A Survey on Evaluation Metrics for Backchannel Prediction Models

Iwan de Kok[1], *Dirk Heylen*[1]

[1]Human Media Interaction, University of Twente, Enschede, The Netherlands

i.a.dekok@utwente.nl, heylen@utwente.nl

Abstract

In this paper we give an overview of the evaluation metrics used to measure the performance of backchannel prediction models. Both objective and subjective evaluation metrics are discussed. The survey shows that almost every backchannel prediction model is evaluated with a different evaluation metric. This makes comparison between developed models unreliable, even beside the other variables in play, such as different corpora, language, conversational setting, amount of data and/or definition of the term backchannel.

Index Terms: backchannel, machine learning, evaluation metrics

1. Introduction

One of the aspects of nonverbal behavior that has been a subject for computational modeling for many years is backchanneling behavior. Originally these backchannel prediction models were developed for spoken dialog systems for telecommunication purposes, but nowadays the aim for these models are virtual humans and robots.

In this paper we will give an overview of metrics and methods used to evaluate the backchannel prediction models developed so far. Table 1 gives an overview of the backchannel prediction models and their evaluation methods. As the table shows there are almost as many evaluation methods as there are backchannel prediction models. This makes a comparison between the different approaches very difficult.

The evaluation methods used can be divided into two categories; objective evaluation or subjective evaluation. The paper is organized to discuss these two evaluation methods separately.

With objective evaluation the performance of the model is compared to (another part of) the corpus that is used for development. The evaluation analyzes how good the model is at reproducing backchanneling behavior of the recorded listener. This type of evaluation has the challenge that people differ in their backchanneling behavior. The responses given by the recorded listener are not the only moments in the conversation where a backchannel is possible or required. Predictions at other times might be just as good. In Section 2 the different measurements and approaches to objectively evaluate the developed backchannel prediction models and deal with this challenge are presented in more detail.

With subjective evaluation observers are used to judge the generated backchanneling behavior of the model. The evaluation analyzes the capability of the model to produce correct and natural backchanneling behavior as perceived by humans. This type of evaluation circumvents the challenges for objective evaluations, but it is more time consuming to perform and is thus unsuited for validating settings of the models and/or rapid prototyping. In Section 3 the different measurements and approaches to subjectively evaluate the backchannel prediction models are presented in more detail.

The paper is concluded with our final thoughts on the subject and recommendations for the future.

2. Objective Evaluations

In objective evaluations of backchannel prediction models the backchannel predictions made by the models are compared to the ground truth. A measure is selected which quantifies the comparison. Measures that are used to report objective evaluations include cross-correlation coefficient [1], precision and recall [2, 3, 4, 7, 16] or F_1 (which is the weighted harmonic mean of precision and recall) [5, 8, 10, 11, 14, 15, 17, 17, 18]. Most authors opt for a measure based on precision and recall, but in three areas differences between measures remain, namely ground truth selection, segmentation and margin of error.

2.1. Ground Truth Selection

The majority of evaluations of backchannel prediction model are performed by comparing the predictions made by the model with the listener in the corpus [2, 4, 5, 7, 8, 11, 14, 15, 16, 17, 18]. As Ward and Tsukahara [4] have noted this is not ideal. When analyzing the performance of their predictive rule they conclude that 44% of the incorrect predictions were cases where a backchanel could naturally have appeared, as judged by one of the authors, but in the corpus there was silence or, more rarely, the start of a turn. Cathcart et al. [5] dealt with this problem by only using high backchannel rate data as test data in order to minimize false negatives.

