Exploring the Structure of Broadcast News for Topic Segmentation

Rui Amaral$^{(1,2,3)}$**, Isabel Trancoso**$^{(1,3)}$

[1]Instituto Superior Técnico
[2]Instituto Politécnico de Setúbal
[3] L^2F - Spoken Language Systems Lab, INESC-ID
{ramaral,imt}@l2f.inesc-id.pt

Abstract

This paper describes our on-going work toward the improvement of the story segmentation module of our alert system for selective dissemination of Broadcast News. We have tried to improve our baseline algorithm by further exploring the typical structure of a broadcast news show, first by training a CART and then by integrating it in a 2-stage algorithm that is able to deal with shows with double anchors. In order to deal with shows with a thematic anchor, a more complex approach is adopted including a topic classification stage. The automatic segmentation is currently being compared with the manual segmentation done by a professional media watch company. The results are very promising so far, specially taking into account that no video information is used.

1. Introduction

Topic segmentation is one of the main blocks of our prototype system for selective dissemination of Broadcast News (BN) in European Portuguese. The system is capable of continuously monitoring a TV channel, and searching inside its news shows for stories that match the profile of a given user. The system may be tuned to automatically detect the start and end of a broadcast news program. Once the start is detected, the system automatically records, transcribes, indexes, summarizes and stores the program. The system then searches in all the user profiles for the ones that fit into the detected topics. If any topic matches the user preferences, an email is send to that user, indicating the occurrence and location of one or more stories about the selected topics. This alert message enables a user to follow the links to the video clips referring to the selected stories.

Although the development of this system started during the past ALERT European project (Neto et al., 2003), we are continuously trying to improve it, since it integrates several core technologies that are within the most important research areas of our group: jingle detection (JD) for excluding areas with publicity; audio pre-processing (APP) or speaker diarization which aims at speech/non-speech classification, speaker segmentation, speaker clustering, and gender and background conditions classification; automatic speech recognition (ASR) that converts the segments classified as speech into text; topic segmentation (TS) which splits the broadcast news show into constituent stories; topic indexation (TI) which assigns one or multiple topics to each story, according to a thematic thesaurus; and summarization, which assigns a short summary to each story.

Except for jingle detection, all the components of this pipeline structure produce information that is stored in an XML (Extendible MarkUp Language) file. At the end, this file contains not only the transcribed text, but also additional information such as the segments duration, the acoustic background classification (e.g. clean/music/noise), the speaker gender, the identification of the speaker cluster, the start and end of each story and the corresponding topics.

The major contributions to the area of topic segmentation come from two evaluation programs: Topic Detection and Tracking (TDT) and TREC Video Retrieval (TRECVID), where TREC stands for The Text Retrieval Conference (TREC), both co-sponsored by NIST and the U.S. Department of Defense. The work described in this paper is closer to the goals of the TRECVID campaigns, in the sense that it examines story segmentation in an archival setting, allowing the use of global off-line information. However, although using the video stream is part of our immediate future plans, the current work does not yet explore video cues.

Our previous work in topic segmentation was based on exploring simple heuristics that failed in many scenarios and motivated us to further explore the typical structure of a BN show, which is commonly found among several TV stations in Portugal. One of the first tasks was thus to extend our original BN corpus which was restricted to several types of BN shows from a single station to include shows from other stations. This extended corpus is briefly described in Section 2.. The next Section is devoted to the presentation of our segmentation methods, starting by the baseline approach, the CART approach and the multi-stage one. Section 4. presents the results for the different approaches, compares the performance of the TS method when the APP and ASR modules are replaced by manual labels and evaluates the contribution of non-news information to the TS task. The final Section concludes and presents directions for future research.

