Learning Influences from Word Use in Polylogue

Tomoharu Iwata, Shinji Watanabe

NTT Communication Science Laboratories
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto, Japan
{iwata.tomoharu,watanabe.shinji}@lab.ntt.co.jp

10.21437/Interspeech.2011-773

Abstract

We propose a probabilistic model for estimating influences among speakers from conversation data with multiple people. In conversations, people tend to mimic their companions' behavior depending on their level of trust. With the proposed model, we assume that the word use of a speaker depends on the word use of previous speakers as well as their own earlier word use and the general word distribution. The influences can be efficiently estimated by using the expectation maximization (EM) algorithm. Experiments on two meeting data sets in Japanese and in English demonstrate the effectiveness of the proposed method.

Index Terms: conversation analysis, influence, latent variable model

1. Introduction

In conversations, people tend to mimic such aspects of their companions' behavior as postures [1], facial expressions [2], lexicon [3, 4], syntax [5], and amplitude [6]. This phenomenon is known in the literature as entrainment, accommodation, adaptation, or alignment [7]. Entrainment is said to indicate that people are trusting, accommodating and empathic [1, 8].

This paper focuses on the entrainment of lexicon in polylogue, or how people are influenced by their companions in terms of word use. The degree to which a person exerts an influence and is influenced by others varies from speaker to speaker. A powerful person is likely to be mimicked by others, and a passive person might often be accommodating to others. The influences also differ between pairs depending on their level of trust. For example, Alice might use words spoken by Bob, but not words spoken by Charlie. The influences therefore have an asymmetric nature.

We propose a simple probabilistic model for estimating influences among speakers from conversation data with multiple people. With the proposed model, we assume that a speaker's word use (word distribution) depends on the preceding word use of other speakers as well as his/her own preceding word use and the general word distribution. We estimate the strength of influence for each pair of speakers using the expectation maximization (EM) algorithm [9].

In recent years, a huge amount of conversation data have been accumulated due to the improvement of recoding devices and automatic speech recognition systems, and there has been great interest in the analysis of conversation. For example, [4] investigated the correlation between task success and similarity of word use, and [8] analyzed the relationship between social game results and word repetition. However, they focused on dyadic conversations, and did not consider the asymmetric nature of influences. On the other hand, we deal with the conversation of multiple people, in which influence and sensitiveness are assumed to depend on the pair of speakers. In addition, since the proposed method is a probabilistic generative model for conversations, we can efficiently estimate inferences in a principled statistical framework, and use it for a language model of the conversation. A number of language models for conversation have been proposed [10, 11]. However, they do not aim to estimate influences between speakers.

2. Proposed Model

Let $\boldsymbol{w} = \{w_1, \cdots, w_t, \cdots\}$ be a word sequence of a polylogue, where w_t represents the tth word, and let $\boldsymbol{s} = \{s_1, \cdots, s_t, \cdots\}$ be its speaker sequence, where s_t indicates the speaker of the tth word. Here, $w_t \in \{1, \cdots, W\}$ and $s_t \in \{1, \cdots, M\}$, where W is the vocabulary size, and M is the number of participants.

In the proposed model, we assume that a speaker's word use depends on the preceding word use of other speakers. The preceding word use of speaker m at position t can be modeled as follows:

$$P_{\mathrm{C}}(w|\boldsymbol{w}_{t-\tau}^{t-1}, m) = \frac{\sum_{t'=t-\tau}^{t-1} \delta(w, w_{t'})\delta(m, s_{t'}) + \beta}{\sum_{w'} \sum_{t'=t-\tau}^{t-1} \delta(w', w_{t'})\delta(m, s_{t'}) + \beta W}, \quad (1)$$

where τ represents the period of the influence, β is a smoothing parameter, and $\delta(x, y)$ is Kronecker's delta, i.e. $\delta(x, y) = 1$ if $x = y$, and 0 otherwise. This probability is proportional to the number of times word w is used by speaker m in the preceding period τ. The smoothing parameter β is introduced to avoid the zero probability problem.

