TRANSFER CROSS ENTROPY FOR FAST SOCIOMETRIC INFERENCE IN LONGITUDINAL COLLECTIONS OF MULTI-PARTY CONVERSATION

Kornel Laskowski

Carnegie Mellon University
Pittsburgh PA, USA

ABSTRACT

A means is proposed of extracting participant-pair interaction measures from a binary representation of behavior in multi-party conversation, by leveraging an extension of transfer entropy. The technique allows for the inexpensive construction of sociomatrices, requiring only a minimum of detection technology. It is expected that the method will tractably enable the application of social network analysis to conversational behavior mined from very large collections of unannotated audio.

***Index Terms*—** Turn-taking, N-gram models, social network analysis, cross entropy, transfer entropy, influence.

1. INTRODUCTION

Social network analysis (SNA) has become an important tool in today's society [1]. It has been used to make inferences about occupational mobility, national prestige, adoption of innovation, and other many-actor concepts. SNA reasoning involves the *manipulation* of matrices of quantified pair-wise relations, known as sociomatrices.

Sociomatrix *computation*, however, is generally considered to lie outside of SNA; a result of this is that SNA is most often applied to datasets in which inter-actor relations are overt or easy to identify (e.g., how frequently two people email each other). For data in which this is not the case, influence may be hypothesized when actors are observed to behave like other actors did before them [2]. However, in multi-actor scenarios, in which it is hard to automatically identify the intended recipient of any particular action, in which mimicry is not a necessary consequence of influence, and in which the space of possible actions is large, measuring influence is more challenging. Multi-party conversation is a good example of such a scenario type.

Words may seem to be a natural choice of conversational action, but their usage is dependent on language, domain, and interaction style. Choosing them as actions for analysis would partition conversational corpora into small but homogenous subgroups, and undermine the potential of applying SNA to conversations in a general sense. Some researchers have therefore turned to records of binary speech activity, or chronograms, as records of action. A chronogram elides the words and retains only the timing information of *when* individual participants were speaking. As the output of speech activity detection or diarization systems, chronograms are readily computable and clearly independent of the availability and performance of higher-level linguistic processing components. First-order influence in two-party conversations has been explored by studying how one participant's immediately preceding speech activity state helps to predict another's current speech activity state [3]. For four-party conversation, [4] proposed to model each individual's influence on an intervening "group state", and the group's influence on each individual, but did not consider direct individual-to-individual relations.

The current work proposes *transfer cross entropy* as a measure of pair-wise influence in conversations with arbitrary numbers of participants, and arbitrarily long Markovian truncations of history. The primary goal is to describe *how* to compute the measure from any conversation, allowing for the subsequent application of SNA methods across conversations of vastly different characteristics. The work's secondary goal is to report on a peculiar finding: when applied to chronograms of conversations by members of a research institute, using a stochastic model of turn-taking, the proposed technique appears to implicitly rank conversants by self-reported education level (a measure of organizational seniority [5]).

2. DATA

The ICSI Meeting Corpus [6] is used to demonstrate the proposed techniques. The corpus is unique in that it contains longitudinal recordings of the same groups of participants, meeting to talk about work; the meetings, it has been claimed, would have occurred even if they were not recorded. This work focuses on one group (`Bmr`) which met to discus the corpus collection project itself. 29 meetings of this group are available; they involve a total of 15 participants, between 3 and 9 per meeting.

3. EXAMPLE OF AN ANNOTATED RELATION

As an illustration of sociomatrix construction, in a setting in which a large amount of manually annotated data is available, consider the concept of the *adjacency pair (AP)*. An AP is a conversation analysis construct in which one speaker's turn forms a "part A" and a subsequent speaker's turn forms a "part B". The two parts may be a question and an answer, or a greeting and a "return" greeting, etc.

A sociomatrix $\mathbf{X}$ is an $N \times N$ matrix, where N is the number of participants in the population under study. $\mathbf{X} \equiv (x_{ij})$, and x_{ij} is a directed relation from participant i to participant j. x_{ij} need not equal x_{ji}, leading to an asymmetric $\mathbf{X}$; x_{ii} is left undefined.

