Sensors and Materials, Vol. 34, No. 5 (2022) 2001–2006
MYU Tokyo

S & M 2949

Picture-to-text Translation-based Assistive Communication System for People with Autism Spectrum Disorder

Yeou-Jiunn Chen*

Department of Electrical Engineering, Southern Taiwan University of Science and Technology,
No. 1, Nan-Tai Street, Yungkang Dist., Tainan City 710301, Taiwan

(Received December 5, 2021; accepted February 15, 2022)

Keywords: assistive communication system, word prediction, picture-to-text translation, autism spectrum disorder

People with autism spectrum disorder (ASD) have different degrees of communication disorder, reducing the quality of their daily lives. Therefore, an assistive communication system could effectively help people with ASD to communicate with other people or with devices. In this study, an assistive communication system with word prediction and picture-to-text conversion is proposed to help people with ASD. To help them use the system, touch sensors are adopted at the operating interfaces. To overcome the problem of the limited number of items that can be displayed by the operating interfaces, recurrent neural networks are adopted to predict the next input picture. To help people understand the meaning of the picture sequences inputted by people with ASD, a picture-to-text translation system, which is based on bidirectional encoder representations from transformers (BERTs), is integrated to convert picture sequences to sentences. Experimental results showed that the proposed assistive communication system can effectively predict the next picture and exactly convert picture sequences to sentences. Thus, the proposed system can effectively help people with ASD communicate with others.

1. Introduction

People with autism spectrum disorder (ASD) may encounter major obstacles in their daily lives. ASD is a complex neurophysiological disorder that causes a person's vision, hearing, and sensation to be different from those of normal people.[1] This may greatly reduce their communication ability, affecting social interactions, communication ability, and behavior. Therefore, training them to communicate with others can effectively improve their quality of life, but this is extremely time-consuming for people with ASD and caregivers. Thus, it is desirable to develop an assistive communication system that can help people with ASD communicate with others and reduce the load of caregivers.

Recently, people with ASD have used word cards or phonetic boards to communicate with others.[2,3] In these approaches, a person can only use a single picture to present a concept or a sentence, making communication very limited and inefficient. However, the purpose of assistive communication is not only for a person with ASD to express their intention, but also to promote

*Corresponding author: e-mail: chenyj@stust.edu.tw
https://doi.org/10.18494/SAM3773

ISSN 0914-4935

their language ability, especially for children. Therefore, an assistive communication system to output complex sentences would be very useful for people with ASD.

In clinical practice, a picture exchange communication system (PECS) has been developed to help people with ASD and successfully used to teach functional communication.[4] The process of the PECS includes six phases: (1) how to communicate, (2) distance and persistence, (3) picture discrimination, (4) sentence structure, (5) responsive requesting, and (6) commenting. Therefore, the PECS showed that pictures are a successful interface for people with ASD. For functional communication, a person can exactly communicate with others by using sentences. Thus, an assistive communication system with sentence expression can help people with ASD exactly present their intentions.

Natural language processing technology has been successfully applied in many applications.[5–7] A recurrent neural-network–based language model (RNN-LM) can be adopted to model the relation between two input words. Thus, according to the current input picture, the RNN-LM can predict the next picture, and the efficiency of user input can thus be greatly improved. Bidirectional encoder representations from transformers (BERTs) have also been successfully used to develop a dialog system and applied to generate user responses.[8] Therefore, BERT can translate the input picture sequence to a complex sentence. Thus, a person with ASD can easily express an intention through a complex sentence. Hence, people with ASD can use an assistive communication system to present an intention by using word prediction and picture-to-text translation.

In this study, an assistive communication system with word prediction and picture-to-text translation is proposed to help people with ASD communicate with others. To make it easy for people with ASD to operate the system, touch sensors are adopted at the operating interface. To increase the efficiency of inputting a picture sequence, an RNN-LM is used for picture prediction, enabling the next input picture to be predicted from the current input picture. To help people with ASD express an intention, BERT-based picture-to-text translation is developed to generate a complex sentence from the picture sequence input by the user.

2. Assistive Communication System

The proposed assistive communication system includes an operating interface, word prediction, and picture-to-text translation. The operating interface is implemented by using a capacitive touch screen. The word prediction and picture-to-text translation are discussed in detail in the following.