The word use of speaker n at position t is then modeled by a mixture of the preceding word use of the participants as follows:

$$P(w|\boldsymbol{w}_{t-\tau}^{t-1}, n) = \sum_{m=1}^{M} \lambda_{nm} P_{\mathrm{C}}(w|\boldsymbol{w}_{t-\tau}^{t-1}, m) + (1 - \sum_{m} \lambda_{nm}) P_{\mathrm{G}}(w), \quad (2)$$

where λ_{nm} represents the influence of speaker m on speaker n, $0 \leq \lambda_{nm} \leq 1$, and $P_{\mathrm{G}}(w)$ is the general word distribution, which does not depend on the preceding conversation. The general word distribution can be obtained by using other corpora. Speaker m who influences the word distribution of speaker n is not observed, and m is a latent variable.

This proposed model is an extension of cache models [12] for multi-speaker conversations. The cache-based language model integrates short-term patterns of word use into the word distribution by means of a cache component. With the proposed

Copyright © 2011 ISCA

model, we build speaker-specific cache components, and set different influences among pairs of speakers. The proposed model can be easily extended for n-gram language models by taking a n-gram sequence as input instead of unigram sequence $\boldsymbol{w}$.

3. Inference

We estimate the influences λ_{nm} based on maximum posterior (MAP) estimation. For simplicity, we rewrite the proposed word distribution in (2) as follows:

$$P(w|\boldsymbol{w}_{t-\tau}^{t-1}, n) = \sum_{m=0}^{M} \lambda_{nm} P_{\mathrm{C}}(w|\boldsymbol{w}_{t-\tau}^{t-1}, m), \tag{3}$$

where we set $P_{\mathrm{C}}(w|\boldsymbol{w}_{t-\tau}^{t-1}, m=0) \equiv P_{\mathrm{G}}(w)$, $\lambda_{n0} \equiv 1 - \sum_{m=1}^{M} \lambda_{nm}$, in which $\lambda_{nm} \geq 0$ and $\sum_{m=0}^{M} \lambda_{nm} = 1$. With this notation, the logarithm of the posterior probability of parameters given the conversation data $\{\boldsymbol{w}_{t=1}^{T}, \boldsymbol{s}_{t=1}^{T}\}$, which is to be maximized, is calculated as follows,

$$L \propto \sum_{t=1}^{T} \log \sum_{m=0}^{M} \lambda_{s_t m} P_{\mathrm{C}}(w_t|\boldsymbol{w}_{t-\tau}^{t-1}, m) + \sum_{n=1}^{M} \log P(\boldsymbol{\lambda}_n|\alpha), \tag{4}$$

where T is the current position, and the second term represents the prior probability for parameters $\boldsymbol{\lambda}_n = \{\lambda_{nm}\}_{m=0}^{M}$. We use the following Dirichlet prior with hyperparameter α:

$$\log P(\boldsymbol{\lambda}_n|\alpha) = \sum_{m=0}^{M} \alpha \log \lambda_{nm}, \tag{5}$$

because it is conjugate to multinomial parameters $\boldsymbol{\lambda}_n$. The inference is made more robust by introducing the priors.

We can efficiently maximize the posterior (4) by using the EM algorithm [9]. The conditional expectation of the complete-data log likelihood with priors is represented as follows:

$$\begin{aligned} Q \quad = \quad & \sum_{t=1}^{T} \sum_{m=0}^{M} P(m|t) \log \lambda_{s_t m} P_{\mathrm{C}}(w_t|\boldsymbol{w}_{t-\tau}^{t-1}, m) \\ & + \sum_{n=1}^{M} \sum_{m=0}^{M} \alpha \log \lambda_{nm}, \end{aligned} \tag{6}$$

where $P(m|t)$ is the probability that the tth word is influenced by speaker m. In the E-step, we compute the probability according to the Bayes rule:

$$P(m|t) = \frac{\lambda_{s_t m} P_{\mathrm{C}}(w_t|\boldsymbol{w}_{t-\tau}^{t-1}, m)}{\sum_{m'=0}^{M} \lambda_{s_t m'} P_{\mathrm{C}}(w_t|\boldsymbol{w}_{t-\tau}^{t-1}, m')}. \tag{7}$$