In multi-party conversation, an $\mathbf{X}$ defined by adjacency describes grounding, and may additionally describe preferred information exchange partners or merely compliance with social norms. Because the ICSI Meeting Corpus contains annotated APs, it is easy to assign to x_{ij} the number of APs in which participant i provides the "part A" of the AP while participant j provides the "part B", normalized by the total time of all `Bmr` meetings in which both participants took part. A *sociogram*, or directed graph, of $\mathbf{X}$ is shown in Figure 1. To improve visualization, only those x_{ij} are shown as arcs whose values exceed $\mu + \sigma$, where μ and σ are the global mean and standard deviation over all x_{ij} that are not undefined[1].

[1] x_{ij} is considered undefined if there are no conversations in which i and j both participated.

978-1-4673-0046-9/12/$26.00 ©2012 IEEE

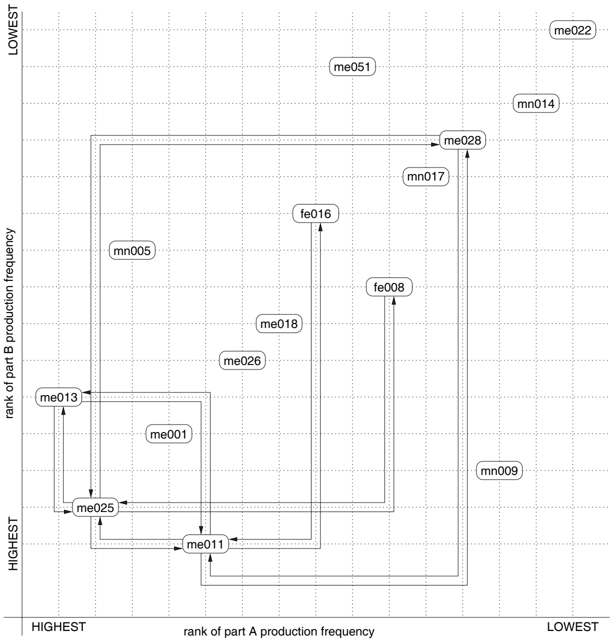

Fig. 1. A directed sociogram showing the dichotomous relation of conversational adjacency. Nodes placed by rank of frequency of "part A" production (along horizontal direction) and of "part B" production (along vertical direction).

What Figure 1 shows is that, once the x_{ij} are thresholded using $\mu + \sigma$, the remaining arcs occur in pairs (the dichotomous relation thus produced is — somewhat surprisingly — symmetric after all). This illustrates that participants exchange "part A" and "part B" roles within dyads: they are more likely to direct their "part As" to those participants who directed to them their own "part As". Adjacency is not uniform across participants; particularly strong ties identify cliques of participants who tend to talk with one another more than with others. For example, the two female group members, `fe008` and `fe016`, appear to prefer adjacency pairing with members `me025` and `me011`, respectively.

Although automatic inference of APs has been attempted [7], it relies not only on automatic speech/non-speech segmentation, but also on automatic speech recognition, automatic dialog act segmentation and classification, and potentially on prosodic analysis. This makes the process difficult to apply efficiently for very large collections of conversations. Also, it is currently limited to domains and languages for which these technologies are mature.

4. TRANSFER CROSS ENTROPY IN CHRONOGRAMS

4.1. Chronogram Construction

A speech/non-speech chronogram $\mathbf{Q}$ of any conversation [8], as commonly understood, is the speech/non-speech segmentation of all K participants to that conversation, discretized at a constant frame step and frame size, and time-aligned. The frame step used here is 100 ms, representing the very shortest verbal productions in the ICSI Meeting Corpus. The resulting $\mathbf{Q}$ is a Markov random field, a matrix $\in \{\square, \blacksquare\}^{K \times T}$. $\square$ and $\blacksquare$ are the absence and presence of speech activity, and T is the number of frames. The tth column of $\mathbf{Q}$, $\mathbf{q}_t$, is the vector concatenation of the states of all parties.

