Generating Natural Language Responses in Robot-Mediated Referential Communication Tasks to Simulate Theory of Mind

Ziming Liu[1](✉), Yigang Qin[2], Huiqi Zou[2], Eun Jin Paek[3], Devin Casenhiser[3], Wenjun Zhou[4], and Xiaopeng Zhao[1]

[1] University of Tennessee, 1506 Middle Drive, Knoxville, TN 37916, USA
zliu68@vols.utk.edu, xzhao9@utk.edu

[2] City University of Hong Kong, 83 Tat Chee Ave., Kowloon Tong, Kowloon 999077, Hong Kong
{yigangqin2-c,huiqizou2-c}@my.cityu.edu.hk

[3] University of Tennessee Health Science Center, 600 Henley Street, Knoxville, TN 37902, USA
{epaek,dcasenhi}@uthsc.edu

[4] University of Tennessee, 916 Volunteer Boulevard, Knoxville, TN 37996, USA
wzhou4@utk.edu

Abstract. With advances in neural network-based computation, socially assistive robots have been endowed with the ability to provide natural conversation to users. However, the lack of transparency in the computation models results in unexpected robot behaviors and feedback, which may cause users to lose their trust in the robot. Theory of mind (ToM) in cooperative tasks has been considered as a key factor in understanding the relationship between user acceptance and the explainability of robot behaviors. Therefore, we develop a dialog system using previously collected data from a robot-mediated cooperative communication task data to simulate natural language smart feedback. The system is designed based on the mechanism of ToM and validated with a simulation test. Based on the result, we believe the designed dialog system bears the feasibility of simulating ToM and can be used as a research tool for further studying the importance of simulating ToM in human-robot communication.

Keywords: Human robot interaction · Theory of mind · Natural language processing

1 Introduction

With advances in Artificially Intelligent (AI) agents and machine comprehension, social robots can be enhanced by intelligent conversational systems to

Supported in part by National Institute of Health under the grant number R01AG077003.

F. Cavallo et al. (Eds.): ICSR 2022, LNAI 13817, pp. 100–109, 2022.
https://doi.org/10.1007/978-3-031-24667-8_9

provide fluid and natural conversations to users in different settings. To translate human behavior into computational algorithms through agent-based modeling, the majority of emerged artificial agents attempt to apply cognitive models to develop human-inspired intelligence. These models mainly rely on neural network based computation, such as machine learning, deep learning, or model-based reinforcement learning. These methods employ nonlinear continuous functions to regulate data and identify patterns. In this process, the system provides little transparency into the internal process of understanding how these machines make these decisions [1]. Therefore, explaining why AI agents exhibit certain behaviors is always a challenge [19]. From a user's viewpoint, the robot's behaviors or feedback may be unexpected. The lack of transparency in these models can impede users' trust since they are unable to understand or predict the robot's behavior. In other words, when users do not understand the cause or function of the robot's behaviors or decisions, they may lose trust in the robot [14]. Trust is a significant and desirable characteristic of human-robot interactions. The lack of trust may further influence the user's acceptance of the robot's input, and reduce the efficacy of developing intuitive interaction [22].

Theory of mind (ToM) is a psychological concept that relates to developing interpretability in human communication. It refers to the ability of an individual to model the mental states to others (e.g., beliefs, goals and desires) [4]. In social assistive robots (SARs), previous studies related to ToM mainly focused on perspective taking and belief management [17]. As machine learning advances, robots can reason about what humans can perceive, and construct their representations of the world. However, ToM seems not only to construct representations of others' perception of the world, but is also critical to predicting and understanding the behaviors of others in social situations [24], which enables humans to successfully communicate and cooperate with each other. Indeed, research suggests that difficulties with ToM underlie (at least in part) the challenges autistic people experience during social interaction [2,8]. In daily activities, in order to communicate efficiently, people must bear in mind the interlocutor's viewpoint and use it as a guide to appropriately shape and interpret the language used to achieve the social interaction goal [7].The main distinction between the two perspectives on ToM is one represents the capacity to understand other's minds, and another is the ability to guide communicative behaviours [16]. The second, which refers to the cooperative mind, is also crucial in human-robot interaction (HRI), but few researchers have investigated this topic.

