for neural network models. In order to train the dialogue model for good performance, the data were corrected manually to remove noise, without affecting grammar, words, and semantics, so that the data were more suitable for the neural network model. For example, incorrect spelling, garbled characters, and repeated phrases when converted from audio files could be changed from "oink@o oink@o" to "oink oink", "let go walk" to "let's go walk", "what do we hafta eat?" to "what do we eat?", and "I dunno" to "I don't know".

3.3.3. Transfer Learning Strategy

Transfer learning is a machine learning method that can reuse the model developed for one task in another different task and serve as the starting point of another task model. Because in computer vision and natural language processing, the development of neural network model requires a large amount of computing and time resources, and the technical span is also relatively large; thus, the pre-trained model is usually reused as the starting point of computer vision and natural language processing tasks [61].

Due to the limitations of the ethics involved in autism research, there are only a handful of open-source-related research resources. Although we have collected and integrated a number of different dialogue corpora of children with autism, it is far from enough to train a well-trained dialogue model. We adopted new data sets to fine-tune the pre-trained model to solve this problem. Figure 6 shows the specific process of training the model through the fine-tuning method. Considering that the new data sets were almost similar to the original data set used for pre-training, the features could be extracted from the new data sets with the same weight. Since we cannot collect large dialogue data sets from autistic children, we used a healthy children's dialogue corpus with enough data sets to train the proposed dialogue model. In order to make this pre-training model understand the language features of autistic children, based on the fine-tuning method, the weight of the pre-training model was fixed, and the proposed model was refined using the dialogue corpus of autistic children.

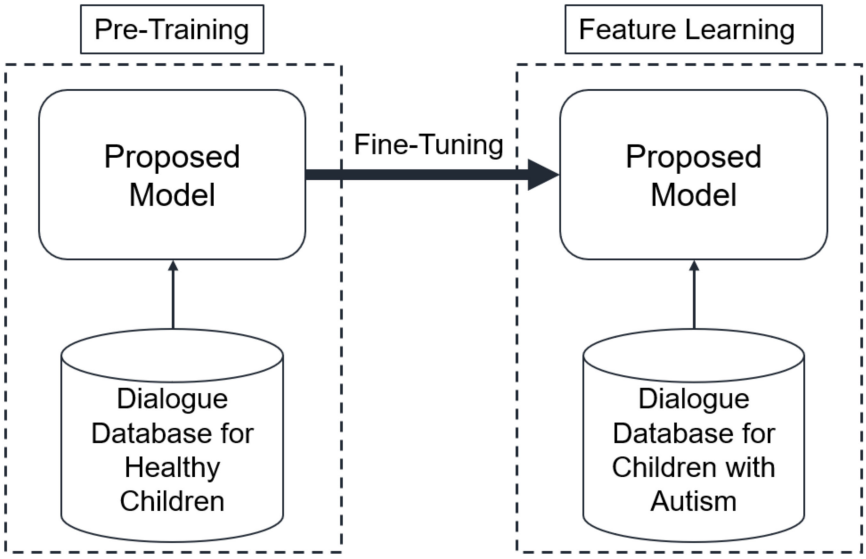


Figure 6. Structural diagram of the fine-tuning model method.

3.4. Baselines

In the experiments, all of the baseline models were not been modified for the task of dialogue for children with autism. We compared our model with the following baselines:

Seq2Seq+Attention: A standard Seq2Seq model with an attention method that is widely used as a baseline in conversation generation tasks. The model did not use beam search. It is hereafter denoted as Seq2Seq.

GCA: GCA model is a new adversarial learning method for generative conversational agents. Adversarial training also yields a trained discriminator that can be used to select the best answer when different models are available. This approach improves the performance on questions not related to the training data. The adversarial method yields significant performance gains over the usual teacher forcing training.

BERT: BERT [62] is, at its core, a transformer language model with a variable number of encoders and self-attention heads. BERT is an unsupervised language representation, and only uses a plain text corpus for pre-training. In this experiment, we used the BERT pre-trained model (24 layers, 1024 hidden layers and 16 heads) to generate text vectors.

3.5. Metrics

In the evaluation of the non-task-oriented dialogue system, the accuracy of utterance selection and utterance generation were evaluated. In other words, we compare the utterance generated by the model with the response in the test data to verify the accuracy of the generated utterance. Another method is to consider the effect of the response sentence through the meaning of each word and to judge the relevance of the test data's response. The word vector is the basis of this evaluation method. The advantage of using a word vector is that it can increase the diversity of answers to a certain extent because most of them are characterized by word similarity, which is much lower than the restriction of requiring identical words in word overlap.

