Topic Segmentation and Indexation in a Media Watch System

Rui Amaral[1,2,3], *Isabel Trancoso*[1,3]

[1]INESC-ID Lisboa, Portugal
[2] Instituto Politécnico de Setúbal
[3] Instituto Superior Técnico

Rui.Amaral@inesc-id.pt, Isabel.Trancoso@inesc-id.pt

Abstract

The goal of this paper is the description of our current work in terms of topic segmentation and indexation and the comparison of their performance with the story boundaries and topics manually chosen by a professional media watch company. The segmentation module explores the typical structure of a broadcast news show, namely by cues provided by the audio pre-processing module, but an improved performance could be achieved by taking its contents into account. The topic indexation module was retrained for the media watch topics. The comparison showed how different criteria, together with manual labeling inconsistencies, may affect the performance.
Index Terms: topic segmentation, topic indexation

1. Introduction

Topic segmentation and indexation play an important role in the prototype system for selective dissemination of Broadcast News (BN) in European Portuguese, developed at INESC-ID. The media watch system was initially built in the context of the ALERT European project [1][2][3] and is the object of continuous improvement in the framework of national project TECNOVOZ. 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 BN show. Once the start is detected, the system automatically records, transcribes, indexes, summarizes and stores the show.

The media watch system integrates several core technologies: jingle detection (JD) for excluding areas with publicity; audio pre-processing (APP) which aims at speech/non-speech classification, gender and background conditions classification, speaker clustering (diarization), and speaker identification; automatic speech recognition (ASR) that converts the segments classified as speech into text; punctuation and capitalization; topic segmentation (TS) which splits the broadcast news show into constituent stories; topic indexation (TI) which assigns one or multiple topics to each story; and summarization, which assigns a short summary to each story.

The first modules of this system were optimized for on-line performance, given their deployment in the fully automatic subtitling system that is running on the main news shows of the public TV channel in Portugal (RTP), since early March[4]. In order to improve the performance of this subtitling system, the lexical and language models of the speech recognizer are daily adapted, and acoustic models are specially trained for very frequent speakers (anchors). The last three modules of the system, on the other hand, are running off-line, exploiting cues that are computed over the whole BN show. 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 explored the typical structure of a BN show, which is commonly found among several TV stations in Portugal. Although this strategy yields fairly good results, improving them can be achieved at the cost of looking not only at the structure, but also at the contents. This means that we need to take into account both the automatic transcription produced by the speech recognizer, and the topics produced by the topic indexation module. Hence, the two modules become interleaved.

In terms of topic indexation, our previous work aimed at classifying the stories according to a hierarchical thematic thesaurus, daily used at RTP for archival purposes. The amount of training material (300h of automatically transcribed, manually segmented and classified stories, collected during 2001) was very unevenly distributed among the 22 top-level topics. This meant that some of them were very badly modeled, and also limited the number of hierarchical levels that we could explore to the first three.

Very recently, however, we faced a new challenge, which was the comparison of the performance of our topic segmentation and indexation modules with the manual topic boundaries and labels done by a professional media watch company. This paper aims at describing the recent work in the two modules and their adaptation to the new task, in order to allow the comparison of their performance with the media watch company results, and assess how far or how close we are to being able to use our topic segmentation and indexation modules in real applications.

The paper starts with a very brief description of our BN media watch corpus that served as a basis for this evaluation in Section 2. The bulk of the paper is devoted to the topic segmentation (section 3) and indexation (section 4) modules. The final Section concludes and presents directions for future research, namely in terms of what can be exploited to improve the performance of our automatic media watch system using video-derived cues.

2. Corpora

The experiments described in this paper use several distinct BN corpora, which justifies a brief description of all the subsets.

SR (Speech Recognition) - This corpus contains around 57h of manually transcribed news shows, collected in 2000 during the ALERT project. It is fully manually transcribed, both in terms of orthographic transcriptions and topic labels. The corpus is subdivided into training (51h) and development (6h) sets. It includes several different types of BN shows, all produced by RTP. 15 of the 33 shows in

this corpus were presented by a single anchor. The others also had a different thematic anchor introducing all the sports stories. This corpus was used for training the topic segmentation module.

TD (Topic Detection) - This corpus contains around 300h of topic labeled news shows, collected during the following 9 months. All the data was manually segmented into stories or headlines, and each story was manually indexed according to the thematic thesaurus adopted by RTP. The corresponding orthographic transcriptions were automatically generated by our ASR module.

