The ESTER evaluation campaign of rich transcription of French broadcast news

G. Gravier (1), J-F. Bonastre (1), E. Geoffrois (2), S. Galliano (2), K. Mc Tait (3), K. Choukri (3)

(1) Association Francophone de la Communication Parlée
(2) Délégation Générale de l'Armement, Centre Technique d'Arcueil
(3) Evaluations and Language resources Distribution Agency

http://www.afcp-parole.org/ester

Abstract

This paper gives an overview of the evaluation campaign ESTER. The aim of this campaign is to evaluate automatic broadcast news transcriptions systems for the French language. The evaluation tasks are divided into three main categories: orthographic transcription, event detection and tracking (*e.g.* speech vs. music, speaker tracking), and information extraction (*e.g.* named entity detection, topic tracking). Each category is evaluated separately. The paper gives detail about the tasks to be performed and the corpus, with a particular emphasis on the manually transcribed reference transcription.

1. Introduction

Objective evaluation of performances in the fields of speech and natural language processing is a major issue in scientific research and technology development. It is however a difficult task as it requires crucial resources, usually manually validated, whose production is hardly accessible to a single laboratory. Moreover, comparing performances can solely be done on a common ground, *i.e.* using standard databases and evaluation metrics.

In the United States of America, a long lasting tradition of evaluation campaigns on speech and natural language technologies permitted the development of large annotated corpora and of well defined evaluation paradigms. Evaluation campaigns organized by NIST and DARPA on automatic transcription (HUB 4, 1998; HUB 4, 1999; RT, 2003), topic retrieval (Wayne, 2000) and named entity detection (ACE, 2001), as well as the NIST campaign on speaker recognition (Martin and Przybocki, 2001) strongly contributed to foster research in those fields.

As far as the French language is concerned, a first wave of evaluation campaigns had been initiated by AUPELF in the nineties. In particular, this effort resulted in a first evaluation campaign on automatic transcription of read speech (Dolmazon et al., 1997). The ESTER campaign[1] is a part of this ongoing effort for developing evaluation campaigns, corpora and evaluation paradigms for the French language. The campaign, organized jointly by the French speaking Speech Communication Association (AFCP), the Ministry of Defense (DGA/CTA) and the Evaluations and Language resources Distribution Agency (ELDA), is part of the EVALDA project dedicated to evaluations on language technologies in the French language[2] . ESTER focuses on the evaluation of rich transcription and indexing of radio broadcasts in French language[3]. The rather recent notion of rich transcription consists in enriching the orthographic transcription with additional information such as thematic indices, speaker turns, or sectioning. The choice for this task was dictated by three main considerations. First, it is a logical progression with respect to the previous AUPELF campaign on read speech transcription. Second, the tasks considered offer a strong application potential. Finally, it complements the NIST Rich Transcription campaign on the English, Arabic and Chinese languages.

This paper describes the goals and organization of the campaign, the tasks considered and the corpora developed in the framework of the evaluation.

2. About the campaign

This section first describes the scientific goals of the campaign before giving details on its implementation.

2.1. Objectives

The objectives of the ESTER campaign are multiple. The first goal is to promote an evaluation environment around speech processing in the French language by setting up a durable evaluation framework. The second one is to develop resources for evaluation on broadcast news material. These resources and related information are meant to be made available to as many laboratories as possible.

From a scientific point of view, an unbiased evaluation of rich transcription system performances is expected from the organization of such a campaign. We also hope to federate research efforts by encouraging laboratories to share information and to collaborate. Workshops are organized throughout the campaign to meet this goal. The expected consequence of all this is a global improvement of transcription performances and new indexing approaches for broadcast news in the French language.

As mentioned previously, one of the objective of this first broadcast news evaluation campaign for the French language is to make available a large annotated corpus for the tasks considered. This corpus, described in more details in section 4., is the main element of the evaluation resource set which will be made available to the scientific community when the campaign is completed, in order to promote

[1] ESTER is the French acronym for Evaluation of Radio Broadcast Rich Transcription System.

[2] The EVALDA project is sponsored by the national program Technolangue.

[3] As of now, evaluation is limited to radio broadcast news.

research activities in this field. Moreover, one of the objective is to distribute the corpus at very low cost for research purposes.

