Social Network Analysis in Multimedia Indexing: Making Sense of People in Multiparty Recordings

Sarah Favre
Idiap Research Institute
CP 592, 1920 Martigny, Switzerland
Ecole Polytechnique Federale de Lausanne
1015 Lausanne, Switzerland
sfavre@idiap.ch

Abstract

This paper presents an automatic approach to analyze the human interactions appearing in multiparty data, aiming at understanding the data content and at extracting social information such as *Which role do people play?*, *What is their attitude?*, or *Can people be split into meaningful groups?*. To extract such information, we use a set of mathematical techniques, namely *Social Networks Analysis* (SNA), developed by sociologists to analyze social interactions. This paper shows that a strong connection can be established between the content of broadcast data and the social interactions of the individuals involved in the recordings. Experiments aiming at assigning each individual to a social group corresponding to a specific topic in broadcast news, and experiments aiming at recognizing the role played by each individual in multiparty data are presented in this paper. The results achieved are satisfactory, which suggests on one side that the application of SNA to similar problems could lead to useful contributions in the domain of multimedia content analysis, and on the other side, that the presented analysis of social interactions could be a significant breakthrough for affective computing.

Keywords: Social Network Analysis, Role Recognition, Story Segmentation, Broadcast data, Meeting Recordings.

1 Introduction

The amount of audio and video material available in digital form is increasing rapidly with the progress of capture and storage technologies, but without effective techniques for structuring its content, it is not possible to make an asset of it. Many research efforts have addressed the problem of extracting information from data for indexing purposes. For example, existing systems extract the information from: automatic speech transcriptions, shot transitions in video, faces and objects in images. However, when I started working on my PhD thesis three years ago, no major efforts were made to automatically analyze social interactions, even if they represent a common subject of multimedia recordings. Human interactions are present everywhere: in movies, television shows, broadcast news, radio programs, meeting recordings, call center conversations, etc. Moreover, psychologists showed that social interactions are one of the main channels through which we understand reality (Kunda, 1999). For these reasons, we decided to investigate automatic approaches for the audio content analysis, based on the extraction of social interactions in multiparty recordings.

We started investigating the audio content analysis on broadcast news, this type of data accounts for large collections of real unconstrained conversations. Our idea was to establish a link between the content of the news, i.e. its structure intended as a sequence of topics, and the social interactions between the individuals who presented the different topics. The rationale of our approach was that individuals involved in the same topic interact more with each other than individuals involved in different topics. To this end, we aimed at identifying *social groups*, i.e. individuals characterized by a high degree of mutual interactions. Preliminary experiments showed that the groups corresponding to the most important (i.e. dominant) topics could be detected effectively (Vinciarelli, 2007). These results suggest that social interactions lead the structure of broadcast data and can facilitate the content analysis of such recordings.

Moreover, we know that individuals involved in news recordings play a specific role or function, and that those roles are governing the structure of the conversations (e.g. an anchorman conducting an interview). We thus extend our original idea that broadcast news are subdivided into social groups, assuming that the interactions between the individuals involved in the social groups are governing the structure of the recordings. We investigated a social-interactions-based approach for the automatic recognition of the roles played by broadcast-news participants. Numerical experiments revealed that around 80 percent of the data-time was correctly labeled in terms of role (Favre, 2008, 2009). This seems to suggest that there is a strong connection between role interactions and content of news data.

Another interesting challenge was to apply our approach to less structured data. In fact, broadcast news follow a predefined structure by assigning specific roles or functions to every person (such as anchorman, guest, interviewer). We hypothesize the existence of a connection between social interactions and content of small-group meetings, where individuals have a position in a given social system and do not follow stable behavioral patterns. Therefore, we investigated an approach for the automatic recognition of the roles played by the participants in the AMI meeting corpus (McCowan, 2005). The results showed that the best recognized role was the *Project Manager*, which acts as a *chairman* and thus follows predictable behavioral patterns. This suggests that the features extracted from social interactions in such data (containing spontaneous interactions) are not sufficient to analyze the structure, and that lexical content is necessary to obtain an effective content analysis (Garg, 2008).

The experiments proposed in this paper aimed at analyzing the audio content, to improve tasks performed in the multimedia content analysis community. However, the results of my work reveal that the approach we propose for the analysis of social interactions could be relevant for the affective community as well. In fact, as soon as people interact, they adopt a specific behavior depending on the social context, depending on how they percieve their interlocutor, and according to their emotions. The analysis of social interactions could thus facilitate the recognition and interpretation of human emotions.

