Broadcast News Story Segmentation Using Social Network Analysis and Hidden Markov Models

Alessandro Vinciarelli
IDIAP Research Institute
CP592 - 1920 Martigny, Switzerland
Ecole Polytechnique Federale de Lausanne
1015 Lausanne, Switzerland
vincia@idiap.ch

Sarah Favre
IDIAP Research Institute
CP592 - 1920 Martigny, Switzerland
Ecole Polytechnique Federale de Lausanne
1015 Lausanne, Switzerland
sfavre@idiap.ch

ABSTRACT

This paper presents an approach for the segmentation of broadcast news into stories. The main novelty of this work is that the segmentation process does not take into account the content of the news, i.e. what is said, but rather the structure of the social relationships between the persons that in the news are involved. The main rationale behind such an approach is that people interacting with each other are likely to talk about the same topics, thus social relationships are likely to be correlated to stories. The approach is based on Social Network Analysis (for the representation of social relationships) and Hidden Markov Models (for the mapping of social relationships into stories). The experiments are performed over 26 hours of radio news and the results show that a fully automatic process achieves a purity higher than 0.75.

Categories and Subject Descriptors: H.3.1 [Content Analysis and Indexing]:

General Terms: Experimentation.

Keywords: Social Network Analysis, Hidden Markov Models, Story Segmentation, Broadcast News.

1. INTRODUCTION

One of the main ways to make the content of a long recording more accessible is to perform a semantic segmentation, i.e. to split the recording into segments which are meaningful from a user point of view [4]. In the case of broadcast news, the segmentation is typically performed in terms of *stories*, i.e. of the single and specific issues that are presented one after each other along a news bulletin. The stories play in broadcast news the same role that the articles play in newspapers. The stories can be thought of as the main building block of broadcast news: any news bulletin can be split into stories and, vice-versa, a sequence of stories can form a news bulletin.

This paper presents a new approach for segmenting news into stories. The main novelty is that the segmentation process does not take into account the content, i.e. what is said, but rather the pattern of the social relationships between the persons participating in the news. The main rationale behind such an approach is that people involved in the same story interact with each other more than people involved in different stories. This means that the stories can be identified by grouping the people that have a high degree of mutual interaction or, in sociological terms, by detecting *social groups*.

The approach proposed in this work is depicted in Figure 1. The process can be split into three major stages: the first performs an unsupervised speaker clustering and splits the audio into segments corresponding to a single voice [1]. The goal of this stage is to detect the persons involved in a news bulletin and the sequence of their interventions. The second stage is the representation of social relationships by means of an *Affiliation Network*, i.e. one of the most common techniques applied in Social Network Analysis [6] to identify individuals with high mutual interaction. The third is the application of Hidden Markov Models (HMM) [3] and Statistical Language Models (SLM) [3] to map social relationships into stories.

The experiments are performed over a corpus of 26 hours of news bulletins provided by Radio Suisse Romande, the French speaking Swiss national broadcasting service. The results show that the average purity (see Section 3) is around 0.75 and that the performance is satisfactory for applications such as *browsing* (the user can quickly find the story of interest out of one hour long recordings), *role recognition* (the distinction between anchormen and other participants), and *semiautomatic multimedia editing* (the user can quickly obtain a perfect story segmentation by simply correcting the automatic one).

To our knowledge, no previous approaches to the story segmentation problem have tried to use social relationships. Most common approaches are based on speech transcriptions or close captions [2], as well as video clues like shot transitions or graphical elements of the images [5]. The approach proposed in this work explores thus an additional source of information that has not been used so far and that could be used in combination with the others.

The rest of this paper is organized as follows: Section 2 presents the segmentation approach, Section 3 presents experiments and results, and Section 4 draws some conclusions.

MM'07, September 23–29, 2007, Augsburg, Germany.

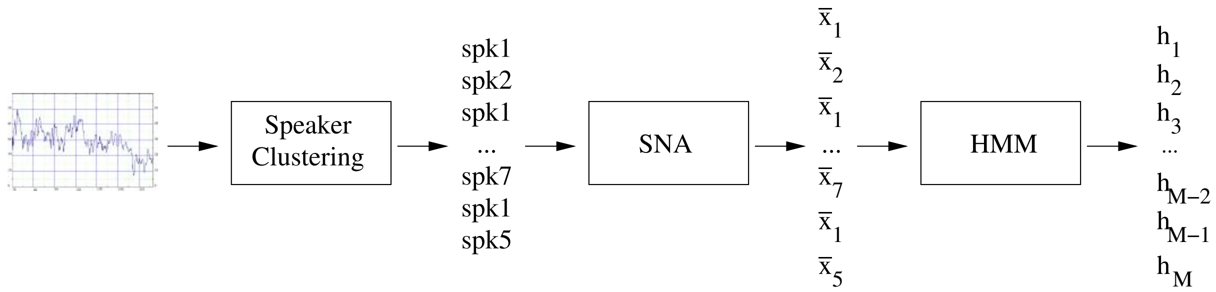

Figure 1: Segmentation approach. This figure shows the three stages of the segmentation approach: the first splits the audio into single speaker segments, the second converts the speakers into vectors using SNA, the third maps the vectors into stories using HMMs.