BLEU [63]: Measuring the consistency of emotional conversation generation without losing the syntax performance can effectively highlight the effect of conversation generation. As for objective syntax evaluation, BLEU, a syntax measure to compute n-gram overlaps between the generated response and the reference response, was also used to measure the syntax of the responses. The n-gram is used to compare the similar proportions of n groups of words between an utterance and a reference. A result of 1-gram represents the number of words in the text that were translated separately, so it reflects the fidelity of the translation. When we calculated more than 2-gram, more often the results reflected the fluency of the translation, and the higher the value, the better the readability of the article. BLEU uses the following formula to calculate the similarity between the generated response utterance

and the reference of the test data based on the number of n-gram matches between two utterances.

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} \omega_n \log p_n\right) \tag{7}$$

p_n compares the generated response sentence with the reference utterance of the entire test data set and calculates the n-gram matching rate. The score is calculated by calculating the geometric mean from 1-gram to N-gram. The BLEU score is represented by a real number from 0 to 1. The higher the value, the better the response generated.

Greedy Matching [64]: The greedy matching method is a matrix matching method based on word level. For each word of the real response, the word with the highest similarity in the generated response is found, and the cosine similarity matching is added and averaged to the maximum extent. The same is performed for the generated response again and the average of the two is taken.

Embedding Average Cosine Similarity [65]: The embedding average method directly uses sentence vectors to calculate the similarity between real response and generated response, while sentence vectors are obtained by the weighted average of each word vector, and then it uses cosine similarity to calculate the similarity between two sentence vectors.

Vector Extrema Cosine Similarity [66]: The utterance vector is calculated by the word vector, and the cosine similarity between the utterance vectors is used to indicate the similarity between the two. The calculation of the speech vector is slightly different; here, the calculation method for the similarity of the speech vector is "Vector Extrema".

Skip-Thought Cosine Similarity [67]: Skip-thoughts vectors is the name given to simple neural network models for learning fixed-length representations of sentences in any natural language without any labeled data or supervised learning. With Word2vec, words can be displayed in distributed expressions, and the processing that takes into account the meaning of words can be performed. Kiros et al. let the model learn distributed expressions by learning to predict the words before and after a word in a sentence. This model is called the Skip-gram model of Word2vec. The model uses large novel texts as a training data set, and the encoder part of the model obtained with the help of the Seq2Seq framework is used as a feature extractor, which can generate vectors of arbitrary sentences. The trained Skip-Thoughts model encodes similar sentences that are close to each other in the embedding vector space. In this experiment, the cosine similarity was calculated based on the vector generated using the Skip-Thoughts model.

When comparing the frequency of words in different texts, the words to be compared are mostly predetermined by the analyst. Such a comparison method can statistically verify the hypothesis put forward by the analyst. However, in actual statistical analysis, it is not certain which word should be paid attention to in advance. Generally, the frequency of all words that appear in the text are compared, and the words with a large frequency difference between the texts are searched. This method does not consider the similarity of words according to their linguistic meanings but is based on the distribution of words appearing around the text set in space. With reference to the word frequency in the test data set, we analyzed the utterances generated by the proposed dialogue model and baselines and calculated their spatial similarity. The evaluation methods used were as follows:

Kullback–Leibler Divergence [68]: The concept of KL divergence comes from probability theory and information theory. The definition of KL divergence is based on entropy. The KL divergence score quantifies how much one probability distribution differs from another probability distribution. In text generation, KL divergence is used to determine the difference between the distribution of the generated text and the reference text data.

Earth Mover's Distance [69]: The earth mover's distance (EMD) is a method to evaluate the dissimilarity between two multi-dimensional distributions in some feature space where a distance measured is between single features. The EMD lifts this distance from individual features to full distributions. Specifically, WordNet is used to define the distance among the index words, and the document similarity is obtained considering the relevance between the index words. In addition, from a linguistic point of view, the value of the distance is used to express the relationship between words from the synonym dictionary that classifies words.

Human Evaluation: We randomly selected 10 scripts from the test data set. The dialogues of each script were independent of each other, and each script had at least 10 utterances and a total of 132 utterances; replies were generated from all of the compared baseline models. We then provided the relevant responses to 12 human annotators for scoring to better understand the quality of the context and responses generated. The score ranged from 1 point to 5 points; a score of 1 point denotes that there was a serious grammar error and was not suitable for the response, while a score of 5 point denotes that it had correct grammar and is suitable as a response for communicating with autistic children. The score calculation was defined as shown in Equation (8):

$$S = \frac{CS}{TS} \times 100\% \tag{8}$$

where S, CS, and TS denote the evaluation score, the conversation score, and total score, respectively.

4. Results

4.1. Automatic Evaluation

The results are given in Table 3. It can be seen from the experimental results that the proposed conversation model was stronger than the other models for all evaluation indicators. Seq2Seq performed rather poorly on nearly all emotion metrics, primarily because it did not consider any affect factor and tended to generate generic responses.

Table 3. Automatic evaluation results. Numbers in bold mean that improvement from the model on that metric is statistically significant over the baseline methods.