In the M-step, we obtain the next estimate of influences $\boldsymbol{\lambda}_n$ by maximizing Q w.r.t. $\boldsymbol{\lambda}_n$ subject to $\sum_{m=0}^{M} \lambda_{nm} = 1$:

$$\lambda_{nm} = \frac{\sum_{t=1}^{T} \delta(n, s_t) P(m|t) + \alpha}{\sum_{m'=0}^{M} \sum_{t=1}^{T} \delta(n, s_t) P(m'|t) + \alpha(M+1)}. \tag{8}$$

Note that the speaker dependent preceding word distribution $P_{\mathrm{C}}(w_t|\boldsymbol{w}_{t-\tau}^{t-1}, m)$ can be calculated by (1) independent of estimating parameters $\boldsymbol{\Lambda} = \{\boldsymbol{\lambda}_n\}_{n=1}^{M}$. The general word distribution $P_{\mathrm{C}}(w_t|\boldsymbol{w}_{t-\tau}^{t-1}, m=0)$ is assumed to be given in advance. By iterating the E-step and the M-step until convergence, we obtain a local optimum solution for influences $\boldsymbol{\Lambda}$.

Table 1: Summary of NTT and RT07 meeting data sets.

	#session	#speakers		#utterance		#vocabulary
		min	max	min	max	
NTT	6	4	4	560	918	2,098
RT07	8	4	6	337	749	3,113

4. Experimental Results

We evaluated the proposed method using the following two real meeting transcription data sets: NTT in Japanese [13] and RT07 in English [14]. Table 1 shows a summary of the NTT and RT07 data sets, and includes the number of sessions, vocabulary size, and the minimum and maximum number of speakers and utterances for a session. With the proposed model, we used $\alpha = 1$ and $\beta = 10^{-8}$ for the hyperparameters, and modeled the preceding word use by using all preceding utterances in the session, or $\tau = \infty$. The general word distribution $P_{\mathrm{G}}(w)$ is learned by using other sessions in each data set.

We estimated the influences between speakers using the proposed model. Figure 1 shows the result. Each node represents a speaker, and the width of the arrow represents the strength of the influence, where only influences with $\lambda_{nm} \geq 0.1$ are shown. The self influence is generally strong, which indicates that the word use depends strongly on the speaker's own preceding word use. This is an intuitive result. There are also many influences between speakers. Some speakers are influential, e.g. speaker 1 in Session 7 in RT07, and some speakers are sensitive to other speakers, e.g. speaker 2 in Session 6 in RT07. Most of the influences are asymmetric. This result indicates that it is important to model the direction of the influences. In all the NTT sessions, speaker 4 was appointed chairperson, and therefore, speaker 4 was influential and not sensitive. The result obtained with the proposed model reveals the influential and non-sensitive characteristics of speaker 4 as shown in Figure 1 (a), where there are more than three arrows from speaker 4 in five out of six sessions, and there is no arrow pointing to speaker 4 from others in all the sessions. In this way, the proposed model can use conversation data to analyze the influences between speakers.

For a quantitative evaluation, we compared the following six models:

- **CC** has a common cache that is shared by all speakers, and a common parameter that control the influence of the common cache. The word distribution is described as follows:

  $$P(w|\boldsymbol{w}_{t-\tau}^{t-1}, n) = \lambda P_{\mathrm{C}}(w|\boldsymbol{w}_{t-\tau}^{t-1}) + (1-\lambda) P_{\mathrm{G}}(w), \tag{9}$$

  where

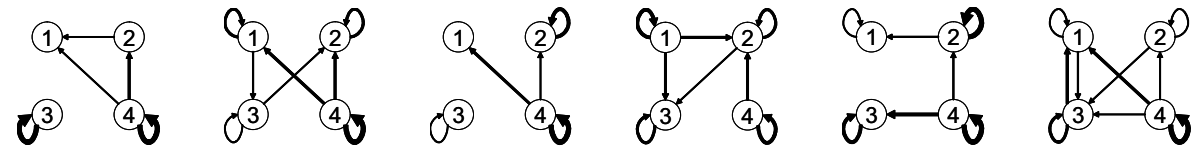
  $$P_{\mathrm{C}}(w|\boldsymbol{w}_{t-\tau}^{t-1}) = \frac{\sum_{t'=t-\tau}^{t-1} \delta(w, w_{t'}) + \beta}{\sum_{w'} \sum_{t'=t-\tau}^{t-1} \delta(w', w_{t'}) + \beta W}, \tag{10}$$

  is the common cache. This is the same as the standard cache language model.