JE (Joint Evaluation) - This corpus contains around 13h, corresponding to two weeks of recordings. It was collected in 2001, still in the scope of the ALERT project, and is fully manually transcribed, both in terms of orthographic and topic labels. Half of the 14 shows were presented by a single anchor. The other half also included a thematic anchor for sports.

EB (Extended BN) - The BE corpus contains around 4h, which were also fully manually transcribed. This corpus was collected during 2006 from a different TV station. The purpose of this extended corpus was the test of our segmentation approaches with more complex BN structures. The first show is presented by a single anchor, but includes very frequent dialogues with a local commentator. The second show is presented by two anchors. The third show includes two anchors and a local commentator.

MW (Media Watch) - This corpus was collected during 2007 and 2008 and was manually segmented and labelled by the media watch company. The orthographic transcriptions were automatically provided by our ASR system. The recordings of the RTP daily evening shows were done independently at our lab and at the company, which implied some synchronization problems. The last 9 months of 2007 were used for training topic models (167 shows); January 2008 was used for development (23 shows) and February/March 2008 (25) were used for testing. Each show has approximately 1h duration, with a single publicity break detected by the JD.

3. Topic segmentation

The goal of TS module is to split the broadcast news show into its constituent stories. This may be done taking into account the characteristic structure of broadcast news shows [5]. 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.

The demand for better segmentation performances 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

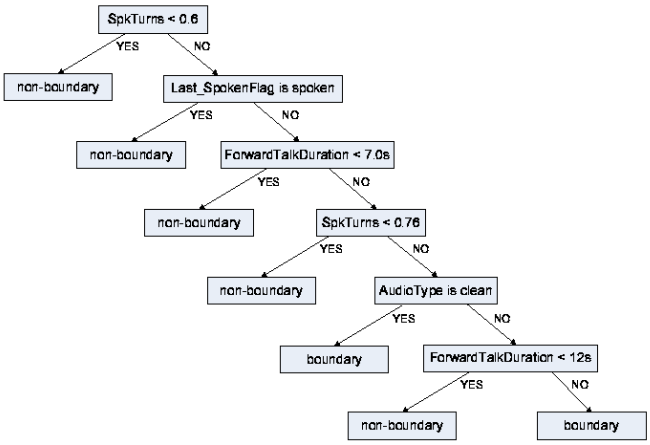

Figure 1: *Diagram of the CART tree. SpkTurns CART feature = the number of turns of the speaker in the whole show; LastSpokenFlag = the presence or absence of speech in the segment; ForwardTalkDuration = the time interval until the next speaker; AudioType = the acoustic background condition.*

these characteristics for the present segment as well as for the previous one. Figure1 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 performed reasonably well for BN shows with a simple structure, i.e. single anchor, but failed with more complex structures, involving 2 anchors, for instance. This led us to adopt a two-stage approach: in a first stage of reclustering, if the BN show is manually 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 application of the CART. This 2-stage approach was only evaluated with the EB corpus.

3.1. Exploring the topic related structure

The CART approach also fails with the presence of a thematic anchor for a certain period of the BN show. In order to deal with this complex structure, a multi-stage approach was adopted, where topic segmentation is interleaved with topic indexation.

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 BN show corresponding to the given theme (sports). 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.

The analysis of the behavior of this multi-stage approach also indicated that whereas the sports news needed to be split

into constituent stories, for other topics, the opposite problem occurred. In fact, false alarms were very frequent for the weather forecast topic, which was typically split into multiple stories, due to the relatively long pauses made by the anchor between the forecasts for each part of the country. In order to deal with this problem, the topic indexation stage was used to detect weather forecast stories. When two or more adjacent weather forecast stories are detected, they are merged into a single one.

3.2. Exploring non-news information

The type of features used in the CART makes our automatic topic segmentation module very dependent on the good performance of the audio-preprocessing module, and mainly on its ability to correctly identify the anchor. On-going work in terms of APP improvement thus has a great influence on the TS module. One of the most recent improvements of the APP module concerns the inclusion of a non-news detector which marks the jingles that characterize the start and end of the BN show, the jingles that delimit publicity segments, and also the ones that signal headlines/teasers.

Just as for jingles, the background music that characterizes headlines typically changes every new season, thus imposing the periodic retraining of the detector. This motivated the collection of the recent MW corpus, which includes the headline music for which the current detector is trained for.

The performance of the previous version of the topic segmentation module was seriously hindered by the presence of headlines. They typically start with a segment containing only music. This initial boundary was thus discarded by the TS module, causing a miss boundary. The moment the anchor starts speaking, with the continuing headline music, was also considered a false alarm. The next story would hence be typically merged with the headline.

