Interspeech 2024
1-5 September 2024, Kos, Greece

Backchannel prediction, based on who, when and what

Yo-Han Park[1], *Wencke Liermann*[1], *Yong-Seok Choi*[1,2], *Seung Hi Kim*[2], *Jeong-Uk Bang*[2], *Seung Yun*[2], *Kong Joo Lee*[1,*]

[1] Chungnam National University, Republic of Korea
[2] Electronics and Telecommunications Research Institute, Republic of Korea

happy115012@cnu.ac.kr, wliermann@o.cnu.ac.kr, yseokchoi@cnu.ac.kr,
{seunghi, jubang0219, syun}@etri.re.kr, kjoolee@cnu.ac.kr

Abstract

Backchannels are fundamental elements within conversations that serve as essential tools for effective communication and interpersonal dynamics. A typical backchannel prediction model primarily utilizes audio signal and text information. But backchanneling can exhibit different patterns depending on who I am, who I talk to, when I talk to them, and what I talk about. Therefore, we propose to employ three related pieces of information to enhance the quality of backchannel prediction models: speaker & listener characteristics, conversation progress, and topic. In our experiments with Korean counseling data, incorporating the suggested information into the model resulted in a performance improvement of 4.1% compared to the baseline model, increasing the F1 score from 50.1% to 54.2% .

Index Terms: backchannel, personality, conversation progress, topic

1. Introduction

Backchannels are conversational elements that the listener employs during the speaker's turn to promote effective communication and foster interpersonal dynamics [1]. Those subtle cues, such as nods, vocalic sounds, or brief remarks, not only indicate active involvement and understanding but also provide valuable feedback to speakers, encouraging them to prolong the conversation [2, 3, 4]. Backchannels also play an important role in cultivating a supportive environment that builds rapport and trust between individuals [5]. Therefore, for a robot or dialogue system to act like a human and increase user satisfaction, it must not only speak fluently but also select and use the appropriate backchannels [6, 7].

Backchannel prediction models generally employ both audio and textual information [8, 9, 10, 11]. The reason for that is simple. On the one hand, backchannels typically occur when there is a change in the speaker's pitch or a short pause [12]. On the other hand, their final surface form depends on the content of the preceding speaker utterance [11], ranging from short generic backchannels such as "Um." to more specific ones aligned with the speaker's words, such as "Really?" or "I see."

Once a defacto default encoding method for audio information, Mel Frequency Cepstral Coefficient (MFCC) [8, 9] have recently been superseded by more powerful hidden representations from pre-trained speech processing models [10, 11]. A similar trend can be observed for text encoding methods. While earlier models fed a fixed size window of 5 to 20 words into a BERT model to obtain a single embedding [8], recent models incorporate much broader context information using sequential or attentive approaches [10, 11].

In addition, building on the theoretical insight that backchannel behavior varies from person to person, [13] aimed at creating a model capable of emulating a specific listener's backchanneling behavior. To achieve this, they introduced listener embeddings trained simultaneously with the backchannel model starting from a listener ID.

However, a preliminary study shows that listener characteristics are not the sole factor determining backchanneling behavior. Instead, some additional external factors seem to be at play. As one possible such factor [14] suggested to include not only information on the listener but also the speaker and their relation. In this paper, we propose to add two more pieces of information that we deem critical for longer ongoing conversations: conversation progress and topic. Our assumption is that as the conversation evolves over time, backchannel patterns shift, influenced by progressively intensifying depth of discussion and relationship development. In a related but distinct manner, the conversation topic should have a significant impact on backchannel usage, with topic-specific intents such as seeking advice or sharing information changing the types of backchannels being used. By incorporating these two elements, we aim to improve backchannel prediction performance.

In summary, our main contributions are as follows:

1. While [14] used learnable embeddings on top of listener & speaker IDs to model conversation partners, we represent individuals not by IDs but through 16 interpretable features capturing an individual's conversation patterns.
2. In addition, we propose to include two more factors, namely conversation progress, and topic, and explore ways to embed them in the model.
3. We show the effectiveness of our approach through an experiment on a Korean counseling dataset, with conversations that each last about 50 minutes. On this data, we observe a 4.1% improvement in performance over a strong baseline.