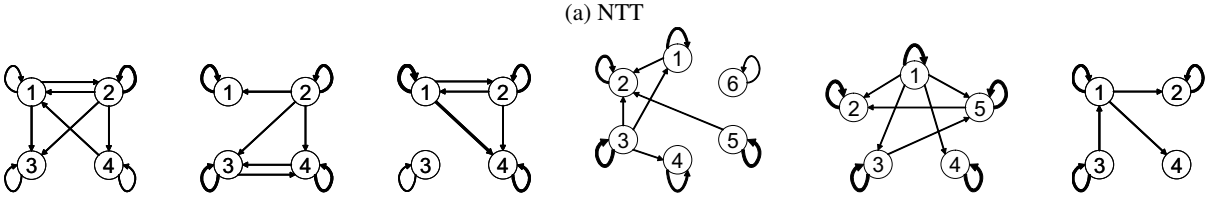
- **OC** has the speaker's own caches, and a common parameter that controls the influence of the speaker's own cache. The word distribution is as follows:

  $$P(w|\boldsymbol{w}_{t-\tau}^{t-1}, n) = \lambda P_{\mathrm{C}}(w|\boldsymbol{w}_{t-\tau}^{t-1}, n) + (1-\lambda) P_{\mathrm{G}}(w). \tag{11}$$

  This model assumes that the word use depends only on the speaker's own preceding word use.



(a) NTT



(b) RT07

Figure 1: Estimated influences by the proposed model in some sessions.

- **IC** has individual caches for each speaker, and a common parameter that controls the influence of the speaker dependent caches. The word distribution is as follows:

$$P(w|\boldsymbol{w}_{t-\tau}^{t-1}, n) = \sum_{m=1}^{M} \lambda_m P_{\mathrm{C}}(w|\boldsymbol{w}_{t-\tau}^{t-1}, m) + (1 - \sum_{m=1}^{M} \lambda_m) P_{\mathrm{G}}(w), \quad (12)$$

  where λ_m represents the influence of speaker m on all speakers including speaker m himself/herself. This model assumes that the strength of the influence depends on the speaker, but the sensitivity does not differ among speakers.

- **CI** has a common cache, and individual parameters that control the influence of the common cache for each speaker as follows:

$$P(w|\boldsymbol{w}_{t-\tau}^{t-1}, n) = \lambda_n P_{\mathrm{C}}(w|\boldsymbol{w}_{t-\tau}^{t-1}) + (1 - \lambda_n) P_{\mathrm{G}}(w). \quad (13)$$

  This model assumes that the word use depends on all speakers' word use and the degree of dependence differs among speakers.

- **OI** has the speaker's own caches, and individual parameters that control the influence depending on the speakers as follows:

$$P(w|\boldsymbol{w}_{t-\tau}^{t-1}, n) = \lambda_n P_{\mathrm{C}}(w|\boldsymbol{w}_{t-\tau}^{t-1}, u) + (1 - \lambda_n) P_{\mathrm{G}}(w). \quad (14)$$

- **II** has individual caches for each speaker, and individual influence parameters. This is our proposed model in (2).

The first letter of a method's name C/O/I represents common/own/individual caches, respectively, and the second letter of the method's name C/I represents common/individual parameters, respectively. Only the proposed model (II) takes the asymmetricity of influences into account. With all the models, we used $\alpha = 1$, $\beta = 10^{-8}$ and $\tau = \infty$.

In each session, we used data until the jth utterance as training data to learn the parameters, and used words after the $(j+1)$th utterance as test data. We evaluated the performance of each model using the perplexity of held-out words:

$$\exp\left(-\frac{\sum_{i=j+1}^{J}\sum_{k=1}^{K_i}\log P(w_{ik}|\boldsymbol{w}_1^j, s_i)}{\sum_{i=j+1}^{J} K_i}\right), \quad (15)$$

Table 2: Average perplexities for each session.