The headline information was used in the TS module to define another story boundary detection rule, since after the headline there is a story boundary defining the beginning of a new report. The use of this information avoids the false boundaries inside headlines and the consequent boundary deletion after the headlines, and avoided also the boundary deletion after the jingle of the program start and after the publicity jingles in the middle of the program.

3.3. Exploring the contents of BN segments

A detailed inspection of the results showed that the main problem still remaining is the false alarm rate due to the anchor interventions in the program, whose duration is long enough to be considered a story introduction.

Some of these anchor segments originate short potential stories. Since we avoid stories too short to be reliably indexed, the present TS algorithm uses the background information to decide which boundary to remove. This decision is critical since the deletion of a correct boundary causes both a false alarm and a miss boundary.

The most frequent false alarm occurs at the end of a story. In fact, although this is not systematically observed for every story, the anchor frequently finishes a story by a short comment or a short acknowledgment of the reporter. These ending segments by the anchor are typically merged with the introduction of the following story, thus creating both a false alarm (when the ending segment starts) and a boundary deletion (when the new story starts).

In order to decrease these false alarms, we explored the automatic transcriptions of the BN shows. In order to improve the merging of short stories with either their left or right neighbors, a CART was trained using the following features: the acoustic background conditions of the left and right stories, the word rate (computed at the first 7s of the short story, which is the minimal time required for a story introduction), the duration of the anchor segment, and the normalized count of matches of unigrams, bigrams and trigrams between the short story and the two neighbors. The matches are computed over the automatic transcripts and the purpose is to detect text similarities between the short story and its neighbors. Hence, if the short story is similar to the one on its left, then the boundary that separates these two stories is deleted.

3.4. Results

The evaluation of the topic segmentation module was done using the standard measures Recall (% of detected boundaries), Precision (% of marks which are genuine boundaries) and F-measure (defined as $2RP/(R+P)$).

The results achieved by the first single-stage CART approach are presented in the first line of Table 1 for the JE corpus. The evaluation of the 2-stage approach could only be done with the EB corpus (second line). The performance of the 3-stage approach is shown in the following line, also for the JE corpus, but only taking the sports topic splitting into account.

The remaining experiments were conducted using the MW corpus. For this purpose, 6 shows during 2007 were randomly chosen. Not all experiments were done using all 6 shows, except the final one. In the table, the index *i* in MWi indicates the number of shows for which results are reported.

Since the recordings of the same BN show were independently done at our lab, and at the media watch company, the recording time may differ. Hence, the first pre-processing task is a computation of the optimum time shift (among a closed set of possible shifts) between the start boundaries that maximizes the F-measure of the TS module. In order to compute the nearest manual boundary, an evaluation window of 10s is used, both to the left and to the right of the detected boundary.

The next two lines compare the performance of the multi-stage approach without and with merging the stories classified as meteorology. The next two lines compare the performance without and with the integration of non-news information. Although the results were achieved with a different BN corpus (with the current headline music), the comparison shows that the use of non-news information in the story segmentation increased the recall and precision values. The next two lines compare the performance of the previous system without and with the integration of the automatic speech recognition results.

A careful analysis of the last results shows that there are still false alarms / miss boundaries pairs that could be avoided if the evaluation window was extended to 2s. The corresponding results are shown in the last two lines of the Table.

The results obtained for the full MW evaluation set were much worse (F-measure=0.81). A close inspection of the results for each show reveals that the different segmentation criteria adopted by the media watch company may be responsible for this behavior. In fact, this company seems to favor merging consecutive stories with the same topic. Whereas the archive staff at RTP marked story boundaries when the anchor introduced a new perspective on the same event, the media watch company prefers merging them into a single story (although there is much inconsistency in this preference). If the BN includes a major event (i.e. a flood in Lisbon), the impact of merging may be very significant, as the number of stories may be reduced to less

Approach	%Rec	%Prec	F-m	corpus
Single-Stage	79.6	69.8	0.74	JE
Two-Stage	81.2	91.6	0.85	EB
Multi-Stage	88.8	56.9	0.69	JE
Multi-Stage	97.1	86.8	0.92	MW1
Multi-Stage (+meteo)	97.1	89.2	0.93	MW1
Multi-Stage	98.9	71.7	0.83	MW3
Multi-Stage (+non-news)	96.8	73.9	0.84	MW3
w/o ASR (eval=1s)	88.0	81.7	0.85	MW6
with ASR (eval=1s)	91.2	83.0	0.87	MW6
w/o ASR (eval=2s)	93.8	87.1	0.90	MW6
with ASR (eval=2s)	97.0	88.2	0.92	MW6

Table 1: Topic segmentation results.

than a half of the normal one in a BN show.