2. Preliminary Study

To investigate whether listener characteristics alone determine backchanneling behavior, we conduct a preliminary study. In this study, data was collected from six listeners who participated in conversations with multiple individuals. As the number of data samples varies across listeners we perform undersampling to create training and test data splits of uniform size for each listener (3,720 and 1,309, respectively). Finally, we fit a backchannel prediction model to the portion of the training data belonging to each listener and then compare their performance when either evaluated on their own test data or test data from other individuals. Figure 1 illustrates the prediction performance across listeners. If backchanneling behavior de-

*Corresponding author.

10.21437/Interspeech.2024-2523

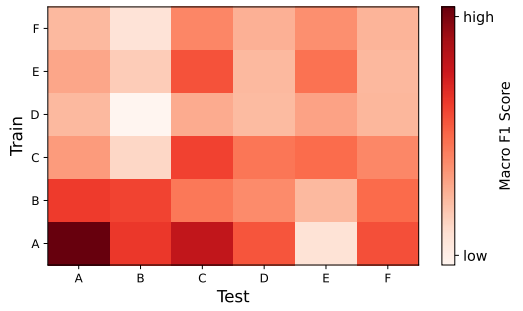

Figure 1: *Model performance across different individuals. The vertical axis shows the individual whose backchannel data was used during training, while the horizontal axis shows the individual for whom we predict backchannels during evaluation. The shading of each cell denotes the F1 score, where darker hues represent higher values.*

pended only on the listener's characteristics, we would expect a model to perform best when trained and tested on data from the same listener, as observed in the case of listener A. However, the given heat map shows no such clear pattern; instead, the best-performing model most often happens to be one that was trained on a different listener. This suggests that in addition to listener characteristics, some other external factors also determine backchanneling behavior.

3. Proposed Method

Our backchannel prediction model is based on the Context-Aware Backchannel Prediction (CABP) model [11]. CABP operates by taking in the audio signal, the current utterance, and the conversational context. It then processes each input to generate four distinct embeddings: current utterance embedding, sequential context embedding, attentive context embedding, and acoustic embedding. We further customize the backchannel prediction process through the addition of person embeddings ("Who am I?", "Who do I talk to?"), a conversation progress embedding ("When do I talk to them?"), and a topic embedding ("What do I talk about?"). Finally, all the generated embeddings are concatenated and input into a classifier to determine the appropriate backchannel type. The overall system architecture is illustrated in Figure 2.

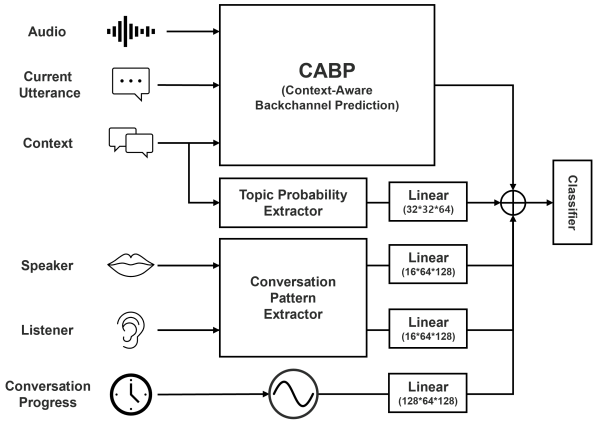

Figure 2: *The overall system model.*

3.1. Person Embedding

Backchanneling behavior is individual and personality dependent [15]. For instance, outgoing and approachable people tend to use backchannels more often. They also tend to use both verbal and nonverbal cues, while introverts typically use only the latter, e.g. nodding to indicate agreement. Therefore, backchannel patterns ultimately vary depending on the personalities of the conversation partners. [14] integrated both the speaker and the listener into the model's encoding. This was accomplished by learning embeddings for each individual. However, relying solely on this approach has its limitations, especially for individuals with fewer conversational interactions or those not present in the training dataset. Therefore, we intend to learn person embedding based on an individual's conversation patterns rather than a discrete, not interpretable index.