2. THE EUROPEAN PORTUGUESE BN CORPUS

Our original European Portuguese Broadcast News corpus, collected in close cooperation with RTP (the public TV station in Portugal), involves different types of news shows, national and regional, from morning to late evening, including both normal broadcasts and specific ones dedicated to sports and financial news. The part of this corpus which is relevant for the current work is divided into 2 main subsets:

- SR (Speech Recognition) - The SR corpus contains around 57h of manually transcribed news shows, collected during a period of 3 months, with the primary goal of training acoustic models and adapting the language models of our large vocabulary speech recognition component of our system. The corpus is subdivided into training (51h) and development (6h) sets. This corpus was also topic labeled manually.
- JE (Joint Evaluation) - The JE corpus contains around 13h, corresponding to two weeks. It was fully manually transcribed, both in terms of orthographic and topic labels. The JE corpus contains a much higher percentage of spontaneous speech (focus conditions F1 + F41) and a higher percentage of speech under degraded acoustical conditions (focus conditions F40 + F41) than our SR training corpus. A potential justification for this fact was that it was recorded during the outbreak of a major war. Figure 1 illustrates the JE contents in terms of focus conditions.

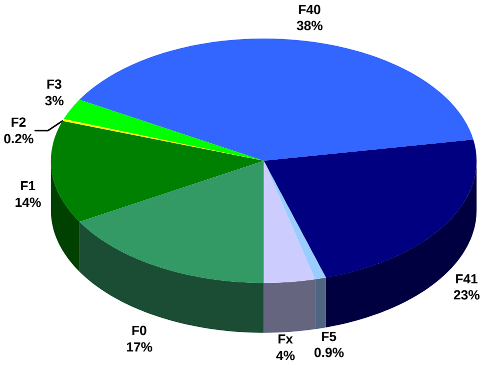

Figure 1: *JE focus conditions time distribution. F0 focus condition = planned speech, no background noise, high bandwidth channel, native speech. F1 = Spontaneous broadcast speech (clean). F2 = Low fidelity speech (narrowband/telephone). F3 = Speech in the presence of background music. F4 = Speech under degraded acoustical conditions (F40 = planned; F41 = Spontaneous). F5 = Non-native speakers (clean, planned). Fx = All other speech (e.g. spontaneous non-native).*

As described above, this corpus was recently complemented with extra BN shows from other TV stations in order to test the performance of our method in different scenarios, namely in terms of the presence of a single anchor, double anchor or a special anchor for a given theme (e.g. sports). This extra BN corpus (EB) contains around 4h and was also fully manually transcribed. Table 1 lists the number of BN shows and stories in each subset of the full corpus which falls within a given category: 1A (single anchor), 2A (double anchor) and TA (thematic anchor).

Note that, on average, one TA show has 30 stories where 28 are presented by the anchor and only 2 are presented by the thematic anchor.

	1A	2A	TA
SR	15 (924)	- (0)	18 (36)
JE	7 (196)	- (0)	7 (14)
EB	1 (30)	2 (60)	- (0)

Table 1: Number of shows (and stories) in each of the three subsets, for the three types of anchor - single, double and thematic.

3. Topic segmentation algorithms

The goal of TS module is to split the broadcast news show into the constituent stories. This may be done taking into account the characteristic structure of broadcast news shows (Barzilay et al., 2000). They typically consist of a sequence of segments that can either be stories or fillers (i.e. headlines / teasers). The fact that all stories start with a segment spoken by the anchor, and are typically further developed by out-of-studio reports and/or interviews is the most important heuristic that can be exploited in this context. Hence, the simplest TS algorithm is the one that starts by defining potential story boundaries in every transition non-anchor / anchor.

Anchor detection could be done by speaker identification, but in order to give the algorithm the potential to process BN shows with previously unknown anchors, the detection was done by finding the speaker with the largest number of turns.