4.2. Smoothing Coupled n-Gram Models

$\mathbf{Q}$ is modeled as the vector-valued output of a Markov process. A model $\mathbf{\Theta}$ provides the likelihood of $\mathbf{Q}$,

$$P_{\mathbf{\Theta}}\left(\mathbf{Q}\right) = \prod_{t=1}^{T} P_{\mathbf{\Theta}}\left(\mathbf{q}_t \mid \ldots, \mathbf{q}_{t-1}\right) , \quad (1)$$

where the ellipsis represents $\mathbf{q}_{t-2}$ and earlier emissions, reflecting the order of the Markov process. The conditioning context is subsequently denoted as $\mathbf{q}_{t-1}^{t-\tau}$, where τ is the number of frames considered. $\mathbf{\Theta}$ is commonly known as a stochastic turn-taking model [9, 10]. As elsewhere, it is assumed that the behavior of all participants is *conditionally independent* (CI), given their joint behavior in $\mathbf{q}_{t-1}^{t-\tau}$; each term on the right-hand side of Equation 1 is thus equal to

$$P_{\mathbf{\Theta}}\left(\mathbf{q}_t \mid \mathbf{q}_{t-1}^{t-\tau}\right) = \prod_{k=1}^{K} P_{\mathbf{\Theta}}\left(\mathbf{q}_t\left[k\right] \mid \mathbf{q}_{t-1}^{t-\tau}\right) . \quad (2)$$

Square brackets index participants.

Although in principle separate models can be inferred for each participant, conditioned on the behavior of specific interlocutors, for computational simplicity the current work makes use of a single model, for all participants, given a collapsed single-scalar description of the conditioning context per history frame. For the joint activity $\mathbf{q}_{t-\tau}$, that scalar is denoted $\|\mathbf{C}_k \cdot \mathbf{q}_{t-\tau}\|$, where $\mathbf{C}_k$ is the $K \times K$ identity matrix with the kth column removed, and $\| \cdot \|$ is the number of entries in the resulting dot product which are $\blacksquare$. In this way, $\|\mathbf{C}_k \cdot \mathbf{q}_{t-\tau}\|$ denotes the number of interlocutors of the kth participant which are in the $\blacksquare$ state at instant $t - \tau$, for some finite τ. A ceiling is applied on this quantity, equal to $K_{max} - 1$, where K_{max} is a parameter to be estimated.

The factors on the right-hand side of Equation 2 then become

$$P_{\mathbf{\Theta}}\left(\mathbf{q}_t\left[k\right] \mid \mathbf{q}_{t-1}^{t-\tau}\right) \doteq P_{\mathbf{\Theta}}\left(\mathbf{q}_t\left[k\right] \mid \left\{\mathbf{q}\left[k\right], \|\mathbf{C}_k\mathbf{q}\|\right\}_{t-1}^{t-\tau}\right)$$

leading to an n-gram model with $n = 2\tau + 1$.

As elsewhere in n-gram modeling, smoothing is employed during model estimation. In this particular setting, since the symbol space consists only of $\square$ and $\blacksquare$, the majority of smoothing techniques used for language modeling (where large Zipf's-law-compliant vocabularies are assumed) are not quite appropriate. Instead, the general and recursive Jelinek-Mercer interpolation [11] is used, with the history-dependent smoothing parameter

$$\lambda\left(h_{t-1}^{t-\tau}\right) = \frac{C\left(h_{t-1}^{t-\tau}\right)}{C\left(h_{t-1}^{t-\tau}\right) + \rho} , \quad (3)$$

where $h_{t-1}^{t-\tau}$ is a particular history and $C\left(h_{t-1}^{t-\tau}\right)$ its count, dependent on a global relevance parameter ρ. Extensive testing reveals that values of approximately 200 for this parameter (for the model orders in this work) lead to highest likelihoods for unseen chronograms. Under the said smoothing arrangement, the effect of extending the conditioning context is shown in Figure 2.

What can be seen from this figure is that longer conditioning histories are beneficial. Furthermore, it is obvious that taking interlocutor behavior into account (by allowing $K_{max} > 1$) is also beneficial: the same cross-entropies are observed using a $K_{max} = 2$ model, at $\tau = 2$, as are observed at $\tau = 5$ with the model for which $K_{max} = 1$ (which ignores interlocutors). Finally, it can be seen that $K_{max} = 3$ is only negligibly better than $K_{max} = 2$. This suggests that, on average, it is not important to know *how many* interlocutors are speaking prior to t, only whether zero or at-least-one are.

