Modeling Wisdom of Crowds Using Latent Mixture of Discriminative Experts

Derya Ozkan and Louis-Philippe Morency
Institute for Creative Technologies
University of Southern California
{ozkan,morency}@ict.usc.edu

Abstract

In many computational linguistic scenarios, training labels are subjectives making it necessary to acquire the opinions of multiple annotators/experts, which is referred to as "wisdom of crowds". In this paper, we propose a new approach for modeling wisdom of crowds based on the Latent Mixture of Discriminative Experts (LMDE) model that can automatically learn the prototypical patterns and hidden dynamic among different experts. Experiments show improvement over state-of-the-art approaches on the task of listener backchannel prediction in dyadic conversations.

1 Introduction

In many real life scenarios, it is hard to collect the actual labels for training, because it is expensive or the labeling is subjective. To address this issue, a new direction of research appeared in the last decade, taking full advantage of the "wisdom of crowds" (Surowiecki, 2004). In simple words, wisdom of crowds enables parallel acquisition of opinions from multiple annotators/experts.

In this paper, we propose a new method to fuse wisdom of crowds. Our approach is based on the Latent Mixture of Discriminative Experts (LMDE) model originally introduced for multimodal fusion (Ozkan et al., 2010). In our Wisdom-LMDE model, a discriminative expert is trained for each crowd member. The key advantage of our computational model is that it can automatically discover the prototypical patterns of experts and learn the dynamic between these patterns. An overview of our approach is depicted in Figure 1.

We validate our model on the challenging task of listener backchannel feedback prediction in dyadic conversations. Backchannel feedback includes the nods and paraverbals such as "uh-huh" and "mm-hmm" that listeners produce as they are speaking. Backchannels play a significant role in determining the nature of a social exchange by showing rapport and engagement (Gratch et al., 2007). When these signals are positive, coordinated and reciprocated, they can lead to feelings of rapport and promote beneficial outcomes in diverse areas such as negotiations and conflict resolution (Drolet and Morris, 2000), psychotherapeutic effectiveness (Tsui and Schultz, 1985), improved test performance in classrooms (Fuchs, 1987) and improved quality of child care (Burns, 1984). Supporting such fluid interactions has become an important topic of virtual human research. In particular, backchannel feedback has received considerable interest due to its pervasiveness across languages and conversational contexts. By correctly predicting backchannel feedback, virtual agent and robots can have stronger sense of rapport.

What makes backchannel prediction task well-suited for our model is that listener feedback varies between people and is often optional (listeners can always decide to give feedback or not). A successful computational model of backchannel must be able to learn these variations among listeners. Wisdom-LMDE is a generic approach designed to integrate opinions from multiple listeners.

In our experiments, we validate the performance of our approach using a dataset of 43 storytelling dyadic interactions. Our analysis suggests three pro-