(a) NTT

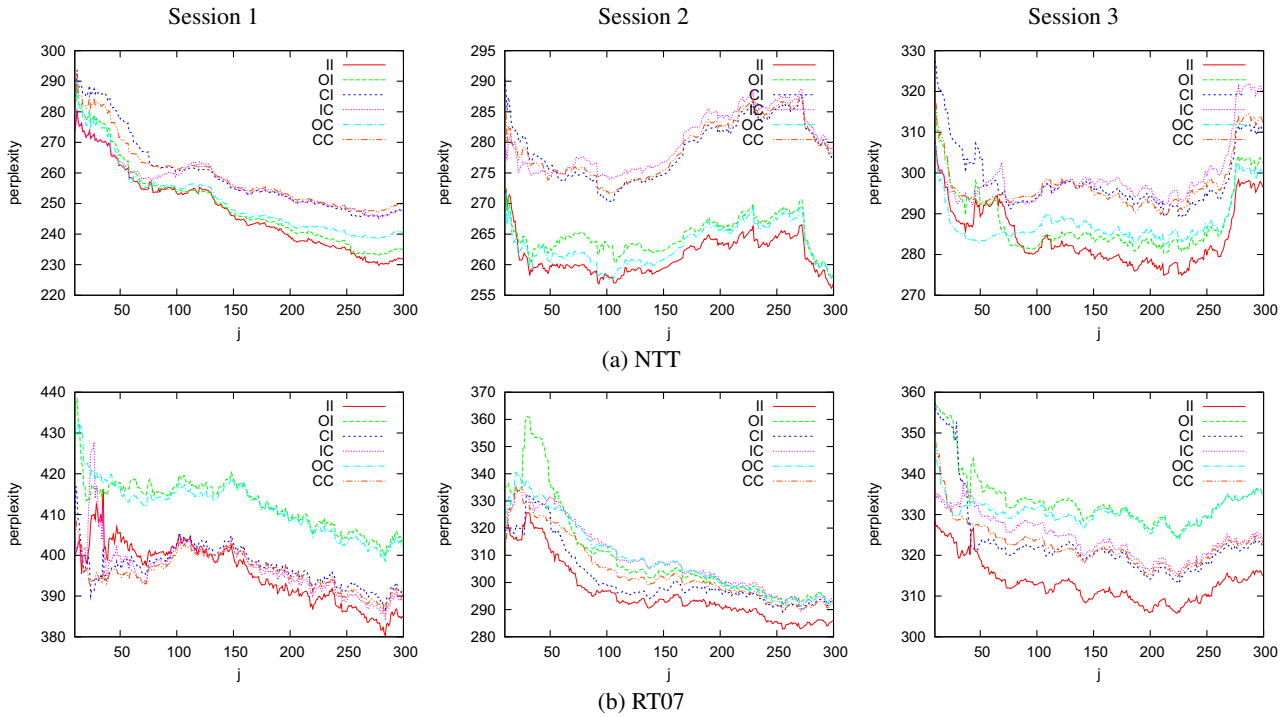
session#	CC	OC	IC	CI	OI	II
1	259.2	251.4	256.5	259.4	249.9	**247.7**
2	279.1	263.2	280.0	278.9	264.7	**261.4**
3	297.1	287.7	298.9	298.3	288.1	**284.4**
4	321.3	307.9	314.9	320.8	309.7	**294.4**
5	332.8	322.6	328.1	332.5	320.3	**313.8**
6	274.4	260.9	277.7	275.9	267.5	**254.9**
average	294.0	282.3	292.7	294.3	283.4	**276.1**

(b) RT07

session#	CC	OC	IC	CI	OI	II
1	395.7	411.8	396.7	397.1	412.7	**395.6**
2	304.3	308.5	308.6	301.6	309.4	**296.1**
3	322.6	330.6	324.4	322.9	333.6	**313.2**
4	373.0	386.1	369.0	377.2	390.7	**368.7**
5	300.5	301.9	299.9	303.2	304.4	**293.4**
6	342.5	343.7	368.2	350.4	352.3	**340.4**
7	340.6	350.7	345.4	355.8	357.4	**332.3**
8	340.7	346.7	345.7	344.9	357.9	**340.3**
average	340.0	347.5	344.8	344.1	352.3	**335.0**

where J is the number of utterances in the session, K_i is the number of words in the ith utterance, w_{ik} is the kth word in the ith utterance, $\boldsymbol{w}_1^j$ is a set of words until the jth utterances, and s_i is the speaker of the ith utterance. A lower perplexity represents higher predictive performance.