In this paper, we describe the approach we proposed for extracting the social interactions through a set of mathematical techniques used by sociologists, namely *Social Network Analysis* (SNA) (Wasserman, 1994). We also describe how we applied machine learning techniques to the social patterns extracted from the data, aiming at classifying individuals into social groups or into roles.

The rest of this paper is organized as follows: Section 2 describes the approach we have applied to extract social groups and roles, Section 3 outlines the experimental setup and the automatic story segmentation results as well as the automatic role recognition results, and finally Section 4 draws some conclusions summarizing the main contributions of this work.

2 Our Approach

The approach we propose includes three main steps: the first is *Speaker Diarization*, which splits the audio into speaker turns, i.e. segments corresponding to single speaker intervals (see Section 2.1). The aim of this first step is to detect the persons involved in the recordings and the sequence of their interventions. The second step is *Feature Extraction*, which applies Social Affiliation Networks (see Section 2.2) (Wasserman, 1994) to represent each person in terms of their relationships with the others. The third step is the actual *Classification* step (see Section 2.3), where each person is assigned to a social group or to a role.

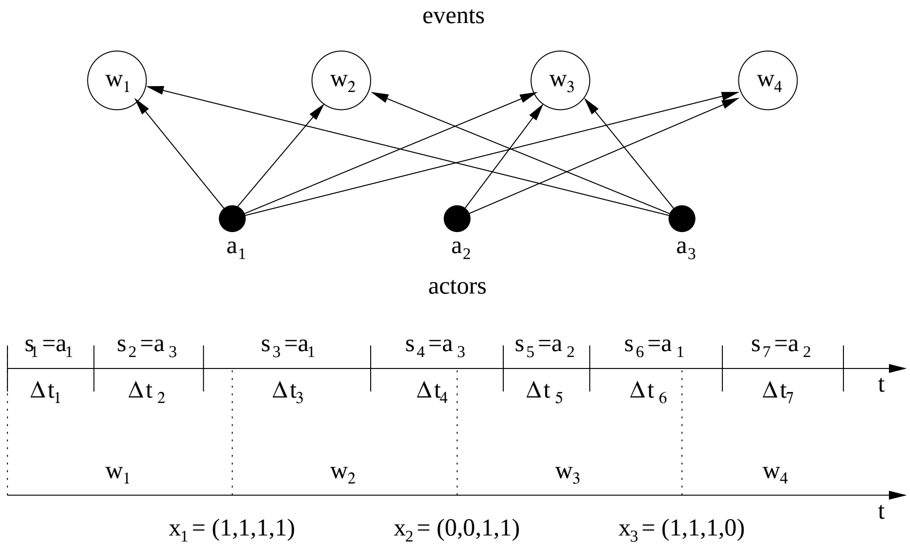

Figure 1: Social Affiliation Network extraction. The events correspond to the segments w_j and the actors are linked to the events when they talk during the corresponding segments. The actors are represented using vectors $\vec{x}_i$ where the components account for the links between actors and events.

2.1 Speaker Diarization

The aim of a speaker diarization system is to segment an audio recording into time intervals during which there is only one speaker talking. The experiments presented in this paper were performed over broadcast data and meeting recordings which requires two different approaches to cope with the characteristics of each type of data. The diarization technique for broadcast data is fully described in Ajmera (2003). A speech/non-speech detection system is used for meeting recordings and is fully described in Dines (2006). The techniques are not described here for space reasons as it is not the main element of interest of this work.

For both kinds of data, the speaker diarization process converts each recording into a sequence of turns $S = \{(s_k, \Delta t_k)\}$, where $k = 1, \ldots, N$, s_k is the speaker label corresponding to the voice detected in the k^{th} turn, and Δt_k is the duration of the k^{th} turn. The label s_k belongs to the set $A = \{a_1, \ldots, a_G\}$ of G unique speaker labels as provided by the speaker diarization process.

