The development of a Portuguese version of a media watch system

Rui Amaral, Thibault Langlois, Hugo Meinedo
João Neto, Nuno Souto, Isabel Trancoso

INESC ID Lisboa / IST
R. Alves Redol, 9, 1000-029 Lisboa, Portugal
http://l2f.inesc-id.pt/
{ramaral,tl,Hugo.Meinedo,Joao.Neto,Nuno.Souto,Isabel.Trancoso}@inesc-id.pt

Abstract

This paper summarizes the work that has been done concerning the Portuguese language in the scope of the ALERT project during its first year. The media watch system that is the goal of this project comprises many different modules, some of them common among the three languages of the project. This paper concentrates on the definition and collection of the necessary linguistic resources for Portuguese, and the development of the speech recognition, topic and jingle detection modules. The first version of the ALERT demo for European Portuguese is also described.

1. Introduction

This paper is intended as a description of the current status of the work on Broadcast News (BN) for European Portuguese that was done in the scope of the ALERT project. The goal of this project of the IST European programme is to develop an alert system for selective dissemination of multimedia information. The project combines the efforts of three types of partners in three different countries: research partners (University of Duisburg (DE), LIMSI (FR) and INESC ID (PT)), integration partners (VECSYS (FR) and Ergoprocesso (now, 4VDO, in PT)), and data providers (Observer (DE), SECODIP (FR) and RTP (national Broadcast News Television, PT)).

The official start of the ALERT project was in January 2000, although the kick-off was in late February. At the end of this 2.5-year project, we intend to have a common demonstration system capable of identifying specific information in multimedia data consisting of video/audio/text streams, using advanced speech recognition and video processing techniques, as well as automatic topic detection algorithms for the three languages of the project (see [1] and [6] for recent progress on French and German, respectively). The system outputs are a set of alert topics to be sent to end users through e-mail or fax, according to stored user profiles. The alert topic set is automatically generated, including metadata for each topic (internal references of non-digital archival copies such as VHS tape number, etc.), transcripts or summaries out of transcripts, or both and a link to source files (picture, video and sound) as URL-address Number. The system is also able to generate batch retrospective alert files, when needed, due to thematic or temporal backwards searches within archived sequences, for specific user needs.

For the Portuguese partners, this project presents the very first opportunity to work on Broadcast News speech recognition, and simultaneously on topic detection. Our previous work on LVCSR was mainly centered on the AUDIMUS system [5], which was trained and evaluated using the BD-PUBLICO corpus [4], a newspaper corpus of similar characteristics to WSJ. Although the state of the art on ASR for our language is such that we are still very far from being able to apply it to automatic captioning, for instance, we believe it will be good enough to derive automatic transcriptions that can be later used for topic detection in a media monitoring system.

As expected, a very significant part of the work we have done during the first year of this project was devoted to defining and collecting the necessary linguistic resources. This work, which is not yet complete, will be described in Section 2. Section 3 briefly summarizes the work done on speech recognition. Due to its relative importance and length, this work will be fully described in a companion paper [3]. Section 4 will be devoted to our preliminary experiments on topic detection. Section 5 describes the work done on jingle detection, another relevant module in the system. The next section describes the ALERT demo, developed by the Portuguese partners and later adapted to the other languages of the project. The final section summarizes the main contributions and our plans for future research.

2. Linguistic Resources

The goal of the workpackage 2 in the ALERT project is to set up similar linguistic resources for the three languages of the project (French, German and Portuguese). We started by defining and collecting a relatively small Pilot corpus for each of the languages (min. 5 hours). The motivation for this collection was to serve as a testbed for accessing the adequacy of the methodologies for data collection, annotation and distribution. This experience later enabled us to define a set of collection guidelines for a multilingual common structure. Since there were substantial differences between the development stages of this type of systems for each language, it was decided to indicate only the minimum requirements (see Table 1) that were considered necessary to develop these systems.

Since the minimum amount defined for textual data was already available for the three languages prior to the project start (see [4], for Portuguese), there was no need to collect this type of data in the scope of the project. However, the newspaper texts that can be daily extracted from the internet constitute a very powerful resource for improving and keeping up-to-date language models and pronunciation lexica. The consortium therefore plans to take advantage of recent newspaper and newswire texts that may potentially benefit the alert capabilities of the system to be developed.