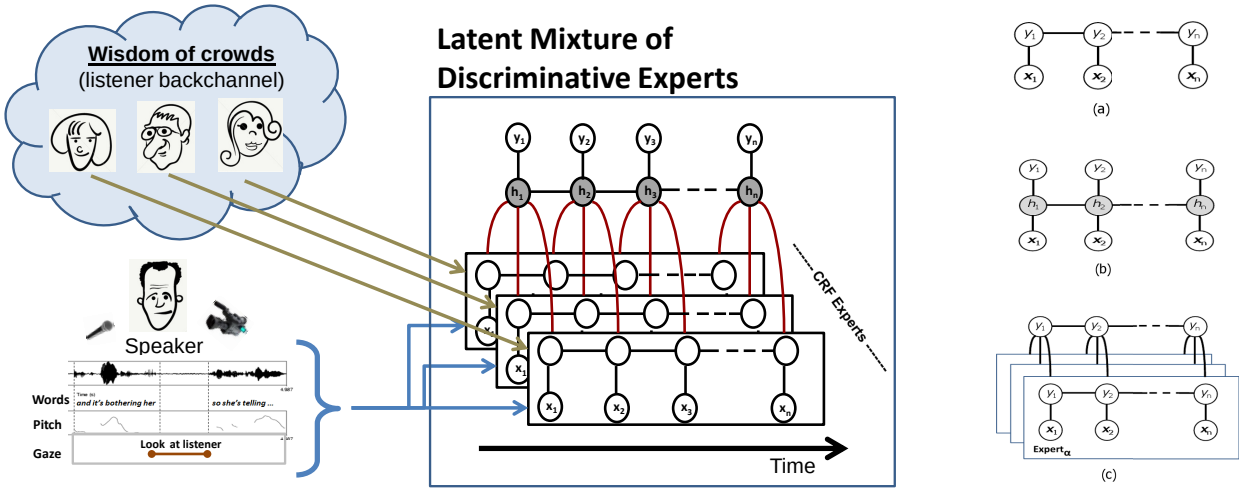

Figure 1: **Left:** Our approach applied to backchannel prediction: (1) multiple listeners experience the same series of stimuli (pre-recorded speakers) and (2) a Wisdom-LMDE model is learned using this wisdom of crowds, associating one expert for each listener. **Right:** Baseline models used in our experiments: **a)** Conditional Random Fields (CRF), **b)** Latent Dynamic Conditional Random Fields (LDCRF), **c)** CRF Mixture of Experts (no latent variable)

totypical patterns for backchannel feedback. By automatically identifying these prototypical patterns and learning the dynamic, our Wisdom-LMDE model outperforms the previous approaches for listener backchannel prediction.

1.1 Previous Work

Several researchers have developed models to predict when backchannel should happen. Ward and Tsukahara (2000) propose a unimodal approach where backchannels are associated with a region of low pitch lasting 110ms during speech. Nishimura et al. (2007) present a unimodal decision-tree approach for producing backchannels based on prosodic features. Cathcart et al. (2003) propose a unimodal model based on pause duration and trigram part-of-speech frequency.

Wisdom of crowds was first defined and used in business world by Surowiecki (2004). Later, it has been applied to other research areas as well. Raykar et. al. (2010) proposed a probabilistic approach for supervised learning tasks for which multiple annotators provide labels but not an absolute gold standard. Snow et. al. (2008) show that using non-expert labels for training machine learning algorithms can be as effective as using a gold standard annotation.

In this paper, we present a computational approach for listener backchannel prediction that exploits multiple listeners. Our model takes into account the differences in people's reactions, and automatically learns the hidden structure among them.

The rest of the paper is organized as follows. In Section 2, we present the wisdom acquisition process. Then, we describe our Wisdom-LMDE model in Section 3. Experimentals are presented in Section 4. Finally, we conclude with discussions and future works in Section 5.

2 Wisdom Acquisition

It is known that culture, age and gender affect people's nonverbal behaviors (Linda L. Carli and Loeber, 1995; Matsumoto, 2006). Therefore, there might be variations among people's reactions even when experiencing the same situation. To efficiently acquire responses from multiple listeners, we employ the Parasocial Consensus Sampling (PCS) paradigm (Huang et al., 2010), which is based on the theory that people behave similarly when interacting through a media (e.g., video conference). Huang et al. (2010) showed that a virtual human driven by PCS approach creates significantly more rapport and is perceived as more believable than the virtual human driven by face-to-face interaction data (from actual listener). This result indicates that the parasocial paradigm is a viable source of information for wisdom of crowds.

In practice, PCS is applied by having participants watch pre-recorded speaker videos drawn from a

Listener1	Listener2	Listener3	Listener4	Listener5	Listener6	Listener7	Listener8	Listener9
pause	POS:NN	pause	pause	pause	POS:NN	Eyebrow up	eye gaze	lowness
label:sub	pause	POS:NN	POS:NN	dirdist:L1	pause	dirdist:L8+	dirdist:R1	eye gaze
POS:NN	label:pmod	label:nmod	low pitch	low pitch	low pitch	POS:NN	POS:JJ	pause

Table 1: Most predictive features for each listener from our wisdom dataset. This analysis suggests three prototypical patterns for backchannel feedback.

dyadic story-telling dataset. In our experiments, we used 43 video-recorded dyadic interactions from the RAPPORT[1] dataset (Gratch et al., 2006). This dataset was drawn from a study of face-to-face narrative discourse ('quasi-monologic' storytelling). The videos of the actual listeners were manually annotated for backchannel feedback. For PCS wisdom acquisition, we recruited 9 participants, who were told to pretend they are an active listener and press the keyboard whenever they felt like providing backchannel feedback. This provides us the responses from multiple listeners all interacting with the same speaker, hence the wisdom necessary to model the variability among listeners.