Several studies have reported the close relationship between ToM and referential communication skills in cooperative tasks [13,18]. Referential communication skills refer to the capacity to verbally transmit the representation of an object, event or idea to a conversational partner to constitute the benchmark of a message [12]. Referential communication tasks (RCTs) are used to evaluate referential communication skills. A traditional RCT is usually conducted with two interlocutors who will act as speaker and listener in turns. Both the speaker and listener need to achieve a collaborative joint goal that ensures that their partner identifies the target referent. During this process, the speaker and listener must establish a shared understanding of the intended referent through verbal communication. Therefore, both interlocutors need to model their partners' viewpoint and adjust their own language accordingly to help each other identify the target referent. Because the task requires understanding the other's point of view, it necessarily involves ToM.

According to previous studies regarding RCTs, ToM skills are related to the communicative behaviors of requesting clarification and giving related information which refers to a communicative strategy called joint review (JR) [21]. Inspired by the association between JR and ToM skills, we aim to **develop a natural language response system to effectively extract the representation of ToM in a robot-mediated RCT**. Based on the theoretical model of JR, the robot needs to understand the user's description and provide human understandable responses, such as requests for clarifications or confirmatory information. Therefore, the robot must extract knowledge from unstructured user's transcripts and providing appropriate feedback based on the knowledge. Because identifying the entities and their semantic relations is a prerequisite for knowledge extraction, we use Bidirectional Encoder Representations from Transformers (BERT) [3], a state-of-the-art neural linguistic model, to extract contextual relations between words. Previous studies have shown the superior performance of BERT on extracting semantic relations in context [10,20]. We aim to apply the semantic relation extracted from BERT to generate a response to ask the user for clarifications or convincing information to simulate JR in RCT. Our proposed response system was validated as having significant performance accuracy in extracting entity relations [26].

In this study, we develop a dialog system based on the robot-mediated RCT task data we collected previously to simulate natural language smart feedback occur in human-human RCT. The system is designed based on the mechanism of ToM and assumed to simulate ToM happened in RCT. The proposed system is believed to apply as a research tool for further studying the importance of simulating ToM in human-robot communication.

2 The Robot-Mediated RCT Experiment and Data Collection

The robot-mediated RCT experiment was conducted by participants interacting with a humanoid robot, Pepper. Each participant went through two phases: a *sorting* phase, and a *testing* phase. During the sorting phase, 12 abstract images were shown on Pepper's tablet (see Fig. 1, **left panel**). The 12 images were created with multiple objective characteristics which can be described with different descriptors. The robot described 3 images out of 12 shown on the screen to the participant, and the participant was asked to tap on the described image accordingly. If the participant selected the wrong image, Pepper would describe the image with a longer description which included more details. If the participant still could not select the correct target image after three rounds of description, Pepper would move on to the next image. The purpose of the sorting phase was

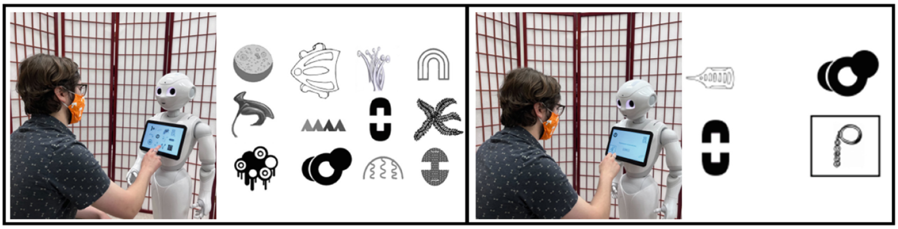

Fig. 1. An example of the sorting phase (left panel) and the testing phase (right panel).

to guide the participant in understanding how to communicate with Pepper in the following testing phase. Participants in each group would identify the same three images in the same order. In the testing phase, Pepper would show four abstract images on the screen for each trial. One image was highlighted by a black box which defined it as the target image (Fig. 1, **right panel**). The participant would organize his/her language to verbally describe the target image to Pepper. All four images could not be easily named or identified with simple labels, but contained different features to be described. Therefore, it was natural to observe participants describing the target image with different words. For example, the target image shown in the right panel can be described as "*keychain*" or "*five connected circles* ." A designed AI-mediated agent [11] would analyze both the transcript from the participant and the four images shown on the screen using a multi-modal vision-and-language analysis model and output four probability scores regarding the possibility of each image that the agent believed was the target image. Once the score for one image was significantly higher than the others, the agent determined that confidence was high enough to select the image with the highest score as the target and say "*I think I found it. Let us move on to the next image*". It would then continue to the next set of four images. If none of the images has a score that is significantly higher than the other images, the robot would ask for more details from the participants by saying "*Could you give me more details?*" The participants would normally have to change their language to describe the same target image based on their predictions of the robot's understanding. Participants perceive the robot as having some level of intelligence since it can understand the description from participants and provide reasonable feedback (e.g., move on to the next one when participant's description is approximate and ask for more details when it is not). The agent analyzed all words the participant used for the current target image as input. Each time that the participant gave the robot a description was counted as one **round**. If Pepper still could not figure out the target image after three rounds, the system would automatically move on to the next trial.

