BPM_MT: Enhanced Backchannel Prediction Model using Multi-Task Learning

Jin Yea Jang[1,2], San Kim[2], Minyoung Jung[2], Saim Shin[2], and Gahgene Gweon[1]

[1]Department of Intelligence and Information, Seoul National University, Republic of Korea
[2]AIRC, Korea Electronics Technology Institute, Republic of Korea
{jinyea.jang,ggweon}@snu.ac.kr
{kimsan0622,minyoung.jung,sishin}@keti.re.kr

Abstract

Backchannel (BC), a short reaction signal of a listener to a speaker's utterances, helps to improve the quality of the conversation. Several studies have been conducted to predict BC in conversation; however, the utilization of advanced natural language processing techniques using lexical information presented in the utterances of a speaker has been less considered. To address this limitation, we present a BC prediction model called BPM_MT (Backchannel prediction model with multitask learning), which utilizes KoBERT, a pre-trained language model. The BPM_MT simultaneously carries out two tasks at learning: 1) BC category prediction using acoustic and lexical features, and 2) sentiment score prediction based on sentiment cues. BPM_MT exhibited 14.24% performance improvement compared to the existing baseline in the four BC categories: continuer, understanding, empathic response, and No BC. In particular, for empathic response category, a performance improvement of 17.14% was achieved.

1 Introduction

Backchannel (BC) is a short and quick reaction, such as "uh-huh" and "yes", of a listener to speaker's utterances in a conversation (Yngve, 1970). Timely and appropriate use of BC in a conversation is important because BC has various functional categories (Cutrone, 2010); and each category has different roles and influences on enriching a conversation (Heinz, 2003; Cohn et al., 2019; Lee et al., 2020). It is important to understand the speaker utterances for the proper use of BC. In this paper, we introduce a method to enhance the utilization of lexical information in utterances for BC category prediction.

Utterances can be expressed by the acoustic and lexical information in spoken dialogue. In early BC prediction studies, there were several cases in which only the acoustic features of utterance were used (Ward and Tsukahara, 2000; Fujie et al., 2005; Poppe et al., 2010). Recently, the additional use of lexical information has improved the performance of the model (Ruede et al., 2017). However, the use of such lexical information was rather simple, including part-of-speech of the last word of utterance (Kawahara et al., 2016) or embedding features at the word level, such as Word2Vec. Ortega et al. (2020) used Word2Vec and reported the state-of-the-art performance level of about 58% accuracy on three types of BC categories, i.e. continuer, assessment, and No BC. Adiba et al. (2021a,b) also selected an approach similar to that of Ortega et al. (2020). However, they focused on addressing the latency of BC responses, not on utilization of lexical information.

In this study, we introduce a BC Prediction Model using Multi-Task Learning (BPM_MT), which learns the main task of BC category prediction and the sub-task of predicting sentiment score simultaneously. The four contributions of this paper are as follows. (1) Our model predicts a novel BC category of "empathic response", which has been identified in existing research, but has not been previously considered. (2) By applying a pre-trained language model (Devlin et al., 2018) to utilize lexical information that captures conversational context, we report a state-of-the art performance level of 76.69% (3) By utilizing sentiment cues, which hasn't been applied in predicting BC categories previously, we showed an increase in prediction performance. (4) Lastly, we report how using different context lengths can contribute to improving the performance of BC category prediction.

As the results of the experiment, BPM_MT showed a performance improvement of 14.24% compared to the baseline model (Ortega et al., 2020), and in particular, empathic response prediction improved by 17.14%. Overall, we demonstrated the effectiveness of enhanced utilization of lexical information on BC prediction.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 3447–3452
November 7–11, 2021. ©2021 Association for Computational Linguistics

Backchannel Category	Labeling Guideline	Example	Occurrence count
Continuer	short ‘yey’, ‘ney’, etc. or its competition	‘yey’, ‘mhm’, ‘ney ney’	5,284
Understanding	expression of agreement or ‘yey’(long), ‘ney’(long) etc	I see, right, ‘ney’(long)	3,900
Empathic response	exclamation or laughter	Huh!, Oh!, (sound of laugh)	1,097
Total	-	-	10,281

Table 1: BC categories, labeling guideline, examples, and occurrence counts (Quotes indicate Korean pronunciation).