Model	BLEU	Greedy Matching	Embedding Average	Vector Extrema	Skip-Thought
Seq2Seq	0.15	0.56	0.70	0.41	0.16
+BERT	0.14	0.54	0.71	0.38	0.22
GCA	0.17	0.57	0.64	0.45	0.33
+BERT	0.16	0.61	0.62	0.49	0.35
Ours	0.23	0.69	0.82	0.55	0.65

Compared with Seq2Seq and GCA, the proposed model had improved BLEU evaluation indicators of 0.08 and 0.06 as can be seen in Figure 7, respectively. Especially in 1-gram, the results of the proposed dialogue model were much better than the results of the other two models. Two-gram was not much greater than the other two models. Starting from 3-gram, the GCA results were greater than the other models. The results verify the effectiveness of taking consistency into account to improve the quality of response. From the experimental results in Table 3, it can be seen that the performance of adding BERT on the BLEU index was not good, and the dialogue generated by the text vector generated by BERT was not consistent in terms of word overlap.

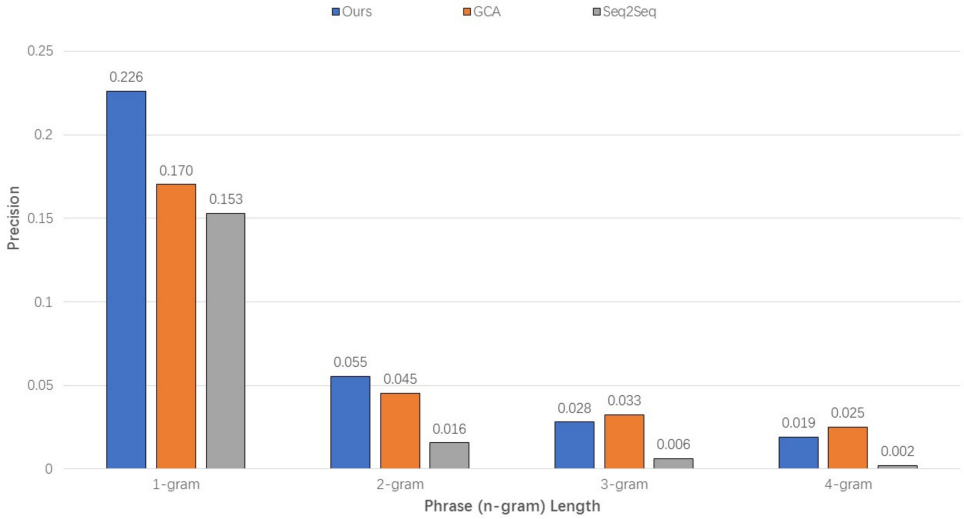


Figure 7. Results of the BLEU evaluation index.

4.2. Similarity Analysis between Word Frequency Distribution

We randomly extracted 1000 sentences from the test data set as a reference and used the NLTK (Natural Language Tool Kit) to segment the response sentences into words. Then, we counted the frequency of the words and made a word frequency table for the test data set. Then, we counted the frequency of words in all utterances based on their appearance frequency and removed low-frequency words from the utterance. The 1000 sentences of conversations were randomly selected from the test data set as input, and the corresponding response texts were obtained from the proposed model and the baselines. With the same method used to create the word frequency table for the test data set, we obtained the word frequency list of the three models. Based on the two evaluation indicators, we respectively calculated the distance from the frequency of the words generated by the three models to the word frequency distribution in the test data set. Table 4 shows the similarity results of the word frequency distributions of the three models.

Table 4. Similarity results of the word frequency distribution.

Model	KL Divergence	EMD
Seq2eq	8.02	14.59
+BERT	7.65	14.12
GCA	6.23	13.05
+BERT	6.12	12.86
Ours	**5.73**	**12.21**

4.3. Human Evaluation

Since there are no standard metrics to evaluate responses, we used human evaluation for the quality of responses to present a better view of the model's performance. Unfortunately, due to research theoretical issues, we could not use robots to actually conduct interactive experiments with autistic children. There were 10 scripts randomly selected from the test data set. The dialogues of each script were independent of each other, and each script had at least 10 sentences for a total of 132 utterances. The utterances of children with autism in each conversation were used as input for the model, and it generated responses for all compared baseline models. The human evaluation was a scoring system conducted in the form of a questionnaire. The questionnaire was evaluated from 1 point (feeling inappropriate) to 5 points (feeling appropriate). A total of 12 people participated in the experiment, and they all had the experience of interacting with children with autism. In order to ensure the accuracy of evaluating the results of the experiment, as much as possible, the human annotators carefully read the information about "conversation skills for

teenagers with autism spectrum disorder" [70] and watched videos related to interactions with autistic children. The evaluation standards we used were as follows:

- The generated response hoped to be encouraging, have positive emotions, and be grammatically correct;
- The generated response sentence was associated with the utterance sentence, and contextual information could be retained in the same script;
- The generated response encourages the continuation of the conversation.

