SEGMENTATION AND INDEXATION OF BROADCAST NEWS

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

[1] IPS/INESC ID Lisboa [2] IST/INESC ID Lisboa
INESC ID Lisboa, Rua Alves Redol, 9,1000-029 Lisboa, Portugal
{Rui.Amaral, Isabel.Trancoso}@inesc-id.pt
http://l2f.inesc-id.pt

ABSTRACT

This paper describes a topic segmentation and indexation system for broadcast news that is integrated in an alert system for selective dissemination of multimedia information. The goal of this work is to enhance the retrieval and navigation through specific spoken audio segments that have been automatically transcribed, using speech recognition. Our segmentation algorithm is based on simple heuristics related with anchor detection. The indexation is based on hierarchical concept trees, containing 22 main thematic domains, for which Hidden Markov models were created. Only the three top levels in this thesaurus are currently used for indexation. On-going work on the identification of some cues related to the structure of TV broadcast news programs is also described.

1. INTRODUCTION

The huge amount of information we can access nowadays in very different formats (audio, video, text) and through distinct channels revealed the necessity to build systems that can efficiently store and retrieve this data in order to satisfy future information needs. This was the framework for the recently finished ALERT European Project (Alert System for Selective Dissemination of Multimedia Information [1]), whose goal was to build a system capable of continuously monitoring a TV channel, and searching inside their news programs for the stories that match the profile of a given client. The system may be tuned to a particular TV channel in order to automatically detect the start and end of a broadcast news program. Once the start is detected, the system automatically records, transcribes, indexes and stores the program. Each of the segments or stories that have been identified is indexed according to a thematic thesaurus. The system then searches in all the client profiles for the ones that fit into the detected categories. If any topic story matches the client preferences, an email is send to that client indicating the occurrence and location of one or more stories about the selected topics. This alert message enables a client to find in the System Website the video clips referring to the selected stories.

This paper concerns only the segmentation and indexation modules of the ALERT system. Such modules may take into account the characteristic structure of broadcast news programs. They typically consist of a sequence of segments which can either be stories or fillers. The first ones describe a certain topic in more or less detail, whereas the second ones summarise the main headlines or an important topic to be presented later in the program. The differences between stories and fillers are not only in terms of the type of contents but include as well audio and visual cues. Stories are most often introduced by the anchor, in studio. After this introductory part, there is often the development of the news which may involve several reporters and include interviews led by the anchor in the studio or by a reporter outside the studio. Fillers, on the other hand, are either spoken by the anchor or an unseen reporter and are often accompanied by special jingles.

The broadcast news corpus and thesaurus used in this work are described in Section 2. The following two sections present our segmentation and indexation algorithms, respectively. Section 5 presents the story segmentation results, using as input stream data that was automatically segmented into sentences together with information about background acoustical environment and speaker identification for each sentence. Section 6 shows the results of an indexation task where the descriptors of the thematic thesaurus were used as indexing keys in stories whose boundaries were manually identified. The last part of the paper describes on-going work that has not yet been integrated in the test prototype and is aimed at

[1] More information about the project may be found at the following URL: http://alert.uni-duisburg.de

distinguishing stories and fillers and, inside each of these blocks, the role of the intervening speakers (anchor, reporter or interviewed). The paper concludes with a discussion of these results and our plans for future research in this area.

2. TOPIC DETECTION CORPUS DESCRIPTION

This section presents the Topic Detection Corpus (TDC) that was used to develop and test our indexation algorithm. This TV Broadcast News Corpus in European Portuguese was manually segmented and indexed using a thesaurus, in cooperation with the national public broadcasting company - RTP (*Rádio Televisão Portuguesa*).

Collected in the scope of the ALERT project over a period of 9 months in 2001, the BN corpus contains around 300 hours of audio data from 133 TV broadcast evening news programs. The corresponding orthographic transcriptions were automatically generated by our speech recognition engine [3]. All the programs were manually segmented into stories or fillers, and each story was also manually indexed according to a thematic, geographic and onomastic thesaurus.