2 Backchannel Data

2.1 Psychiatric Counseling Data

The data used in this study[1] were actual psychological counseling/treatment conversations between doctors and patients in the Department of Mental Health, CHA Bundang Medical Center in the Republic of Korea. This data was suitable to address our research question of BC prediction because a large portion of doctors’ utterances, 84% in our data, in counseling sessions are BC utterances. In counseling conversations, doctors use BC as a way to create a comfortable atmosphere and induce deep conversations with patients (Sadock and Sadock, 2011). Accordingly, an existing study (Kawahara et al., 2016) also had staged a counseling environment to collect BC data.

The data included a total of 51 audio files of counseling conversations in Korean language between a doctor and a patient. The total data duration was 37 hours 45 minutes 38 seconds, and all audio recordings were transcribed in text form.

2.2 Unit of Analysis for BC Labeling

It is important to provide an annotator with a unit of analysis for consistent BC labeling. Figure 1 depicts an example of a unit of analysis in an audio file of the counseling data we used, which consists of the patient’s utterance (PU) and the doctor’s utterance (DU). Three annotators we recruited were asked to check the DU and label the most appropriate BC while considering the previous PU and the next PU of the DU. They were instructed to exclude the DU as BC if turn-taking occurs after the DU. In conversation, a speaker and a listener do not change roles even after BC occurs (Yngve, 1970).

2.3 Backchannel Labeling

The labeling of BC was performed in two steps. For the first step, we decided whether the spoken utterance was BC or not following the guideline (Young and Lee, 2004). For the second step, we performed the labeling of BC categories by referring to the BC categories presented in (Cutrone, 2010). The labeling guidelines for each category were adjusted to reflect the characteristics of counseling data between doctors and patients. Finally, three BC categories were labeled: "continuer", an expression of concentration on patient utterance, "understanding", an expression of understanding and consent, and "empathic response", which expresses empathy and emotion. The corresponding labeling guidelines and examples are listed in Table 1.

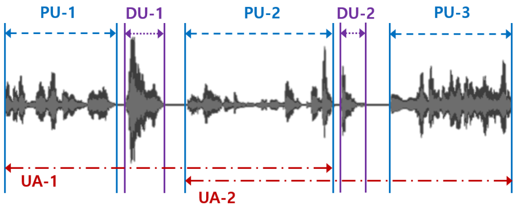

Figure 1: Examples of the unit of analysis: Patient’s Utterance (PU), Doctor’s Utterance (DU), and Unit of Analysis (UA).

The reliability of the agreement of annotators was measured. The Fleiss’ Kappa (Fleiss, 1971) score for BC identification was 0.99, and the Free Marginal Kappa (Randolph, 2005) score for the BC category was 0.90. The results confirmed the reliability of the labeling guideline (Landis and Koch, 1977). As a result of labeling, 10,281 out of 12,240 doctors’ utterances were labeled as BC. The frequency for each BC category is listed in Table 1.

3 Backchannel Prediction Model

The main task of the model is to take a patient’s utterance as input and predict a category of BC generated by a doctor. According to the presence of a sub-task that learns the sentiment score of the utterance, two types of models were designed for the performance experiment. One model was the single task model (BPM_ST) that included only the main task. The other model was the multi-task model (BPM_MT) that included both the main and

[1] This study was approved by the Institutional Review Board (IRB)