Others have dealt with this problem by collecting multiple perspectives on appropriate times to provide a

Authors	Subjective	Objective	Objective Metric	Ground Truth	Segmentation	Margin of Error
Watanabe & Yuuki (1989) [1]		✓	Cross-Correlation Coef.	Multiple (Nodding)	Continuous	-
Okato et al. (1996) [2]		✓	Precision / Recall	Single	Continuous	-100/500ms
Noguchi et al. (1998) [3]		✓	Precision / Recall	Multiple (Keyboard)	Pause-Bound Phrases	-
Ward & Tsukahara (2000) [4]		✓	Precision / Recall	Single	Continuous	-500/500ms
Cathcart (2003) [5]		✓	F_1	Single	Words	-
Fujie et al. (2004) [6]	✓		-	-	-	-
Takeuchi (2004) [7]	✓	✓	Precision / Recall	Single	100ms Pause Frames	-
Kitaoka et al. (2005) [8]	✓	✓	F_1	Multiple (Keyboard)	100ms Pause Frames	-
Nishimura et al. (2007) [9]	✓		-	-	-	-
Morency et al. (2008) [10]		✓	F_1	Single	Continous	0/1000ms
De Kok et al. (2010) [11]		✓	F_1 / $F_{Consensus}$	Multiple (Parallel)	Continous	-500/500ms
Huang et al. (2010) [12]	✓		-	-	-	-
Huang et al. (2010) [13]	✓		-	-	-	-
Ozkan & Morency (2010) [14]		✓	F_1	Single	Continous	0/1000ms
Ozkan & Morency (2010) [15]		✓	F_1	Single	Continous	0/1000ms
Poppe et al. (2010) [16]		✓	F_1	Single	Continous	-200/200ms
De Kok et al. (2012) [17]	✓	✓	F_1	Single	Continous	-500/500ms
Ozkan & Morency (2012) [18]		✓	F_1 / UPA	Single	Continous	0/1000ms

Table 1: Overview of the corpus based backchannel prediction models developed so far.

backchannel. This was either by asking multiple people to press a key on a keyboard at times they would give a backchannel in reaction to a recorded speaker [3, 12, 13], asking multiple people to intentionally nod [1] or by recording multiple listeners in parallel who were led to believe the only listener [11].

Recently two measures have been proposed that are specifically aimed at being applied to such multiple perspective data, $F_{Consensus}$ [11] and User-Adaptive Prediction Accuracy [18].

De Kok et al. [11] recorded 3 listeners in parallel interaction with the same speaker. Each listener was unaware of the other two listeners. Combining the three 'versions' of the ground truth, moments are identified where one, two or three listeners responded. Following the reasoning that the moments where more listeners performed a backchannel are more important for a model to predict, but a prediction should only be regarded as being false if it is at a moment where none of the listeners performed a backchannel they proposed the $F_{Consensus}$ metric. In this metric precision is calculated using all the moments a listener performed a backchannel as ground truth, while recall is calculated using only the moments where the majority of listeners performed a backchannel as ground truth. The weighted harmonic mean is taken as the final performance measure.

Ozkan and Morency [18] have proposed User-Adaptive Prediction Accuracy as an evaluation metric for backchannel prediction models. For this measure the model is asked for n most likely backchannel moments in reaction to a speaker, where n is the number of backchannel given by the ground truth listener. This measure allows evaluation of the ability of the model to adapt to different listeners. Some listeners may backchannel frequently, while others backchannel only a limited number of times.

2.2. Segmentation

With regards to segmentation the majority of models are evaluated on continuous data [1, 2, 4, 10, 11, 14, 15, 16, 17, 18]. This means that a prediction for a backchannel can be made at any time during the interaction, usually at a 10ms interval. However, some models have limitations that segment the interaction in bigger chunks of data.

Noguchi and Den [3] use pre-delimited pause-bounded phrases as data. The proposed backchannel prediction model predicts for each such segment whether it is followed by a backchannel or not. Cathcart et al. [5] make a similar decision after each word.

Both Takeuchi et al. [7] and Kitaoka et al. [8] have proposed a model that classify frames with no speech from the speaker. These pauses were split into segments of 100ms. For each of these segments the pause was classified as either 'making a backchannel', 'taking the turn', 'waiting for the speaker to continue' or 'waiting to make a backchannel or take the turn'.