This baseline algorithm has several pitfalls. It fails to detect a boundary between two stories when the first story is all spoken by the anchor and produce false boundaries in every anchor intervention after the story introduction. The considerable amount of miss and false alarm problems led us into further explore the typical structure of a BN show, by adding further heuristics, such as eliminating stories that are too short to put a label on. Rather than hand-tuning these heuristics, we decided to train a CART (Classification and Regression Tree) with potential characteristics for each segment boundary such as: the number of turns of the speaker in the whole show; the total amount of time for that speaker in the whole show; the segment duration (close to a sentence-like unit); the speaker gender; the acoustic background condition; the presence or absence of speech in the segment; the time interval until the next speaker; and the insertion of the segment within an interview region (i.e. with alternating speakers). Each feature vector has these characteristics for the present segment as well as for the previous one.

Figure 2 depicts the characteristics automatically selected by the CART in our development corpus. It is interesting to notice how the CART manages to discard potential story boundaries in non-anchor/anchor transitions in interviews, for instance, by discarding short segments by the anchor.

The CART approach performs reasonably well for BN shows with a simple structure, i.e. single anchor, but fails with more complex structures, involving 2 anchors, for instance, leading us to adopt a two-stage approach: in a first stage of re-clustering, if the BN show is labeled as having two anchors, the two speaker ids with the most frequent turns are clustered into a single label. This stage works as a pre-processing stage, which is then followed by the CART stage.

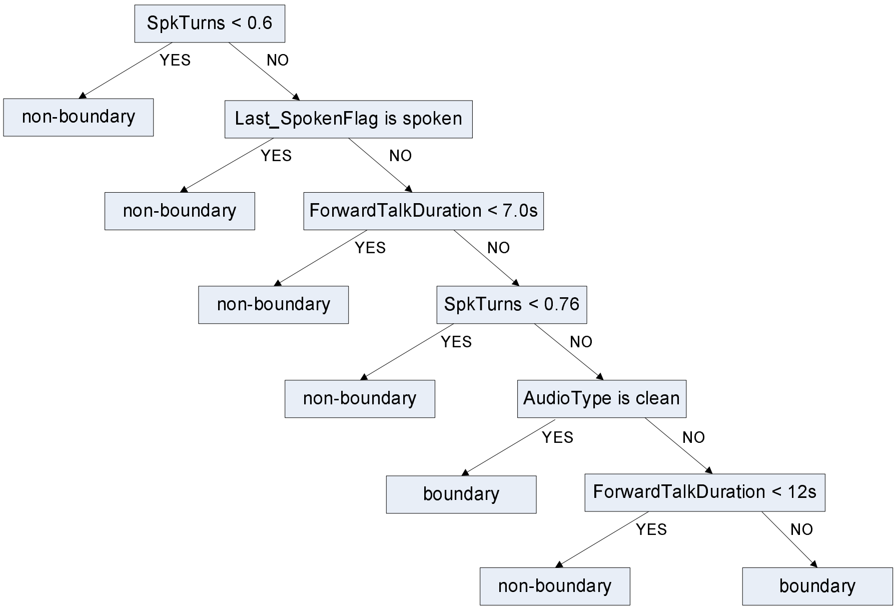

Figure 2: *Diagram of the CART tree. SpkTurns CART feature = the number of turns of the speaker in the whole show; LastSpokenFlag = the presence or absence of speech in the segment; ForwardTalkDuration = the time interval until the next speaker; AudioType = the acoustic background condition*

In order to deal with an even more complex structure involving a thematic anchor (e.g. sports) for a certain period of the BN show, a multi-stage approach had to be adopted. The first stage uses a simple rule to identify potential story boundaries in every non-speech/anchor transitions. The second stage applies the topic indexation module to isolate the portion of the XML file corresponding to the given theme. This stage allows potential story boundaries to appear within the given theme, instead of creating one huge story, with all sports events grouped together. After these initial two stages which create potential story boundaries, a third stage of boundary removal is applied. This final stage uses the same type of rules adopted by the CART to remove boundaries inside interview segments, or boundaries which would create a too short introduction by the anchor, or boundaries that would create a too short story. The relatively short number of stories introduced by the thematic anchor prevented us from training a new CART that would learn these rules automatically.