First, we define a set of conversation patterns. These patterns include, among others, the average duration of a speaking turn, the frequency of backchannel responses, and the distribution of specific backchannel types. Table 1 shows all 16 conversation patterns we identified.

Table 1: *Conversation Patterns.*

Average/Standard deviation number of filler per turn
Average/Standard deviation duration per turn
Average/Standard deviation number of words per turn
Average number of backchannels per turn
Percentage of Continuers/Understanding/Empathetic
Average number of Continuer/Understanding/Empathetic per minute
Average number of Continuer/Understanding/Empathetic per turn

Patterns are extracted from the conversation that a backchannel prediction is to be made for, normalized based on their maximum value observed across the data set, and then input into a linear layer to generate person embeddings.

More precisely, like [14], we distinguish between an individual's embedding when in the listener role compared to when in the speaker role. For this purpose, we apply the respective linear layer, depending on the role.

3.2. Conversation Progress Embedding

Participants typically begin a conversation with greetings and small talk before moving on to the main topic. As the conversation progresses, people gradually reveal more about themselves and establish rapport. Figure 3 suggests that backchannel patterns also evolve over time. It shows fluctuations in the density of backchannel usage over the course of a conversation for three unique individuals. Initially, backchannel usage is minimal, but as the conversation progresses, the density steadily increases before decreasing again towards the end. To capture this dynamic, we incorporate information on conversation progress in our backchannel prediction model.

To encode this information, we adopt fixed positional embeddings in the form of sinusoidal functions like they are commonly applied in the transformer position embeddings [16]. Sinusoidal embeddings (SE) are defined as:

$$SE_{(t,2i)} = sin(t/10000^{2i/d_{emb}})$$
$$SE_{(t,2i+1)} = cos(t/10000^{2i/d_{emb}}) \quad (1)$$

where t represents a point in time, while d_{emb} and i are the embedding dimension and the index position to be calculated, respectively.

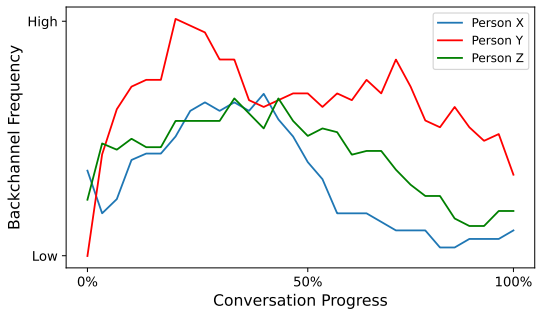

Figure 3: *Backchannel frequency in relation to conversation progress.*

However, sinusoidal embeddings alone are not suitable for representing continuous time. Therefore, we adopt a strategy of discretizing time into intervals, where each interval corresponds to a predefined duration. We settled on 30-second intervals as the result of a hyperparameter study with values of 10, 30, and 60 seconds. By segmenting time into these intervals, we can assign each point in time an integer t, which is used to extract the corresponding sinusoidal embedding (with $d_{emb} = 128$). This embedding is then put through a linear layer to generate the final conversation progress embedding.

3.3. Topic Embedding

Backchannel behavior reflects a listener's comprehension of and interest in specific topics. Backchannels can indicate agreement with the topics being discussed or confusion when encountering unfamiliar topics. Similarly, in counseling, backchannels may signal involvement or support, the density of which naturally increases with the sensitivity of the topic. Conversations like this, about emotions, may evoke empathetic responses, while discussions focused on practical problem-solving may rather evoke expressions of understanding. For this reason, we include a conversation topic embedding in our backchannel prediction model.

We utilize BERTopic [17] to extract topics from conversations. BERTopic employs a document-topic modeling approach, clustering documents using BERT embeddings and class-based tf-idf. However, considering an entire conversation as a single document is computationally expensive and lacks precision when considering topic change and development. Therefore, we instead define three consecutive turns as fragments of conversation to which the topic model will be applied.