The Thematic Thesaurus is hierarchically organized into 22 top domains which are: Justice and Rights (JR), Defence and Security (DS), Society and Sociology (SS), Political Organisation (PO), Sports and Leisure (SL), Transportation (TR), Science and Technology (ST), Communication and Documentation (CD), Work and Employment (WE), Economy and Finance (EF), Health and Feeding (HF), Religion and Ethics (RE), Arts and Culture (AC), House and Living (HL), Industry (IN), Environment and Energy (EE), Agriculture (AG), European Union (EU), History (HI), Weather Forecast (WF), Events (EV) and Education (ED). On the whole, the Thesaurus contains 7781 descriptors distributed among 10 levels as shown in Table 1.

Table 1. Descriptor distribution among thesaurus levels

Thesaurus level	Descriptors %
1st-level	0.21%
2nd -level	7.62%
3rd -level	48.32%
4th -level	26.47%
5th-level	11.83%
6th-level	3.84%
7th-level	0.86%
8th-level	0.63%
9th-level	0.19%
10th-level	0.03%

The onomastic and geographic thesauri have 1765 and 1890 entries, respectively. The first ones include institution names, as well as person names. These entries are currently used by the manual annotators to identify the story speakers, and not the persons who are the subject of the story.

The topic detection corpus was divided into 3 subsets, covering different periods in time, for training, development and evaluation purposes. The training corpus includes 85 programs, corresponding to 2451 report segments and 530 fillers. The development corpus includes 21 programs, corresponding to 699 report segments and 144 fillers. The evaluation corpus includes 27 programs, corresponding to 871 report segments, and 134 fillers. Very frequently, a report segment is classified into more than one topic. Such report segments will originate multiple stories which justifies that, for instance, the development corpus includes 6073 stories. The distribution of the thematic domains among the stories is shown in Figure 2 for the 3 subsets.

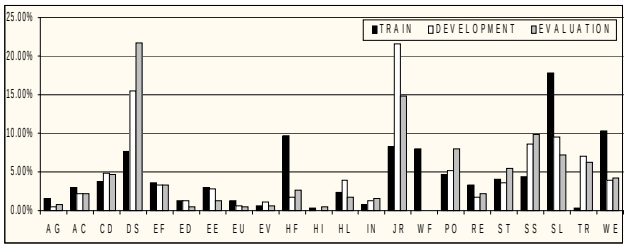

Fig. 2. Topic distribution in the topic detection corpus.

3. STORY SEGMENTATION

The input to the segmentation algorithm is a stream of audio data, which was automatically segmented into sentences (or rather "transcript segments", defined by pauses), and later transcribed by our automatic speech recognition (ASR) system. Each transcript segment contains as well some information related to the background acoustic environment, the speaker gender, and the speaker identification. All this metadata is also automatically extracted from the speech signal.

The speaker identification is of particular importance to the segmentation algorithm, namely, the classification as anchor or non-anchor. In fact, in our broadcast news programs, the anchors are responsible for introducing most stories. They are also the speakers whose id-numbers appear most often during the whole program, independent of the duration of each talk.

Our segmentation algorithm is based on a very simple heuristic derived from the above assumptions. It identifies the transcript segments belonging to the most frequent speaker-id number (the anchor), and defines potential

story boundaries in every transition "non-anchor transcript segment/anchor transcript segment".
In the next step, we try to eliminate stories that are too short (containing less than 3 spoken transcript segments), because of the difficulty of assigning a topic with so little transcribed material. In these cases, the short story segment is merged with the following one.
The next stage, following this two-step segmentation algorithm, is indexation, as described in the next section. After this classification stage, a post-processing segmentation step may be performed, in order to merge all the adjacent segments classified with the same topic.