3 Modeling Wisdom of Crowds

Given the wisdom of multiple listeners, our goal is to create a computational model of backchannel feedback. Although listener responses vary among individuals, we expect some patterns in these responses. Therefore, we first analyze the most predictive features for each listener and search for prototypical patterns (in Section 3.1). Then, we present our Wisdom-LMDE that allows to automatically learn the hidden structure within listener responses.

3.1 Wisdom Analysis

We analyzed our wisdom data to see the most relevant speaker features when predicting responses from each individual listener. (The complete list of speaker features are described in Section 4.1.) We used a feature ranking scheme based on a sparse regularization technique, as described in (Ozkan and Morency, 2010). It allows us to identify the speaker features most predictive of each listener backchannel feedback. The top 3 features for all 9 listeners are listed in Table 1.

This analysis suggests three prototypical patterns. For the first 3 listeners, pause in speech and syntactic information (POS:NN) are more important. The next 3 experts include a prosodic feature, low pitch, which is coherent with earlier findings (Nishimura et al., 2007; Ward and Tsukahara, 2000). It is interesting to see that the last 3 experts incorporate visual information when predicting backchannel feedback. This is in line with Burgoon et al. (Burgoon et al., 1995) work showing that speaker gestures are often correlated with listener feedback. These results clearly suggest that variations be present among listeners and some prototypical patterns may exist. Based on these observations, we propose new computational model for listener backchannel.

3.2 Computational Model: Wisdom-LMDE

The goals of our computational model are to automatically discover the prototypical patterns of backchannel feedback and learn the dynamic between these patterns. This will allow the computational model to accurately predict the responses of a new listener even if he/she changes her backchannel patterns in the middle of the interaction. It will also improve generalization by allowing mixtures of these prototypical patterns.

To achieve these goals, we propose a variant of the Latent Mixture of Discriminative Experts (Ozkan et al., 2010) which takes full advantage of the wisdom of crowds. Our Wisdom-LMDE model is based on a two step process: a Conditional Random Field (CRF, see Figure 1a) is learned for each wisdom listener, and the outputs of these expert models are used as input to a Latent Dynamic Conditional Random Field (LDCRF, see Figure 1b) model, which is capable of learning the hidden structure within the experts. In our Wisdom-LMDE, each expert corresponds to a different listener from the wisdom of crowds. More details about training and inference of LMDE can be found in Ozkan et al. (2010).

[1] http://rapport.ict.usc.edu/

4 Experiments

To confirm the validity of our Wisdom-LMDE model, we compare its performance with computational models previously proposed. As motivated earlier, we focus our experiments on predicting listener backchannel since it is a well-suited task where variability exists among listeners.

4.1 Multimodal Speaker Features

The speaker videos were transcribed and annotated to extract the following features:

Lexical: Some studies have suggested an association between lexical features and listener feedback (Cathcart et al., 2003). Therefore, we use all the words (i.e., unigrams) spoken by the speaker.

Syntactic structure: Using a CRF part-of-speech (POS) tagger and a data-driven left-to-right shift-reduce dependency parser (Sagae and Tsujii, 2007) we extract four types of features from a syntactic dependency structure corresponding to the utterance: POS tags and grammatical function for each word, POS tag of the syntactic head, distance and direction from each word to its syntactic head.

Prosody: Prosody refers to the rhythm, pitch and intonation of speech. Several studies have demonstrated that listener feedback is correlated with a speaker's prosody (Ward and Tsukahara, 2000; Nishimura et al., 2007). Following this, we use downslope in pitch, pitch regions lower than 26th percentile, drop/rise and fast drop/rise in energy of speech, vowel volume, pause.

Visual gestures: Gestures performed by the speaker are often correlated with listener feedback (Burgoon et al., 1995). Eye gaze, in particular, has often been implicated as eliciting listener feedback. Thus, we encode the following contextual features: speaker looking at listener, smiling, moving eyebrows up and frowning.

Although our current method for extracting these features requires that the entire utterance to be available for processing, this provides us with a first step towards integrating information about syntactic structure in multimodal prediction models. Many of these features could in principle be computed incrementally with only a slight degradation in accuracy, with the exception of features that require dependency links where a word's syntactic head is to the right of the word itself. We leave an investigation that examines only syntactic features that can be produced incrementally in real time as future work.