2.1 Design of word prediction

In this study, an RNN-LM is applied for word prediction, where the structure of the RNN is shown in Fig. 1.[6] An RNN has current input x_t and previous RNN activation x_{t-1}. Therefore, the active function a_t can be defined as

$$a_t = \tanh\left(w_x x_t + w_a x_{t-1}\right), \tag{1}$$

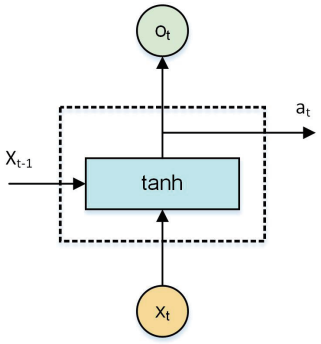

Fig. 1. (Color online) Structure of RNN.

where w_x and w_a are the weights of x_t and x_{t-1}, respectively. The output o_t is defined as

$$o_t = soft\max\left(w_o a_t\right), \tag{2}$$

where w_o is the weighting matrix.

2.2 Design of picture-to-text translation

The architecture of BERT is shown in Fig. 2. When the picture sequence $X = x_1x_2\ldots x_N$ is input, the following algorithm of BERT is applied to transform it into the text sequence $T_1T_2\ldots T_N$.

Step 1: Given the input sequence X, BERT first encodes it into a word-embedding representation, $E_1E_2\ldots E_N$.

Step 2: Let h_i^l denote the hidden representation of the lth layer in the encoder and h_i^0 be the word-embedding representation of the input sequence. Then, the attention layer is adopted to map the input sequence to the output sequence A_i^l defined as

$$A_i^l = attn_s\left(h_i^{l-1}\right), \tag{3}$$

where $attn_s$ is the multi-head attention function.

Step 3: Each of the layers contains a fully connected feedforward network (FFN) and consists of two linear transforms with rectified linear unit activation. Therefore, the output of the FFN can be defined as

$$s_i^l = \max\left(0, A_i^l W_1 + b_1\right)W_2 + b_2, \tag{4}$$

where W_1 and W_2 are the weighting matrices and b_1 and b_2 are the bias vectors.

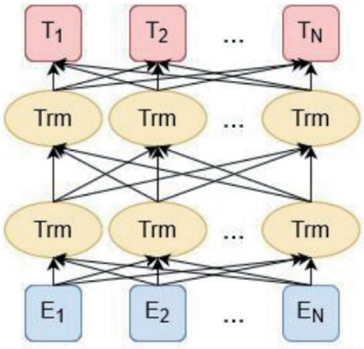

Fig. 2. (Color online) Architecture of BERT.

Step 4: Finally, a linear transformation and softmax layer are used to map s_t^L to obtain the tth predicted word T_t. The decoding process continues until the end-of-sentence token is reached.

3. Experimental Results

In this study, 18 normal people (13 males and five females, aged 20 to 27 years) were asked to collect a bilingual corpus. In this corpus, 111k sentences including 2126k words were used to train the proposed approaches. The experimental results are discussed in the following.

3.1 Results of word prediction

Five normal people and three people with ASD were asked to participate in experiments to evaluate the performance of word prediction. Thirty testing picture sequences, representing sentences often used in daily life, were randomly selected, and the experimental subjects were asked to input the picture sequences. The performance of word prediction was evaluated using the information transmission rate (ITR, pictures/minute), and the results for normal people and people with ASD are shown in Tables 1 and 2, respectively.

The experimental results showed that the proposed approach can greatly increase the ITR for normal people and people with ASD. However, the improvement for people with ASD was lower than that for normal people. The reason is that people with ASD usually find it more difficult to concentrate on operating assistive communication systems.

3.2 Results of picture-to-text translation

Seven normal people (five males and two females, aged 20 to 26 years) were asked to participate in experiments to evaluate the performance of picture-to-text translation. In this experiment, 100 sentences, which are usually used in daily life, were randomly selected from the corpus. Subjects were asked to familiarize themselves with these sentences. Each subject was asked to design 30 sentences with a similar style as their testing sentences and use the designed assistive communication system to express these sentences. After comparing the output sentence

Table 1
Performance (pictures/minute) of word prediction for normal people.

Person No.	With word prediction	Without word prediction
N1	13.25	11.23
N2	12.34	9.74
N3	13.40	10.02
N4	12.57	8.91
N5	14.61	11.41
Average	13.23	10.26

Table 2
Performance (pictures/minute) of word prediction for people with ASD.