The testing phase had 24 trials. Among the 24 trials, the three abstract images used in the sorting phase were included as target images. All the images shown in the testing phase were presented in a pre-determined order. Therefore, all participants in the same group saw the same sequence of 24 trials. For each trial, a participant may have 1, 2, or 3 rounds. Participants' speech in each round during each trial was audio-recorded.

The dataset with 96 young adults' speech transcripts during a robot-mediated RCT was applied in this study. All participants were native speakers of English and recruited from a large engineering course offered at a large state university in the southeast US. The participants were randomly and evenly divided into two groups. In each group, a robot-mediated RCT was conducted with the same protocol but different image sets. The study protocol was approved by the Institutional Review Board (IRB) of the University of Tennessee, Knoxville (UTK IRB-21-06631-XM).

3 Proposed Method

In this study, we developed a system to generate natural language responses in a robot-mediated RCT. Inspired by ToM, we created a dialog system that allows robots to effectively communicate and engage with human users. The overall workflow of the designed dialog system is demonstrated in Fig. 2.

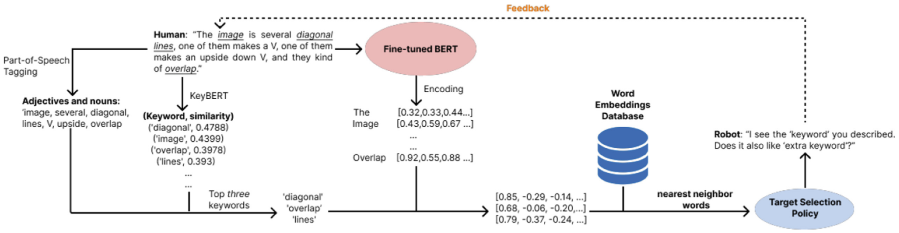
Feedback
Human: "The image is several diagonal lines, one of them makes a V, one of them makes an upside down V, and they kind of overlap."
Fine-tuned BERT
Part-of-Speech Tagging
KeyBERT
Encoding
Adjectives and nouns: 'image, several, diagonal, lines, V, upside, overlap
(Keyword, similarity)
('diagonal', 0.4788)
('image', 0.4399)
('overlap', 0.3978)
('lines', 0.393)
...
...
The Image [0.32,0.33,0.44...] [0.43,0.59,0.67 ...]
...
...
Overlap [0.92,0.55,0.88 ...]
Word Embeddings Database
Robot: "I see the 'keyword' you described. Does it also like 'extra keyword'?"
Top three keywords
'diagonal' 'overlap' 'lines'
[0.85, -0.29, -0.14, ...]
[0.68, -0.06, -0.20,...]
[0.79, -0.37, -0.24, ...]
nearest neighbor words
Target Selection Policy

Fig. 2. Human-robot dialog system workflow

The dialogue module analyzes the input from the users and produces the corresponding response. The transcript is firstly analyzed via the designed keyword matching approach to identify keywords which contain core semantics and are significant to the sentence. The word embeddings of the three most significant keywords are encoded by BERT and compared with every word in the word-embedding corpus using cosine similarity to find the three words (**extra keywords**) sharing the most relevant semantic meaning. The list serves as a reference for providing feedback includes "information the robot has understood" and "information the robot requests". For example, I see the "**keyword**" you described. Does it look like "**extra keyword**"?

The dialog architecture contains two key components: (1) keyword matching (Sect. 3.1) and (2) representation construction (Sect. 3.2). The architecture is trained with the 96 participants' dataset collected in a robot-mediated RCT.