2.2. Implementation

The ESTER campaign is divided into two phases. Phase 1 is a pilot evaluation on a small subset of the final corpus while Phase 2 corresponds to the evaluation campaign itself. Each phase is followed by a workshop.

The Phase 1 pilot evaluation, started in June 2003 and completed in January 2004, aimed at validating and improving the evaluation paradigms and metrics from the participating sites feedback. About ten sites, academic and industrial laboratories, participated with various implication levels. Only transcription and segmentation tasks were implemented in Phase 1 (see section 3. for a detailed description of the tasks). For the transcription task, five sites returned results with word error rates ranging from 10 to 50 percent.

For most sites, the pilot study consisted in getting acquainted with the broadcast news transcription task and in developing a system. Indeed, many sites had a previous background in read speech transcription, but very few had a background in planned and spontaneous speech, mainly because of the lack of available transcribed corpora.

Phase 2 evaluation will be conducted on a larger corpus, for training as for testing. This corpus is described in details in section 4.. This phase will enable to measure progress made with respect to phase 1 due to the availability of a larger corpus and of a structured evaluation environment (*i.e.* well defined paradigms, relatively short time-frame, cooperation and competition between sites). The Phase 2 development stage officially started in March 2004.

Participation in the ESTER campaign is opened to all interested participants on a voluntary basis. Participation is free and remains possible until the official start of the test phase, scheduled in December 2004. During the campaign, participating sites have access to the entire evaluation resource set. Sites actually participating in the final test stage, *i.e.* sites submitting results, will be allowed to keep all the data at no additional cost for research purposes. Evaluation data will be made available by ELRA/ELDA to non participating sites shortly after the end of the test phase via different licenses, ranging from the 'evaluation package' meant to reproduce the evaluation to the unlimited use of the data. A low cost package will be proposed to enable academic laboratories to work on such data.

3. Evaluated tasks

The ESTER evaluation implements three tasks, namely transcription (T), segmentation (S) and information extraction (E). The two first tasks constitutes the core of the campaign while the 'information extraction' task is more prospective. Each task is in turn divided into categories, summarized in table 1.

Though not independent in practice, each task is evaluated separately with the adequate paradigm, in order to best characterize the various components of a radio broadcast indexing system.

category	description
T / TRS	orthographic transcription
T / TTR	real time transcription
S / SES	sound event tracking
S / SRL	speaker diarization
S / SVL	speaker tracking
S / SIL	interactive speaker tracking
E / EN	named entity detection
E / SD	document segmentation
E / ST	topic detection and tracking
E / QR	information retrieval (question answering)

Table 1: Tasks and categories evaluated in the campaign

3.1. Transcription

The transcription task is the classical task which consists in producing the (normalized) orthographic transcription from the waveform. This task is evaluated in terms of word error rates. Systems are divided into two categories: systems operating in real-time or less (TTR) and other systems (TRS).

The use of resources other than the distributed ones is authorized, provided the additional resources are prior to the test corpus (prior to March 2003). Participants using additional data are encouraged to submit contrastive results solely based on the official training data.

3.2. Segmentation

The segmentation task groups together categories aiming at detecting, tracking and grouping together audio events, priorly known and identified or not. Four categories are considered, namely sound event tracking, speaker diarization, speaker tracking and interactive speaker tracking.

Sound event tracking (SES) consists in detecting portions of the document containing a particular event known beforehand. In this evaluation, sound events considered are speech and music. The task is therefore to identify, on the one hand, parts of the document containing music, whether in the foreground or in the background, and, on the other hand, parts of the document containing speech, possibly with background music.

Speaker diarization (SRL) aims at segmenting documents into speaker turns and to group together portions of the document uttered by the same speaker. Speakers are not known beforehand and identification is not required. Systems must return a segmentation of the document with an eventual arbitrary speaker identifier for each segment.

Speaker tracking (SVL) is somewhat similar to sound event tracking with speakers being the events to track. The task consists in detecting portions of the document that have been uttered by a given speaker known beforehand and for which training data are available before the test stage.

Finally, interactive speaker tracking (SIL) is a variant of the speaker tracking category where a system may ask questions to disambiguate decisions, thus simulating interaction with a human operator. For example, systems may ask whether two segments where uttered by the same speaker or not. Results will be evaluated as a function of the number of questions asked.

