ESTER, une campagne d'évaluation des systèmes d'indexation automatique d'émissions radiophoniques en français.

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

La problématique de l'évaluation objective des performances dans le domaine du traitement automatique de la parole et du langage naturel est à la fois intrinsèquement liée à la recherche scientifique et un enjeu majeur pour les développements technologiques. C'est cependant une problématique complexe qui requiert des efforts importants. En effet, le développement comme l'évaluation de nouvelles techniques nécessitent des ressources importantes dont la production est difficilement accessibles aux laboratoires. De plus, les métriques d'évaluation doivent prendre en compte tous les aspects de la tâche à évaluer tout en restant simples et homogènes pour ne pas nuire à la comparaison des résulats.

Aux Etats-Unis d'Amérique, une longue tradition de campagne d'évaluation dans le domaine des technologies vocales et du traitement du langage naturel a permis de mettre à la disposition de la communauté scientifique des corpus de grande taille et des protocoles d'évaluation fiables. Les campagnes d'évaluation DARPA sur la transcription [1, 2, 4], la détection thématique [5] et la détection d'entités (ACE) ou les campagnes du NIST sur la reconnaissance du locuteur [3] ont fortement contribué à dynamiser la recherche dans ces domaines.

Concernant la langue française, une première vague de campagnes d'évaluation avait été lancée dans les années 1990, à l'initiative de l'AUPELF, notamment concernant la transcription automatique de la parole [6].

La campagne d'Evaluation des Systèmes de Transcription enrichie d'Emissions Radiophoniques (ESTER) s'inscrit dans la continuité de cette logique d'évaluation et de développement de corpus et protocoles. La campagne ESTER est dédiée à la transcription enrichie et à l'indexation de journaux radiophoniques de langue française. La notion de transcription enrichie, relativement récente, consiste à ajouter à la transcription orthographique de la parole des informations complémentaires, comme les tours de parole et/ou le locuteur courant. Les émissions radiophoniques ont été retenues pour trois raisons. D'une part, ce type de données présente une progression en terme de difficulté comparativement aux précédentes campagnes de langue française. D'autre part, les tâches choisies montrent un potentiel intéressant en terme d'applications pratiques. Enfin, ce choix permet une comparaison directe avec la campagne NIST Rich Transcription (RT).

La campagne ESTER est organisée conjointement par l'Association Francophone de la Communication Parlée (AFCP), le Centre Technique d'Arcueil de la Délégation Générale pour l'Armement (DGA/CTA) et ELDA (Evaluations and Language resources Distribution Agency) dans le cadre du projet EVALDA[1].

Les objectifs principaux de cette campagne sont de promouvoir une dynamique de l'évaluation, autour du traitement de la parole de langue française, de mettre en place une structure pérenne d'évaluation et de diffuser le plus largement possible les informations et les ressources concernées par ces évaluations. Sur le plan scientifique, les résultats attendus sont bien évidemment de mesurer objectivement les performances des systèmes de transcription enrichie en français mais également de fédérer les efforts et d'inciter les initiatives de recherche collaborative. Une amélioration générale des performances est attendue, comme une conséquence des points précédents.

Par ailleurs, cette première évaluation a également pour but le développement d'un corpus annoté conséquent adapté à la tâche visée. L'ensemble des ressources nécessaires à l'évaluation, dont ce corpus annoté constitue l'élément essentiel, sera mis à la disposition de la communauté scientifique à l'issue de la campagne d'évaluation (l'accessibilité financière de cette ressource est un des objectifs du projet), permettant ainsi de nombreuses activités de recherche dans ce domaine.

Nous décrivons dans cet article le fonctionnement de la campagne, les tâches envisagées avec les protocoles expérimentaux associés, ainsi que les spécifications du corpus développé.

[1] Projet générique de campagnes d'évaluation francophones autour des technologies du langage, financé par le programme interministériel français Technolangue.

2. Fonctionnement

La campagne d'évaluation ESTER est organisée suivant deux phases : une première phase de mise en place et de réglage, dite phase de « test à blanc », et une phase de test officiel, correspondant à la campagne elle-même. Chaque phase est suivie d'un atelier de travail permettant de faire le point sur le déroulement de la campagne et sur les éléments scientifiques abordés.