The guidelines do not cover the type and amount of meta-

Authors in alphabetical order

Table 1: *Multilingual common structure guidelines.*

Type of data	MPEG-1 layer 2 or 3
Audio format	encoding quality at least as good as 16KHz sampling rate, 32 kbps, mono, collected from antenna
Video format	MPEG-1
Annotation	manual or automatic, based on the Transcriber tool following LDC guidelines with English tags
Thematic orientation	news and interviews shows, being a representative sample of the shows that the consortium wants to deal with
Pilot corpus	min. 5 hours (manually annotated)
Speech recognition corpus	min. 50 hours (training); 3 hours (development); 3 hours (evaluation) separate periods for the three sets (automatically annotated)
Topic detection corpus	min. 300 hours (automatically annotated + topic labelled)
Text Corpus	min. 100 millions words

data that is either available from the data providers or can be automatically extracted from the spoken corpora (e.g. speaker gender, speech/non-speech, bandwidth, etc.). Another issue not covered is the database structure for future archiving and search. The fact that each data provider currently has its own metadata structure which best serves its client needs means that there will be no attempt at arriving at a common metadata structure. These issues will be discussed in a later stage of the project.

The corpus collection for Portuguese resulted from the collaboration between the three Portuguese partners. In local meetings, we defined the shows that we want to include in our data collection process from a variety of shows presented in the two main channels of RTP (Channel 1 and Channel 2) and the recording conditions. RTP as data provider was responsible for collecting the data at their premises and for making the annotation. INESC was responsible for defining a schedule for the recordings, helping training the annotator, verifying the annotation and for packaging the data. The board "MPEG Movie Maker Plus" (Optibase)[1] was used for MPEG-1 video and audio encoding. For the pilot corpus, the audio was recorded at 44.1 KHz at 16 bits/sample. The final corpus was recorded at 32 KHz. Both were later downsampled to 16 kHz. For each show we have 4 files: the audio file, the video MPEG-1 file (only for the pilot corpus), the transcription file and a worksheet file with a synthetic summary for each story.

The pilot corpus was recorded during one week in April 2000, amounting to 5.5 hours. The training data of the speech recognition corpus was recorded during October and November of 2000 (61 hours). The development data was recorded in one week in December (8 hours) and the evaluation data during one week in January (6 hours). The topic detection corpus will comprise close to 300 hours of recordings on a daily basis and over a period of 5 months, starting in February 2001.

All the corpora will be (at least) automatically annotated. The pilot corpus was manually transcribed from scratch[2]. The automatic transcriptions of the speech recognition corpus will be manually corrected (the process is not yet complete). The topic detection corpus will be automatically transcribed.

Despite the fact that topic labelling was only mandatory for the topic detection corpus, we decided to topic label the pilot corpus as well. Each show will be manually segmented into stories and each story will be manually classified according to a thematic, geographic and onomastic (names of persons, companies and institutions) thesaurus. Commercial breaks and fillers will be annotated as non-news data.

The thesaurus is currently structured into 20 thematic areas, each of them hierarchically divided, totalling 7,781 descriptors and 1,615 non-descriptors. The structure of this thesaurus follows rules which are generally adopted within EBU (European Broadcast Union). One problem is the fact that the use of this thesaurus by RTP is fairly recent. A large part of RTP's archive since 1956 comprises analytical and synthetic summaries for each story together with descriptors (also thematic, geographic, and onomastic), without any type of hierarchical organization, and, unfortunately, any diacritics. Excluding person names, they exceed 70,000 different descriptors. Queries using this past archive have been made by journalists, when writing their pieces, but no records have been kept. Hence, we do not have yet examples of user profiles as in a real media watch system.

3. Speech recognition

The AUDIMUS LVCSR system which is on the basis for this work is a hybrid system, combining the temporal modeling capabilities of Hidden Markov Models (HMMs) with the pattern discriminative classification capabilities of multilayer perceptrons (MLPs). The system was previously trained on the BD-PUBLICO corpus, with a vocabulary of 27K words, achieving a WER of 14.7% on the corresponding test set.