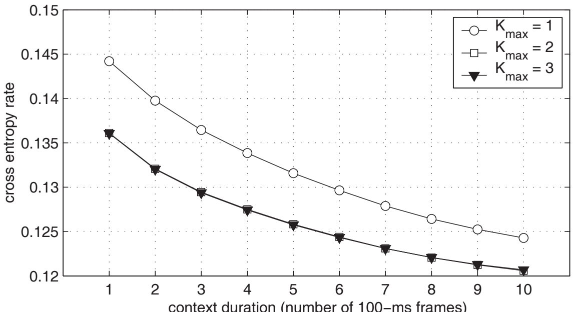

Fig. 2. Agglomerated leave-one-meeting-out cross-entropy rate (in bits per frame, along y-axis) as a function of the duration of the conditioning history (in 100-ms frames, along x-axis).

4.3. What n-Gram Models Learn

Given the behavior of models in Figure 2, it is of interest in this work to understand *why* cross entropy reductions accompany a sensitivity to interlocutor behavior. Some light is shed on this matter in Figure 3, demonstrating the impact of conditioning using interlocutors' states at $t-1$ on the learned model probabilities.

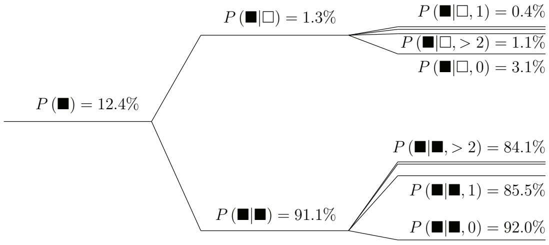

Fig. 3. Unigram $P(\mathbf{q}_t[k])$, uncoupled bigram $P(\mathbf{q}_t[k]\,|\mathbf{q}_{t-1}[k])$, and coupled bigram $P(\mathbf{q}_t[k]\,|\mathbf{q}_{t-1}[k], \|\mathbf{C}_k\mathbf{q}_{t-1}\|)$ probabilities of speaking (■) in the ICSI corpus, in ascending order from top to bottom.

The figure depicts an key aspect of conversational speech deployment. When someone is not already speaking, their probability of beginning to do so is largest when the number of speaking interlocutors is zero (3.1%), and smallest when that number is one (0.4%). When more than one interlocutor is speaking, the probability of any non-speaking participant starts speaking (1.1%) is larger than when only one interlocutor is speaking (0.4%). When a participant is already speaking, on the other hand, their probability of continuing to do so monotonically decreases as the number of speaking interlocutors increases (92.0%, 85.5%, 84.1%). Evidently, one-at-a-time speaking is the most stable joint configuration.

4.4. Measuring Influence

The turn-taking model descrinbed above can now be used to measure the relationship between any pair of participants.

It should be noted that the likelihood of any row k of a particular $\mathbf{Q}$, corresponding to the behavior of participant k, is given by Equation 2, assuming that $\mathbf{\Theta}$ encodes the norms of speech deployment [10] in the corpus. Some rows can be expected to be more likely than others, depending on how closely the participants in question conform to those norms. To quantify the influence that an interlocutor j has on k, the jth row is eliminated from all instants of the conditioning history, when estimating the likelihood of the kth row. If j in fact conditions k's behavior, then this modified cross entropy will be higher than that in Equation 2, in which j's row is left untouched. That is, it will become harder to predict k's behavior, and more bits will be necessary to encode the kth row.

More consisely, the proposed measure of influence of j on k, $\mathrm{Infl}_{j\to k}$, is hereby given as

$$\begin{aligned} \mathrm{Infl}_{j\to k} \quad \equiv \quad & -\sum_{t=1}^{T} \log P_{\mathbf{\Theta}}\left(\mathbf{q}_t[k] \,|\, \{\mathbf{q}[k], \|\mathbf{C}_{kj}\mathbf{q}\|\}_{t-1}^{t-\tau}\right) \qquad (4) \\ & + \sum_{t=1}^{T} \log P_{\mathbf{\Theta}}\left(\mathbf{q}_t[k] \,|\, \{\mathbf{q}[k], \|\mathbf{C}_{k}\mathbf{q}\|\}_{t-1}^{t-\tau}\right) \ , \end{aligned}$$

where $\mathbf{C}_{kj}$ is identical to $\mathbf{C}_k$, except *both* the kth and the jth column are eliminated, rather than just the kth.