4. Results

The evaluation of the topic segmentation was done using the standard measures Recall (% of detected boundaries), Precision (% of marks which are genuine boundaries) and F-measure (defined as $2RP/(R+P)$). Table 2 shows the TS results. Every time a story boundary was inserted in the beginning of a filler segment, this was counted as a false alarm. These results together with the field trials we have conducted (Trancoso et al., 2003), show that we have a boundary deletion problem when a story is finished back in the studio. In those situations, the TS algorithm frequently produces a false boundary and if the next story boundary is to close to this one, the later will be deleted. A boundary deletion can also occur if the anchor presence is detected inside a filler segment. Since the filler segments are very short, a false boundary produced inside the filler segment will be to close to the next story boundary. When this happens, the story boundary is deleted (19% of the program events are fillers). The false alarm problems are mainly due to the anchor interventions during the development of a story by a reporter or during an interview.

Approach	%Recall	%Precision	F-measure	Corpus
Baseline	58.8	47.3	0.52	JE
CART	79.6	69.8	0.74	JE
2-Stage	81.2	91.6	0.85	EB
3-Stage	88.8	56.9	0.69	JE

Table 2: Topic segmentation results for manually annotated data.

The comparison of the segmentation results achieved for the JE corpus shows that the CART performance is better than the 3-stage algorithm. In fact, the increase in the recall score attained with the 3-stage algorithm is achieved by finding new story boundaries introduced by the thematic anchor, but at the cost of generating some false alarms which consequently lowers the precision value. For the EB corpus, the CART with the pre-processing stage achieved very good results which indicates that the shows in the EB corpus have a similar structure.

4.1. Dependence on previous APP and ASR stages

The experiments described above were done using manually annotated data. When this manual annotation is replaced by our automatic audio pre-processing stage, we achieve comparable results. In fact, the F-measure for the baseline approach is even higher with the APP module (0.57), whereas for the 3-stage approach, we obtain almost the same (0.68). We were expecting a larger degradation, given that the approaches were trained with manually annotated data.

The most relevant role of our APP module in terms of topic segmentation is speaker clustering. The diarization error rate is 26.1%, a fairly good value if we take into account the very high number of speakers in a typical show of the JE corpus (close to 80), and the fact that we are using a low latency method that does not perform any global optimization on all the speakers of each show (Amaral et al., 2006). The current algorithm still produces more than one cluster per speaker sometimes, specially when the background conditions differ for the same speaker. This never compromise the anchor detection, as the speaker with the higher number of turns. However, some of the segments spoken by the anchor may be attributed to a different cluster. When this happens at the beginning of a story we have a miss, but when this happens in the middle of a story (for instance in a sequence involving anchor - journalist 1 - anchor - journalist 2) this clustering error will avoid a false alarm.

Despite the different corpus, a close look at the best results achieved in TRECVID story segmentation task (F=0.7) (Smeaton et al., 2004) show our good results, specially considering the lack of video information in our approach.

The automatic identification of the type of BN show (single anchor, double anchor or thematic anchor) was not yet dealt with. In fact, so far, Portuguese TV channels have adopted a very regular pattern for each show. This is the type of problem, however, where video information can play a very important role.

4.2. Impact of non-news information

Recently our APP module started to include additional information on the XML file indicating non-news segments inside the broadcast news program. The non-news segments correspond to fillers, jingles and advertising segments. That information was used in the TS task to define another story boundary detection rule, since after the filler there is a story boundary defining the beginning of a new report. A preliminary experiment was done using the non-news information in the story segmentation of three automatically annotated broadcast news programs (3h). Table 3 shows the results.

Approach	%Recall	%Precision	F-measure
CART	98.9	71.7	0.83
3-Stage	96.8	73.9	0.84

Table 3: Topic segmentation results using non-news information.