Using BERTopic, we then compute the topic probabilities for each conversation fragment up to the current utterance. Finally, the overall topic probability is obtained as the weighted moving average of the topic probabilities of all previous conversation fragments:

$$p_j^{overall} = (1 - q) * p_{j-1}^{overall} + q * p_j \quad (2)$$

where p_j is the topic probability of the jth conversation fragment. q regulates how much importance is attributed to past and present topics, with higher values assigning more importance to the present. We use $q = 0.5$. The computed topic probabilities in the form of a vector are then entered into a linear layer to create the final topic embedding.

4. Experiments

4.1. Dataset

In our experiments, we employ a small private dataset of Korean counseling sessions collected by ETRI[1]. The dataset consists of 85 conversations, each lasting about 50 minutes and totaling 72 hours. They constitute first-time interactions between a counselor and counselee. There are a total of 91 participants, including 6 counselors and 85 counselees. Thus, each counselee participates in one conversation, while each counselor interacts with at least 7 counselees.

Backchannel annotations cover the following categories: Continuer (CONT), Understanding (UND), and Empathetic (EMPH). A Continuer is a brief vocalic cue or repetition; signalling continued attention to the speaker. Understanding covers longer cues that indicate that a speaker's statement has been understood. Empathetic responses involve actively mirroring the listener's emotions, such as admiration, surprise, or agreement.

We enrich the data with additional "NoBC" instances, which we retrieve through two different approaches. First, we adopt the approach used by [13], which assigns the position two seconds before each backchannel occurrence the label "NoBC". Second, we collect positions of speaker change, i.e. points of turn transition, as "NoBC" instances, because in dialogue systems it is important to distinguish when to take the next turn and when to produce a backchannel [18]. For both methods, instances are removed if they fall within 1 second of an existing backchannel instance. As a result, we obtain a data set of 65,789 instances, the statistical breakdown of which is shown in table 2.

Table 2: *Korean Counseling Backchannel Data Statistics*

Category	# of Data	Ratio
Continuer (CONT)	18,847	28.6%
Understanding (UND)	8,659	13.1%
Empathetic (EMPH)	4,643	7.1%
NoBC	33,640	51.2%
Total	65,789	100%

4.2. Baseline

As our base model, we employ a modified version of [11], making it more capable of exploiting past backchanneling patterns through its contextual embedding. As is, [11] stores only utterances in its context memory and, therefore, lacks insight into recent backchannel usage. To overcome this limitation, we store past backchannels like regular utterances and increase the memory size from 7 to 10. To distinguish between utterance and backchannel, we prepend one of four learnable backchannel embeddings to each utterance before passing it to BERT, i.e. "[CLS] [CONT|UND|EMPH|NOBC] [Counselor|Counselee] Text$_T$". Finally, we also increased the length of the audio signal input from 1.5 to 3 seconds.

4.3. Experimental Setup

The CABP model employs wav2vec 2.0[2] and KoBERT[3] as pretrained models. The model was trained for 20 epochs with early

[1]Electronics and Telecommunications Research Institute
[2]https://huggingface.co/facebook/wav2vec2-base
[3]https://aiopen.etri.re.kr/

Table 3: *Ablation study results as averaged across 5 random seeds (± one standard deviation). Asterisks (∗) and cross marks (†) mean significantly better than the baseline model and person model, respectively ($p < 0.05$). **Bold** represents the highest score.*

Person	Progress	Topic	Params	Macro-F1	Continuer	Understanding	Empathetic	NoBC
-	-	-	206M	50.1 (±0.3)	59.6 (±1.9)	36.9 (±1.3)	23.3 (±1.7)	80.8 (±0.4)
-	+	-	+148k	50.7 (±0.7)	60.0 (±2.6)	39.1 (±1.1)*	23.3 (±3.6)	80.5 (±0.7)
-	-	+	+69K	49.5 (±1.8)	59.2 (±1.8)	38.4 (±1.2)	20.1 (±6.8)	80.2 (±0.6)
-	+	+	+216K	49.9 (±1.7)	58.9 (±2.9)	38.5 (±1.1)*	21.4 (±5.9)	80.9 (±0.5)
+	-	-	+281K	53.5 (±0.5)*	**64.4** (±**0.8**)*	**42.1** (±**1.8**)*	26.5 (±2.4)	80.8 (±0.4)
+	+	-	+428K	**54.2** (±**0.1**)†	64.3 (±1.3)	41.8 (±3.0)	29.9 (±3.9)†	**81.0** (±**0.8**)
+	-	+	+349K	53.3 (±0.5)	**64.4** (±**0.5**)	41.5 (±1.9)	26.8 (±2.4)	80.6 (±1.0)
+	+	+	+497K	**54.2** (±**0.6**)	64.2 (±1.0)	41.1 (±4.0)	**30.3** (±**4.6**)†	**81.0** (±**0.3**)