Table 5 is an example of a questionnaire survey among 10 scripts randomly selected from the test data set in order to evaluate the response text generated from a part of a single sentence.

Table 5. Example response generated using the proposed model.

Query	Model	Response	Score
Hi.	Reference	Hi! what's that?	-
	Seq2Seq	What do you like to do?	2
	GCA	Hi! mister potato.	3
	Ours	Hi, what is this one for me?	3
Butterfly.	Reference	There's a butterfly and a puppy.	-
	Seq2Seq	You know what I need to buy	1
	GCA	Butterfly.	2
	Ours	Do you want to do that?	3
Puppy.	Reference	What's this? Is it the boy?	-
	Seq2Seq	What is it?	2
	GCA	What kind of puppy?	4
	Ours	Oh, lovely puppy.	4
Boy.	Reference	What's that?	-
	Seq2Seq	Have we seen a boy?	3
	GCA	Well, it is a boy.	3
	Ours	Have we seen a boy?	3
Owie.	Reference	Yeah, let's look at the book, what do you see?	-
	Seq2Seq	I forgot.	2
	GCA	What color is the owie?	2
	Ours	Do you want to do that?	3
Bear.	Reference	Bear, let's count them, one, two	-
	Seq2Seq	OK.	1
	GCA	What's this?	2
	Ours	Good boy, what is this?	3
Three.	Reference	Three, that's right.	-
	Seq2Seq	Yes.	1
	GCA	What's this?	2
	Ours	Yeah, three.	3

Table 6 is a questionnaire survey that scores the performance of three models in 10 dialogue scripts based on the contextual coherence of the generated response text. The conversation script of each scene contained more than ten pairs of dialogue utterances. This result indicates that the human annotators scored the context of the utterances generated by the three models.

Table 6. Example of scoring 10 dialogue scripts.

Script Number	Ours	GCA	Seq2Seq
I	3	2.5	2
II	3.5	2.5	1.5
III	3.5	2	1.5
IV	2.5	2	1.5
V	3	2.5	1.5
VI	3	3	1
VII	2.5	2.5	1.3
VIII	2.5	2	1.3
IX	3	2	1
X	3	3	2.5

The results of the human evaluation experiment are shown in Table 7. Numbers in bold mean that improvement in the model based on human judgement was statistically significant over the baselines.

Table 7. Human experimental evaluation results.

Model	Single Sentence Average Score	Script Average Score
Seq2Seq	1.89	1.83
GCA	2.82	2.87
Ours	**3.05**	**3.23**

5. Discussion

Autism spectrum disorder is a lifelong illness, and no cure has yet been found. Autism begins early in childhood and lasts throughout a person's life. Children with ASD have a unique set of characteristics, but most would have difficulty socializing with others, communicating verbally or non-verbally, and behaving appropriately in a variety of settings. Left untreated, an individual with ASD may not develop effective or appropriate social skills. If a child is not making friends, sustaining a conversation, able to play in an imaginative way, inflexible with routines, or overly preoccupied with certain objects, it is important to learn the cause of these behaviors and obtain support and services to help. Through early intervention, many things can be done to improve the quality of life of children. Robots are one of the better choices for accompanying people with autism who are in childhood. However, for most robots, the dialogue system uses traditional techniques to produce responses. This requires a limited response based on a large number of conversational databases. Robots cannot produce meaningful answers when the conversations have not been recorded in the database. Therefore, the purpose of the research was to improve the language ability of the dialogue system when interacting with children with autism so that it can generate a good response to short text input. The developed dialogue model based on the encoder–decoder structure was trained on the dialogue data of healthy children, and it learned how to have a dialogue with autistic children through the transfer learning method. We conducted experiments based on automatic evaluation indicators and human evaluation indicators.

Word overlap evaluation index: In the field of machine translation, 1-gram becomes an indicator of the correctness of word translation, and high-order n-gram is an indicator of translation fluency. Since n-grams only have the same words in the utterance, even synonyms would be regarded as different, thereby reducing the results. The proposal model had the best performance in the generation of the same single word, but compared to multiple words, it was not as good as GCA. Among all the results, the BLEU result was the lowest value overall. The response space in the dialogue system often diverged. BLEU did not care about grammar, only the distribution of content, which is suitable for measuring the performance of the data set, and it could not play a good role at the sentence level. In terms of evaluating non-task-oriented dialogue systems, it is difficult to

say that BLEU was the best evaluation index. Therefore, it is of great significance to study appropriate evaluation indicators.