Due to the large effort involved in manual transcriptions of BN data, this task is not yet complete. Using the manually transcribed data so far, we have built a relatively small training corpus of 13 hours (including the pilot corpus and a subset of the speech recognition corpus) and a test corpus of 82 minutes. Our first experiments in recognition of broadcast news were therefore based on the AUDIMUS system trained on the BD-PUBLICO corpus. As expected, when this system was tested with the BN test corpus, the performance was better for the F0 focus condition (studio, planned, native, clean). Many speech recognition errors could be easily attributed to the fact that the language model was trained using exclusively written newspaper texts.

With the new system trained on the BN corpus, the system achieved 25.9% WER when tested with F0, and 45.6% when tested in all conditions. Given the relatively small size of the BN corpus used for training, the improvements that can be expected merely through training with a large annotated corpus of Broadcast News could not therefore be fully accomplished yet. Full details will be presented in [3], including experiments with larger vocabularies.

4. Topic detection

The preliminary work that we have been doing on topic detection assumed the existence of pre-defined story boundaries. It also assumed the existence of a training corpus manually classi-

[1] http://www.visiblelight.com/mall/moviemaker_plus/

[2] http://www.ldc.upenn.edu/kkarins/convs.paul.main.html

fied into topics. For each topic, a language model is built from the corresponding stories. These topic models use unigrams statistics and are smoothed versions of a global unigram model obtained with all the training text. After topic models construction, the next step concerns the specification of HMM topology for each topic. In this work, each topic is modeled by a unique state with a "self-loop" and a transition probability. The estimation of these probabilities is obtained from the text corpus. Once the HMM model is defined, the last phase corresponds to the search of the best hypothesis, performed with the Viterbi algorithm.

Since the collection of the topic detection corpus is still ongoing, in order to develop and test our approach with a manually topic labelled corpus, we have adopted the BD-PUBLICO corpus [4], for some preliminary experiments. The corpus contains texts from the daily Portuguese newspaper "PÚBLICO" collected almost on a daily basis from 1995 to 1998. A subset of this corpus was recorded by 120 speakers, but since it only comprises isolated sentences, we shall only describe the experiments performed with the text corpus. This corpus is topic-labelled using 9 broad categories: education (Ed), culture (Cl), sports (Sp), economy (Ec), science (Sc), international issues (In), politics (Po), media (Me) and society (So).

For our experiments we selected a set of stories with a maximum of 2,000 words and a minimum of 100, corresponding to the 28-month period from September 95 to December 97. The first 16 months were used to train the system. This training material has about 23 million words spread across 42,000 articles and a 560-word per story average length. The remainder of the corpus was equally divided into development and evaluation sets, each containing about 3 million words. The topic annotation results on the evaluation corpus showed 89.9% correctness and an average of 260 out-of-vocabulary words (OOV). The topic confusion matrix is presented in Table 2. It shows us that the media topic is the harder to identify. It is often confused with the culture and society topics, which is reasonable since they are strongly related. The table also shows that the society topic is the second best choice for 7 of the other 8 topics.

Table 2: *Confusion matrix*

	Ed	Cl	Sp	Ec	Sc	In	Po	Me	So
Ed	88.7	3.2	0.0	0.0	0.0	0.2	2.1	0.0	5.8
Cl	0.4	95.7	0.0	0.6	0.0	0.9	0.2	0.5	1.7
Sp	0.3	0.2	95.2	0.4	0.0	0.0	1.6	0.6	1.9
Ec	0.4	0.3	0.0	88.2	0.0	0.7	3.3	0.0	7.0
Sc	1.9	1.9	0.0	1.5	87.3	0.0	0.4	0.0	6.9
In	0.0	0.9	0.1	0.6	0.0	93.5	1.8	0.3	2.8
Po	0.2	0.7	0.0	2.2	0.0	3.8	88.2	0.0	5.0
Me	1.4	17.0	1.7	4.0	1.1	1.7	2.6	59.5	10.9
So	0.4	2.9	0.1	1.7	3.5	2.7	2.1	0.5	86.0

In order to experiment with the BN corpus, we selected 52 stories from the material already manually transcribed. For 36 of these stories, we could find corresponding texts in the online "PÚBLICO" newspaper which allowed us to extract the corresponding classification into 9 topics. For the remaining 16, we used our best judgment in classifying them.