Table 2 shows the average perplexities for the NTT and RT07 data sets, in which the number of training utterances ranges from $j = 10$ to $j = 300$. The proposed model achieved the lowest perplexities in all sessions. This result indicates that it is important to estimate the asymmetric influences between speakers, which only the proposed model considers. Figure 2 shows the perplexities with different numbers of training utterances for some sessions. Generally speaking, the perplexity decreased as the number of training utterances increased because the estimation accuracy of the influences and preceding word use improves. In some sessions, for example Session 3 in the NTT data set, the perplexity increased because of the change of topics. Except when the number of training utterances was small, the perplexity of the proposed model (II) steadily achieved the lowest perplexities.



(a) NTT

(b) RT07

Figure 2: Perplexities with different numbers of training utterances for some sessions. The horizontal axis represents the number of training utterances.

The average computational time for learning parameters in the proposed model with 300 training utterances was 0.01 and 0.02 seconds for the NTT and RT07 data sets, respectively. The proposed model is very efficient, and it can be used in real time applications [13].

5. Conclusion

We have proposed a probabilistic model for learning influences from conversation data with multiple speakers. We have confirmed experimentally that the proposed model can extract influences between speakers and learn conversation's word use.

Although our results have been encouraging to date, our model can be further improved in a number of ways. First, we would like to estimate influences using other behaviors, such as nonverbal speech acts, posture and eye movement, as well as word use. Second, the proposed model can be extended by modeling the dynamics of influences although in this paper we assume that the influences do not change over time. Third, we must determine the period of the influence automatically. Finally, we would like to evaluate the proposed model in an automatic speech recognition system.

6. References

[1] T. L. Chartrand and J. A. Bargh, "The chameleon effect: the perception-behavior link and social interaction," *Journal of Personality and Social Psychology*, vol. 76, no. 6, pp. 893–910, 1999.

[2] U. Dimberg, "Facial reactions to facial expressions," *Psychophysiology*, vol. 19, no. 6, pp. 643–647, 1982.

[3] S. Brennan, "Lexical entrainment in spontaneous dialog," in *ISSD '96*, 1996, pp. 41–44.

[4] A. Nenkova, A. Gravano, and J. Hirschberg, "High frequency word entrainment in spoken dialogue," in *ACL '08: HLT*, 2008, pp. 169–172.

[5] D. Reitter, F. Keller, and J. D. Moore, "Computational modelling of structural priming in dialogue," in *NAACL '06*, 2006, pp. 121–124.

[6] R. Coulston, S. Oviatt, and C. Darves, "Amplitude convergence in children's conversational speech with animated personas," in *ICSLP '02*, vol. 4, 2002, pp. 2689–2692.

[7] H. Giles, J. Coupland, and N. Coupland, *Accommodation theory: Communication, context, and consequence*. Cambridge University Press, 1991.

[8] L. E. Scissors, A. J. Gill, and D. Gergle, "Linguistic mimicry and trust in text-based cmc," in *CSCW '08*, 2008, pp. 277–280.

[9] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[10] G. Ji and J. Bilmes, "Multi-speaker language modeling," in *HLT-NAACL '04*, 2004, pp. 133–136.

[11] M. Purver, T. L. Griffiths, K. P. Körding, and J. B. Tenenbaum, "Unsupervised topic modelling for multi-party spoken discourse," in *ACL '06*, 2006, pp. 17–24.

[12] R. Kuhn and R. D. Mori, "A cache-based natural language model for speech recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 6, pp. 570–583, 1990.

[13] T. Hori, S. Araki, T. Yoshioka, M. Fujimoto, S. Watanabe, T. Oba, A. Ogawa, K. Otsuka, D. Mikami, K. Kinoshita, T. Nakatani, A. Nakamura, and J. Yamato, "Real-time meeting recognition and understanding using distant microphones and omni-directional camera," in *SLT '10*, 2010, pp. 412–417.

[14] J. G. Fiscus, J. Ajot, and J. S. Garofolo, "The rich transcription 2007 meeting recognition evaluation," in *Multimodal Technologies for Perception of Humans*, 2008, pp. 373–389.