Inspection of Equation 4 reveals that it is a cross entropy rate, which can be rewritten as

$$\begin{aligned} \mathrm{Infl}_{j\to k} \quad \equiv \quad & -\sum P_{ML}\left(\mathbf{q}_t[k], \{\mathbf{q}[k], \|\mathbf{C}_{k}\mathbf{q}\|\}_{t-1}^{t-\tau}\right) \qquad (5) \\ & \times \log \frac{P_{\mathbf{\Theta}}\left(\mathbf{q}_t[k] \,|\, \{\mathbf{q}[k], \|\mathbf{C}_{kj}\mathbf{q}\|\}_{t-1}^{t-\tau}\right)}{P_{\mathbf{\Theta}}\left(\mathbf{q}_t[k] \,|\, \{\mathbf{q}[k], \|\mathbf{C}_{k}\mathbf{q}\|\}_{t-1}^{t-\tau}\right)} \ , \end{aligned}$$

where the sum is over all joint events inside $P_{ML}(\cdot)$, their maximum likelihood (ML) estimate in the chronogram $\mathbf{Q}$ under study.

The cross entropy rate in Equation 5 is a *conditional cross entropy rate*; the elimination of some conditioning information — in the form of the jth row from $t-1$ to $t-\tau$ — makes it a *conditional transfer cross entropy* (an extension of transfer entropy [12]). Values above zero indicate the cost, in number of bits per frame, incurred by assuming that j is irrelevant to the prediction of the kth row. Note that the conditional transfer cross entropy can be negative, since participants may deviate from the global norm $\mathbf{\Theta}$.

4.5. Sociomatrix Computation

Next, the required sociomatrix $\mathbf{X}$ is computed. It should be noted that, in any chronogram, a participant j may appear to influence participant k under the proposed model $\mathbf{\Theta}$ in three main ways. First, if j merely speaks a lot, then j will eliminate putative opportunities for k to speak (cf. Figure 3). Second, j may frequently pause within utterance. This could render two interlocutors, j and j', have unequal influence even if their total time spent in ■ is identical, and would suggest that participants orient themselves not so much to the production of speech by others at any specific instant, but perhaps to production in the vicinity of that instant. Third, and of most interest here, participants may wish to follow specific interlocutors, once those interlocutors are finished, or almost finished, speaking.

To eliminate the first effect, x_{ij}, the entries of sociomatrix $\mathbf{X}$, are assigned the values

$$x_{ij} \quad \equiv \quad \frac{\mathrm{Infl}_{j\to k}}{P_{ML}\left(\mathbf{q}_t[j] = ■\right)} \ . \qquad (6)$$

The equation normalizes the influence $\mathrm{Infl}_{j\to k}$ by the amount of time that j is speaking in $\mathbf{Q}$. The second aspect is not addressed in the current work, treating intra-utterance pausing as noise. The proposed form of Equation 6 is therefore assumed to grossly account for the

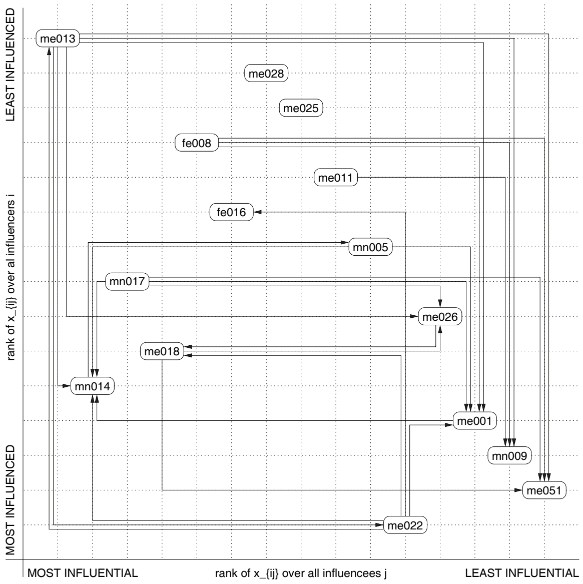

Fig. 4. A directed sociogram showing the dichotomous relation of "Infl$_{j\rightarrow k}$ normalized by the speaking time of j" across speech/non-speech chronograms. Nodes placed by rank of average influence (along horizontal direction) and of average influenceabilty (along vertical direction).

third aspect described above, namely preference for following specific interlocutors, and for doing so carefully in order to conform to the turn-taking norms described by model $\mathbf{\Theta}$.