stopping and a batch size of 32. The learning rate was set to 3e-4 for the pre-trained models and 1e-5 for all other parameters. AdamW optimization was used along with a cosine annealing schedule and warm-up ratios of 0.3 and 0.1 for the pre-trained model and other parameters, respectively.

For our experiments, we divide the train-validate-test dataset using a 3:1:1 ratio per conversation. As evaluation metrics, we chose the F1 score for each backchannel category as well as their macro-average F1. Each model was trained on 5 randomized seeds, and the performance comparison is based on the average of these trials.

5. Results and Discussion

5.1. Experiment Results

Table 3 shows the results of an ablation study that investigates the effectiveness of each single embedding type and their combination. The introduction of the conversation progress embedding increased prediction performance from a macro F1 of 50.1% to 50.7% compared to the baseline. In particular, a major improvement of 2.2% (from 36.9% to 39.1%) could be observed for the category "Understanding". In contrast, the topic embedding hurt prediction performance, causing the macro F1 to drop to 49.5%. Finally, the person embedding proved most effective with a significant improvement of 3.4%, reaching an F1 score of 53.5% and improving the F1 across all categories except "NoBC" by each more than 3%.

Building on this observation, we now add the conversation progress and/or topic embedding to a model with the person embedding. The combination of person and progress embedding yielded the highest result, with an F1 score of 54.2%. Notably, there was a significant improvement from 26.5% to 29.9% for the category "Empathetic". Again, the addition of the topic model resulted in a slight performance decrease of 0.2%. Finally, using all three features returned an F1 score of 54.2%, identical to the model using only person and time embeddings.

In addition, we compare our proposed approach with other related work that builds on person embeddings. [14] used speaker and listener embeddings as inputs to a neural tensor network [19] to derive an embedding of their relationship. We apply this approach directly to our model, adding only the neural tensor network. Experimental results show that this model yields a macro F1 score of 53.3%, which is similar to the performance of the model that uses only person embeddings. One possible reason for this result could be that this embedding is supposed to capture the relationship between conversation partners. However, in a counseling setting, this relationship is already predetermined and fixed as counselor-counselee.

5.2. Discussion

The results presented in Table 3 indicate that incorporating a topic embedding does not provide the expected improvements; rather, it appears to have either no effect or even degrade performance. This could indicate, among other things, that the model already captures vague notions of topic, which is possible as CABP captures context in a sequential and attentive manner, extending beyond the current utterance. To investigate this hypothesis, we conduct a small experiment that evaluates the effectiveness of topic embeddings when integrated into a BERT-only model that relies solely on the current utterance as input. We observed that this BERT-only model could benefit from the addition of topic embeddings, yielding an increase in macro F1 of 1.9% (from 36.2% to 38.1%).

Topic embeddings incorporate information from previous utterances. Similarly, CABP uses conversation context to predict backchannels. Thus, it appears that the ineffectiveness when used in tandem with CABP is because this model implicitly encodes topic information.

6. Conclusion

We proposed a new promising way to encode speaker & listener information in a backchannel prediction model using 16 interpretable features that capture an individual's conversation patterns. Additionally, we showed that adding a conversation progress embedding can further improve prediction performance, while an explicit topic embedding seems redundant in the presence of contextual text embeddings. Person and progress embeddings, when applied together, yield a significant improvement in macro F1 of 4.1% (from 50.1% to 54.2%) compared to a strong baseline.

7. Limitation

We proposed to generate person embeddings on top of 16-dimensional feature vectors capturing different conversation patterns. However, such patterns may be unknown in a real-world setting, especially for new conversation partners. Future research should look into methods to bridge this gap. For instance, one might experiment with retrieving pre-defined patterns based on a person's first impression, a preliminary questionnaire, or a 10-minute mock conversation.