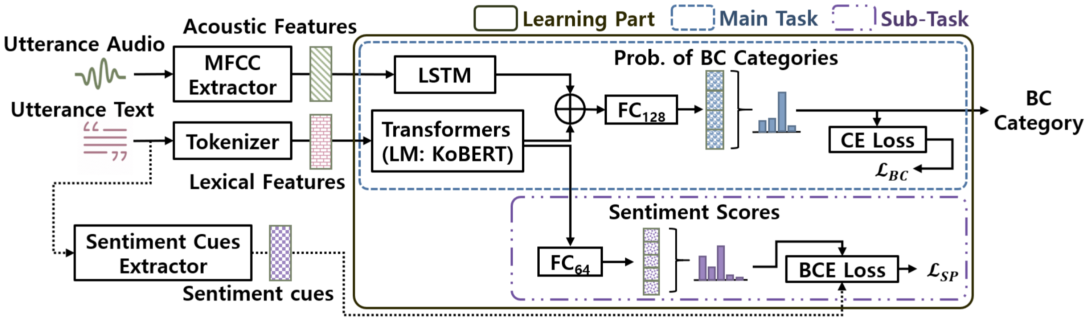

Figure 2: Backchannel (BC) prediction Model with Multi-task Learning.

sub-tasks. Both models used audio and text as input data. Figure 2 illustrates the models.

3.1 BPM_ST

In BPM_ST, only the operation corresponding to the main task is performed in the learning part of Figure 2. The overall learning process was as follows. First, the Mel-frequency Cepstrum Coefficient (MFCC) feature was extracted from utterance audio. The MFCC Extractor generated acoustic feature vectors consisting of 13 coefficients per 25ms unit, which are the input of LSTM layers. LSTM was used to consider the sequential nature of the audio data. LSTM layers have a bi-directional structure, and the acoustic hidden representation vector was calculated by summing the results of the bidirectional calculation.

The utterance text data were converted into a lexical feature in the form of an embedding vector for each token using the language model tokenizer. The pre-trained model was an encoder structure of Transformers (Vaswani et al., 2017), that received a lexical feature vector as an input and calculated a hidden representation vector for each token using a multi-head self-attention mechanism and a feed-forward neural network. Among them, the vector corresponding to the first token (CLS) was used as the lexical hidden representation for the next step. Acoustic and lexical hidden representation vectors were merged into one concatenated representation vector, that was fed into layers in a fully connected FC_{128} (128 nodes). The softmax layer of FC_{128} outputs the probability value for each BC category; thus, the cross-entropy (CE) loss $\mathcal{L}_{BC}$ was calculated.

3.2 BPM_MT

In BPM_MT, the main and the sub-tasks were simultaneously learned; therefore, all operations in the learning part of Figure 2 were performed. This is the same as BPM_ST regarding the operation process of the main task.

The sub-task was inspired by the studies that learn to predict dialogue acts and sentiment simultaneously (Li et al., 2020; Qin et al., 2020). We expected that sentiment cue could be helpful in performance improvement because both BC and the dialogue act prediction task were based on understanding the speaker's utterances. We also refer to the method of reinforcing sentiment knowledge of a pre-learned language model introduced in Tian et al. (2020).

The target sentiment score required for the sub-task of BPM_MT was extracted using the Korean MPQA (Park et al., 2018) dictionary. In the dictionary, there are words corresponding to each of the five sentiments: strong positive, positive, neutral, negative, and strong negative. When a word in the dictionary was found in the input utterance, sentiment information was extracted by counting each sentiment. After generating count information of five sentiments for the input utterance unit, a normalized sentiment score vector was constructed by dividing it by the number of words in the input utterance.

In the sub-task, the lexical hidden representation vector created from the Transformers was entered as FC_{64} (64 nodes). The output of FC_{64} was the value for each sentiment score obtained using the sigmoid function, and the average binary cross-entropy (BCE) loss $\mathcal{L}_{SP}$ for the five sentiment scores was calculated. The final loss for training the model $\mathcal{L}_{Total}$ was computed as follows: $\mathcal{L}_{Total} = (1 - \lambda)\mathcal{L}_{BC} + \lambda\mathcal{L}_{SP}$. It was important to set the application ratio of the two losses to not have a negative effect on the prediction of the BC category, which is the main task. Through an experiment, λ was set to 0.1.