Each category results in a segmentation of the document in terms of presence or absence of a particular event, hence the task name. The performance measure for such tasks is the classification error rate, computed on time marks. For the tracking categories (SES, SVL and SIL), the considered measure is a weighted sum of the false acceptance and false rejection rates, relative to the classification rate of a "dummy" system that does not detect anything. A specific performance measure is considered for the diarization task in order to take into account deletions and insertions of speech in addition to speaker substitutions, after optimal matching between true and arbitrary speaker names (see (RT, 2003) for details).

3.3. Information extraction

The information extraction task groups categories which aims at extracting "high level" information useful for indexing or document retrieval purposes, with an application oriented question/answering category. Categories for this task are named entity detection, document segmentation, topic tracking and question answering.

Named entity detection (EN) is the task of detecting in an audio document occurrences of an identified entity. In the framework of ESTER, we limit ourselves to the detection of direct mentions of person, location, organization and event (historical, social, etc) names as well as dates and physical measures (*i.e.* followed by a unit). Indirect mentions are not considered in the scope of the current evaluation campaign. Performance will be evaluated based on the (automatic) transcription by counting the number of (correct) words correctly tagged as named entity after alignment of the automatic transcription with the reference one. Alternate performance measures based on time rather than words will be explored.

Broadcast material is structured in terms of shows, reports and topics, possibly with advertisement between and within shows. The document segmentation (SD) category aims at retrieving such structure from the audio stream. This category is limited to document structure analysis and systems are not required to give any information on the (thematic) content of the different sections. However, we plan to also evaluate in this category systems that cluster together reports on the same topic (without topic identification).

In a way similar to speaker tracking, topic tracking (ST) aims at detecting portions of the document that matches a given topic. We will limit ourselves to broad and general topics in the scope of the current evaluation. Examples of broad topics are 'sport' or 'politics' with corresponding general topics 'volley-ball' and 'Gulf war'.

The last category (QR) is dedicated to the evaluation of complete question answering systems. The goal is to answer a question formulated in natural language. As this is a prospective category, only a few questions will be considered and performances will be evaluated by human experts.

4. Corpora

Three main resources are distributed to participating sites, two of them being created in the framework of the ESTER project. A first resource is the broadcast news corpus which consists of manually transcribed radio broadcast news shows. Text resources are also given for language modeling purposes. A less traditional resource consists of large amount (about 2000h) of non transcribed broadcast news material intended to explore research issues in unsupervised training and adaptation.

4.1. Audio resources

The main resource for the ESTER evaluation is a corpus containing 100h of manually transcribed radio broadcast news shows from various French speaking radio stations. The layout of the corpus is summarized in table 2.

The training and development parts of the corpus contain material from four radio stations, namely Radio France International (RFI), France Inter, France Info and Radio Télévision Marocaine (RTM). The three first stations are French national radio stations while the last one is a Moroccan radio station. Only broadcast news shows were recorded, including advertisements.

The corpus is divided into three separate parts for training, developing and testing respectively. The training part contains 82 hours of shows and the development part 8 hours. The test part contains 10 hours, 2 hours from each of the above mentioned sources plus 2 hours from a different source, unknown to the participating sites[4]and thus labeled 'surprise' in table 2. The unseen source in the test data is meant to evaluate the impact of the knowledge of the document source on performances.

The transcription divides the show into sections that roughly correspond to the development of one of the headline news, with separate (non transcribed) sections for advertisements. Topic indices are associated to sections. Note that the section structure corresponds to the news broadcast structure considered in the document segmentation task. Sections are divided into speaker turns. For each speaker turn, the speaker identity (or possibly the speaker identities for multiple speaker turns), the channel and bandwidth and the detailed orthographic transcription are provided. The orthographic transcription is synchronized with breath pauses and provides information on the pronunciation (for example for acronyms or person names). It also includes non linguistic events, such as lip noises or laughters, and named entity tags. Independently of sections or speaker turns, background events such as the presence of music are also indicated.

Annotations were carried out using Transcriber software and more details on the annotation guidelines can be found in the software documentation[5].