2. STORY SEGMENTATION APPROACH

This section presents in detail the story segmentation approach depicted in Figure 1. The next sections present each stage of the process.

2.1 Speaker Clustering

For space reasons, this section does not describe the speaker clustering approach (a full presentation is available in [1]), but rather its results on the data.

The speaker clustering works directly over the raw audio data and splits the recordings into single speaker segments. Since the process is unsupervised, speakers are assigned labels that do not correspond to their identity, but simply distinguish one voice, thus one speaker, with respect to the others. In principle, the same voice is given always the same label, but the process is affected by errors and it happens that the same speaker is given different labels in different moments (this is due in particular to background noises), or that different voices are given the same label (this happens in particular when several persons talk through the phone and their voices become similar). The effects of the errors over the segmentation results are shown in Section 3.

The result of the clustering is a sequence of pairs $(s_j, \Delta t_j)$:

$$L = \{(s_1, \Delta t_1), \dots, (s_M, \Delta t_M)\} \tag{1}$$

where s_j is the label assigned to the speaker voice detected in the j^{th} segment of the audio, and Δt_j is the duration of the same segment. The same label can be assigned to several segments meaning that the same speaker talks several times (see lower part of Figure 2).

2.2 Social Relationships Representation

Social Network Analysis (SNA) is a corpus of mathematical techniques that sociologists use to study the relationships between individuals sharing a common environment [6]. This work uses the so-called *Affiliation Networks*, i.e. graphs where there are two kinds of nodes (*actors* and *events*) and only nodes of different kind can be connected (see upper part of Figure 2). The actors correspond to the persons in the social environment under investigation. The events can be defined as actual gatherings (e.g. conferences, meetings, concerts, etc.) or through the proximity in time and space (e.g. living in the same part of the town, taking the bus at the same time, etc.). The rationale behind Affilitation Networks (AN) is that people participating in the same events are more likely to interact with each other, thus ANs provide an effective representation of relationship patterns.

In the case of this work, the events are defined using the proximity in time (see lower part of Figure 2): the news bulletins are split into N uniform non-overlapping windows w_j and each one of these is an event. An actor a_i is said to participate in event e_j when he/she talks during window w_j. In this way, each actor is represented with a vector $\vec{y}_i = (y_{i1}, \dots, y_{iN})$ where the component y_{ik} accounts for the presence (or absence) of actor a_i in event e_k. In the case of this work, $y_{ik} = 1$ if actor a_i talks during window w_k, and $y_{ik} = 0$ otherwise. Since the number of events can be rather high (up to 20 in this work) the dimensionality of the vectors $\vec{y}$ is reduced using the Principal Component Analysis (PCA). The resulting vectors $\vec{x}_i$ are used as input for the next step of the process.

2.3 Story Segmentation

After the representation stage, each recording is converted into a sequence $X = (\vec{x}_1, \dots, \vec{x}_M)$ of D-dimensional vectors, where M is the number of single speaker segments detected at the speaker clustering step, and D is the number of Principal Components retained after the application of the PCA to the $\vec{y}_i$ vectors (see previous section). The goal of the story segmentation is to assign each vector $\vec{x}_i$ a label h_i which can be either the number of a story (e.g. *story 2*, or *story 7*) or the *anchorman* role, i.e. the activity of the journalists that manage the bulletins and participate in most of the stories.

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

$$H^* = \arg\max_{H \in \mathcal{H}} p(H|X)p(H) \tag{2}$$

where $\mathcal{H}$ is the set of all possible H sequences. The term $p(H|X)$ can be estimated using a fully connected Hidden Markov Model (HMM) 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:

$$p(H) = \prod_{k=3}^{M} p(h_k|h_{k-1}, h_{k-2}). \tag{3}$$

Once the sequence H^* is found, each vector $\vec{x}_i$ is assigned to a story or to the anchorman role. Since each vector corresponds to a segment of the recording, finding H^* results