In tests using the manual transcripts, our topic detection system obtained 61.5% correct results. Using the transcripts automatically generated using the AUDIMUS system trained with the BD-PUBLICO corpus, the score was 59.6%. Using the system trained with the BN corpus, the score reached 63.5%. It is important to notice that 83.3% of the errors may be attributed to the high confusability between the society topic and the others. It is also worth noticing that some of the errors may be attributed to wrong manual classification. In fact, if we exclude the 16 stories which we have classified ourselves, the scores increase significantly (66.7%, 61.1% and 66.7%, respectively). The high score obtained with the transcripts produced by the BN-trained recognizer may be justified by the fact that the manual transcripts show a large OOV rate. In fact, the topic language models were created using the same lexicon as in the AUDIMUS system trained with the BD-PUBLICO corpus.

In parallel, we have also been working on topic segmentation. Following the same paradigms of TDT [2], we have used a clustering based approach, similar to the one described in [7]. It is a two-stage unsupervised clustering based on nearest neighbor search and the Kullback-Leibler distance measure. Results will be described in a future paper.

5. Jingle detection

A subject closely related with topic segmentation is jingle detection. Since the final goal of the project is to build a system that continuously monitors TV channels, performing ASR and topic detection on selected programs, a module for TV program identification and segmentation is obviously important.

Our first attempt was to build a jingle detection module based only on audio signals. A model of each program's jingle is build. The audio stream is processed by the various models that trigger when the beginning of the corresponding program is played.

The audio stream is preprocessed by extracting PLP coefficients using 40ms windows with 10ms overlap. A set of nine successive windows forms a pattern. The models are represented by Radial Basis Function Networks with Gaussian units. The number of units may vary from model to model but, until now, every model has between 25 and 50 units. Each model is trained using a specific training set. The training phase consists in using a technique of gradient descent of the quadratic error with respect to every parameters of the Gaussian units (center, width, and weight).

The training sets are made of around 150,000 patterns equally distributed between positive and negative examples i.e. patterns that do belong to the class and patterns that do not. The negative examples have been chosen from samples of other jingles, conversation, noisy conversations and advertising.

So far, only three models have been built: for the "TELEJORNAL" (evening news on RTP channel), for "RTP ECONOMIA" (a financial news program) and "REMATE" (a sports news program). The test has been performed by processing a typical sequence of pre-recorded TV shows (news, advertising, weather news, football game, etc...)

The preliminary results obtained are described below:

- TELEJORNAL: The jingle is always correctly recognized. Two false alarms of a duration inferior to 4 seconds were detected when processing the jingle of the sports news program.
- RTP ECONOMIA: This model is the one that performs worse. The model was confused by a part of the evening news that showed demonstrators shouting in the streets. The jingle was properly recognized.
- REMATE: This model revealed to be very accurate. The jingle was always recognized, and only one false alarm was observed.

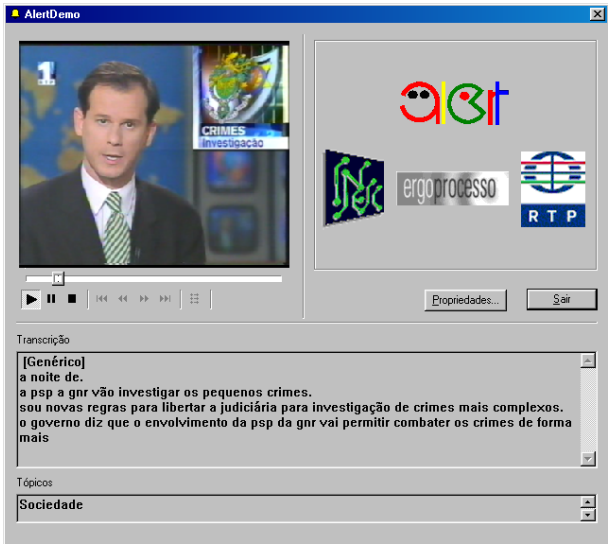

Figure 1: *The first demo of ALERT for European Portuguese.*

Future work will consist in refining the models and build new models for other programs. We will consider the use of video data for the recognition of jingles if that reveals to be necessary. We will apply this technique to the identification of acoustic environments that will serve the ASR part of the system.