2.2 Feature Extraction

The turn sequence S generated by the speaker diarization system is used to build a Social Affiliation Network (SAN), a graph with two types of nodes (*actors* and *events*) where links are not allowed between nodes of the same type (see Figure 1) (Wasserman, 1994). In our experiments, the actors correspond to the speakers as detected in the diarization process, and the events correspond to D uniform non-overlapping segments spanning the whole length of the recording (see lower part of Figure 1). The rationale is that actors participating in the same events (i.e. participants speaking during the same time intervals) are likely to interact with each other. Therefore, the SAN extracts the evidence of interactions in terms of: *who talks to whom and when.*

One of the main advantages of this representation is that each actor a_i can be represented with a D-dimensional vector $\vec{x}_i$, where the component x_j accounts for the presence or the absence of each actor a_i in the different events j. The component x_j is set to 1 if the actor a_i talks during the j^{th} segment and 0 otherwise (the corresponding vectors are shown at the bottom of Figure 1). The persons that interact more with each other tend to talk during the same segments and are represented by similar vectors.

The number D of events was set through crossvalidation during the experiments.

2.3 Classification

We applied our approach to two different tasks: the first is *the story segmentation*, i.e. the segmentation of radio and television news into the different topics presented one after the other, corresponding to social groups (more details can be found in Vinciarelli (2007)). The second is *the role recognition*, i.e. the automatic recognition of the roles played by the individuals participating in broadcast news or multiparty meetings (see (Favre, 2008, 2009; Garg, 2008) for more details). The next two sections outline the machine learning techniques used to perform these two tasks.

2.3.1 The Story Segmentation Approach

This section presents an approach to make the content of a long recording more accessible: an automatic segmentation in terms of topics, i.e. *stories.*

The main idea behind the approach presented here is that people involved in the same story interact more with each other than people involved in different stories. This means that the stories can be identified by grouping the people having a high degree of mutual interaction or, in sociological terms, by detecting *social groups.*

The goal of the story segmentation is to assign the sequence of speakers talking during a conversation (represented by vectors $\vec{x}_i$ as explained in Section 2.2), a sequence of labels h_i corresponding to the stories presented one after the other during the recordings.

This corresponds to finding the sequence $H^* = (h_1, \ldots, h_M)$ which maximizes the *a-posteriori* probability:

$$H^* = \arg \max_{H \in \mathcal{H}} p(X|H)p(H) \tag{1}$$

where $\mathcal{H}$ is the set of all possible H sequences. The term $p(X|H)$ can be estimated by using a fully connected Hidden Markov Models (HMMs) (Rabiner, 1989) with S +1 states, where S is the maximum number of stories that can be observed. In fact, S states account for stories and one state accounts for the anchorman role. The emission probability function for each state is a mixture of Gaussians. The term $p(H)$ can be estimated using a tri-gram statistical language model (SLMs) (Rosenfeld, 2000).

2.3.2 The Role Recognition Approach

This section presents an approach for content analysis using the roles of the person involved in broadcast data and group meetings.

The idea of the approach is that the interactions between the roles played by the persons involved in the recordings, are governing the structure of the data.

We have considered two different approaches for the role classification: the first assigns a specific role to each speaker voice involved in the recordings using *Bayesian classifiers.* The second approach considers the sequence of speakers talking during a conversation, taking into account the dynamics of the conversation, and aligns the sequence of speakers with a sequence of roles applying *probabilistic sequential models.*

The Role Recognition Approach based on Bayesian Classifiers Section 2.2 has shown that the interaction patterns of every speaker i can be represented by a vector $\vec{x}_i$. Furthermore, every speaker i talks during a fraction τ_i of the total time of a recording. We can thus represent every speaker by a vector $\vec{y}_i = (\vec{x}_i, \tau_i)$. Consider the vector $\vec{r} = (r_1, \ldots, r_G)$, where r_i is the role of speaker i, and the vector of observation $Y = \{\vec{y}_1, \ldots, \vec{y}_G\}$, where $\vec{y}_i$ is the vector representing speaker i. The problem of assigning the role to all speakers can be thought of as the maximization of the *a-posteriori* probability $\mathrm{p}(\vec{r}\,|\,Y)$. By applying Bayes Theorem and by taking into account that $p(Y)$ is constant during recognition this problem is equivalent to finding $\vec{\hat{r}}$ such that:

$$\vec{\hat{r}} = \arg \max_{\vec{r} \in \mathcal{R}^G} \mathrm{p}(Y\,|\,\vec{r})\,\mathrm{p}(\vec{r}), \tag{2}$$

where $\mathcal{R}$ is the set of the predefined roles.