Word vector evaluation index: Compared with Seq2Seq, the proposed model made 0.13, 0.12, 0.14, and 0.49 improvements in greedy matching, embedding average cosine similarity, vector extrema cosine similarity, and skip-thought cosine similarity metrics, respectively. Likewise, for GCA, we also achieved 0.12, 0.18, 0.1, and 0.32 improvements, respectively. The word vector evaluation index focuses on comparing the semantic similarity between the generated sentence and the actual sample, but it is difficult to capture long-distance semantic information based only on the word vector. Intuitively, words with special meaning in the text should have a higher priority than the commonly used expressions. Since most texts show tendencies to a greater or lesser extent, if this method calculates the similarity in the vector space, higher-order general sentences would be output first. After adding the text vector generated by BERT, the overall results of each indicator were slightly improved. Because a context-free model (such as word2vec or GloVe) generates a word vector representation for each word in the vocabulary, it is prone to word ambiguity. However, the results of Seq2Seq+BERT on greedy matching and vector extrema dropped slightly, and similarly, the results of GCA+BERT on embedding average also decreased. We believe that the multi-head attention mechanism used in BERT did not place the position information of the text sequence in an important position. It can be seen from the results that our model's choice of key information was due to the other two models, and our model could generate utterances corresponding to the key information of the input utterance.

Similarity index of word frequency distribution: Compared with Seq2Seq+BERT, the proposed model resulted in improvements of 1.92 and 1.91 in Kullback–Leibler divergence and earth mover's distance metrics, respectively. Likewise, for GCA+BERT, we also achieved 0.39 and 0.65 improvements, respectively. We found that the text vector generated based on BERT resulted in a good performance improvement over the original model. However, due to the lack of information about short text input and the lack of important location information generated by BERT in the two baseline models, the performance of the model was not as good as expected. In the process of word segmentation, although the abbreviation and complete form of the word could be unified, the same word did not change according to the different tenses such as "do" and "did". This is believed to be due to the fact that the NLTK toolkit does not integrate the morphology of the same word well during the process of word segmentation. In addition, stop words refer to words that are excluded from the processing target because of reasons such as general uselessness in natural language processing. Function words, such as auxiliary words and auxiliary verbs, appeared frequently but had no key information, which adversely affected the amount of calculations and performance. On the other hand, when calculating the frequency distribution similarity of the Kullback–Leibler divergence and earth mover's distance, the weights of each word were set to be the same.

Human evaluation: For the scores of 396 individual response sentences, the average score of the proposal dialogue model was 3.05, the average score of the GCA model was 2.82, and the average score of the Seq2Seq model was 1.89. For the scoring of the contextual script, the average score of the proposal dialogue model was 3.23, the average score of the GCA model was 2.84, and the average score of Seq2Seq was 1.83. From the overall results, we can see that the average score of the proposed model was above 3 points. Especially in regard to the evaluation's results of the contextual script, the average score was 3.2 points. Due to the architecture of the proposed model, the context could be learned, and we believe that the proposed model can predict the before and after responses. Furthermore, we observed that the generated response text had positive emotions. Although the grammar of the response sentence generated by the GCA model was fluent, the response sentence generated by this model was not strongly related to the expected context when the input sequence was short. We think this was because the model structure was not adjusted according to the language characteristics of children with autism. For the Seq2Seq model,

large security response texts, such as "whit is it" and "ok", were generated, and the model generated large sentences that lacked fluency and sentences irrelevant to the query. We believe that this was not only because of the poor quality of the autism corpus, but also because the Seq2Seq model did not consider any influencing factors and tended to generate general responses.

6. Conclusions

In this article, we proposed a chat robot for children with autism that could obtain contextual information and improve sensitivity to context. We collected the dialogue information of children with autism and modified and integrated them to solve the current situation of the insufficient corpus in this field and used the method of transfer learning to make the model learn the characteristics of dialogue of autistic children. We adopted the NAO robot as an experimental platform to import the proposed dialogue model into it, and we designed its corresponding actions to facilitate interaction with autistic children. We conducted large automatic evaluation index experiments and human evaluation experiments on the proposed model. Extensive experimental results showed that our new chatbot performed favorably on both content coherence and user satisfaction against other models. It successfully learned the discourse characteristics in the training conversation with autistic children as well as the ability to adapt and respond to short text input properly.