2.3. Margin of Error

For the models evaluated using precision and recall based measures on continous data another discriminating factor applies, namely the margin of error. Precision and recall based measures rely on the evaluation whether a prediction is 'at the same time' as the ground truth. The definition of 'at the same time' differ between evaluations. Okato et al. [2] use a margin of error of -100ms to +300ms from the onset of the ground truth backchannel, Ward and Tsukahara [4] and De Kok et al. [11, 17] use a margin of error of -500ms to +500ms, Poppe et al. [16] use a margin of -200ms to +200ms, and Morency et al. [10] and Ozkan et al. [14, 15, 18] use a margin of error of 0ms to +1000ms.

3. Subjective Evaluations

When it comes to subjective error measures several strategies have been used to establish the performance of the models. The approaches used so far either evaluate a general impression of the backchannel behavior or individual backchannels.

Fujie et al. [6] made a pair-wise comparison between models in which the general impression of the backchannel behavior is measured. A subject interacted twice with a conversation robot system which backchanneling behavior was driven by two different models. After these interactions the subject was asked on a 5 point scale, which system they preferred, with 1 being system A, 5 being system B and 3 being no preference.

Huang et al. [12] also evaluated the general impression of the backchannel behavior. They generated virtual listeners in response to recorded speakers and presented these interactions to 17 subjects. Similar to Fujie et al. [6] the subjects were presented with three different virtual listeners each driven by a different backchannel prediction model. After each interaction the subject was asked 7 questions about their perceived experience with regard to the timing of backchannels. On a 7-point Likert scale the subjects rated the virtual listeners on 'closeness', 'engrossment', 'rapport', 'attention', 'amount of inappropriate backchannels', 'amount of missed opportunities' and 'naturalness'.

Poppe et al. [16] also let participants evaluate virtual listeners in interaction with recorded speakers. They asked participants for each fragment "How likely do you think the listener's backchannel behavior has been performed by a human listener". The participants made their judgement by setting a slider that corresponded to a value between 0 and 100.

Kitaoka et al. [8] had 5 subjects rate each generated backchannels individually. The data presented to the subjects were 16 to 18 samples of single sentences followed

by a backchannel. Each generated backchannel rated was rated on a 5-point scale ranging from 'early' to 'late', with an extra option for 'outlier'. They did this process for backchannels generated at times predicted by their model and times as found in the corpus. They accumulated the counts of the 5 subjects and reported the percentage of ratings in the "good" category (rating 3). The same approach was used by Nishimura et al. [9].

De Kok et al. [17] evaluated their models in a similar fashion as Kitaoka et al. [8]. Subjects judged individual backchannels on their appriopriateness. Contrary to Kitaoka et al. the process was done in real time and over the course of multiple conversational moves keeping the backchannels in context. Subjects would hit the spacebar on a keyboard when they would see an inappropriately timed backchannel. As an evaluation metric they presented the percentage of backchannels that were not judged as inappropriate by any of the judges.

4. Conclusion

As this survey has shown, a wide variety of evaluation metrics have been used in the past. This makes comparing different methods in terms of performance even more complicated than it already is. Most models are trained and testing on different corpora, which differ in language, type of conversations, amount of data and exact definition of backchannel. This already makes a comparison between reported values unreliable. On top of these differences the evaluation methods used also differ from each other. Some evaluation measures are used often (such as F_1), but even then a direct comparison is not always fair because of differences in segmentation or margin of error. Also in subjective evaluations differences are there.

Development of backchannel prediction models would benefit from a unified way to evaluate performance. It would give more insight into the performance of the model in comparison to previous work. A benchmark corpus would be the ideal for this purpose, but a unified evaluation metric would be a start.

5. References

[1] T. Watanabe and N. Yuuki, "A Voice Reaction System with a Visualized Response Equivalent to Nodding," in *Proceedings of the third international conference on human-computer interaction, Vol.1 on Work with computers: organizational, management, stress and health aspects*, 1989, pp. 396–403.