Model	Context Length	All	Continuer	Understanding	Empathic Response	No BC
Kawahara	5 words	52.90	31.79	35.70	12.72	74.69
Ortega		54.63	45.76	34.38	12.00	71.98
BPM_ST		66.21 (+11.58)	52.91 (+ 7.15)	48.68 (+14.30)	28.49 (+16.49)	83.87 (+11.89)
BPM_MT		68.87 (**+14.24**)	58.51 (+12.75)	50.55 (+16.17)	29.03 (+17.03)	85.65 (+13.67)
Kawahara	10 words	56.41	39.07	33.19	14.65	78.68
Ortega		58.82	51.73	37.25	14.17	75.57
BPM_ST		69.61 (+10.79)	56.18 (+ 4.45)	49.46 (+12.21)	30.26 (+16.09)	88.49 (+12.92)
BPM_MT		72.64 (+13.82)	60.75 (+ 9.02)	52.45 (+15.20)	31.31 (**+17.14**)	90.99 (+15.42)
Kawahara	20 words	63.17	47.16	30.87	15.12	88.83
Ortega		70.91	60.88	49.64	16.51	90.18
BPM_ST		75.40 (+ 4.49)	62.83 (+ 1.95)	52.18 (+ 2.54)	31.85 (+15.34)	95.45 (+ 5.27)
BPM_MT		**76.69** (+ 5.78)	**65.89** (+ 5.01)	**53.48** (+ 3.84)	**32.86** (+16.35)	**95.89** (+ 5.71)

Table 2: BC category prediction results of four categories: F1 weighted score for 'All' and F1 score for each BC category (numbers in parentheses indicate the difference in values from that of the model by Ortega et al. (2020)).

3.3 Experimental Setup

Two types of BC prediction experiments were conducted: 1) BC prediction with four categories; continuer, understanding, empathic response, and No BC, and 2) BC prediction with three categories; continuer, understanding, and No BC. We conducted the second experiment using three categories prediction because the number of empathic responses in the data was relatively less than the other categories.

Models of Kawahara et al. (2016) and Ortega et al. (2020) were selected as baselines. Kawahara et al. (2016) used logistic regression for BC prediction. Ortega et al. (2020) used CNN as a neural network structure for both acoustic and lexical information.

In the training data, audio input was set in units of 1,500 ms and text input was set in units of 5 words according to Ortega et al. (2020). The inputs are the past speech audio and text based on the moment BC occurs. Additionally, data for 10 and 20 words were also organized to observe the performance trend according to the length of the text context. The context length means the maximum number of words. The training data, including the negative sample (No BC), were also configured to be equal to the total number of BCs according to Ortega et al. (2020).

The pre-training language model used in BPM was KoBERT [2], and ReLU was used as the activation function for each hidden layer; the dropout rate was 0.3. The batch size was 64, and the number of epochs was 60. Optimization was performed using SGD as the parameter of the Transformers and Adam for the other parameters. The learning rate was 0.0005. The data were divided into training, validation, and test sets at a ratio of 3:1:1. The best model [3] was saved by early stopping regularization based on the validation result.

Model	All	Continuer	Under-standing	No BC
Kawahara	58.17	33.01	49.07	75.46
Ortega	59.46	46.96	47.42	71.74
BPM_ST	71.26	51.99	60.26	86.49
BPM_MT	**73.97**	**58.91**	**63.68**	**86.70**

Table 3: BC category prediction results of three categories (context length at 5 words): F1 weighted score for 'All' and F1 score for each BC category.