Person No.	With word prediction	Without word prediction
A1	9.29	7.27
A2	9.57	7.66
A3	9.47	7.36
Average	9.44	7.41

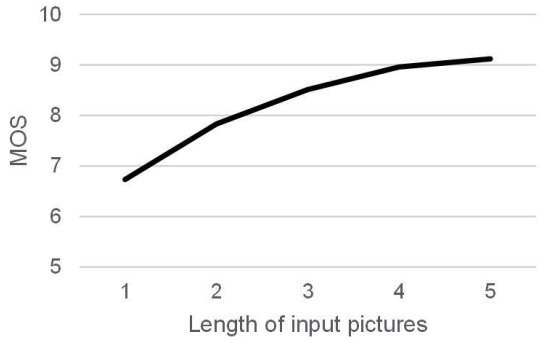

Fig. 3. MOS of picture-to-text translation in terms length of input picture sequence.

and the designed sentence, the subject gave a mean opinion score (MOS) ranging from 10 (excellent expression) to 1 (unsatisfactory expression).

The experimental results are shown in Fig. 3. According to the results, the MOS reaches an acceptable level when the number of input pictures is three. This is because the input picture sequences with three pictures were always in the sentence structure of "subject + verb + object", which is sufficient to express the user's intention. In the case of longer input picture sequences, the subjects usually used pictures representing adjectives and adverbs to appropriately express their intention. According to the results of MOS, people with ASD can produce sentences with a satisfactory level of expression by using three pictures.

4. Conclusions

In this study, an assistive communication system was successfully developed to help people with ASD. Its feature of word prediction can effectively help users to quickly find the next input

pictures, and the use of picture-to-text translation enables people to exactly express their intentions. The experimental results showed the efficiency of the proposed approaches for helping people express themselves in daily life. In the future, the proposed approach should be examined by carrying out tests on a larger number of people.

Acknowledgments

The author would like to thank the Ministry of Science and Technology, Taiwan for the financial support (MOST 110-2221-E-218-004) and the Higher Education Sprout Project of the Ministry of Education, Taiwan.

References

1 C. Lord, M. Elsabbagh, G. Baird, and J. Veenstra-Vanderweele: The Lancet **392** (2018) 508. https://doi.org/10.1016/S0140-6736(18)31129-2
2 G. Hornero, D. Conde, M. Quilez, S. Domingo, M. Pena Rodriguez, B. Romero, and O. Casas: IEEE Access **3** (2015) 1288. https://doi.org/10.1109/ACCESS.2015.2466110
3 N. M. Franco, R. N. Fidalgo, E. A. Silva, T. F. Cacalcante, and P. H. S. Brito: Proc. 2014 IEEE Int. Conf. e-Health Networking, Applications and Services (Healthcom) 335. https://doi.org/10.1109/HealthCom.2014.7001864
4 A. S. Bondy and L. A. Frost: Focus Autism Other Develop. Disabilities **9** (1994) 1. https://doi.org/10.1177/108835769400900301
5 T. Mikolov, M. Karafit, L. Burget, J. Cernock, and S. Khudanpur: Proc. 2010 Int. Conf. Int. Speech Communication Association (INTERSPEECH, 2010) 26.
6 T. Mikolov and G. Zweig: Proc. 2012 IEEE Int. Conf. Spoken Language Technology Workshop (IEEE SLT, 2012) 234. https://doi.org/10.1109/SLT.2012.6424228
7 T. Mikolov, S. Kombrink, L. Burget, J. Cernocky, and S. Khudanpur: Proc. 2011 IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP, 2011) 5528. https://doi.org/10.1109/ICASSP.2011.5947611
8 J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova: Proc. 2019 Int. Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT, 2019) 4171.

About the Author

Yeou-Jiunn Chen received his B.S. degree in mathematics from Tatung Institute of Technology, Taipei, Taiwan, and his Ph.D. degree from the Institute of Information Engineering, National Cheng Kung University, Tainan, Taiwan, in 1995 and 2000, respectively. He was with the Advanced Technology Center, Computer and Communications Laboratories, Industrial Technology Research Institute, from 2001 to 2005 as a researcher. He is currently a professor at the Department of Electrical Engineering, Southern Taiwan University of Science and Technology, Tainan, Taiwan. His research interests include biomedical signal processing, spoken language processing, and artificial intelligence. Dr. Chen is a member of the Biomedical Engineering Society, Taiwan Rehabilitation Engineering and Assistive Technology Society, and Association for Computational Linguistics and Chinese Language Processing. (chenyj@stust.edu.tw)