[2] Y. Okato, K. Kato, M. Kamamoto, and S. Itahashi, "Insertion of interjectory response based on prosodic information," *Proceedings of IVTTA '96. Workshop on Interactive Voice Technology for Telecommunications Applications*, pp. 85–88, 1996.

[3] H. Noguchi and Y. Den, "Prosody-based detection of the context of backchannel responses," in *Fifth International Conference on Spoken Language Processing*, 1998.

[4] N. Ward and W. Tsukahara, "Prosodic features which cue back-channel responses in English and Japanese," *Journal of Pragmatics*, vol. 32, no. 8, pp. 1177–1207, 2000.

[5] N. Cathcart, J. Carletta, and E. Klein, "A shallow model of backchannel continuers in spoken dialogue," *European ACL*, pp. 51–58, 2003.

[6] S. Fujie, K. Fukushima, and T. Kobayashi, "A conversation robot with back-channel feedback function based on linguistic and non-linguistic information," in *Proc. Int. Conference on Autonomous Robots and Agents*, 2004, pp. 379–384.

[7] M. Takeuchi, N. Kitaoka, and S. Nakagawa, "Timing detection for realtime dialog systems using prosodic and linguistic information," *International Conference on Speech Prosody*, pp. 529–532, 2004.

[8] N. Kitaoka, M. Takeuchi, R. Nishimura, and S. Nakagawa, "Response Timing Detection Using Prosodic and Linguistic Information for Human-friendly Spoken Dialog Systems," *Transactions of the Japanese Society for Artificial Intelligence*, vol. 20, pp. 220–228, 2005.

[9] R. Nishimura, N. Kitaoka, and S. Nakagawa, "A spoken dialog system for chat-like conversations considering response timing," in *Proceedings of the 10th International Conference on Text, Speech and Dialogue*. Springer, 2007, pp. 599–606.

[10] L.-P. Morency, I. de Kok, and J. Gratch, "Predicting Listener Backchannels: A Probabilistic Multimodal Approach," in *Intelligent Virtual Agents*, 2008, Conference proceedings (article), pp. 176–190.

[11] I. de Kok, D. Ozkan, D. Heylen, and L.-P. Morency, "Learning and Evaluating Response Prediction Models using Parallel Listener Consensus," in *Proceeding of International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, 2010.

[12] L. Huang, L.-P. Morency, and J. Gratch, "Parasocial Consensus Sampling: Combining Multiple Perspectives to Learn Virtual Human Behavior," in *Proceedings of Autonomous Agents and Multi-Agent Systems*, Toronto, Canada, 2010, pp. 1265–1272.

[13] ——, "Learning Backchannel Prediction Model from Parasocial Consensus Sampling : A Subjective Evaluation," in *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2010, pp. 159–172.

[14] D. Ozkan and L.-P. Morency, "Concensus of Self-Features for Nonverbal Behavior Analysis," in *Human Behavior Understanding*, 2010.

[15] D. Ozkan, K. Sagae, and L.-P. Morency, "Latent Mixture of Discriminative Experts for Multimodal Prediction Modeling," in *Proceedings of the 23rd International Conference on Computational Linguistics*. Association for Computational Linguistics, 2010, pp. 860–868.

[16] R. Poppe, K. P. Truong, D. Reidsma, and D. Heylen, "Backchannel Strategies for Artificial Listeners," in *Intelligent Virtual Agents*, Philadelphia, Pennsylvania, USA, 2010, pp. 146–158.

[17] I. de Kok, R. Poppe, and D. Heylen, "Iterative Perceptual Learning for Social Behavior Synthesis," Centre for Telematics and Information Technology University of Twente, Tech. Rep., 2012.

[18] D. Ozkan and L.-P. Morency, "Latent Mixture of Discriminative Experts," *Accepted for publication in ACM Transaction on Multimedia*.