In order to simplify the problem, we make the assumption that the observations are mutually conditionally independent given the roles. In the case we are considering, it seems also reasonable to assume that the observation $\vec{y}_i$ of speaker i only depends on their role r_i and not on the roles of the other speakers. To further simplify the problem, we assume that the interaction vectors $\vec{x}_i$ and the speaking time τ_i are statistically independent given the role, and thus Equation (2) can be rewritten as:

$$\vec{\hat{r}} = \arg \max_{\vec{r} \in \mathcal{R}^G} \mathrm{p}(\vec{r}) \prod_{k=1}^{G} \mathrm{p}(\vec{x}_k \,|\, r_k)\, \mathrm{p}(\tau_k \,|\, r_k). \tag{3}$$

We estimated $\mathrm{p}(\vec{x} \,|\, r)$ using mixtures of Bernoulli distributions (Bishop, 2006). Probability $\mathrm{p}(\tau \,|\, r)$ was estimated using Gaussian distributions. The *a-priori* probability of the roles $\mathrm{p}(r)$ was estimated making the assumption that the roles are independent. We modeled $\mathrm{p}(r)$ as the fraction of speakers in the training set labeled with the role r, and thus we do not take into account the constraints that the role distribution across different participants in a given recording must respect, e.g. there is only one *Anchorman* in a talk-show, or there is only one *Project Manager* in a meeting, etc. (see Favre (2008) for more details).

Role Recognition Approach based on Probabilistic Sequential Models In this approach (see Favre (2009) for more details), we consider the sequence of speakers, taking into account the dynamics of the conversation. We have seen in Section 2.2 that the interaction patterns of every speaker i can be represented by a vector $\vec{x}_i$. Furthermore, every speaker i talks during a fraction τ_i of the total time of a recording. We can thus represent every speaker by a vector $\vec{y}_i = (\vec{x}_i, \tau_i)$. Therefore, each recording can be represented by a sequence $Y = (\vec{y}_1, \ldots, \vec{y}_N)$, where N is the number of turns detected at the speaker diarization step.

The role recognition can be thought of as finding the role sequence R^* satisfying the following equation:

$$R^* = \arg \max_{R \in \mathcal{R}^N} p(Y|R)p(R), \tag{4}$$

where $R = (r_1, \ldots, r_N)$ is a sequence of roles of length N, $r \in \mathcal{R}$ ($\mathcal{R}$ is a predefined set of roles), and $\mathcal{R}^N$ is the set of all role sequences of length N.

In our experiments, the likelihood $p(Y|R)$ is estimated with a fully connected, ergodic, Hidden Markov Models (HMMs) (Rabiner, 1989) where each state corresponds to a role $r \in \mathcal{R}$. The *a-priori* probability $p(R)$ is estimated using a 3-gram statistical language model (Rosenfeld, 2000):

$$p(R) = \prod_{k=3}^{N} p(r_k | r_{k-1}, r_{k-2}). \tag{5}$$

3 Data and Results

This section outlines the experimental setup and the results.

3.1 Story Segmentation Results

The story segmentation experiments were performed over two corpora: the first, referred to as C2 in the following, contains 27 one hour long talk-shows broadcasted by *Radio Suisse Romande* (the French speaking Swiss National broadcasting service) during February 2005. Each bulletin is managed by two anchormen that start and stop the stories by giving the floor to different participants. The average number of participants is 25. The second corpus is the largest existing database of news video, namely TRECVID (TRECVID, 2003), which consists of 229 news video of 30 minutes provided by ABC and CNN.

We used a leave-one-out approach to train HMMs models and achieved a performance in terms of purity (Ajmera, 2003) of 0.80 and 0.64 over broadcast data (C2) and the television news (TRECVID) respectively, showing that the groups corresponding to the most important stories are detected effectively (by important stories we mean the most dominant stories). These results show

Table 1: Role recognition performance for C1 and C2. The table reports both the overall accuracy and the accuracy for each role.

	overall (σ)	AM	SA	GT	IP	HR	WM
Results over C1							
C1 Bayes	82.5 (6.9)	98.0	3.6	91.8	8.0	64.6	79.9
C1 HMMs + 3-gram	79.7 (9.3)	97.8	10.3	81.4	25.7	59.5	78.0
Results over C2							
C2 Bayes	82.6 (6.9)	75.0	88.3	91.6	N/A	18.3	6.7
C2 HMMs + 3-gram	86.1 (6.8)	74.4	91.9	92.0	N/A	72.8	30.5

Table 2: Role recognition performance for C3. The table reports both the overall accuracy and the accuracy for each role.

	overall (σ)	PM	ID	ME	UI
C3 Bayes	43.5 (23.9)	75.3	15.1	15.1	40.0
C3 HMMs + 3-gram	46.9 (24.9)	61.8	25.0	39.4	33.8

that there is a link between the social interactions and the content of news data. The proposed approach enables to extract social groups that can be used to analyze the content of the data.