4 Results and Discussion

The experimental results of four categories are listed in Table 2. Each value is an average obtained using 5-fold cross-validation. Overall, BPM_MT exhibited a higher performance improvement compared to BPM_ST and the baseline models, indicating that the use of sentiment cues was helpful. Regarding the context length, there was an improvement in the performance of up to a minimum of 5.78% (20 words) and a maximum of 14.24% (5 words) compared to the baseline model of Ortega. The BPM_MT model outperforms the baseline models even when the input context is limited to 5 words. Therefore, the BPM_MT can be used when less context information is available with a small number of contextual word data is available as input. Among the BC categories, the performance improvement for the empathic response category was the greatest in comparison with the baseline models. When the context length was 10, the performance increased by 17.14% compared to the Or-

[2] https://github.com/monologg/KoBERT-Transformers

[3] NVIDIA TITAN RTX was used for the experiments

tega model. This emphasizes the importance of understanding the speaker's sentiment in utterances or the flow of conversation to express the signal of empathy. Compared to BPM_ST, BPM_MT showed the highest improvement (5.6%) for the continuer category.

Table 3 summarizes the experimental results of three categories on the context length of 5 words. BPM_ST and BPM_MT still outperformed the baseline models in BC prediction without empathic response.

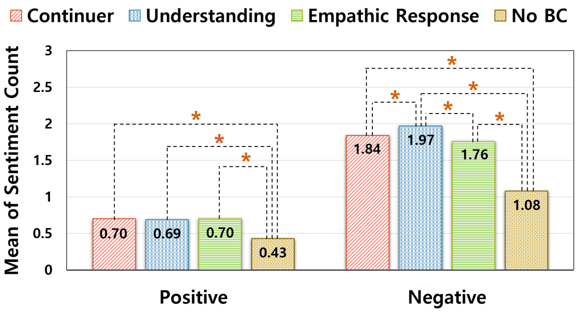

Figure 3: Distribution of positive and negative sentiment words by BC category in the context length equals to 5 ($*p < .05$).

Figure 3 shows the distribution of positive and negative sentiment words depending on BC categories. We used the data of the context length equals to 5, because setting the model with that length showed the highest performance improvement when the sub-task was applied. To compare the results by BC category, we ran ANOVA with the Tukey's posthoc test. The result shows significant differences of the means among categories (F(3, 20549) = 2013.7, p < 0.05 for the positive sentiment; F(3, 20549) = 5318.6, p < 0.05 for the negative sentiment). More cases of significant differences were shown in the negative sentiment. This indicates the importance of negative sentiment information that exists in patient utterances in BC category prediction. When comparing the categories of No BC and BCs, richer sentiment information induced BC generation in both positive and negative sentiments. These indicate that additional learning about sub-tasks was an appropriate approach to improving BC performance.

5 Conclusions

We presented a BC category prediction model that achieved a performance of 76.69% by utilizing lexical information with sentiment cues. Across the three types of BC categories, the empathic response category, which hasn't been previously measured, showed a performance improvement of 17.14% compared to the model of Ortega. This increase implies that utilization of lexical information helped in predicting empathy function.

The limitation of this study is the imbalance in the number of data according to the BC category. Because this may be a characteristic of the counseling data used in the study, we may have to check other types of BC data. In the future, we plan to further improve the performance of the BC prediction model using the attention mechanism between different modality information (Tsai et al., 2019) by simultaneously receiving multi-modality information as input.

Ethical Considerations

This study was approved by the Institutional Review Board (IRB) at the CHA medical center (CHAMC 2020-07-046). To preserve the privacy of patients and doctors in the counseling data, we applied two methods: anonymization and data deletion. The personally identifiable information of patients and doctors was fully anonymized. We deleted information that can be used to infer personal information. Data were securely managed. Only the researchers participating in this study had access to the data for both labeling and experimental work.

Acknowledgements

This work was supported by Institute for Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2021-0-00354, Artificial intelligence technology inferring issues and logically supporting facts from raw text)

References

Amalia Istiqlali Adiba, Takeshi Homma, Dario Bertero, Takashi Sumiyoshi, and Kenji Nagamatsu. 2021a. Delay mitigation for backchannel prediction in spoken dialog system. In *Conversational Dialogue Systems for the Next Decade*, pages 129–143. Springer.