4. STORY INDEXATION

Story indexation is performed in two steps. We start by detecting the most probable story topic, using automatically transcribed text. Our decoder is based on the HMM (Hidden Markov Model) methodology and the search for the best hypothesis is accomplished with the Viterbi algorithm [9]. The topology used to model each of the 22 thematic domains is single-state HMMs with self-loops, transition probabilities, and either unigram or bigram language models [7]. Models were computed from automatically transcribed stories with manually placed boundaries which were post-processed in order to remove function words and lemmatize the remaining ones. Lemmatization was performed using a subset of the SMORPH dictionary with 97524 entries [4]. Smoothed bigram models were built from this processed corpus with an absolute discount strategy and a cutoff of 8.
In the second step, we find for the detected domain all the second and third level descriptors that are relevant for the indexation of the story. To accomplish that, we count the number of occurrences of the words corresponding to the domain tree leafs and normalize these values with the number of words in the story text. Once the tree leaf occurrences are counted, we go up the tree accumulating in each node all the normalized occurrences from the nodes below [2]. The decision of whether a node concept is relevant for the story is made only at the second and third upper node levels, by comparing the accumulated occurrences with a pre-defined threshold. The decision to restrict indexation to the second and third node levels was made taking into account the ALERT project goals and the data sparseness at the thesaurus lower levels.

5. SEGMENTATION RESULTS

For the evaluation of our segmentation algorithm, we adopted the metric used in the 2001 Topic Detection and Tracking (TDT 2001) benchmark NIST evaluation [5]. The judgment was made according to Table 2, using a sliding evaluation window of 50 words.

Table 2. Segmentation judgement for each window

Judgement	Situation
Correct	There is a computed and a reference boundary inside the evaluation window.
Correct	Neither a computed nor a reference boundary is inside the evaluation window.
Miss	No computed boundary is inside the evaluation window that contains a reference boundary.
False Alarm	A computed boundary is inside the evaluation window that does not contain a reference boundary.

The cost segmentation function C_{Seg} is defined as:
$(C_{Seg})_{Norm} = C_{Seg} / \min(C_{Miss} \times P_{Target} , C_{FA} \times P_{Non\text{-}Target})$
where
$C_{Seg} = C_{Miss} \times P_{Miss} \times P_{Target} + C_{FA} \times P_{FA} \times P_{Non\text{-}Target}$
and
C_{Miss}: cost of a miss.
P_{Miss}: conditional probability of a miss
P_{Target}: a priori target probability
C_{FA}: cost of a false alarm
P_{FA}: conditional probability of a false alarm
$P_{Non\text{-}Target}$: a priori non-target probability (1- P_{Target})

Using the values of C_{miss} and C_{FA} adopted in TDT2001 [6] (1 and 0.3, respectively), we achieved a normalized value for the segmentation cost of 0.84 for a P_{Target} of 0.8. A slightly higher value (0.86) was obtained without the post-processing stage that merged adjacent story segments if their topic classification is equal. However, the segmentation cost value did not reach 0.9, which was state-of-the-art in TDT2001 for this task. Several critical problems were detected: one of the reasons for boundary deletion is related to anchor detection in filler segments. Filler segments are very short segments spoken by the anchor and usually followed by a new story introduced by the anchor. In this scenario, and since all potential story boundaries are located in transitions "non-anchor transcript segment/anchor transcript segment", the boundary mark will be placed at the beginning of the filler region and no more boundary marks will be placed.
Another reason for boundary deletion is the presence of multiple anchors. Some of the programs in our corpus had in fact two anchors, one of which was responsible only for the sports stories. Our simple heuristic does not yet support multiple anchors. The story boundaries introduced by the latter will all be missing.

6. INDEXATION RESULTS

To measure the performance of the indexation algorithm, an experiment was done using the stories of the evaluation corpus and ignoring all the filler segments. In order to discard the influence of segmentation errors, this experiment was done using manually placed story boundaries and automatically transcribed texts.

In the evaluation of the indexation algorithm, we had to take into account the fact that there are stories that were manually indexed with more than one thematic domain (39% of the stories). We considered a hit every time the topic decoded is present in the topics manually identified in the story by the human annotators.

Our first set of experiments considered only the classification into the 22 hierarchical domains. The correctness achieved in the evaluation corpus was 73.80% using the bigram model, and 73.53% using the unigram model. The proximity of the results indicates that the amount of training data is not enough to build robust bigram models. Figure 6 shows the confusion matrix that can be obtained using only the subset of the evaluation corpus corresponding to stories that were manually topic annotated with a single topic.

