Evaluation of an alert system for selective dissemination of broadcast news

Isabel Trancoso[*], *João Neto*[*], *Hugo Meinedo*[*], *Rui Amaral*[†]

[*] INESC-ID Lisboa/IST [†] INESC-ID Lisboa/IPS
L^2F - Spoken Language Systems Lab
INESC ID Lisboa, Rua Alves Redol, 9,1000-029 Lisboa, Portugal
{Isabel.Trancoso, Joao.Neto, Hugo.Meinedo, Rui.Amaral}@l2f.inesc-id.pt
http://www.l2f.inesc-id.pt

Abstract

This paper describes the evaluation of the system for selective dissemination of Broadcast News that we developed in the context of the European project ALERT. Each component of the main processing block of our system was evaluated separately, using the ALERT corpus. Likewise, the user interface was also evaluated separately. Besides this modular evaluation which will be briefly mentioned here, as a reference, the system can also be evaluated as a whole, in a field trial from the point of view of a potential user. This is the main topic of this paper. The analysis of the main sources of problems hinted at a large number of issues that must be dealt with in order to improve the performance. In spite of these pending problems, we believe that having a fully operational system is a must for being able to address user needs in the future in this type of service.

1. Introduction

This paper describes the last stage of our involvement in the ALERT European project[1]. The goal of this project was to build a system capable of continuously monitoring a TV channel, and searching inside their news programs for stories that match the profile of a given user. The system may be tuned to automatically detect the start and end of a broadcast news program. Once the start is detected, the system automatically records, transcribes, indexes, summarizes and stores the program. The system then searches in all the user profiles for the ones that fit into the detected topics. If any topic matches the user preferences, an email is send to that user indicating the occurrence and location of one or more stories about the selected topics. This alert message enables a user to follow the links to the video clips referring to the selected stories. The last stage of this project was the field trial of the alert system.

The system includes three main blocks: a *CAPTURE* block, responsible for the capture of each of the programs defined to be monitored, a *PROCESSING* block, responsible for the generation of all the relevant markup information for each program, and a *SERVICE* block, responsible for the user and database management interface. A simple scheme of semaphores is used to control the overall process [1].

In the *CAPTURE* block we have access to a list of programs to monitor and their time schedule (expected starting and ending time). This time information is the input to a capture program that through a direct access to a TV cable network starts the recording of the specified program at the defined time. This is done using a normal TV capture board (Pinnacle PCTV Pro). This capture program generates an MPEG-1 video and audio encoding file, with the audio at 44.1KHz, 16 bits/sample and stereo. When the recording process is finished, an MPEG-1 file is generated, together with the corresponding semaphore signal that will initialize the next block.

In the *PROCESSING* block, the audio stream extracted from the MPEG-1 file is processed through several stages that successively segment, transcribe and index it, compiling the resulting information into an XML file.

The *SERVICE* block deals with the user interface, through a set of web pages, and database management of user profiles and programs. Each time a new program is processed, an XML file is generated and the database is updated. The matching between the new stories and the user profiles generates a list of alerts which are sent to the users through an e-mail service.

The system has been implemented on a network of 3 machines. These are ordinary PCs running Windows 2000 and Linux. The present implementation of the system is focused on demonstrating the usage and features of this system for the 8 o'clock evening news broadcasted by RTP (Radio Televisão Portuguesa). RTP, the Portuguese user partner in the ALERT project, is interested on indexing every story and not only the stories according to certain profiles. To accomplish this indexing task, we based our topic concept in a thematic thesaurus definition that was used at RTP in their manual daily indexing process. This thesaurus follows rules which are generally adopted within EBU (European Broadcast Union) and has an hierarchical structure with 22 thematic areas in the first level. Although each thematic area is subdivided into (up to) 9 lower levels, we implemented only 3 in our system. In fact, it is difficult to represent the knowledge associated with a deeper level of representation due to the relative small training data in our automatic topic indexation method. This structure, complemented with geographic (places) and onomastic (names of persons and organizations) descriptors, makes our topic definition. The use of this hierarchically organized structure makes the Portuguese system significantly different from the one developed by the French [2] and German [3] partners of the project, and from the type of work involved in the TREC SDR Track [4].

Each component of the *PROCESSING* block was evaluated separately, using the ALERT corpus specifically collected for training, development and evaluation purposes during different periods in 2000 and 2001. Likewise, a separate evaluation of the user interface itself also took place, conducted by RTP. The results of the modular evaluation will be briefly mentioned here in Sections 2, for the *PROCESSING* block, and 3, for the user interface. The main topic of this paper is the description of our first field trials, which will be addressed in Section 4.

[1] http://alert.uni-duisburg.de

2. Evaluation of the main components of the PROCESSING block

This section tries to give an overview of the most relevant results of the main components of the *PROCESSING* block. For the sake of space, the technical description of these components will be omitted.

