Topic Indexing of TV Broadcast News Programs

Rui Amaral[1] and Isabel Trancoso[2]

[1] EST-Setúbal – Inst. Politécnico Setúbal / INESC-ID Lisboa
Rua Vale de Chaves, Estefanilha
2914-508 Setúbal, Portugal
ramaral@est.ips.pt

[2] IST/INESCD-ID Lisboa
Rua Alves Redol
1000-029 Lisboa, Portugal
Isabel.Trancoso@inesc-id.pt
http://l2f.inesc-id.pt

Abstract. This paper describes a topic segmentation and indexation system for TV broadcast news programs spoken in European Portuguese. The system is integrated in an alert system for selective dissemination of multimedia information developed in the scope of an European Project. The goal of this work is to enhance the retrieval of specific spoken documents 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 (thesaurus), containing 22 main thematic domains, for which Hidden Markov models and topic language models were created. On-going experiments related to multiple topic indexing are also described, where a confidence measure based on the likelihood ratio test is used as the hypothesis test.

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 is the framework for our European Project, 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.

N.J. Mamede et al. (Eds.): PROPOR 2003, LNAI 2721, pp. 219–226, 2003.

This alert message enables a client to find in the System Website the video clips referring to the selected stories. This report concerns only the segmentation and indexing modules of the described system.

The broadcast news corpus and thesaurus used in this work are described in Sect. 2. The following two sections present our segmentation and indexing 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, and describes ongoing work on detection of multiple topics. The report concludes with a discussion of these results and our plans for future research.

2 Topic Detection Corpus Description

This section presents the Topic Detection Corpus (TDC) that was used to develop and test our indexing 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 European 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 [5]. 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 %	Thesaurus level	Descriptors %
1st-level	0.21%	6th-level	3.84%
2nd -level	7.62%	7th-level	0.86%
3rd -level	48.32%	8th-level	0.63%
4th -level	26.47%	9th-level	0.19%
5th -level	11.83%	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 used to identify the story speakers, and not the persons who are the subject of the story.

The topic detection corpus was divided into three 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 Fig. 1 for the three subsets.

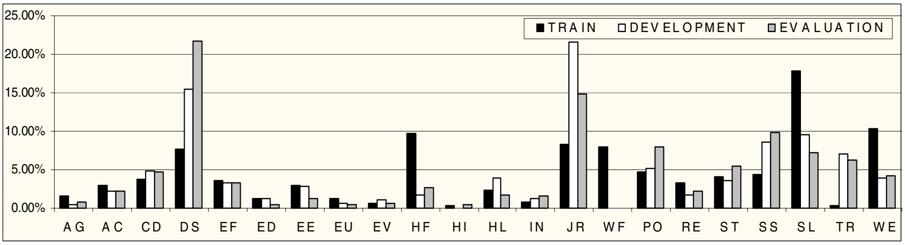

Fig. 1. Topic distribution in the topic detection corpus

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 (tagged with id-numbers). 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 fact, the classification of shorts stories is very "noisy". This type of situation occurs every time a sequence of transcript segments spoken by the anchor is interrupted by one or more transcript segments spoken by someone else. In these cases, the short story segment is merged with the following one. The next stage, following this two-step 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 Indexing

Story indexing is performed in two steps. We start by detecting the most probable story topic, using the automatically transcribed text for each story. Our decoder is based on the HMM (Hidden Markov Model) methodology and the search for the best hypothesis is accomplished with the Viterbi algorithm. The topology used to model each of the 22 thematic domains is single-state HMMs with self-loops, transition probabilities, and bigram language models. For each of the 22 domains, a smoothed bigram model was built with an absolute discount strategy and a cutoff of 8, meaning that bigrams occurring 8 or fewer times are discarded. The referred models built from the training corpus, give the state observation probabilities. The statistics for each domain were computed from automatically transcribed stories with manually placed boundaries. The corresponding text was post-processed in order to remove all function words (527) and lemmatizing the remaining ones. Lemmatization was performed using a subset of the SMORPH dictionary with 97524 entries [6]. Smoothed bigram statistics were then extracted from this processed corpus using the CMU-Cambridge Statistical Language Modeling Toolkit v2 [3].

In the second step, we find for the detected domain all the second and third level descriptors that are relevant for indexing the story. To do 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 [4]. 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 indexing to these upper levels was made taking into account the ALERT project goals and the data sparseness at the thesaurus lower levels.

Ongoing work related to the topic indexing, concerns the multiple topic indexing of the broadcast news stories. This procedure allows a more realistic indexing since the great majority of the news stories covers more than on topic domains (39% of the Topic Detection Corpus). To produce the multiple indexing we used unigram topic models and a confidence measure based on the likelihood ratio test (*LLR*). The comparison score is the log likelihood ratio between the topic likelihood $p(W/T_i)$ and the non-topic likelihood $p(W/\overline{T})$.

$$LLR = \log\left(\frac{p(W/T_i)}{p(W/T)}\right)$$

where W is the words sequence (story) and T_i a specific topic. The output of the hypothesis test is a score. When the comparison score is higher than a certain threshold the hypothesis is accepted, otherwise is rejected.