Figure 4 suggests that closely dovetailing speech production with the turn terminations of specific interlocutors is not a symmetric relation in general. Only the pairs (me013,me022), (me018,me026) and (mn014,mn005) have arcs which survive thresholding in both directions. The main observation is that participants in the top left corner of the figure are helpful in predicting the behavior of participants in the bottom right. It turns out that only the five right-most participants did not have a doctorate at the time of recording (as evidenced in the corpus meta-data); meanwhile, me013, at the top left, was the only professor in the Bmr meetings. Some observations related to this can be found in [5].

5. DISCUSSION

The technique proposed in this work has not been quantitatively evaluated on any particular task. A most natural candidate would be validation against human-produced self-reports about pair-wise relationships with other conversants. To the author's knowledge, no such annotated corpus has yet been collected and made available.

It should be noted that the presented modeling techniques can operate on *any* binary chronogram, of any type or subset of vocal activity. It is straightforward to extend the models to more choices than $\{\square, \blacksquare\}$. It is also possible to construct predictors of one chronogram type using other chronogram types for interlocutors. In short, the current work should be treated as merely a preliminary effort. Easy-to-construct relations include the conditioning of laughter on laughter, which may indicate "liking" [13].

6. CONCLUSIONS

A technique has been presented for inferring dichotomous sociometric relations from collections of simple representations of individual multi-party conversations. These representations require only automatic vocal activity recognition, rather than conventional linguistic processing involving speech recognition, parsing, and discourse structure analysis. The inference technique is therefore readily deployable, today. The current work has argued for the technique's theoretical soundness, has presented some preliminary results, and has visually explored an example of a dichotomous relation. The applications of the technique to social psychology and to social network media appear promising, and the technique is sufficiently mature for empirical validation using manually labeled relations.

7. REFERENCES

[1] S. Wasserman and K. Faust, *Social Network Analysis*, Cambridge University Press, Cambridge, UK, 1994.

[2] A. Goyal, F. Bonchi, and L. V. S. Lakshmanan, "Learning influence probabilities in social networks," in *Proc. WSDM*, New York NY, USA, 2010, pp. 241–250.

[3] T. Choudhury, B. Clarkson, S. Basu, and A. Pentland, "Learning communities: Connectivity and dynamics of interacting agents," in *Proc. IJCNN*, Portland OR, USA, 2003, pp. 2797–2802.

[4] D. Zhang, D. Gatica-Perez, S. Bengio, and D. Roy, "Learning influence among interacting Markov chains," in *Proc. NIPS*, Vancouver BC, Canada, 2005.

[5] K. Laskowski, M. Ostendorf, and T. Schultz, "Modeling vocal interaction for text-independent participant characterization in multi-party meetings," in *Proc. SIGDial*, Columbus OH, USA, 2008, pp. 148–155.

[6] A. Janin et al., "The ICSI Meeting Corpus," in *Proc. ICASSP*, Hong Kong, China, 2003, pp. 364–367.

[7] M. Galley, K. McKeown, J. Hirschberg, and E. Shriberg, "Identifying agreement and disagreement in conversational speech: Use of Bayesian networds to model pragmatic dependenciess," in *Proc. ACL*, Barcelona, Spain, 2004.

[8] E. Chapple, "The Interaction Chronograph: Its evolution and present application," *Personnel*, vol. 25, no. 4, pp. 295–307, 1949.

[9] T. Wilson, J. Wiemann, and D. Zimmerman, "Models of turn-taking in conversational interaction," *J Language and Social Psychology*, vol. 3, no. 3, pp. 159–183, 1984.

[10] K. Laskowski, "Modeling norms of turn-taking in multi-party conversation," in *Proc. ACL*, Uppsala, Sweden, 2010, pp. 999–1008.

[11] F. Jelinek and R. Mercer, "Interpolated estimation of Markov source parameters from sparse data," in *Proc. Workshop on Pattern Recognition in Practice*, Amsterdam, The Netherlands, 1980.

[12] T. Schreiber, "Measuring information transfer," *Physica D: Nonlinear Phenomena*, vol. 142, no. 3-4, pp. 346–382, 2000.

[13] P. Glenn, *Laughter in Interaction*, Cambridge University Press, Cambridge, UK, 2003.