2.1. La phase de « test à blanc »

La phase de test à blanc a débutée en juin 2003 pour s'achever en janvier 2004. Cette phase a permis la validation d'une partie des protocoles d'évaluation sur un corpus de taille restreinte. Une dizaine de laboratoires ont participé à cette première phase, avec des degrés d'investissement variés. Pour la majorité de ces participants, la première phase a été l'occasion de découvrir la problématique que constitue un corpus d'émissions radiophoniques. En effet, si de nombreux laboratoires français ou francophones ont pu travailler sur des corpus de parole lue, notamment grâce à la campagne d'évaluation AUPELF, jusqu'à récemment la quasi-absence de corpus de parole en situation réelle transcrite, en français, a pénalisé les recherches dans ce domaine.

La partie principale de la phase de test à blanc a permis aux participants de rendre, en condition d'évaluation (protocoles fixés, planning imposé...) une soumission à une ou plusieurs des tâches ouvertes. Cette phase doit s'achever par un atelier d'analyse des résultats en mars 2004.

2.2. La campagne d'évaluation

La deuxième phase de la campagne constitue la campagne d'évaluation proprement dite. Cette phase repose sur un corpus de taille plus conséquente, aussi bien pour l'apprentissage que pour le test. Le corpus complet est décrit en 4. Cette phase va permettre de mesurer les progrès autorisés par la convergence de deux facteurs : la mise à disposition d'un corpus et l'élément structurant induit par la campagne (des protocoles définis, un planning inscrit dans un laps de temps relativement court, la coopération et l'émulation inter-participants). Le début de la phase 2 est fixé à mars 2004 avec une campagne de test en décembre 2004.

2.3. Participation à ESTER

La participation à ESTER est ouverte à tous les acteurs du domaine, aussi bien académiques qu'industriels, sur la base d'une participation bénévole des laboratoires. Une partie des frais de mission directement induits par la campagne sont cependant pris en charge pour les laboratoires académiques. L'inscription à ESTER reste ouverte jusqu'à la date d'envoi des données de test (novembre 2004) [2].

Pendant la durée de la campagne, les participants inscrits ont accès à l'ensemble des données d'apprentissage et de développement. À l'issue de la campagne, les participants ayant effectivement soumis des résultats conservent le droit d'utiliser et d'exploiter ces données (à des fins de recherche uniquement). Après la campagne, l'ensemble des ressources sera disponible sur le catalogue de ELRA/ELDA. Différentes formules seront proposées aux acteurs extérieurs au projet, depuis la possibilité de recréer les conditions de test et de comparer les performances de leur système par rapport à ceux testés dans la campagne ESTER, jusqu'aux droits d'utilisation non limités des ressources. Une formule intermédiaire dédiée aux laboratoires académiques permettra un accès aux ressources, limité à des fins de recherche, pour un tarif accessible.

Tab. 1: Récapitulatif des tâches et catégories

catégorie	description
T / TRS	transcription orthographique
T / TTR	transcription temps réel
S / SES	suivi d'événements sonores
S / SRL	segmentation et regroupement de locuteurs
S / SVL	suivi de locuteurs
S / SIL	segmentation de locuteurs interactive
E / EN	détection d'entités nommées
E / SD	segmentation thématique de document
E / ST	suivi de thèmes
E / QR	recherche d'information (question/réponse)

3. Description des tâches

La campagne ESTER s'organise autour de trois tâches : la transcription (T), la segmentation (S) et l'extraction d'informations (E). Les deux premières tâches constituent le « noyau dur » de la campagne tandis que la tâche *extraction d'informations* regroupe des thèmes plus prospectifs, qui ne seront évalués que lors de la deuxième phase de la campagne. Chaque tâche est elle-même divisée en catégories avec, pour chaque tâche, une catégorie de systèmes orientée application. Les différentes catégories associées à chacune des tâches sont résumées dans le tableau 1 et détaillées ci-dessous.

Bien que n'étant pas indépendantes dans la pratique, les tâches sont évaluées séparément avec une métrique propre, afin de caractériser au mieux les performances des différents niveaux d'un système complet d'indexation d'émissions radiophoniques.