For each of the 22 topic domains a non-topic language model was created using all the training stories manually classified as belonging to other topics and not related to the topic in question. A confidence measure is computed using topic and non-topic probabilities. The detection of any topic in a story occurs every time the correspondent confidence level is above a predefined threshold. The threshold is different for each topic in order to account for the differences in the modeling quality of the topics, and the values computed using the development corpus were evaluated with the accuracy metric.

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 [7]. 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:

$$(C_{Seg})_{\text{Norm}} = C_{Seg} / \min(C_{Miss} \text{ x } P_{Target}, C_{FA} \text{ x } P_{Non\text{-}Target})$$

where

$$C_{Seg} = C_{Miss} \text{ x } P_{Miss} \text{ x } P_{Target} + C_{FA} \text{ x } P_{FA} \text{ x } P_{Non\text{-}Target}$$

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 [7] (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, a first 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, taking into account the fact that many stories were manually indexed with more than one topic, 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 using a single topic per story. The correctness achieved in the evaluation corpus using our bigram model was 73.80%. Figure 2 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.

The rightmost column of the matrix indicates the number of stories accounted for. By observing this matrix, 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. Examples are: "defense and security" which is confused in 22% of the cases with "society and sociology" (32% of "defense and security" stories were also classified as "society and sociology" stories in the training corpus), and "economy and finance" which is confused in 17% of the cases with "political organization" (16% of "economy and finance" stories were also classified as " political organization " stories in the training corpus).

	AC	AG	CDI	DS	ED	EE	EF	EU	EV	HF	HL	HI	IN	JR	PO	RE	SL	SS	ST	TR	WE	WF	TOTAL
AC			25%								25%				25%			25%					4
AG																		100%					1
CDI			**33%**	67%																			3
DS				**70%**										3%	3%			22%		2%			63
ED					**25%**					25%					25%			25%					4
EE	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0
EF							**67%**							8%	17%						8%		12
EU	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0
EV	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0
HF										**100%**													1
HL																	100%						1
HI	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0
IN	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0
JR				8%										**65%**	15%			8%			4%		26
PO			3%	7%			10%			3%	3%				**69%**			3%					29
RE	100%																						1
SL			1%														**98%**	1%					93
SS			10%	10%										20%				**60%**					10
ST											14%							14%	**43%**	29%			7
TR																		14%		**71%**	14%		7
WE							33%							33%	33%								3
WF																						**100%**	27

Fig. 2. Confusion matrix for a subset of the Evaluation Corpus

This experiment was also repeated using unigram topic models, yielding a correctness value of 73.53%. The proximity of the results indicates that the amount of training data is not enough to build robust bigram models.

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%). It is important to notice that this evaluation was performed only on those stories whose top level domain was correctly identified. Given the nature of the algorithm, the descriptor search is restricted to a specific domain identified in an earlier stage of the decoding process.

Our last experiments were aimed at evaluating the decoder performance using multiple topics. The accuracy and correctness results obtained were 92% and 94%, respectively. The results should be compared with the baseline result of 90% of accuracy that would be achieved if the topic decoder misses the detection of all the topic targets in all stories.

Another experiment was done by building new models for some topics, using only the stories that were classified as belonging to that single topic. This modeling approach is only possible for those topics for which we have a considerable amount of single-topic stories. Once again, new thresholds were calculated for each topic and the indexation experiment yielded an accuracy of 91.6% and a correctness of 93.0%.

Comparing the results of both experiments, we conclude that the first experiment yielded the best performance although they were quite close. The second experiment results were expected given the considerable reduction on the amount of the training data used to build the unigram statistic for some topics.

7 Conclusions and Future Work

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

Our current work is aimed at improving the story segmentation method. We intend to explore some information related to the speaker role inside the news programs. The anchors usually introduce stories and conduct the news program. Journalists usually develop the introduced stories where guests can appear in an interview sequence. We believe that the knowledge of the news program structure will enhance the precision of the segmentation task.

As future work in terms of indexing, we also intend to collect more data in order to build better bigram language models, because the ones used in this work were built using a high cutoff value. In addition, we intent to continue the ongoing work with the multiple topic indexing because this is indeed the situation of 39% of the Topic Detection Corpus.

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. 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.
3. Clarkson, P., Rosenfeld, R., "Statistical Language Modeling using the CMU-Cambridge Toolkit", in Proc. EUROSPEECH 97, Rhodes, Greece, 1997.
4. 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.
5. 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.
6. C. Hagège: SMORPH: um analisador/gerador morfológico para o português., Lisboa, Portugal, 1997.
7. NIST Speech Group: The 2001 Topic Detection and Tracking (TDT2001) Task Definition and Evaluation Plan. ftp://jaguar.ncsl.nist.gov//tdt/tdt2001/evalplans/TDT01.Eval.Plan.v1.2.ps, 15 November 2002.
8. Ng, K., "Survey of Approaches to Information Retrieval of Speech Messages" Technical report, Spoken Language Systems Group, MIT, February 1996.