3.1 Keyword Matching

To develop a dialog system that can provide natural language response, the agent needs to understand the users' conceptual interpretation of the image. Keywords represent the specific semantics directly as a minimum knowledge unit. Moreover, they are valid and timely for tracking the information exchange among knowledge barriers [25]. Therefore, attention mechanism of keywords are widely applied in conversation understanding [15]. In the current study, we aim to extract context-based keywords as the representation of semantics to prove robots' understanding level of users.

KeyBERT is a state-of-the-art keyword extraction method that uses BERT embeddings to extract keywords that are the most representative of the underlying text document [6]. As shown in Fig. 2, the user's transcript is firstly analyzed via KeyBERT to identify keywords that contain more semantic meanings than other words and connote the main idea of the sentence. After obtaining the document-level representation (i.e., the sentence embedding) from BERT, KeyBERT extracts word embeddings and calculates their cosine similarity with the sentence. A term with the highest value is considered the one best representing the subject of the input.

According to Kennedy's research [9], vagueness complicates the processing of linguistic reasoning. To extend linguistic reasoning to vague description, we manually operationalized a contextually-determined threshold for selecting keywords. Due to the experimental settings in RCT, participants would only provide a description of the target image. The context would naturally include the semantic representations of the target image. Since all the target images are black & white abstract images, the shape (e.g., circle) and object words (e.g., keychain) contain more conspicuous semantic features than other tokens in the sentence. As nouns and adjectives contain the most information about shape and object, we filtered the transcript input with part-of-speech (POS) tagging. Only the top three nouns and adjectives with the highest significance by cosine similarity with the sentence embedding were selected as keywords.

3.2 Representation Construction

Simply demonstrating that the robot can identify the correct target image is not sufficient to simulate ToM. Based on the concept of ToM, the robot needs to incorporate information from the user's description in its responses in order to establish a common vocabulary for understanding and effectively simulate a JR strategy [4]. In the context of the RCT, ToM requires that the robot construct a representation of users' description from their point of view. In other words, the robot needs to provide extra information relevant to user's description. For example, the user describes: "*It is a keychain.*" With a JR strategy, the robot is expected to produce utterances such as: "*Does the keychain* (related to user's description) *have a circle shape* (extra information)?" To allow the robot to provide extra information, we generate a corpus containing semantic feature of transcripts collected before using fine-tuned BERT embeddings. The purpose of the corpus is to compare with the extracted keywords and select the word from the corpus with the most similar semantic representation as **extra keyword**.

Word embedding numerically captures the semantic relations between words. Words with similar meanings are proximate in the embedding space [5]. The nearest neighbors of a word indicate the meaning of the word in the context. Alternatively, they collectively represent a form of knowledge. Therefore, such a corpus allows us to calculate and compare the semantic correlation between existing transcripts in the corpus and the user's input in the RCT. Due to the advantage of semantic awareness, as shown in Fig. 2, we fine-tuned and applied BERT to extract word embeddings from transcripts in the dataset. We applied two state-of-the-art BERT fine-tuning approaches: 1) BERT-ITPT-FIT (within-task-pre-trained and then fine-tuned) and 2) BERT-FIT (direct fine-tuned) [23].

To mimic the natural communication in RCT, transcripts of each round are used as inputs. To ensure the applicability of this corpus under different settings, the two sets of transcripts were combined and collapsed as one dataset. By manually selecting words that are objects and shapes from the transcripts, we ensured that the word embeddings contained informative semantics about the target image. Only these words' embeddings are saved in the word-embedding corpus. The word embeddings serve as vector representations of their semantics.

4 System Validation and Results

The dialog system was evaluated based on how well the system can determine additional keywords. The testing dataset was applied as transcript inputs into the designed dialog system. If one of the three relevant words determined by the system exists in the transcripts from the training dataset which described the same target image, it would count as a match. Otherwise, it was not a match. The proportion of simulation transcripts that contain at least one match (referred to as match ratio hereafter) was used as the criteria of comparison.

Due to the assumption of object and shape words, we had planed to validate the system by calculating the values when the extracted keywords with and without any shape and object words (refers to normal and unexpected situation, respectively). The shape and object list was split from the manual selection from transcripts in the dataset for each training set to maintain the consistency in the training-simulation data partition. Each model and shape/object list being used by that model was trained and generated from a subset of the whole data so that the simulation transcripts were not accessible in the training process and served as unseen sentences merely for validation.

Method of Token Representation. When BERT is fine-tuned for a downstream task, token representation is one of the salient factors affecting its performance since different layers of the BERT model output different semantic features [3]. We tested two approaches to represent every token: (1) only using the output features (hidden state of the BERT encoder) from the last layer and (2) summing all the output features from the last four layers.