2.1. Audio segmentation, classification and clustering

The first stage of the *PROCESSING* block extracts the audio file from the MPEG-1 stream downsampling it to 16kHz, disregarding, for the time being, any information that could be derived from the video stream. The resulting file is then processed by a Jingle Detector, where we select the program's precise start and ending time, and cut the commercial breaks.

The new audio file containing only the relevant parts of the program is then fed through an Audio Segmentation module [5]. This module segments the audio file into homogeneous regions according to background conditions, speaker gender and special speaker id (anchors).

The audio segmentation module achieved a miss ratio of 14% in terms of segment boundary detection and an insertion ratio of 18%. Each segment is then classified into speech/non-speech, achieving an error rate of 4.4% for tagging speech segments as non-speech. Gender classification is used next as a mean to improve speaker clustering. The misclassification rate was 7.1%. Background classification turned out to be a very hard task, namely because there are many segments in the training material with music plus noise. Segments that were labeled as containing speech are then divided into sentences by an energy endpoint detector, that crudely assumes that a speech pause will correspond to an end-of-sentence boundary. Unfortunately, the news reporters and news anchors do not always make a breath pause at the end of sentences. Presently, this is the major source for incorrect sentence boundaries, causing a significant degradation at the output of the recognizer [6].

The next module of this block tries to cluster all speech segments that were produced by the same speaker. The clusters can then be used for an acoustic model adaptation in order to improve the speech recognition rate, namely for frequent speakers, such as anchors. Speaker cluster information can also be used by topic detection and story segmentation algorithms to determine speaker roles inside the news show. Our method achieves a cluster purity greater than 97%. The anchor identification module that follows this clustering stage includes speaker models for the current three anchors of the 8 o'clock news, achieving a deletion rate of 9%, and an insertion rate of 2%.

2.2. Speech recognition

The current version of the AUDIMUS speech recognition system, that corresponds to the second stage of the *PROCESSING* block, has a vocabulary of 57k and was retrained with over 45 hours of BN data [7]. The corresponding OOV rate is 1.4%. For the F0 focus condition (planned speech, no background noise, high bandwidth channel, native speech), the system achieved a word error rate of 14.8%. The average WER for all conditions was 26.5%. These results were obtained at 7.6 xRT on a Pentium III 1GHz computer running Linux with 1Gb RAM.

The fact that our audio segmentation module produces more sentences than it should (when compared with manual segmentation) and usually shorter and more subdivided has a negative impact in terms of language model errors near incorrect sentence boundaries.

2.3. Topic segmentation and indexation

The third stage of the *PROCESSING* block tries to cluster the transcribed segments into stories and assign them one or more topics within the thesaurus [8]. This is done taking into account the characteristic structure of broadcast news programs. They typically consist of a sequence of segments which can either be stories or fillers. Our topic segmentation algorithm is based on a very simple heuristic that assumes that all new stories start with a segment spoken by a speaker identified as anchor. 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, merging them with the following segments. Our algorithm achieved a normalized value for the segmentation cost of 0.84 for a priori target probability of 0.8 (using cost values of miss and false alarms as adopted in TDT2001 [9]).

One of the reasons for boundary deletion is related to filler segments, introduced by the anchor. Using the crude heuristics described above, the boundary mark will be placed at the beginning of the filler region and no additional boundary marks will be placed in the beginning of the following story.

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.

The next stage is topic indexation. To measure the performance of our algorithm, an experiment was done using the stories of the ALERT evaluation corpus with manually placed story boundaries and automatically transcribed texts. We ignored all filler segments. For the first level of the thesaurus, the algorithm achieved 73.8% correctness. For the second and third levels, the results achieved a precision of 76.4% and 61.8%, respectively, but the accuracy is rather low given the high insertion rate (order of 20%).

After this classification stage, a post-processing segmentation step is performed, in order to merge all the adjacent segments classified with the same topics.

One of the most relevant issues in terms of indexation is multiple topics. In fact, most of the stories in the ALERT corpus have been classified into more than one topic. Very recently, we have improved our algorithm in order to assign multiple topics per story [10]. This improvement, however, was not yet in place during the evaluation period. The post-processing strategy to be used with multiple topics still needs to be discussed.

The last stage of the *PROCESSING* block involves the generation of a title and a summary. The current version assumes that the opening sentence of the story is a good approximation for its title, and that the first few sentences provide a good summary. This simple solution turned out satisfactory, in spite of being completely dependent on the story segmentation process.

3. Evaluation of the user interface

On the user interface, there is the possibility to sign up the service, which enables the user to receive alerts on future programs, or to search on the current set of programs for a specific topic. The field trials so far concerned only the first possibility.