3.1. Transcription

La tâche de transcription consiste à produire une transcription orthographique du document, à partir du signal audio. Les transcriptions sont évaluées en terme de taux de mots erronés (*Word Error Rate*). Pour cette tâche, on distinguera dans une catégorie à part les systèmes de transcriptions pour lesquels le temps de réponse total du système est limité à une fois le temps réel. Cette dernière catégorie vise à faire émerger de nouvelles méthodes permettant une transcription rapide.

3.2. Segmentation

La tâche de segmentation consiste à détecter, suivre et regrouper des événements sonores, préalablement identifiés ou pas. Cette tâche se divise en quatre catégories.

Le suivi d'événements sonore (SES) vise à détecter dans un document les plages correspondant à un événement

[2] Les modalités d'engagement sont détaillées dans le contrat disponible sur le site Internet `http://www.afcp-parole.org/ester`.

sonore donné, préalablement connu. Dans le cadre de l'évaluation, nous nous limitons aux deux événements parole et musique. La catégorie consiste donc à détecter d'une part les plages contenant de la musique et, d'autre part, les plages contenant de la parole.

La segmentation et regroupement en locuteur (SRL) a pour but de découper le flux audio en tours de parole et de regrouper les plages associées à un même locuteur (sans chercher à l'identifier). Dans cette catégorie, un système retourne donc une segmentation du document spécifiant les plages de silence et un identifiant arbitraire de locuteur pour chaque segment supposé correspondre à un tour de parole.

La catégorie suivi de locuteur (SVL) consiste à détecter les zones du document où un locuteur donné, connu à l'avance (pour lequel on dispose de données d'apprentissage), est présent. Bien que portant également sur le locuteur, cette catégorie est très différente de la catégorie SRL pour laquelle aucune identité n'est fournie. En fait, la catégorie SVL se rapproche dans sa métrique plus de SES, en considérant comme événement sonore un locuteur donné.

Enfin, la segmentation interactive du locuteur (SIL) reprend la catégorie SRL mais se rapproche d'une situation applicative, en simulant les interactions avec un opérateur. Les systèmes pourront, grâce au recours à un oracle, lever certaines ambiguïtés (e.g. répondre à la question : est-ce que ces deux segments proviennent d'un même locuteur ?). Les résultats seront mesurés en fonction des requêtes adressées à l'oracle.

Pour toutes les catégories liées à cette tâche, le résultat est une segmentation du document en terme de présence ou absence d'un événement, d'où la dénomination de la tâche. La métrique associée est le taux d'erreur de classification. Pour les catégories SES et SVL, ce taux est calculé comme la somme pondérée des taux de fausse acceptation et de faux rejet, ces derniers taux étant calculés sur l'ensemble des événements considérés (l'ensemble des locuteurs pour SVL, parole et musique pour SES). Une métrique spécifique est utilisée pour évaluer les systèmes dans la catégorie SRL (et SIL) de manière à prendre en compte l'omission ou la détection abusive des zones de parole en plus des confusions entre locuteurs, après appariement entre les noms de locuteurs et les noms arbitraires fournis par le système.

3.3. Extraction d'information

Cette partie de l'évaluation propose des tâches plus prospectives, proches d'applications réelles, permettant de juger de l'utilité d'une transcription au-delà de la simple mesure du taux d'erreur de mot. Dans ce cadre, plusieurs catégories sont envisagées, permettant l'enrichissement des transcriptions avec des informations de plus haut niveau et la recherche d'information.

La détection d'entités nommées (EN) consiste à détecter dans un document (audio) les occurrences d'entités identifiées. Dans le cadre de ESTER l'évaluation porte sur les noms de personnes, de lieux, d'organisations, d'événement (politique, historique, social, etc.), d'objets ainsi que sur les dates et les valeurs (mesure avec unité). L'évaluation se limite pour l'instant à la détection des occurrences de telles entités et ne porte pas sur la détection de l'ensemble des références (références directes et co-références) à une entité.

Le flux audio provenant des émissions radiophoniques contient des informations structurées par émissions, rubriques et sujets. Le but de la segmentation thématique de document (SD) est de reconstituer cette structure à l'aide d'indices acoustiques et linguistiques, afin de segmenter le flux audio en sections cohérentes, sans pour autant avoir à fournir une indication sur le contenu (thématique) des différentes sections (l'évaluation de méthodes de regroupement visant à grouper les sections traitant d'un même sujet est également envisagée).