Finally, to encourage works on the use of raw, non transcribed, audio data for unsupervised training and adaptation, the audio corpus contains an additional part of about 2000 hours of non transcribed broadcast news shows. Participating sites willing to use this corpus are required to make available all transcriptions, whether manual or automatic, they may produce, thus enabling to create a non

[4]For obvious reasons, the authors cannot communicate here the name of this 'surprise' source!

[5]http://www.etca.fr/CTA/gip/Projets/Transcriber

source	phase 1		phase 2		
	train/dev	test	train/dev	non-trans	test
France Inter	19h40/2h40	2h40	8h/2h	300h	2h
France Info	–	–	8h/2h	1000h	2h
RFI	11h/2h	2h	8h/2h	500h	2h
RTM	–	–	18h/2h	100h	2h
surprise	–	–	–	–	2h
total	40h		50h	2000h	10h
time frame	1998–2000		2003	2004	2004

Table 2: Content of the training (train), development (dev) and test (test) sets for the two phases of the campaign.

controlled transcribed corpus at very low cost. The resulting corpus will be distributed by ELRA/ELDA at the end of the evaluation campaign.

4.2. Text resources

Two text corpora intended for language modeling are provided. The first one consists of articles from the French newspaper "Le Monde". Articles cover the time period from 1987 to 2003 and contains about 300 million words plus topic tags for each article. The second corpus consists of transcriptions of the European Council debates. This corpus, known as MLCC, contains 5.5 million words. Note that the debates transcription are edited transcriptions, that is elaborated transcriptions which reflects the content of the debates, rather than exact transcriptions of (each word of) the debates. These two corpora are completed with the manual transcription of the audio corpus described above.

Text resources are intended to be used for language modeling in transcription tasks but also as training material for topic characterization related tasks.

4.3. Other resources

Other, unofficial, resources such as grapheme to phoneme conversion softwares, silence detectors are made available by participating sites to other participating sites for sake of convenience. Such resources are listed on the campaign web site and most of them are freely accessible to non participating sites.

Furthermore, most current participants agreed to make available resources derived from their development work such as word graphs and automatic transcriptions of the development and test part of the audio corpus, or phonetic alignments on the training part of the corpus. These resources will be distributed with the audio corpus at the end of the campaign. They will also be made available on the campaign web site.

5. Conclusion

We described the organization of the ESTER evaluation campaign on the rich transcription of French radio broadcast news. The recent completion of the pilot evaluation campaign outlined its success with the participation of most of the actors of the domain in France. This first phase justifies the choices made from the beginning. On the one hand, associating in the organizing committee an independent center to assess performances, a scientific association to take care of scientific questions and a corpus specialist permitted to meet the goals defined. On the other hand, many sites have been participating enthusiastically and actively in spite of the lack of funding for their effort.

In the future, we hope that this logic of ongoing evaluations will help create a strong and dynamic community in the field of spoken document transcription and indexing in the French language and that new techniques will emerge from these evaluations. One of our goal is also to enlarge this community to other speech related fields such as phonetics and linguistics. Making derived resources such as phonetic alignments, word graphs or automatic transcriptions available is a first step toward this goal and the organizing committee welcomes any request or suggestion in this direction.

6. References

ACE, 2001. *ACE 6-Month Meeting*. http://www.nist.gov/speech/tests/ace/phase2/doc/nyu-meeting.htm.

Dolmazon, Jean-Marc, Frédéric Bimbot, Gilles Adda, Marc El-Bèze, J. C. Caërou, Jérôme Zeilinger, and Martine Adda-Decker, 1997. Organisation de la première campagne AUPELF pour l'évaluation des systèmes de dictée vocale. In *Journées Scientifiques et Techniques du Réseau Francophone d'Ingénierie de la Langue de l'AUPELF-UREF*.

HUB 4, 1998. *Broadcast News Transcription and Understanding Workshop*. http://www.nist.gov/speech/publications/darpa98.

HUB 4, 1999. *Broadcast News Workshop*. http://www.nist.gov/speech/publications/darpa99.

Martin, Alvin and Mark Przybocki, 2001. The NIST Speaker Recognition Evaluations: 1996-2001. In *Speaker Odyssey*.

RT, 2003. *Spring 2003 Rich Transcription Workshop*.

Wayne, C., 2000. Multilingual topic detection and tracking: Successful research enabled by corpora and evaluation. In *Language Resources and Evaluation Conference*.