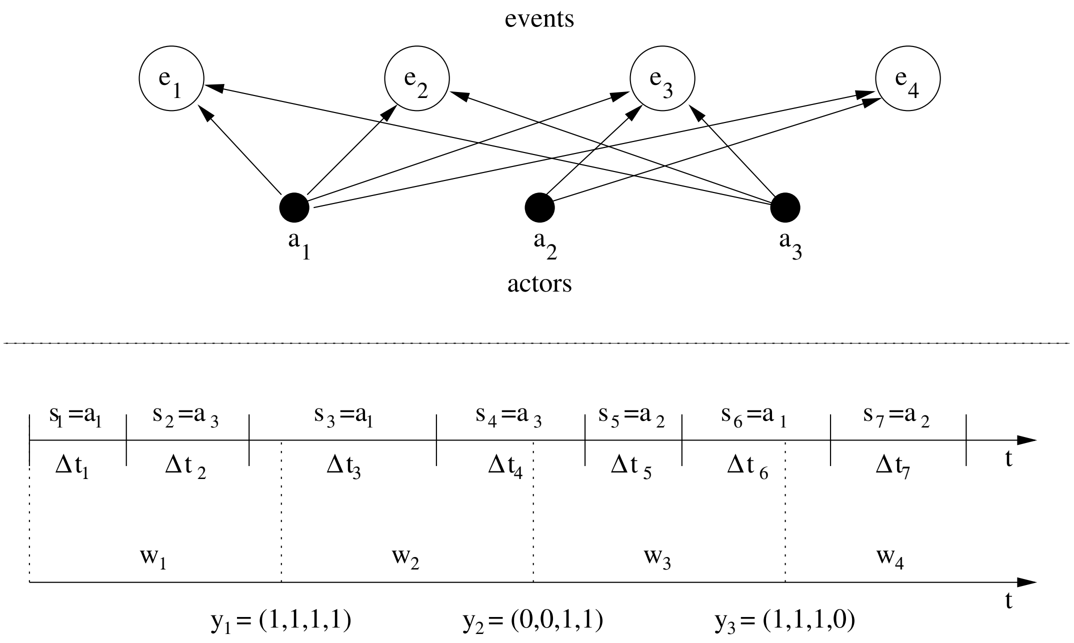

Figure 2: Interaction representation. The figure shows how the speakers are converted into vectors by representing their proximity in time.

into segmenting the recording into stories and anchorman interventions.

3. EXPERIMENTS AND RESULTS

The experiments of this work have been performd over a corpus of 26 news bulletins provided by Radio Suisse Romande, the Swiss national broadcasting service. Each bulletin is one hour long and it is composed of a sequence of stories presented by different persons. Moreover, each bulletin is managed by two anchormen that start and stop the stories by giving the floor to different people. The next sections present the performance metric used in this work and the results obtained at different stages.

3.1 Purity

The results of this work are presented in terms of *purity* π, a performance metric commonly applied in segmentation problems. Given a recording, consider a groundtruth segmentation $S = \{(s_1, \Delta t_1), \dots, (s_{N_g}, \Delta t_{N_g})\}$ and an automatic segmentation $S^* = \{(s_1^*, \Delta t_1^*), \dots, (s_{N_a}^*, \Delta t_{N_a})\}$. The purity π is:

$$\pi = \left(\sum_{i=1}^{N_g} \frac{\tau(s_i)}{T} \sum_{j=1}^{N_a} \frac{\tau^2(s_i, s_j^*)}{\tau^2(s_i)} \right) \cdot \left(\sum_{j=1}^{N_a} \frac{\tau(s_j^*)}{T} \sum_{i=1}^{N_g} \frac{\tau^2(s_i, s_j^*)}{\tau^2(s_j^*)} \right)$$

where $\tau(s_i, s_j^*)$ is the length of the intersection between the time interval corresponding to segment s_i and the time interval corresponding to segment s_j^*, $\tau(s_i)$ is the length of the time interval corresponding to segment s_i, T is the total length of the segmented recording. In each parenthesis, the first term is the fraction of recording a segment accounts for, and the secon term is a measure of how much a given segment is split into smaller fragments. The terms $\tau(s_i)$ at the numerator and $\tau^2(s_i)$ at the denominator are left explicit for the sake of clarity.

The purity value is bounded between 0 and 1, the closer it is to 1, the better it is the segmentation. When the segmentation is perfect, i.e. $S = S^*$, the value of π is 1.

3.2 Story Segmentation Performance

The first step of the process is the speaker clustering. The goal of this stage is to identify the different voices involved in each bulletin in order to reconstruct the pattern of the social relationships. The average purity of the clustering process is 0.77. The average number of speakers involved in the bulletins is 30, but the average number of speakers detected through the clustering is 36.9. This means that the speaker clustering process tends to overestimate the number of speakers and this is due mainly to different background noises that tend to be interpreted as different voices. The impact of the clustering errors on the story segmentation performance is shown below.