Author Contributions: Conceptualization, both authors; methodology, both authors; software, T.S.; data preprocessing, T.S.; data analysis, both authors; supervision, T.S.; writing—original draft preparation, T.S. Both authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by the Research Clusters program of Tokushima University (No. 2003002).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data are available in a publicly accessible repository. Publicly available data sets were analyzed in this study; these data can be found in [71].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. National Institute of Mental Health. Autism Spectrum Disorder. Available online: https://www.nimh.nih.gov/health/topics/autismspectrum-disorders-asd/index.shtml (accessed on 13 July 2018).
2. Cabibihan, J.J.; Javed, H.; Ang, M., Jr.; Aljunied, M.S. Why Robots? A Survey on the Roles and Benefits of Social Robots in the Therapy of Children with Autism. *Int. J. Soc. Robot.* **2013**, *11*, 593–618. [CrossRef]
3. Ren, F.; Bao, Y. A Review on Human-Computer Interaction and Intelligent Robots. *Int. J. Inf. Technol. Decis. Mak.* **2020**, *19*, 5–47. [CrossRef]
4. Ren, F. Affective Information Processing and Recognizing Human Emotion. *Electron. Notes Theor. Comput. Sci.* **2009**, *225*, 39–50. [CrossRef]
5. Ren, F.; Zhou, Y. CGMVQA: A New Classification and Generative Model for Medical Visual Question Answering. *IEEE Access* **2020**, *8*, 50626–50636. [CrossRef]
6. Ren, F.; Kang, X.; Quan, C. Examining Accumulated Emotional Traits in Suicide Blogs with an Emotion Topic Model. *IEEE J. Biomed. Health Inf.* **2015**, *20*, 1384–1396. [CrossRef]
7. Emanuel, R.; Weir, S. Using LOGO to catalyse communication in an autistic child. In Proceedings of the 2nd Summer Conference on Artificial Intelligence and Simulation of Behaviour, Edinburgh, UK, 12–14 July 1976; pp. 118–129.
8. Libin, A.; Libin, E. Person-Robot Interactions from the Robopsychologists' Point of View: The Robotic Psychology and Robotherapy Approach. *Proc. IEEE* **2004**, *11*, 1789–1803. [CrossRef]
9. Ohkubo, E.; Negishi, T.; Oyamada, Y.; Kimura, R.; Naganuma, M. Studies on necessary condition of companion robot in the RAA application. In Proceedings of the 2003 IEEE International Symposium on Computational Intelligence in Robotics and Automation, Kobe, Japan, 16–20 July 2003.
10. Dautenhahn, K.; Billard, A. Games Children with Autism can Play with Robota, a Humanoid Robotic Doll. In Proceedings of the 1st Cambridge Workshop on Universal Access and Assistive Technology, Cambridge, UK, 25–27 March 2002. Issue 1.
11. Kostrubiec, V.; Kruck, J. Collaborative research project: Developing and testing a robot-assisted intervention for children with Autism. *Front. Robot. AI* **2020**, *7*, 1–16. [CrossRef] [PubMed]