6. The ALERT Demo

Together with 4VDO (formerly Ergoprocesso), we designed the first version of the ALERT demo for Portuguese, in which 3 main windows could be simultaneously visualized (1): the video window (with typical Play, Stop, Pause, Rewind and Forward buttons), the Transcription window, containing the output of the recognition module, and the Topic window, containing the ouput of the topic detection module. Since we have not yet developed a fully operative segmentation module, the topic detection module was applied to a sliding window over the audio signal. The window length and updating frequency are adjustable (currently 50s and 10s, respectively). This first demo uses results from the speech recognition and the topic detection modules obtained off-line.

7. Conclusions and Future Work

This paper reports on the work done during the first year within the scope of the ALERT project for European Portuguese. A very large part of the effort was devoted to building the language-specific resources. The existence of a BN corpus, although not yet complete, enabled significant improvements in terms of speech recognition results (described in [3]) and laid the basis for preliminary experiments on topic detection, topic segmentation and jingle detection.

Future work will concentrate on all these areas, obviously, but will also involve joint collaboration with our colleagues working on video processing.

As a follow-up of this project, we have submitted a national proposal for enlarging the present scope to other varieties of Portuguese, namely those spoken in Africa and America. This will also enable us to address a question which is frequently asked to speech researchers in our community: whether a single Portuguese speech recognizer will be able to deal with the different varieties of Portuguese spoken in European, American and African countries.

Given the large differences in terms of acoustic models, pronunciation lexica and language models, we anticipate that is worth building different recognizers for European and American Portuguese. But what about African (or even Asian) countries, in which for many speakers, Portuguese is a second language? We would like to know what degradation can be obtained by testing a recognizer trained for European Portuguese with speech from African countries and, if this degradation justifies it, train accent-specific recognizers. These recognizers will be the building blocks of a multi-accent recognition system with an accent identification module as a front-end.

In order to address the problems of accent identification and accent-specific speech recognition, the proposed project will start by building a multi-accent broadcast news corpus for Portuguese, which in itself, will constitute a major resource not only for training recognition systems, but also for linguistic studies on the different accents.

8. Acknowledgements

The authors would like to thank their ALERT partners, namely João Sequeira, Nuno Guimarães, António Vaz, Alexandre Mendes, Orlando Gonçalves, César Mendes, Hilário Lopes, Sandra Seabra and Conceição Santos. This work was partly funded by the IST project ALERT[3]. INESC ID Lisboa had support from the POSI Program of the "Quadro Comunitário de Apoio III".

9. References

[1] Barras, C., Lamel, L. and Gauvain, J.-L., "Automatic transcription of compressed broadcast audio", Proc. of ICASSP'2001, Salt Lake City , Utah, July 2001.

[2] Fiscus, J., Doddington, G., Garofolo, J., Martin, A., "NIST's 1998 Topic Detection and Tracking Evaluation (TDT2)", in Proceedings of DARPA Broadcast News Workshop, USA, February 1999.

[3] Meinedo, H., Souto, N. and Neto, J., "Speech Recognition of Broadcast News for the European Portuguese Language" submitted to EUROSPEECH'2001.

[4] Neto, J., Martins, C., Meinedo, H., and Almeida, L., "The Design of a Large Vocabulary Speech Corpus for Portuguese", Proc. EUROSPEECH'97, Rhodes, Greece, September 1997.

[5] Neto, J., Martins, C., and Almeida, L., "A Large Vocabulary Continuous Speech Recognition Hybrid System for the Portuguese Language", Proc. of ICSLP'98, Sydney, Australia, December 1998.

[6] Iurgel, U., Meermeier, R., Eickeler, S. and Rigoll, G., "New approaches to audio- visual segmentation of TV news for automatic topic retrieval", Proc. of ICASSP'2001, Salt Lake City , Utah, July 2001.

[7] Yamron, J. P., Carp, I., Gillick, L., and Lowe, S., "A Hidden Markov Model Approach to Text Segmentation and Event Tracking", Proc. of ICASSP'98, Seattle, May 1998.

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