When entering for the first time (registration), or when logging into the system, users have access to two different web pages. One page is a form for entering personal information;

another is a form for choosing the topics that define the user profile. The profile definition is based on a thematic indexation with three hierarchical levels, just as used in the *PROCESSING* block. Additionally, a user can further restrict his/her profile definition to the existence of onomastic and geographical information or a free text string. The profile definition results from an AND logic operator on these four kinds of information. For instance, a user may choose the following combination of third-level thematic descriptor AND onomastic descriptor:
(Sports and Leisure (Sports Modality (Sports with Ball))) AND Sporting

A user can simultaneously select a set of topics, by multiple selection in a specific thematic level, or by entering different individual topics. The combination of these topics can be done through an "AND" or an "OR" boolean operator.

The alert email messages include 4 different fields. The first includes the title of the news broadcast, the date, the time the system finished processing and sent the email and a percentage score indicating how well the story matched the chosen profile. E.g.: Telejornal + 2003-03-11 + 00:07:40 + 65%

The second field includes the title of the story and a URL where one could find the corresponding RealVideo stream. The third field of the alert message summarizes its main contents, and the fourth one discriminates the topic categories that were matched in the presented story. The system groups in the same email to the user all the alert messages corresponding to the matched topics for that user for a given broadcast show.

The user interface was evaluated according to a form distributed by RTP in which several aspects such as suitability for the task, self-descriptiveness, controllability, conformity with user expectations, and learnability were scored on a 5-point scale. The worst scores were assigned to self-descriptiveness, an aspect which in fact we had little time to improve.

4. Field Trials

This section starts by describing the contents of the evaluation forms that were distributed among the field trial participants. It then presents our first results and discusses the feedback obtained from these tests.

4.1. Global Evaluation form

The global evaluation forms should be filled independently for each selected topic. The form includes three parts. The first part is supposed to be filled while the evaluator is watching the TV Broadcast. Every time a story is presented about the chosen topic, the evaluator should take down the date of the broadcast and provide a short description of the story. The second part is supposed to be filled after the usual time for receiving alert messages. It includes, for each story actually observed, the following questions:

- Did you receive an alert message? Yes/No
- If yes, classify
  - start -2; -1; 0; +1; +2
  - end -2; -1; 0; +1; +2
  - subdivided into several alert messages? Yes/No
- Observations

The first question indicates if there was a hit or a miss. The 5-level timing scale, used in case of a hit, should indicate whether the system segmented the story correctly (0 for start and 0 for end), started too soon (negative start) or too late (positive start), ended too soon (negative end) or too late (positive end). The 2 levels on each side are subjective indications of more acceptable (+/-1) or less acceptable (+/-2) story boundaries. With this 5-level scale we were not able to catch the situations where a story appears subdivided into several alert messages, so asked the users to mark these occurrences. The third part of the form should also be filled after receiving the alert messages. For each topic, the participant should list all false alarms received (including the date and title of the story).

4.2. Field trial results

The system was initially tested by 4 researchers who were part of the INESC ID Alert team and helped debugging the system. After this pilot phase, all the other researchers of the group were invited to register as potential users. The global evaluation involved only 8 other evaluators, so the results are not yet significant. In fact, having tried not to influence the potential users in their profile definition, we have found that some researchers chose sub-sub-topics which were too rare, which caused too few evaluation forms to be received. The evaluation period started in the beginning of February, but the impossibility of maintaining the system at the original RTP location originated its reinstallation at INESC ID, which caused a significant interruption.

Altogether, the potential users filled short descriptions of 47 stories and reported only 1 false alarm. Table 1 shows our still very preliminary results:

No. stories		47
No. hits (total)		31
No. hits with	Start = -2	2
	Start = -1	5
	Start = 0	19
	Start = +1	4
	Start = +2	1
No. hits with	End = -2	2
	End = -1	4
	End = 0	18
	End = +1	5
	End = +2	2
No. subdivided stories		0

Table 1: Global results.

4.3. Discussion

The analysis of the different sources of error reported by the evaluators already allowed the improvement of different aspects of the system. One of the aspects concerns the performance when only free strings are specified in a profile vs. specifying a thematic domain or sub-domain. As expected, free string matching is more prone to speech recognition errors, specially when involving only a single word that may be erroneously recognized instead of another. Onomastic and Geographic classification, for the same reason, is also currently error prone. It is based on simple word matching, with no attention to context, nor elaborate named entity extraction methods. This can cause, for instance, the detection of the onomastic descriptor *Pena* in a sentence like *vale a pena* (it is worth).

Thematic matching is more robust in this sense. However, the thesaurus classification using only the top levels is not self-

evident for the untrained user. For instance, should a user who chose Science and Technology as a topic consider an alert message about the occurrence of an earthquake as a hit? In order to verify if a top level has been correctly assigned, one may have to search the corresponding lower level descriptors.