The impact of headlines is also significant. Whereas in the previous corpora manually marked by RTP, the start of the headline music marked the boundary, the media watch company marked the boundary as the start of the speech.

We believe that these results can be improved by adapting the segmentation algorithm to the new segmentation criteria.

4. Topic indexation

The number of topics used by the media watch company has evolved during the last year. Whereas during the first months the main distinction was between national and international news, these two broad categories were further subdivided in the following months. In the subdivision, one could distinguish 9 new topics that could be further subdivided, although this hierarchical structure was not consistently used, and was not very visible. In fact, it was provided to us as an HTML file, in which all topics of each story were written in the same line, separated by a punctuation sign. Table 2 shows the amount of stories for each topic in our training, development and test sets. The topic "meteorology" (or weather forecast) was very rarely identified as topic. In fact, the stories on this topic were classified as "national", but included weather forecast as the title of the piece. Because of the importance of this topic for our segmentation module, we extracted the topic information from the title.

Topic	Train	Dev	Test	%Acc
National	3558	526	518	83.10
International	1859	227	233	87.13
Economy	946	194	149	90.65
Education	196	17	52	96.22
Environment	235	34	33	94.91
Health	315	90	50	95.69
Justice	496	63	91	94.26
Meteorology	69	7	6	99.41
Politics	1838	285	357	87.24
Security	1037	138	158	87.99
Society	1455	193	260	74.43
Sports	719	118	98	96.86

Table 2: Number of stories in each topic in the training, development and test sets of the MW corpus, and corresponding accuracy.

For each of the 12 classes, topic and non-topic unigram language models were created using the stories of the MW corpus which were pre-processed in order to remove function words and lemmatize the remaining ones. Topic detection is based on the log likelihood ratio between the topic likelihood $p(W/T_i)$ and the non-topic likelihood $p(W/\overline{T_i})$. The detection of any topic in a story occurs every time the correspondent score is higher than a predefined threshold. The threshold is different for each topic in order to account for the differences in the modeling quality of the topics.

4.1. Results

The last column of Table 2 shows the accuracy values obtained for the media watch test corpus. As expected, best results were obtained for the meteorology topic. The bad results obtained for the society topic can be justified by the fact that it is some sort of a miscellaneous topic. The difficulty in assigning the national and international topics may be also partly justified by the fact that stories on how the country is viewed abroad may be manually classified with both topics. This type of indexation will be specially hard to implement automatically.

5. Conclusions and future work

This paper described our on-going work on the topic segmentation and indexation modules for broadcast news, ending by the comparison of their results with the ones of a media watch company. This evaluation showed how different segmentation and indexation criteria, together with manual labeling inconsistencies may affect the performance.

Taking advantage of the daily newspaper collection that is currently done to update language and lexical models is part of our future plans, in order to improve topic models as well.

Our collaboration with experts in video segmentation and shot representation in the framework of European project VIDI-VIDEO allows us to discuss the feasibility of using video derived cues for the task of topic segmentation. A preliminary experiment with a single recent BN show revealed potential advantages, namely in terms of detecting video cues such as anchor, double news-anchor, news studio, and split screen.

6. Acknowledgments

The present work is part of Rui Amaral's PhD thesis, initially sponsored by a FCT scholarship. This work was partially funded by PRIME National Project TECNOVOZ number 03/165, and by the European project Vidi-Video.

7. References

[1] Neto, J., Meinedo, H., Amaral, R., Trancoso, I., A system for selective dissemination of multimedia information resulting from the ALERT project, Proc. MSDR '2003, Hong Kong (April 2003)

[2] Lo, Y., Gauvain, J., The LIMSI topic tracking system for TDT 2002, Proc. DARPA Topic Detection and Tracking Workhsop, Gaithersburg, USA, November 2002.

[3] Werner, S., Iurgel, U., Kosmala, A., Rigoll, G., Tracking topics in broadcast news data, Proc.ICME'2002, Lausanne, Switzerland, September 2002.

[4] Martins, C., Teixeira, A. and Neto, J. , "Dynamic language modeling for a daily broadcast news transcription system", Proc. ASRU 2007, Kyoto, Japan.

[5] Barzilay, R., Collins, M., Hirschberg, J., Whittaker, S., The rules behind roles: Identifying speaker role in radio broadcast, Proc. AAAI 2000, Austin, USA, July 2000.