The story segmentation process involves two hyperparameters, the first is the number N of windows used to split the recordings (see Section 2) and the second is the amount of variance retained after the application of the PCA. The experiments have been performed using N values between 10 and 20, and keeping at least 70% of the variance. Table 1 shows the performance for different values of the hyperparameters: the purity is always around 0.75 and no major changes are observed when increasing the number of windows or the amount of retained variance (at least in the observed ranges). This seems to suggest that the system is stable with respect to the choice of the above parameters. The results of Table 1 have been obtained using a leave-one-out approach: the HMMs have been trained over 25 recordings using the 26^{th} as test set and each recording has been used alternatively as test set.

On average, the number of stories in the bulletins is 25.2, but the average number of stories detected by the system is 16.5. This means that the nost common error consists in

Table 1: Story segmentation performance. The table reports the purity as a function of the number of windows (win) and the amount of variance retained.

	variance fraction			
win	70%	80%	90%	100%
10	0.74	0.76	0.76	0.78
12	0.74	0.76	0.76	0.78
14	0.74	0.76	0.76	0.77
16	0.76	0.74	0.78	0.78
18	0.74	0.78	0.78	0.79
20	0.75	0.77	0.78	0.79

Table 2: Effect of the speaker clustering errors. The results have been obtained using 14 windows over both manual and automatic speaker segmentations.

	variance fraction			
speak. segm.	70%	80%	90%	100%
manual	0.80	0.80	0.80	0.82
automatic	0.74	0.76	0.76	0.77

grouping different stories rather than in splitting singles stories into smaller segments. The main reason is that speakers involved in different stories, but talking in the same window tend to be represented with similar vectors, thus tend to be attributed to the same story. This apply in particular to shorter stories (less than two minutes) that often follow each other in some specific moments of the bulletins. Another cause of error is that the anchormen tend to talk about different stories in the same intervention and the corresponding story changes cannot be detected by the system presented in this work.

Table 2 shows the effect of the speaker clustering errors on the segmentation performance. The results correspond to $N = 14$, but they are similar to those obtained for all other values of N. The first line of Table 2 shows the purity achieved using the groundtruth speaker segmentation, the second line shows the performance achieved using the speaker clustering. The differences are rather low and the impact of the clustering errors on the segmentation performance seem to be negligible.

Since one of the HMM states corresponds to the anchormen, the segmentation has as a side effect the discrimination between the two journalists managing the bulletin and the rest of the speakers. Although this is not the goal of the work, still it represents a useful information. The anchormen are detected with an accuracy of 74.7%. In other words, 74.7% of the time labeled as anchormen by the system, actually corresponds to such role.

4. CONCLUSIONS

This paper has presented a new approach for the segmentation of broadcast news into stories. The main novelty of this work is that the segmentation is based on the social relationships between the different participants rather than on the actual content, i.e. on what is being said. To our knowledge, no other approaches have been based on the same idea.

The experiments are performed over a corpus of 26 one hour long bulletins and show that a purity of around 0.75 can be achieved. Such a performance can be considered satisfactory for tasks like fast browsing (where the goal is to quickly reach a point of interest in a long recording), semi-automatic data editing (where the goal is to manually adjust the automatic segmentation in order to achieve fully correct results) or role recognition (where the goal is to distinguish the anchormen from the rest of the speakers).

The main limit of the approach is that it works only when the stories are actually reported by different persons and this is not always the case in broadcast news. On the other hand, speaker clustering is a technology simpler and easier to use than other techniques to extract the content of the data (e.g. speech recognition), thus the overall approach is rather robust.

The nexts major step will be the use of alternative HMM topologies (e.g. left-right models with different numbers of stories). The experiments will be extended to other news corpora, in particular TrecVid.

Acknowledgements

This work is supported by the Swiss National Science Foundation through the National Center of Competence in Research on Interactive Multimodal Information Management (IM2). The authors wish to thank Guillermo Aradilla Zapata and Mirko Hannemann for their help.

5. REFERENCES

[1] J. Ajmera. *Robust Audio Segmentation*. PhD thesis, École Polytechnique Fédérale de Lausanne (EPFL), 2004.

[2] J. Gauvain, L. Lamel, and G. Adda. The LIMSI broadcast news transcription system. *Speech Communication*, 37(1-2):89–108, 2002.

[3] F. Jelinek. *Satistical methods for speech recognition*. MIT Press, 1994.

[4] K. Koumpis and S. Renals. Content-based access to spoken audio. *IEEE Signal Processing Magazine*, 22(5):61–69, 2005.

[5] A. Merlino, D. Morey, and M. Maybury. Broadcast news navigation using story segmentation. In *Proceedings of ACM Conference on Multimedia*, pages 381–391, 1997.

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