A frequent criticism concerned the merging of several news into a single one. In fact, this occurred mainly with the topic *sports with ball* which proved to be very popular. Being a third-level topic, several news about football involving different teams were grouped in the same alert message. This is in accordance with the current limitations of the system to third-level topics, because of the lack of topic annotated material to adequately train lower level models. However, the amount of training material is significantly different from topic to topic, so a different strategy, allowing lower level definition whenever a sufficient amount of training material is available, might be more useful in the long term. In fact, the large miss percentage is mostly due to the lack of training material for certain topics.

When the evaluation was done using only a single topic for each story, errors were much more frequent, as there were frequent oscillations. For instance, the first part of a story on a pedophilia case that is about to go court may be classified under ethics and the second under law. On the other way, false alarms in anchor detection may cause the splitting of a story in those cases where there are topic oscillations along the story.

Missing the anchor detection may have too much negative impact on the system, as this causes several stories to be grouped together. This has called our attention to the fact that this component has been fine tuned using a too small development corpus and also to the fact that we need a more elaborate segmentation strategy.

The fact that the partial evaluation of each component was done using a test corpus that did not significantly differ in time from the training and development corpus, contributed to a certain adequacy of the chosen lexicon and language models that is not observed any longer, almost two years after.

A major world wide event, such as war in Iraq, has also a great impact on the performance. The duration of the news show greatly exceeds the normal recording times, which has caused missing the last part of the broadcast, even when using fairly large tolerance values for the recording times. The percentage of the broadcast that is devoted to this topic is obviously very large. Rather than being classified as a single story, it is typically subdivided into multiple stories on the different aspects of the war at a national and international level, which again shows the difficulty of achieving a good balance between grouping under large topics or subdividing into smaller ones.

5. Concluding remarks

The above evaluation results and subsequent discussion has hinted at a large number of issues that must be urgently dealt with in order to improve the performance of the system.

- Retraining models for additional anchors;
- Expanding and periodically updating the vocabulary;
- Periodically updating language models;
- Improving the topic segmentation strategy using more robust features;
- Retraining topic models with additional topic annotated data, eventually using the feedback from the evaluation forms about the matched stories;
- Improving multiple topics indexation and segmentation;
- Improving the sentence boundary detection strategy;
- Introducing named entity extraction methods to enhance onomastic and geographic indexation;
- Improving summarization techniques.

Despite this large *shopping list*, we would like to stress the fact that having a fully operational system is a must for being able to address user needs in the future in this type of service.

Our small panel of potential users was unanimous in finding such type of system very interesting and useful, specially since they were often too busy to watch the full broadcast and with such a service they had the opportunity of watching only the most interesting parts. In spite of the frequent interruptions of the system, due to the fact that we are actively engaged in its improvement, the reader is invited to try it by registering at http://sagres.inesc-id.pt.

6. Acknowledgments

The authors would like to thank Alexandre Mendes from 4VDO for his support in the development of the ALERT demo. This work was partially funded by IST-HLT European programme project ALERT and by FCT project POSI/33846/PLP/2000. INESC-ID Lisboa had support from the POSI Program of the "Quadro Comunitário de Apoio III". Hugo Meinedo is sponsored by a FCT scholarship (SFRH/BD/6125/2001).

7. References

[1] J. Neto, H. Meinedo, R. Amaral, and I. Trancoso, "A system for selective dissemination of multimedia information resulting from the alert project," in *Proc.MSDR'2003*, Macau, China, April 2003.

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

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

[4] J. Garofolo, G. Auzanne, and E. Voorhees, "The TREC spoken document retrieval track: A success story," in *Proc. RIAO'2000*, Paris, France, April 2000.

[5] H. Meinedo and J. Neto, "Audio segmentation, classification and clustering in a broadcast news task," in *Proc. ICASSP'2003*, Hong Kong, China, April 2003.

[6] ——, "Automatic speech annotation and transcription in a broadcast news task," in *Proc.MSDR'2003*, Hong Kong, China, April 2003.

[7] H. Meinedo, D. Caseiro, J. Neto, and I. Trancoso, "Audimus+ a broadcast news speech recognition system for the european portuguese language," in *Proc. PROPOR '2003*, Faro, Portugal, Junho 2003.

[8] R. Amaral and I. Trancoso, "Segmentation and indexation of broadcast news," in *Proc. MSDR'2003*, Hong Kong, China, April 2003.

[9] N. S. Group, "The 2001 topic detection and tracking (tdt2001) task definition and evaluation plan," 2001.

[10] R. Amaral and I. Trancoso, "Indexing broadcast news," in *Proc. NDDL'2003 - Third Int. Workshop on New Developments in Digital Libraries*, Angers, France, April 2003.