Figure 3 gives a summary of evaluation results regarding the dialog simulation performance. Overall, although BERT-ITPT-FIT yielded slightly performance than BERT-FIT for text classification, the difference is not statistically significant ($t_{30} = 1.01$, $p > .05$ in normal and worst situations; $t_{14} = .35$, $p > .05$ with the sum of the last four layers to represent tokens; and $t_{14} = 1.37$, $p > .05$ with the last layer. Therefore, we have not found a significant contribution of within-task pre-training for classification in our case. One explanation could be the small volume of the training data which made it inefficient transferring the BERT language model to this specific domain.

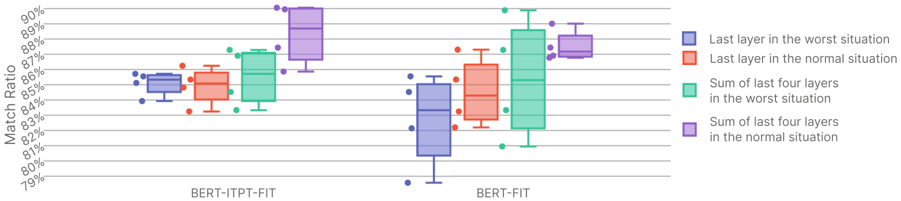

Fig. 3. Simulation results for the normal and worst situations

Despite limited performance improvement by using within-task pre-training in our design, we observed significant differences in match ratios between outcomes with different token representations. Experiments using the sum of the last four layers' features produced significantly higher mean match ratio than only using the last layer's features in normal and worst situations altogether ($t_{30} = 2.87$, $p = .0075$). The BERT-ITPT-FIT model with the sum of the last four layers' output features as the token representation attained the highest match ratio of 90.05% (matched 172/189 transcripts). The significant difference also holds when comparing the mean match ratio for each model training approach: BERT-ITPT-FIT ($t(14) = 2.13$, p=.0518) and BERT-FIT ($t(14) = 2.01$, p = .0636). Our simulation results are consistent with the conclusion that the sum of the last four layers captures richer semantic meanings in a variety of levels than the last layer alone [3]. Based on previous results and those of our simulation, a hypothesis would be that representing the tokens by concatenating the last four layers' output features could improve the match ratio in simulation and the overall performance in field experiments, which could be tested in future work.

The dialog simulation results evidenced the system's capacity to find relevant and coherent words to form the response. Moreover, we found that representation of the tokens had an effect on performance in the dialog simulation. To summarize, our simulation preliminarily confirmed the validity of the proposed dialog system in finding relevant words to facilitate a jointly reviewed conversation in RCT. The effect of model-training approach and token representation was analyzed. More training data and alternative token representation methods will be explored in the future study. Based on the results, the designed dialog system bears the feasibility to simulate human's JR strategy in RCT. Regarding the close relationship between ToM and JR in RCT, we believe the designed dialog can simulate ToM during RCT.

5 Discussion and Conclusion

We developed a robot dialog system for RCT based on the mechanism of ToM applied in daily human-human communication. The aim of the proposed dialog system is to enhance the user's understanding on robot's intention, and further improve users' trusts towards the robot. Regarding the results from the validation test, the designed system contains the ability to determine extra keywords necessary for clarification. Therefore, we believe the designed system bears an acceptable performance to conduct the proposed dialog and can be a research tool for the future studies in human-robot communication. Further field study is needed to test ecological validity of the dialog system to understand how it impacts the trust level of users in real-life conditions and interactions between AI agents and humans.