The data set applied in this study was chosen because it covers longer conversations (∼50min) compared to the publicly available Switchboard Corpus data that is limited to short exchanges (∼5min). However, we cannot openly share it with the community, as it contains sensitive information about real individuals talking about their personal experiences.

8. Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2022-0-00608, A development of artificial intelligence technology of multimodal interaction for empathetic and social conversations with humans)

9. References

[1] A. Amer, C. Bhuvaneshwara, G. K. Addluri, M. M. Shaik, V. Bonde, and P. Müller, "Backchannel detection and agreement estimation from video with transformer networks," 2023.

[2] R. Ishii, X. Ren, M. Muszynski, and L.-P. MorencY, "Multimodal and multitask approach to listener's backchannel prediction: Can prediction of turn-changing and turn-management willingness improve backchannel modeling," in *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents (IVA)*, 2021.

[3] T. Kawahara, T. Yamaguchi, K. Inoue, K. Takanashi, and N. G. Ward, "Prediction and generation of backchannel form for attentive listening systems." in *Interspeech*, 2016, pp. 2890–2894.

[4] A. Morikawa, R. Ishii, H. Noto, A. Fukayama, and T. Nakamura, "Determining most suitable listener backchannel type for speaker's utterance," in *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*, 2022, pp. 1–3.

[5] L. Huang, L.-P. Morency, and J. Gratch, "Virtual rapport 2.0," in *Intelligent Virtual Agents: 10th International Conference, IVA 2011, Reykjavik, Iceland, September 15-17, 2011. Proceedings 11*. Springer, 2011, pp. 68–79.

[6] K. Inoue, D. Lala, K. Yamamoto, S. Nakamura, K. Takanashi, and T. Kawahara, "An attentive listening system with android erica: Comparison of autonomous and woz interactions," in *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2020, pp. 118–127.

[7] A. I. Adiba, T. Homma, and T. Miyoshi, "Towards immediate backchannel generation using attention-based early prediction model," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7408–7412.

[8] J. Y. Jang, S. Kim, M. Jung, S. Shin, and G. Gweon, "Bpm_mt: Enhanced backchannel prediction model using multi-task learning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 3447–3452.

[9] R. Ruede, M. Müller, S. Stüker, and A. Waibel, "Enhancing Backchannel Prediction Using Word Embeddings," in *Proc. Interspeech 2017*, 2017, pp. 879–883.

[10] W. Liermann, Y.-H. Park, Y.-S. Choi, and K. Lee, "Dialogue act-aided backchannel prediction using multi-task learning," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 15 073–15 079.

[11] Y.-H. Park, W. Liermann, Y.-S. Choi, and K. J. Lee, "Improving backchannel prediction leveraging sequential and attentive context awareness," in *Findings of the Association for Computational Linguistics: EACL 2024*, 2024, pp. 1689–1694.

[12] K. P. Truong, R. W. Poppe, and D. K. Heylen, "A rule-based backchannel prediction model using pitch and pause information," in *11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010*. International Speech Communication Association (ISCA), 2010, pp. 3058–3061.

[13] D. Ortega, C.-Y. Li, and N. T. Vu, "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)*. IEEE, 2020, pp. 8064–8068.

[14] D. Ortega, S. Meyer, A. Schweitzer, and N. T. Vu, "Modeling speaker-listener interaction for backchannel prediction," 2023.

[15] P. Blomsma, G. Skantze, and M. Swerts, "Backchannel behavior influences the perceived personality of human and artificial communication partners," *Frontiers in Artificial Intelligence*, vol. 5, p. 835298, 2022.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 6000–6010.

[17] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," 2022.

[18] N. G. Ward, D. Aguirre, G. Cervantes, and O. Fuentes, "Turn-taking predictions across languages and genres using an lstm recurrent neural network," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 831–837.

[19] R. Socher, D. Chen, C. D. Manning, and A. Ng, "Reasoning with neural tensor networks for knowledge base completion," *Advances in neural information processing systems*, vol. 26, 2013.