4.2 Baseline Models

Consensus Classifier In our first baseline model, we use consensus labels to train a CRF model, which are constructed by a similar approach presented in (Huang et al., 2010). The consensus threshold is set to 3 (at least 3 listeners agree to give feedback at a point) so that it contains approximately the same number of head nods as the actual listener. See Figure 1 for a graphical representation of CRF model.

CRF Mixture of Experts To show the importance of latent variable in our Wisdom-LMDE model, we trained a CRF-based mixture of discriminative experts. This model is similar to the Logarithmic Opinion Pool (LOP) CRF suggested by Smith et al. (2005). Similar to our Wisdom-LMDE model, the training is performed in two steps. A graphical representation of a CRF Mixture of experts is given in the Figure 1.

Actual Listener (AL) Classifiers This baseline model consists of two models: CRF and LDCRF chains (See Figure 1). To train these models, we use the labels of the "Actual Listeners" (AL) from the RAPPORT dataset.

Multimodal LMDE In this baseline model, we compare our Wisdom LMDE to a multimodal LMDE, where each expert refers to one of 5 different set of multimodal features as presented in (Ozkan et al., 2010): lexical, prosodic, part-of-speech, syntactic, and visual.

Random Classifier Our last baseline model is a random backchannel generator as desribed by Ward and Tsukahara (2000). This model randomly generates backchannels whenever some pre-defined conditions in the prosody of the speech is purveyed.

4.3 Methodolgy

We performed hold-out testing on a randomly selected subset of 10 interactions. The training set contains the remaining 33 interactions. Model parameters were validated by using a 3-fold cross-validation strategy on the training set. Regulariza-

Model	Wisdom of Crowds	Precision	Recall	F1-Score	T-test
Wisdom LMDE	Yes	0.2473	0.7349	**0.3701**	-
Consensus Classifier (Huang et al., 2010)	Yes	0.2217	0.3773	0.2793	p=0.0021
CRF Mixture of Experts (Smith et al., 2005)	Yes	0.2696	0.4407	0.3345	p=0.3605
AL Classifier(CRF)	No	0.2997	0.2819	0.2906	p=0.0707
AL Classifier(LDCRF) (Morency et al., 2007)	No	0.1619	0.2996	0.2102	p=0.0014
Multimodal LMDE (Ozkan et al., 2010)	No	0.2548	0.3752	0.3035	p=0.0251
Random Classifier	No	0.1277	0.2150	0.1570	p=0.0055

Table 2: Comparison of our Wisdom-LMDE model with previously proposed models. The last column shows the paired one tailed t-test results comparing Wisdom LMDE to each model.

tion values used are 10k for k = -1,0,..,3. Numbers of hidden states used in the LDCRF models were 2, 3 and 4. We use the hCRF library[2] for training of CRFs and LDCRFs. Our Wisdom-LMDE model was implemented in Matlab based on the hCRF library. Following (Morency et al., 2008), we use an encoding dictionary to represent our features. The performance is measured by using the F-score, which is the weighted harmonic mean of precision and recall. A backchannel is predicted correctly if a peak happens during an actual listener backchannel with high enough probability. The threshold was selected automatically during validation.

4.4 Results and Discussion

Before reviewing the prediction results, is it important to remember that backchannel feedback is an optional phenomena, where the actual listener may or may not decide on giving feedback (Ward and Tsukahara, 2000). Therefore, results from prediction tasks are expected to have lower accuracies as opposed to recognition tasks where labels are directly observed (e.g., part-of-speech tagging).

Table 2 summarizes our experiments comparing our Wisdom-LMDE model with state-of-the-art approaches for behavior prediction (see Section 4.2). Our Wisdom-LMDE model achieves the best F1 score. Statistical t-test analysis show that Wisdom-LMDE is significantly better than Consensus Classifier, AL Classifier (LDCRF), Multimodel LMDE and Random Classifier.

The second best F1 score is achieved by CRF Mixture of experts, which is the only model among other baseline models that combines different listener labels in a late fusion manner. This result supports our claim that wisdom of clouds improves learning of prediction models. CRF Mixture model is a linear combination of the experts, whereas Wisdom-LMDE enables different weighting of experts at different point in time. By using hidden states, Wisdom-LMDE can automatically learn the prototypical patterns between listeners.

One really interesting result is that the optimal number of hidden states in the Wisdom-LMDE model (after cross-validation) is 3. This is coherent with our qualitative analysis in Section 3.1, where we observed 3 prototypical patterns.