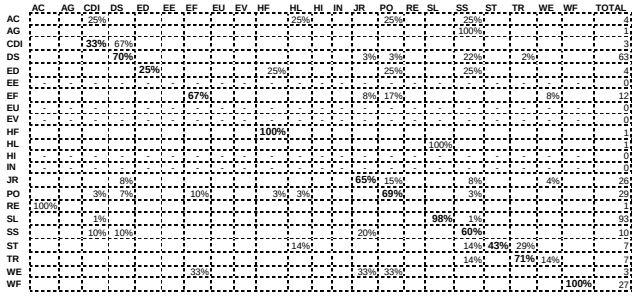

Fig. 6. Confusion matrix for a subset of the Evaluation Corpus.

The rightmost column of the matrix indicates the number of stories accounted for. We see that the least confusable topic is "weather forecast" which is never confused in a one-to-one classification. Some of the substitution errors are easily understood, given the topic proximity.

In terms of the second and third level descriptors, the results achieved a precision of 76.39% and 61.76%, respectively, but the accuracy is rather low given the high insertion rate (order of 20%).

7. STORY/FILLER AND SPEAKER ROLE IDENTIFICATION

Errors in filler detection and / or anchor detection have been mentioned in a previous section as the reason for most segmentation problems. In this section, we try to extract some cues about the structure of TV broadcast news programs that may be useful to improve the segmentation and detection tasks. The filler detection will allow us to isolate the fillers segments (such as headlines) where the acoustic background conditions are mostly responsible for erroneous transcriptions. The identification of the speaker role (anchor, journalist or a speaker interviewed) will hopefully allow us better identify the transcripts segments related to the anchors, which are responsible for the beginning of every news story. We started this effort of discovering relevant cues using manually transcribed data. Hence, instead of using the ALERT topic detection corpus, we used the speech recognition (SR) corpus collected in the same project for retraining acoustic and language models. The main reason for using this manually transcribed corpus was not the absence of speech recognition errors, but rather the accompanying metadata including story/filler classification, speaker role, acoustic background, etc. The training subset of this SR corpus includes 30 programs corresponding to 760 stories and 195 fillers and a number of segments classified as *no-trans*. The development subset includes 3 programs, with 87 stories and 16 fillers. On the whole, this subset includes 533 consecutive segments, each characterized by a different speaker role, of which only 18 correspond to fillers. It is important to notice that if all segments in this subset were classified as stories, we would obtain a correction rate of 96.62%. Likewise, if all speakers were classified as anchors (the most frequent classification), we would obtain a correction rate of 42.64%.

Following the type of method described in [6], we decided to train 2 different CARTs [8], one for story/filler classification and another for speaker role classification. Since these CARTs were trained with transcriptions and metadata that were manually produced, our goal with these experiments was simply to know what would be the best possible performance of such decision trees.

Both CARTs were built using characteristics such as overlap between speakers, segment duration, proportion of duration of clean/noise/music parts, proportion of context words, number of repetitions and filled pauses, proportion of questions, and language models for each of the three speaker roles. Some of these characteristics refer not only to the values obtained in the present segment, but also to the values obtained in the preceding one. The story/filler CART used the speaker role as one of the characteristics and vice-versa. The story/filler CART yielded a correction rate of 98.32% and the speaker role one yielded 83.71%.

A very recent experiment was done using a CART trained with the manual annotations of the SR training set corpus to classify each transcript segment boundary as "story boundary/non-story-boundary" (SB/NSB). The training subset of this SR corpus includes 1112 boundaries classified as SB and 19073 classified as NSB. The development subset includes 124 boundaries classified as SB and 2315 classified as NSB. The test subset includes

25 boundaries classified as SB and 500 classified as NSB. It is important to notice that if all boundaries in the development and test subsets were classified as non-story - boundaries, we would obtain a correction rate of 94.92% and 95.24%, respectively (reference classification).

The CART was built using characteristics such as the speaker role (anchor/non-anchor), segment duration and proportion of duration of clean/noise/music parts. Some of these characteristics refer not only to the values obtained in the segment to the right of the boundary, but also to the segment to the left. The SB/NSB CART yielded a correction rate of 97.21% in the development set. The selection of the best features for the development set improved the correctness to 97.95%. After this tuning step, an experiment with the SR test set corpus yielded a correction rate of 97.33%.

