# Indexing 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 (stories) 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. The broadcast news corpus that is the basis for this work was collected for European Portuguese in the scope of the European Project ALERT.

## 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 the 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 refer-

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

ring to the selected stories. This paper concerns only the segmentation and indexation modules of the Alert system.

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 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*).

### 2.1 Broadcast News Corpus

Collected in the scope of the ALERT project over a period of 9 months (February - October 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 [1]. 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 manual segmentation also identified the so called filler segments, containing either headlines or short story descriptions presented to draw the audience attention to stories that will be presented later in the program. Filler segments were not indexed.

### 2.2 Thematic, Geographic and Onomastic Thesaurus.

Manual indexation was done using a thematic, geographic and onomastic (names of persons and organizations) thesaurus by RTP trained annotators.
The Thematic Thesaurus contains 21 hierarchical concept trees whose domains 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).

The Thematic Thesaurus contains 7781 descriptors and 1615 non-descriptors whose relations are established via:

- hierarchical relations of the type general term (GT) and specific term (ST);
- associative relations such as related term (RT);
- equivalence relations such as use for (UF) or use in case (UIC).

Besides the above relations, there is a tag called explanation node (NE), which gives some contextual indications for the use of the corresponding descriptor.

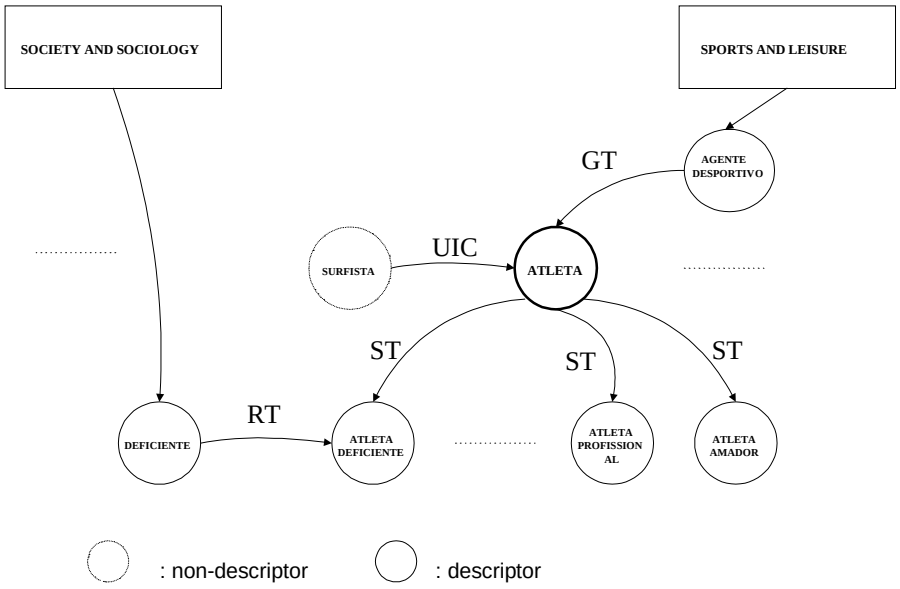


**Fig. 1.** Hierarchical concept tree - Thematic Thesaurus

Figure 1 illustrates the thesaurus structure. Let us focus our attention in the thesaurus descriptor *atleta* (athlete). This descriptor has a general term called *agente desportivo* (sports agent), which is its immediate upper node of the tree, and belongs to the Sports and Pleasure thematic domain. The descriptor *atleta* has at least three specific terms, namely: *atleta deficiente*, *atleta amador* and *atleta profissional* (handicapped, amateur and professional). That means that any of these three terms have a higher degree of specification than the upper node for the current domain. In the example, we can also see the non-descriptor term *surfista* (surfer) that when present should be replaced by the thesaurus descriptor *atleta*. This descriptor has an explanation note that indicates the situations where the descriptor should be used. The descriptor *atleta deficiente* has at least one related term called *deficiente* (handicapped) meaning that there is at least one related descriptor in another thematic domain tree (Society and Sociology).

The distribution of the descriptors among the several levels is represented in table 1.

**Table 1.** Descriptors 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 and person names. These entries are used to identify the story speakers, and not the persons who are the subject of the story.

### 2.3 Training, Development and Evaluation Subsets

The topic detection corpus was divided into three subsets for training, development and evaluation purposes. The corpus division accounted for the need of enough material to train the statistical approaches, as well the time span between the training and the evaluation stories to reflect the real field conditions.