5 Conclusions

In this paper, we proposed a new approach called Wisdom-LMDE for modeling wisdom of crowds, which automatically learns the hidden structure in listener responses. We applied this method on the task of listener backchannel feedback prediction, and showed improvement over previous approaches. Both our qualitative analysis and experimental results suggest that prototypical patterns exist when predicting listener backchannel feedback. The Wisdom-LMDE is a generic model applicable to multiple sequence labeling tasks (such as emotion analysis and dialogue intent recognition), where labels are subjective (i.e. small inter-coder reliability).

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 0917321 and the U.S. Army Research, Development, and Engineering Command (RDE-COM). The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

[2]http://sourceforge.net/projects/hrcf/

References

Judee K. Burgoon, Lesa A. Stern, and Leesa Dillman. 1995. *Interpersonal adaptation: Dyadic interaction patterns*. Cambridge University Press, Cambridge.

M. Burns. 1984. Rapport and relationships: The basis of child care. *Journal of Child Care*, 4:47–57.

N. Cathcart, Jean Carletta, and Ewan Klein. 2003. A shallow model of backchannel continuers in spoken dialogue. In *European Chapter of the Association for Computational Linguistics. 51–58*.

Aimee L. Drolet and Michael W. Morris. 2000. Rapport in conflict resolution: Accounting for how face-to-face contact fosters mutual cooperation in mixed-motive conflicts. *Journal of Experimental Social Psychology*, 36(1):26–50.

D. Fuchs. 1987. Examiner familiarity effects on test performance: Implications for training and practice. *Topics in Early Childhood Special Education*, 7:90–104.

J. Gratch, A. Okhmatovskaia, F. Lamothe, S. Marsella, M. Morales, R.J. Werf, and L.-P. Morency. 2006. Virtual rapport. *Proceedings of International Conference on Intelligent Virtual Agents (IVA), Marina del Rey, CA*.

Jonathan Gratch, Ning Wang, Jillian Gerten, and Edward Fast. 2007. Creating rapport with virtual agents. In *IVA*.

L. Huang, L.-P. Morency, and J. Gratch:. 2010. Parasocial consensus sampling: combining multiple perspectives to learn virtual human behavior. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*.

Suzanne J. LaFleur Linda L. Carli and Christopher C. Loeber. 1995. Nonverbal behavior, gender, and influence. *Journal of Personality and Social Psychology. 68, 1030-1041*.

D. Matsumoto. 2006. *Culture and Nonverbal Behavior*. The Sage Handbook of Nonverbal Communication, Sage Publications Inc.

L.-P. Morency, I. de Kok, and J. Gratch. 2008. Predicting listener backchannels: A probabilistic multimodal approach. In *Proceedings of the Conference on Intelligent Virutal Agents (IVA)*.

Ryota Nishimura, Norihide Kitaoka, and Seiichi Nakagawa. 2007. A spoken dialog system for chat-like conversations considering response timing. *International Conference on Text, Speech and Dialog. 599-606*.

D. Ozkan and L.-P. Morency. 2010. Concensus of self-features for nonverbal behavior analysis. In *Human Behavior Understanding in conjucion with International Conference in Pattern Recognition*.

D. Ozkan, K. Sagae, and L.-P. Morency. 2010. Latent mixture of discriminative experts for multimodal prediction modeling. In *International Conference on Computational Linguistics (COLING)*.

Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, Linda Moy, and David Blei. 2010. Learning from crowds.

Kenji Sagae and Jun'ichi Tsujii. 2007. Dependency parsing and domain adaptation with LR models and parser ensembles. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 1044–1050, Prague, Czech Republic, June. Association for Computational Linguistics.

A. Smith, T. Cohn, and M. Osborne. 2005. Logarithmic opinion pools for conditional random fields. In *ACL*, pages 18–25.

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast - but is it good? Evaluating non-expert annotations for natural language tasks.

James Surowiecki. 2004. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. Doubleday.

P. Tsui and G.L. Schultz. 1985. Failure of rapport: Why psychotheraputic engagement fails in the treatment of asian clients. *American Journal of Orthopsychiatry*, 55:561–569.

N. Ward and W. Tsukahara. 2000. Prosodic features which cue back-channel responses in english and japanese. *Journal of Pragmatics. 23, 1177–1207*.