Another experiment was done with the same corpus but now using automatic annotations. Once again, a CART was built using the same characteristics as the above mentioned one. In the development set, the CART yielded a correction rate of 98.14% after selecting the best features. For the SR test set, the correctness rate was 97.82%.

Using the CART built with the manual data in the classification of the automatic annotated segments the results are 97.95% and 97.91% in the development and test corpus. The above results are summarized in Table 3.

Table 3. Correctness rate in the story boundary/non-boundary classification (SR test set.).

	Manual annotations	Automatic annotations
Reference	95.24%	98.00%
CART manual	97.33%	97.91%
CART auto	-	97.82%

Table 4. Correctness rate for each category in the story boundary and non-boundary classification (SR test set).

	Manual annotations		Automatic annotations	
	SB	NSB	SB	NSB
Reference	0%	100%	0%	100%
CART manual	60.00%	99.02%	34.78%	99.20%
CART auto	-	-	39.13%	99.02%

The discriminated results presented in Table 4, however, give us more insight into the problem. The CART trained with automatically annotated data shows a better performance in the story boundary detection task than the one trained with the manual annotations over the same material. The most important conclusion that we can derive when comparing the correctness results with the ones obtained with a reference procedure is probably the need to find more characteristics to improve the identification of the story boundary segments and not so much the non-boundary ones.

8. CONCLUSIONS AND FUTURE WORK

This paper presented a topic segmentation and detection system for performing the indexation of broadcast news stories that have been automatically transcribed. Despite the limitations described in the paper, the complete system is already in its test phase at RTP and at INESC ID.

Our current work is aimed at improving the story segmentation method. Our preliminary experiments with story/filler and speaker role detection based on manually annotated characteristics yielded promising results, but much remains to be done in order to tune these trees and automatically extract the intervening characteristics.

As future work in terms of indexation, we intend to collect more data in order to build better bigram language models. In addition, we plan to index stories with multiple topics associated with their confidence values. This will enable us to allocate more than one topic per story, which is indeed the situation of 39% of the Topic Detection Corpus.

9. REFERENCES

[1] Fiscus, J., Doddington, G., Garofolo, J., Martin, A., "NIST'S 1998 Topic Detection and Tracking Evaluation (TDT2)". In Proc. DARPA Broadcast News Workshop, Feb. 1999.

[2] Alexander Gelbukh, Grigori Sidorov and Adolfo Guzmán-Arenas: Document Indexing With a Concept Hierarchy. In: New Developments in Digital Libraries. Proceedings of the 1st International Workshop on New Developments in Digital Libraries (NDDL - 2001). ICEIS PRESS, Setúbal, 2001.

[3] H. Meinedo, N. Souto, J. Neto: Speech Recognition of Broadcast News for the European Portuguese language. Proceedings ASRU'2001 - IEEE Automatic Speech Recognition and Understanding Workshop, Madonna di Campiglio, Italy, December 2001.

[4] C. Hagège: SMORPH: um analisador/gerador morfológico para o português., Lisboa, Portugal, 1997.

[5] NIST Speech Group: The 2001 Topic Detection and Tracking (TDT2001) Task Definition and Evaluation Plan, 15 November 2002.

[6] R. Barzilay, M. Collins, J. Hirschberg, and S. Whittaker: The Rules Behind Roles: Identifying Speaker Role in Radio Broadcast. In Proc. AAAI 2000.

[7] Clarkson, P., Rosenfeld, R., "Statistical Language Modeling using the CMU-Cambridge Toolkit", in Proc. EUROSPEECH 97, Rhodes, Greece, 1997.
[8] P. Taylor R. Caley, A. Black, S. King, "Edinburgh Speech Tools Library", System Documentation Edition 1.2, 15th June 1999.
[9] Yamron, J. P., Carp, I., Gillick, L., Lowe, S., "A Hidden Markov Model Approach to Text Segmentation and Event Tracking", in Proceedings of ICASSP-98, Seattle, May 1998.