The comparison of the results in table 3 with the ones presented in table 2 shows that the use of non-news information in the story segmentation increases the recall and precision value. The use of non-news information avoids the false boundaries inside fillers with the consequent boundary deletion after the fillers, and avoids also the boundary deletion after the jingle of the program start and after the advertising in the middle of the program. The performance achieved with both approaches is comparable, because there is no thematic anchor in the three programs used in the evaluation. A close inspection of the results, shows that the main problem still the false alarm rate due to the anchor interventions in the program, whose talk duration is long enough to be a story introduction. Although these cases have de-creased, the anchor interventions at the end of the story are critical because they usually produce a false alarm and a boundary deletion. We intend to explore the automatic transcriptions in order to decrease the number false boundary cases, due to the anchor interventions after the story introduction.

5. Conclusions and Future Work

This paper presented our on-going work toward the improvement of the story segmentation module of our alert system. We have tried to improve our baseline algorithm by further exploring the typical structure of a broadcast news show, first by training a CART and then by integrating it in a 2-stage algorithm that is able to deal with shows with double anchors. In order to deal with shows with a thematic anchor, a more complex approach was adopted including a topic classification stage.

For the past two weeks, the informal cooperation with a media watch company has allowed us to assess the precision of our topic segmentation module as compared to manual labeling. This assessment was not yet formally done for a significant number of shows, but the results are very promising. We obtained a recall of 92.3% and a precision of 71.6%. In this case, we adopted the same strategy of the media watch company of marking filler boundaries as normal story boundaries. Fillers can still cause some false alarms when the typical music background is not properly detected by the APP module, leading it to be processed in the further stages. The false boundary cases due to the anchor interventions in the middle of the program is a critical problem and we intend to explore the automatic transcriptions to decrease the number of cases.

The fact that the manual mark is inserted by looking at the video and audio signals simultaneously justifies a small delay in the automatic story labels which only use the audio signal. In fact, the automatic mark is typically inserted at the start of the clean background segment by the anchor which may be preceded in the same story by some short segment marked as non-speech. In fact, some of the stories were correctly detected although their automatic boundaries were delayed by an average of 2s, which exceeds the test window of plus or minus 1s used to evaluate the number of hits. If these delayed boundaries were counted as hits, the recall would be 95.2% and the precision 73.9%.

The correction of this short delay is currently being addressed, namely by combining audio and video cues. This fusion may also lead to a better anchor identification rate and is one of our major tasks in the VIDI-Video European project.

6. Acknowledgments

The present work is part of Rui Amaral's PhD thesis, initially sponsored by a FCT scholarship. INESC-ID Lisboa had support from the POSI program of the "Quadro

Comunitário de Apoio III". The authors would like to acknowledge the continuing support of our former ALERT project colleagues - João Neto, Hugo Meinedo and Diamantino Caseiro.

7. References

Amaral, R., H. Meinedo, D. Caseiro, I. Trancoso, and J. Neto, 2006. Automatic vs. manual topic segmentation and indexation in broadcast news. In *Proc. IV Jornadas en Tecnologia del Habla*. Zaragoza, Spain.

Barzilay, R., M. Collins, J. Hirschberg, and S. Whittaker, 2000. The rules behind roles: Identifying speaker role in radio broadcast. In *Proc. AAAI 2000*. Austin, USA.

Neto, J., H. Meinedo, R. Amaral, and I. Trancoso, 2003. A system for selective dissemination of multimedia information resulting from the alert project. In *Proc. MSDR '2003*. Hong Kong.

Smeaton, Alan F., Paul Over, and Wessel Kraaij, 2004. Trecvid: evaluating the effectiveness of information retrieval tasks on digital video. In *MULTIMEDIA '04: Proceedings of the 12th annual ACM international conference on Multimedia*. New York, NY, USA: ACM Press.

Trancoso, I., J. Neto, H. Meinedo, and R. Amaral, 2003. Evaluation of an alert system for selective dissemination of broadcast news. In *Proc. Eurospeech '2003*. Geneva, Switzerland.