12. She, T.; Kang, X.; Nishide, S.; Ren, F. Improving LEO robot conversational ability via deep learning algorithms for children with autism. In Proceedings of the 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS), Nanjing, China, 23–25 November 2018; pp. 416–420.
13. Available online: https://www.healthline.com/health/signs-of-autism-in-3-year-old#severity-of-symptoms (accessed on 1 September 2021).
14. Cashin, A.; Barker, P. The triad of impairment in autism revisited. *J. Child Adolesc. Psychiatr. Nurs.* **2009**, *22*, 189–193. [CrossRef] [PubMed]
15. Wall, K. *Autism and Early Years Practice*; Sage: Thousand Oaks, CA, USA, 2009.
16. Brookdale Care. Specialist Triad of Impairments. Available online: http://www.brookdalecare.co.uk/what-is-autism#triad (accessed on 1 September 2021).
17. Johnson, C.P.; Myers, S.M. Identification and evaluation of children with autism spectrum disorders. *Pediatrics* **2007**, *120*, 1183–1215. [CrossRef] [PubMed]
18. Kozima, H.; Michalowski, M.P.; Nakagawa, C. Keepon: A playful robot for research, therapy, and entertainment. *Int. J. Soc. Robot.* **2009**, *1*, 3–18. [CrossRef]
19. Welch, K.C.; Lahiri, U.; Warren, Z.; Sarkar, N. An approach to the design of socially acceptable robots for children with autism spectrum disorders. *Int. J. Soc. Robot.* **2010**, *2*, 391–403. [CrossRef]
20. Schiavone, G.; Formica, D.; Taffoni, F.; Campolo, D.; Guglielmelli, E.; Keller, F. Multimodal ecological technology: From child's social behavior assessment to child–robot interaction improvement. *Int. J. Soc. Robot.* **2011**, *3*, 69–81. [CrossRef]
21. Kalchbrenner, N.; Blunsom, P. Recurrent continuous translation models. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, WA, USA, 19–21 October 2013; pp. 1700–1709.
22. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to sequence learning with neural networks. In Proceedings of the 27th International Conference on Neural Information Processing Systems, Bangkok, Thailand, 18–22 November 2014; Volume 2, pp. 3104–3112.
23. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
24. Luong, T.; Sutskever, I.; Le, Q.V.; Vinyals, O.; Zaremba, W. Addressing the rare word problem in neural machine translation. *arXiv* **2014**, arXiv:1410.8206.
25. Jean, S.; Cho, K.; Memisevic, R.; Bengio, Y. On using very large target vocabulary for neural machine translation. *arXiv* **2014**, arXiv:1412.2007.
26. Vinyals, O.; Kaiser, L.; Koo, T.; Petrov, S.; Sutskever, I.; Hinton, G. Grammar as a foreign language. *arXiv* **2014**, arXiv:1412.7449.
27. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. *arXiv* **2014**, arXiv:1411.4555.
28. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
29. Cho, K.; van Merrienboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H. Learning phrase representation using rnn encoder-decoder for machine translation, computation and language. *arXiv* **2014**, arXiv:1406.1078.
30. Song, Y.; Luximon, Y. The face of trust: The effect of robot face ratio on consumer preference. *Comput. Hum. Behav.* **2021**, *116*, 106620. [CrossRef]
31. Song, Y.; Luximon, A.; Luximon, Y. The effect of facial features on facial anthropomorphic trustworthiness in social robots. *Appl. Ergon.* **2021**, *94*, 103420. [CrossRef] [PubMed]
32. Loth, E.; Garrido, L.; Ahmad, J.; Watson, E.; Duff, A.; Duchaine, B. Facial expression recognition as a candidate marker for autism spectrum disorder: How frequent and severe are deficits? *Mol. Autism* **2018**, *9*, 1–11. [CrossRef] [PubMed]
33. Robins, B.; Dautenhahn, K.; Dickerson, P. From isolation to communication: A case study evaluation of robot assisted play for children with autism with a minimally expressive humanoid robot. In Proceedings of the 2nd International Conference on Advances in Computer–Human Interactions, Cancun, Mexico, 1–7 February 2009; IEEE Press: New York, NY, USA, 2009; Volume 2, pp. 205–211.
34. Kozima, H.; Nakagawa, C.; Yasuda, Y. Children–robot interaction: A pilot study in autism therapy. *Prog. Brain Res.* **2007**, *164*, 385–400.
35. Simut, R.E.; Vanderfaeillie, J.; Peca, A.; Van de Perre, G.; Vanderborght, B. Children with autism spectrum disorders make a fruit salad with probo, the social robot: An interaction study. *J. Autism Dev. Disord.* **2016**, *46*, 113–126. [CrossRef]
36. Valadao, C.T.; Goulart, C.; Rivera, H.; Caldeira, E.; Bastos Filho, T.F.; Frizera-Neto, A. Analysis of the use of a robot to improve social skills in children with autism spectrum disorder. *Res. Biomed. Eng.* **2016**, *32*, 161–175. [CrossRef]
37. Golestan, S.; Soleiman, P.; Moradi, H. Feasibility of using sphero in rehabilitation of children with autism in social and communication skills. In Proceedings of the 2017 International Conference on Rehabilitation Robotics (ICORR), London, UK, 17–20 July 2017; pp. 989–994. [CrossRef]
38. Yun, S.-S.; Choi, J.; Park, S.-K.; Bong, G.-Y.; Yoo, H. Social skills training for children with autism spectrum disorder using a robotic behavioral intervention system: Social skills training for children with ASD. *Autism Res.* **2017**, *10*, 1306–1323. [CrossRef]
39. Bharatharaj, J.; Huang, L.; Krägeloh, C.; Elara, M.R.; Al-Jumaily, A. Social engagement of children with autism spectrum disorder in interaction with a parrot-inspired therapeutic robot. *Proc. Comput. Sci.* **2018**, *133*, 368–376. [CrossRef]
40. Pour, A.G.; Taheri, A.; Alemi, M.; Meghdari, A. Human-robot facial expression reciprocal interaction platform: Case studies on children with autism. *Int. J. Soc. Robot.* **2018**, *10*, 179–198. [CrossRef]