3.2 Role Recognition Results

The experiments on the role recognition task are performed over three different corpora. The first, referred to as C1 in the following, contains 96 news bulletins broadcasted by *Radio Suisse Romande* (the French speaking Swiss National broadcasting service) during February 2005. The average length of C1 recordings is 11 minutes and 50 seconds, and the average number of participants is 12. The second corpus is C2 (see Section 3.1). The third corpus, referred to as C3 in the following, is the AMI meeting corpus (McCowan, 2005), a collection of 138 meeting recordings involving 4 persons each and with an average length of 19 minutes and 50 seconds.

The roles of C1 are *Anchorman* (AM), *Second Anchorman* (SA), *Guest* (GT), *Interview Participant* (IP), *Headline Person* (HP), and *Weather Man* (WM). Roles with the same name are played in C2 (with the exception of IP that appears only in C1), but they correspond to different functions (e.g., AM are not expected to deliver the news but to entertain in talk-shows). The roles of C3 are *Project Manager* (PM), *Marketing Expert* (ME), *User Interface Expert* (UI), and *Industrial Designer* (ID).

Table 1 reports the results achieved on C1 and C2, and in Table 2, those obtained on C3. The performance is measured in terms of *accuracy*, intended as the percentage of time correctly labeled in terms of roles in the test set. We used a leave-one-out approach to train our models and to select the number D of segments used to split the recordings (see Section 2.2). For each corpus, the first line reports the results achieved with the approach based on a Bayesian classifier (Bayes) (see Section 2.3.2), and the last row reports the results when using probabilistic sequential models (HMMs + 3-gram) (see Section 2.3.2). Each accuracy value is accompanied by the standard deviation of the accuracies achieved over the different recordings of each corpus.

The results show that the overall role recognition accuracy is above 80 percent for both C1 and C2. This highlights the strong connection between the interactions of the speakers playing roles and the content of the recordings in broadcast data. However, in meeting recordings, the roles are recognized with a lower accuracy. The explanation is that the roles in meetings are *informal*, i.e. they correspond to a position in a given social system and do not correspond to stable behavioral patterns like in the case of the *formal* roles in broadcast data. In meetings, the only role recognized

C1	HMM C	HMM W
Bayes C	77.1	2.7
Bayes W	5.4	14.8
C2	HMM C	HMM W
Bayes C	81.1	5.0
Bayes W	1.5	12.4
C3	HMM C	HMM W
Bayes C	27.1	14.0
Bayes W	11.1	47.7

Table 3: Diversity assessment. The table reports the percentage of data-time where the Bayes and HMM based role recognition approach are both correct (C), both wrong (W), or one wrong and the other correct.

approach	overall	PM	ID	ME	UI
SNA	43.1	75.7	13.4	16.4	41.2
lexical	67.1	78.3	53.0	71.9	38.1
SNA+lexical	67.9	84.0	50.1	69.8	38.1

Table 4: Role recognition results for C3. The table reports both the overall accuracy and the accuracy for each role.

with a high accuracy is the *Project Manager* (PM). The reason is because the PM also acts as a *chairman*, playing thus a more formal role than the domain experts ID, ME, and UI.

The comparison between the approach based on HMMs and the one based on Bayesian classifiers results in a significant degree of *diversity* (see Table 3). The probabilistic sequential approach results in an improved recognition of less frequent roles, which are typically penalized by Bayesian classifiers because of their low *a-priori* probability. The combination of the two approaches could thus lead into significant performance improvements, and will be the subject of future work.