The training corpus was collected from March to mid August 2001. It includes 85 programs, corresponding to 2451 report segments and 530 fillers. The report segments involve 6073 stories with 312 words each, on average. Very frequently, a report segment is classified into more than one topic (for instance, Sports and Leisure and Society and Sociology, as in the example above). Such report segments will originate multiple stories, which justifies the difference between the number of report segments and stories. In the above case, a single report segment will originate one story used for building the SL language model and another for building the SS language model. The distribution of the thematic domains among the story programs is shown in figure 2.

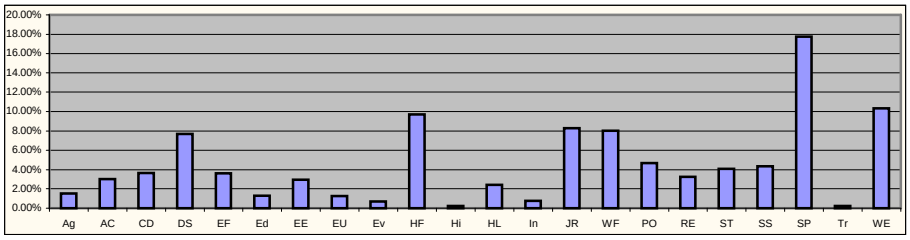


**Fig. 2.** Topic representation in the training corpus.

The development corpus was collected in September 2001 and includes 21 programs, corresponding to 699 report segments and 144 fillers. The report segments involve 1172 stories, whose thematic distribution is shown in figure 3.

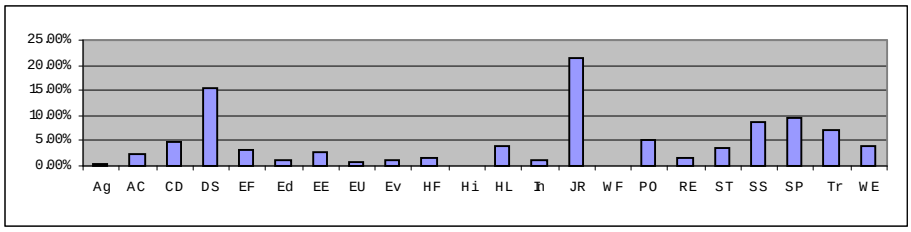


**Fig. 3.** Topic representation in the development corpus.

The evaluation corpus was collected in October 2001. It includes 27 programs corresponding to 871 report segments, and 134 fillers. The segments involve 1528 stories, whose thematic distribution is shown in Figure 4.

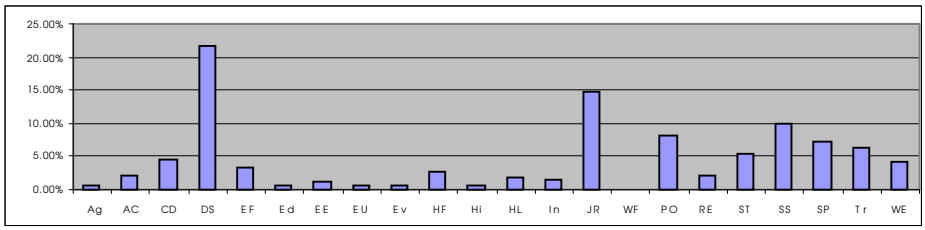


**Fig. 4.** Topic representation in the evaluation corpus.

## 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 fre-

quent 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 unknown. 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 Indexation

Story indexation 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 [2]. The topology used to model each of the 22 thematic domains is single-state HMMs with self-loops, transition probabilities, and bigram language models [3]. For each of the 22 domains, a smoothed bigram model was built with an absolute discount strategy and a cutoff of 8 [3], 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 [4]. Smoothed bigram statistics were then extracted from this processed corpus using the CMU-Cambridge Statistical Language Modeling Toolkit v2 [5].

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 [6]. 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 simple segmentation algorithm, we adopted the metric used in the 2001 Topic Detection and Tracking (TDT 2001) benchmark NIST evaluation [7]. In this Evaluation Plan, the evaluation performance is defined in terms of probability of miss and false alarm errors (PMiss and PFA). A miss is considered when the algorithm fails to identify an existing boundary. A false alarm occurs when the algorithm outputs a non-existing boundary, according to the reference boundaries. To evaluate the outputted boundaries produced by the algorithm, an evaluation window of 50 words (same size as adopted in TDT2001 [7]) was used, and for each window translation, a judgment was done according to table 2.