41. Costa, A.P.; Charpiot, L.; Lera, F.R.; Ziafati, P.; Nazarikhorram, A.; Van Der Torre, L. More attention and less repetitive and stereotyped behaviors using a robot with children with autism. In Proceedings of the 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), Nanjing, China, 27–31 August 2018; pp. 534–539. [CrossRef]
42. Chalmers, C. Robotics and computational thinking in primary school. *Int. J. Child-Comput. Interact.* **2018**, *17*, 93–100. [CrossRef]
43. Pliasa, S.; Fachantidis, N. Can a robot be an efficient mediator in promoting dyadic activities among children with Autism Spectrum Disorders and children of Typical Development? In Proceedings of the 9th Balkan Conference on Informatics—BCI'19, Sofia, Bulgaria, 26–28 September 2019; ACM Press: New York, NY, USA, 2019; pp. 1–6. [CrossRef]
44. Lebersfeld, J.B.; Brasher, C.; Biasini, F.; Hopkins, M. Characteristics associated with improvement following the SAM robot intervention for children with autism spectrum disorder. *Int. J. Pediatr. Neonatal Care* **2019**, *5*, 151. [CrossRef]
45. Clabaugh, C.; Mahajan, K.; Jain, S.; Pakkar, R.; Becerra, D.; Shi, Z. Long-term personalization of an in-home socially assistive robot for children with autism spectrum disorders. *Front. Robot. AI* **2019**, *6*, 110. [CrossRef] [PubMed]
46. Yoshikawa, Y.; Kumazaki, H.; Matsumoto, Y.; Miyao, M.; Kikuchi, M.; Ishiguro, H. Relaxing gaze aversion of adolescents with autism spectrum disorder in consecutive conversations with human and android robot—A preliminary study. *Front. Psychiatry* **2019**, *10*, 370. [CrossRef] [PubMed]
47. Diehl, J.J.; Schmitt, L.M.; Villano, M.; Crowell, C.R. The clinical use of robots for individuals with Autism Spectrum Disorders: A critical review. *Res. Autism Spectr. Disord.* **2012**, *6*, 249–262. [CrossRef] [PubMed]
48. Ludwig, O. End-to-end Adversarial Learning for Generative Conversational Agents. *arXiv* **2017**, arXiv:1711.10122.
49. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the Empiricial Methods in Natural Language Processing, Doha, Qatar, 25–29 October 2014.
50. Available online: http://doc.aldebaran.com/2-1/software/choregraphe/choregraphe_overview.html (accessed on 1 September 2021).
51. MacWhinney, B. *The CHILDES Project: Tools for Analyzing Talk*, 3rd ed.; Lawrence Erlbaum Associates: Mahwah, NJ, USA, 2008.
52. Gelman, S.A.; Taylor, M.G.; Nguyen, S. Mother-child conversations about gender: Understanding the acquisition of essentialist beliefs. *Monogr. Soc. Res. Child Dev.* **2004**, *69*, I-142. [CrossRef]
53. Moyle, M.J.; Ellis Weismer, S.; Lindstrom, M.; Evans, J. Longitudinal relationships between lexical and grammatical development in typical and late talking children. *J. Speech Lang. Hear. Res.* **2007**, *50*, 508–528. [CrossRef]
54. Demetras, M. *Working Parents' Conversational Responses to Their Two-Year-Old Sons*; University of Arizona: Tucson, AZ, USA, 1989.
55. Brown, R. *A First Language: The Early Stages*; Harvard University Press: Cambridge, MA, USA, 1973.
56. Braunwald, S.R. The development of because and so: Connecting language, thought and social understanding. In *Processing Interclausal Relationships in the Production and Comprehension of Text*; Costermans, J., Fayol, M., Eds.; Lawrence Erlbaum Associate: Hillsdale, NJ, USA, 1997.
57. Stine, E.L.; Bohannon, J.N., III. Imitations, interactions, and language acquisition. *J. Child Lang.* **1983**, *10*, 589–603. [CrossRef]
58. Bloom, L. *One Word at a Time: The Use of Single-Word Utterances before Syntax*; Mouton: The Hague, The Nethlands, 1973.
59. Bloom, L.; Lightbown, P.; Hood, L. Structure and variation in child language. *Monogr. Soc. Res. Child Dev.* **1975**, *40*, 160. [CrossRef]
60. Bliss, L. The development of modals. *J. Appl. Dev. Psychol.* **1988**, *9*, 253–261. [CrossRef]
61. Pan, S.J.; Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 1345–1359. [CrossRef]
62. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pretraining of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
63. Papineni, K.; Roukos, S.; Ward, T. BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; pp. 311–318.
64. Rus, V.; Lintean, M. A comparison of greedy and optimal assessment of natural language student input using word-to-word similarity metrics. In Proceedings of the 7th Workshop on Building Educational Applications Using NLP, Stroudsburg, PA, USA, 7 June 2012; pp. 157–162.
65. Wieting, J.; Bansal, M.; Gimpel, K. Towards universal paraphrastic sentence embeddings. *arXiv* **2015**, arXiv:1511.08198.
66. Forgues, G.; Pineau, J.; Larcheveque, J.M. Bootstrapping dialog systems with word embeddings. In Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue, Cambridge, UK, 30 April–1 May 2004.
67. Kiros, R.; Zhu, Y.; Salakhutdinov, R.R.; Zemel, R.; Urtasun, R.; Torralba, A.; Fidler, S. Skip-thought vectors. In Proceedings of the 28th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 3276–3284.
68. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [CrossRef]
69. Rubner, Y.; Tomasi, C.; Guibas, L. The Earth Mover's Distance as a Metric for Image Retrieval. *Int. J. Comput. Vis.* **2000**, *40*, 99–121. [CrossRef]
70. Available online: https://raisingchildren.net.au/autism/communicating-relationships/communicating/conversationskills-for-teens-with-asd (accessed on 1 September 2021).
71. MacWhinney, B. Understanding spoken language through TalkBank. *Behav. Res. Methods* **2019**, *51*, 1919–1927. [CrossRef]