Amalia Istiqlali Adiba, Takeshi Homma, and Toshinori Miyoshi. 2021b. Towards immediate backchannel generation using attention-based early prediction model. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7408–7412. IEEE.

Michelle Cohn, Chun-Yen Chen, and Zhou Yu. 2019. A large-scale user study of an alexa prize chatbot: Effect of tts dynamism on perceived quality of social dialog. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 293–306.

Pino Cutrone. 2010. The backchannel norms of native english speakers: A target for japanese l2 english learners. *Language Studies Working Papers*, 2:28–37.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Shinya Fujie, Kenta Fukushima, and Tetsunori Kobayashi. 2005. Back-channel feedback generation using linguistic and nonlinguistic information and its application to spoken dialogue system. In *Ninth European Conference on Speech Communication and Technology*.

Bettina Heinz. 2003. Backchannel responses as strategic responses in bilingual speakers' conversations. *Journal of pragmatics*, 35(7):1113–1142.

Tatsuya Kawahara, Takashi Yamaguchi, Koji Inoue, Katsuya Takanashi, and Nigel G Ward. 2016. Prediction and generation of backchannel form for attentive listening systems. In *Interspeech*, pages 2890–2894.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Sangwon Lee, Naeun Lee, and Young June Sah. 2020. Perceiving a mind in a chatbot: effect of mind perception and social cues on co-presence, closeness, and intention to use. *International Journal of Human–Computer Interaction*, 36(10):930–940.

Jingye Li, Hao Fei, and Donghong Ji. 2020. Modeling local contexts for joint dialogue act recognition and sentiment classification with bi-channel dynamic convolutions. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 616–626.

Daniel Ortega, Chia-Yu Li, and Ngoc Thang Vu. 2020. Oh, jeez! or uh-huh? a listener-aware backchannel predictor on asr transcriptions. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8064–8068. IEEE.

Sang-Min Park, Chul-Won Na, Min-Seong Choi, Da-Hee Lee, and Byung-Won On. 2018. Knu korean sentiment lexicon: Bi-lstm-based method for building a korean sentiment lexicon. *Journal of Intelligence and Information Systems*, 24(4):219–240.

Ronald Poppe, Khiet P Truong, Dennis Reidsma, and Dirk Heylen. 2010. Backchannel strategies for artificial listeners. In *International Conference on Intelligent Virtual Agents*, pages 146–158. Springer.

Libo Qin, Wanxiang Che, Yangming Li, Mingheng Ni, and Ting Liu. 2020. Dcr-net: A deep co-interactive relation network for joint dialog act recognition and sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8665–8672.

Justus J Randolph. 2005. Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa. *Online submission*.

Robin Ruede, Markus Müller, Sebastian Stüker, and Alex Waibel. 2017. Enhancing backchannel prediction using word embeddings. In *INTERSPEECH*, pages 879–883.

Benjamin J Sadock and Virginia A Sadock. 2011. *Kaplan and Sadock's synopsis of psychiatry: Behavioral sciences/clinical psychiatry*. lippincott williams & wilkins.

Hao Tian, Can Gao, Xinyan Xiao, Hao Liu, Bolei He, Hua Wu, Haifeng Wang, and Feng Wu. 2020. Skep: Sentiment knowledge enhanced pre-training for sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4067—4076.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Nigel Ward and Wataru Tsukahara. 2000. Prosodic features which cue back-channel responses in english and japanese. *Journal of pragmatics*, 32(8):1177–1207.

Victor H Yngve. 1970. On getting a word in edgewise. In *Chicago Linguistics Society, 6th Meeting, 1970*, pages 567–578.

Richard F Young and Jina Lee. 2004. Identifying units in interaction: Reactive tokens in korean and english conversations. *Journal of Sociolinguistics*, 8(3):380–407.