**Table 2.** Segmentation judgement for each window translation

| 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 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 [7] (1 and 0.3, respectively), we achieved a normalized value for the segmentation cost of 0.835 for a $P_{Target}$ of 0.8.

The segmentation cost value did not reach 0.9, which was state-of-the-art in TDT2001 for this task. One potential reason for this low value is the post-processing step, in which adjacent story segments are merged if their topic classification is equal. Our next segmentation experiments were hence aimed at studying the influence of the merging criterion in this post-processing stage. In fact, we compared this post-processing stage that was based on merging adjacent stories with the same domain

classification (*1st-level descriptors*), with 3 others: a post-processing stage based on merging adjacent stories with the same second level descriptors (*2nd-level*), a post-processing stage based on merging adjacent stories with the same third level descriptors (*3rd-level*), and no post-processing stage at all (Non). The results are shown in Figure 5.

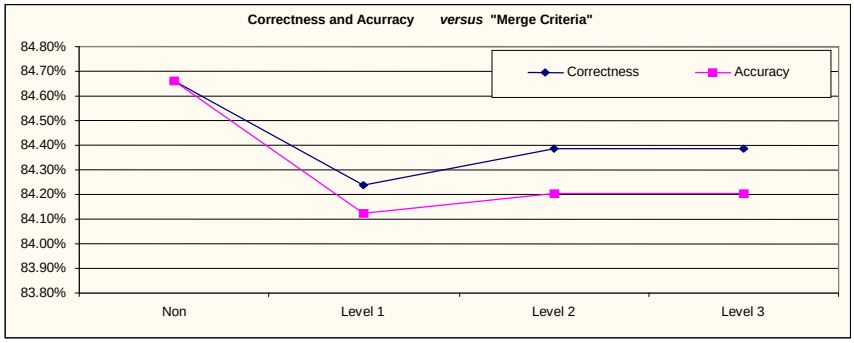


**Fig. 5.** Experiment results using different merge criteria in the segmentation process.

We note that the correction and accuracy values are very close indicating that there is no real advantage in using any merging stage.

Even without this stage, however, the "miss" rate is still very high, which motivated a closer look at the segmentation results. 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. To make the problem even more complex, filler segments are often partially corrupted by music, which makes them difficult to transcribe correctly.

Another reason for boundary deletion is the presence of multiple anchors in a broadcast news program. Some of the broadcast news programs in our corpus had in fact two anchors, one of which was responsible only for the sports stories. Our simple heuristic was based on defining a single anchor as the speaker that appeared more often, independently of the talk duration. Using this criterion, we got only the main anchor and not the sports anchor. The story boundaries introduced by the latter will all be missing. This obviously calls for a more refined anchor detection procedure.

## 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 using our bigram model was 73.80%. 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.

| | 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. 6.** Confusion matrix for a subset of the Evaluation Corpus.

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).

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.

## 7 Conclusions and Future Work

This paper presented a topic segmentation and detection system for performing the indexation of broadcast new stories that have been automatically transcribed. Despite the limitations described in the paper, 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 indexation, 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 also plan to allow the decoder to output all the domain scores associated with confidence values. This procedure will enable us to allocate more than one topic per story. This is indeed the situation of 39% of the Topic Detection Corpus.

## Acknowledgements

The present work is part of Rui Amaral's PhD thesis, initially sponsored by a FCT scholarship (PRAXIS XXI/BD/15844/98). This work was also partially funded by IST-HLT European program project ALERT. INESC-ID Lisboa had support from the POSI program of the "Quadro Comunitário de Apoio III".

## References

1. 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).
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. Xuedong Huang, Alex Acero, Hziao-Wuen Hon : Spoken Language Procesing - A guide to theory, algorithm and system development, Prentice Hall (2001).
4. C. Hagège: SMORPH: um analisador/gerador morfológico para o português., Lisboa, Portugal (1997).
5. Clarkson, P., Rosenfeld, R., "Statistical Language Modeling using the CMU-Cambridge Toolkit", in Proc. EUROSPEECH 97, Rhodes, Greece (1997).
6. 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).
7. NIST Speech Group: The 2001 Topic Detection and Tracking (TDT2001) Task Definition and Evaluation Plan, 15 November (2002).