References

1. Calder, M., Craig, C., Culley, D., De Cani, R., Donnelly, C.A., Douglas, R., Edmonds, B., Gascoigne, J., Gilbert, N., Hargrove, C., et al.: Computational modelling for decision-making: where, why, what, who and how. Royal Soc. Open Sci. **5**(6), 172096 (2018)
2. Chiu, H.M., et al.: Theory of mind predicts social interaction in children with autism spectrum disorder: A two-year follow-up study. J. Autism Dev. Disord. 1–11 (2022). https://doi.org/10.1007/s10803-022-05662-4
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2018). https://doi.org/10.48550/ARXIV.1810.04805
4. Foss, N., Stea, D.: Putting a realistic theory of mind into agency theory: implications for reward design and management in principal-agent relations. Eur. Manage. Rev. **11**(1), 101–116 (2014)
5. Fu, R., Guo, J., Qin, B., Che, W., Wang, H., Liu, T.: Learning semantic hierarchies via word embeddings. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1199–1209 (2014)
6. Grootendorst, M.: Keybert: Minimal keyword extraction with bert. (2020). https://doi.org/10.5281/zenodo.4461265
7. John, A.E., Rowe, M.L., Mervis, C.B.: Referential communication skills of children with williams syndrome: understanding when messages are not adequate. Am. J. Intell. Dev. Disab. **114**(2), 85–99 (2009)
8. Jones, C.R., et al.: The association between theory of mind, executive function, and the symptoms of autism spectrum disorder. Autism Res. **11**(1), 95–109 (2018)
9. Kennedy, C.: Vagueness and grammar: the semantics of relative and absolute gradable adjectives. Linguist. Philos. **30**(1), 1–45 (2007)
10. Lin, C., Miller, T., Dligach, D., Bethard, S., Savova, G.: A bert-based universal model for both within-and cross-sentence clinical temporal relation extraction. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop, pp. 65–71 (2019)

11. Liu, Z., et al.: A demonstration of human-robot communication based on multi-skilled language-image analysis. In: 2021 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE), pp. 126–127. IEEE (2021)
12. Liu, Z., Paek, E.J., Yoon, S.O., Casenhiser, D., Zhou, W., Zhao, X.: Detecting alzheimer's disease using natural language processing of referential communication task transcripts. J. Alzheimer's Disease **86**(3), 1–14 (2022)
13. Maridaki-Kassotaki, K., Antonopoulou, K.: Examination of the relationship between false-belief understanding and referential communication skills. Eur. J. Psychol. Educ. **26**(1), 75–84 (2011)
14. Miller, T.: Explanation in artificial intelligence: insights from the social sciences. Artif. Intell. **267**, 1–38 (2019)
15. Mou, L., Song, Y., Yan, R., Li, G., Zhang, L., Jin, Z.: Sequence to backward and forward sequences: a content-introducing approach to generative short-text conversation. arXiv preprint arXiv:1607.00970 (2016)
16. Nilsen, E.S., Fecica, A.M.: A model of communicative perspective-taking for typical and atypical populations of children. Dev. Rev. **31**(1), 55–78 (2011)
17. O'Reilly, Z., Silvera-Tawil, D., Tan, D.W., Zurr, I.: Validation of a novel theory of mind measurement tool: the social robot video task. In: Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction, pp. 89–93 (2021)
18. Paal, T., Bereczkei, T.: Adult theory of mind, cooperation, machiavellianism: the effect of mindreading on social relations. Personality individ. Differ. **43**(3), 541–551 (2007)
19. Rai, A.: Explainable AI: From black box to glass box. J. Acad. Mark. Sci. **48**(1), 137–141 (2020)
20. Shi, P., Lin, J.: Simple bert models for relation extraction and semantic role labeling. arXiv preprint arXiv:1904.05255 (2019)
21. Sidera, F., Perpiñà, G., Serrano, J., Rostan, C.: Why is theory of mind important for referential communication? Curr. Psychol. **37**(1), 82–97 (2018)
22. Song, Y., Luximon, Y.: Trust in AI agent: a systematic review of facial anthropomorphic trustworthiness for social robot design. Sensors **20**(18), 5087 (2020)
23. Sun, C., Qiu, X., Xu, Y., Huang, X.: How to fine-tune BERT for text classification? In: Sun, M., Huang, X., Ji, H., Liu, Z., Liu, Y. (eds.) CCL 2019. LNCS (LNAI), vol. 11856, pp. 194–206. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32381-3_16
24. Whiten, A., Byrne, R.W.: The machiavellian intelligence hypotheses (1988)
25. Xu, J., Bu, Y., Ding, Y., Yang, S., Zhang, H., Yu, C., Sun, L.: Understanding the formation of interdisciplinary research from the perspective of keyword evolution: a case study on joint attention. Scientometrics **117**(2), 973–995 (2018). https://doi.org/10.1007/s11192-018-2897-1
26. Zhang, Z., Wu, Y., Zhao, H., Li, Z., Zhang, S., Zhou, X., Zhou, X.: Semantics-aware bert for language understanding. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 9628–9635 (2020)