The limitation of our approach is represented by the low values of accuracy obtained on the meeting recordings. This suggests that the social interaction based role recognition approach is not well suited for unconstrained data characterized by spontaneous interactions (such as multiparty meetings). Table 4 shows the role recognition accuracy for the C3 corpus combining a SNA and lexical based role recognizer. The first line reports the accuracies obtained by using only SNA, the second line those obtained using only the lexical approach, and the last line those obtained using the combination of the two. The lexical approach appears to be a more reliable cue for the recognition of the roles in such meeting recordings (more details can be found in (Garg, 2008)).

4 Conclusion and Contributions

This paper has presented automatic approaches aiming at analyzing the human interactions appearing in multiparty data in order to indexing multimedia content. The idea developed in this paper is that a strong connection can be established between the content of multiparty material and social interactions.

We demonstrated that this assumption is relevant in the case of data characterised by structured human interactions, such as broadcast news. We investigated a story segmentation task where the social interactions allow to index the content into social groups corresponding to the topics presented in the news. Moreover, we performed a role recognition task where the social interactions allow to assign a role to each individual, and thus to structure the content of the recordings. In the case of more spontaneous interactions (like meetings), the experiments show that our proposed approach is not sufficient and that lexical content analysis is necessary.

This paper suggests that the analysis of social interactions in combination with other behavioral cues extracted from audio (e.g. prosodic and lexical features) and video (e.g. gestures, visual focus of attention) could lead to useful contributions in the domains of multimedia content analysis and affective computing. In fact, the analysis of social groups could be further developed into discriminating between *positive* vs *negative* stories, or *sad* vs *happy* stories.

Moreover, following Kelly (2001), we can hypothesize that the social groups are characterized by a composition of the moods, emotions, and sentiments brought by group members. The analysis of social groups, combined with lexical and prosodic features, could help in understanding group affect, e.g. *Can we define group mood?*, or *Do groups share emotions as do individuals?*.

Acknowledgements

This work is supported by the Swiss National Science Foundation (under the National Centre of Competence in Research on Interactive Multimodal Information Management), and by the European Community's Seventh Framework Programme (FP7/2007-2013), under grant agreement no. 231287 (SSPNet). The author wishes to thank Alessandro Vinciarelli and Alfred Dielmann.

References

Ajmera, J. and Wooters, C. (2003). A robust speaker clustering algorithm. In *Proceedings of IEEE Workshop on Automatic Speech Recognition Understanding*, pages 411–416.

Bishop, C.M. (2006). *Pattern Recognition and Machine Learning.* Springer Verlag.

Dines, J. and Vepa, J. and Hain, T. (2006). The segmentation of multi-channel meeting recordings for automatic speech recognition. In *Proceedings of Interspeech*, pages 1213–1216.

Favre, S. and Salamin, H. and Dines, J. and Vinciarelli, A. (2008). Role Recognition in Multiparty Recordings using Social Affiliation Networks and Discrete Distributions. In *Proceedings of the ACM International Conference on Multimodal Interfaces*, pages 29–36.

Favre, S. and Dielmann, A. and Vinciarelli, A. (2009). Automatic Role Recognition in Multiparty Recordings Using Social Networks and Probabilistic Sequential Models. To Appear in *Proceedings of the 2009 ACM International Conference on Multimodal Interfaces.*

Garg, N. and Favre, S and Salamin H. and Hakkani-Tur, D. and Vinciarelli, A. (2008). Role Recognition for Meeting Participants: an Approach Based on Lexical Information and Social Network Analysis. In *Proceedings of the ACM International Conference on Multimedia*, pages 693–696.

Kelly, Janice R and Barsade, Sigal G. (2001). Mood and Emotions in Small Groups and Work Teams. In *Organizational Behavior and Human Decision Processes*, vol.86, no.1, pages 99–130.

Kunda, Z. (1999). Social Cognition. MIT Press.

McCowan, I. and al. (2005). The AMI meeting corpus. In *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, page 4.

Rabiner, L.R. (1989). A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In *Proceedings of the IEEE*, vol.77, pages 257–286.

Rosenfeld, R. (2000). Two decades of statistical language modeling: Where do we go from here? In *Proceedings of the IEEE*, vol.88, no.8, pages 1270–1278.

Vinciarelli, A. and Favre, S. (2007). Broadcast News Story Segmentation Using Social Network Analysis and Hidden Markov Models. In *Proceedings of ACM International Conference on Multimedia*, pages 261–264.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis.* Cambridge University Press.

TREC Video retrieval Evaluation (2003). http://www-nlpir.nist.gov/projects/tv2003/tv2003.html.