L'objectif du suivi thématique (ST) est de détecter les zones du flux sonore où un thème donné est abordé. Le principe est similaire à celui du suivi de locuteur (SVL), la cible étant ici un thème au lieu d'un locuteur. Dans le cadre de la campagne ESTER 2004, le suivi thématique reste limité à des catégories thématiques relativement larges (rugby, guerre du Golfe, etc.).

Une quatrième et dernière catégorie, appelée question/réponse (QR), est envisagée, en lien avec une autre campagne du projet EVALDA appelée EQUER. Elle consiste à retourner non pas un document ou une portion de document pertinent par rapport à une requête mais directement la réponse à la question posée.

4. CORPUS

Trois types de ressources sont fournis dans le cadre du projet ESTER. D'une part, les ressources classiques, soit les ressources acoustiques (corpus de parole transcrite) et textuelles (journaux ou transcriptions approchées de débats officiels). D'autre part, une ressource originale de parole non transcrite est proposé. Il s'agit de corpus de parole sans transcription associé mais en grande quantité (environ 2000h), destiné à explorer la possibilité d'un apprentissage non supervisé.

4.1. Corpus d'émissions transcrites

La campagne se base sur l'utilisation d'un corpus acoustique de 100 heures d'émissions radiophoniques. Ces émissions sont transcrites orthographiquement et incluent les tours de parole avec l'identité des locuteurs, les conditions acoustiques et divers événements tels que la présence de bruits. Dans leur version finale elles seront aussi étiquetée en entités nommées et en thèmes. Ces annotations sont basées sur le logiciel Transcriber[3].

L'organisation du corpus est décrite dans le tableau 2. Il est composé de 4 radios différentes : Radio France International (RFI), France Inter, France Info et la Radio Télévision Marocaine (RTM, en langue française). Il se décompose en 82 heures d'apprentissage, 8 heures de développement et 10 heures de test. Environ la moitié du corpus d'apprentissage a été mis à disposition par la DGA afin de pouvoir démarrer la campagne dès le début du projet. Le reste du corpus est produit dans le cadre du projet, et pris en charge par ELDA en collaboration avec la DGA. Les données d'apprentissages et de développement seront distribuées

[3]Voir http://www.etca.fr/CTA/gip/Projets/Transcriber/ pour plus de détails sur le logiciel et sur les instructions de transcription.

TAB. 2: Répartition des corpus acoustiques

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
période	1998–2000		2003	2004	2004

aux participants au début de la phase 2 (mars 2004), et les données de tests pour la campagne elle-même (novembre 2004).

Le corpus de test, transcrit aussi tardivement que possible, se composera de 2 heures de chacune des radios précédentes, plus 2 heures provenant d'une autre radio « surprise » dont l'identité restera inconnue des participants, pour mesurer si la connaissance préalable des sources influence les performances.

Afin de favoriser les travaux portant sur l'utilisation de données acoustiques brutes (non transcrites) pour l'apprentissage, un corpus d'environ 2000 heures de parole non transcrites sera également fourni pour la 2^{e}phase de la campagne aux participants qui en feront la demande. Ces derniers devront fournir en échange les transcriptions réalisées, automatiques ou manuelles, à l'issue de l'évaluation. La ressource ainsi créée pourra ensuite être distribuée par ELDA, selon les règles du marché.

4.2. Ressources textuelles

En plus des émissions transcrites précédemment citées, deux corpus textuels seront fournis aux participants : le corpus du journal « Le Monde », des années 1987 à 2003 (l'année 2003 sera distribuée dès qu'elle sera disponible), et la transcription de débats du Conseil Européen (corpus MLCC, 5,5 millions de mots). L'ensemble des corpus textuels cités ci-dessus est ou sera disponible via ELDA.

4.3. Autres ressources

L'organisation d'une campagne d'évaluation telle qu'ESTER implique de la définition et la mise à disposition de nombreuses autres ressources, plus difficiles à énumérer ici. La partie la plus visible est constituée par l'ensemble des protocoles d'évaluation des performances et des scripts de mesure associés. D'autres éléments, comme des phonétiseurs, des segmenteurs parole/non parole, etc. sont mis à la disposition de tous par des participants. Ces ressources sont toutes référencées sur le site de la campagne.

5. CONCLUSION

En conclusion, cet article a décrit les principaux aspects de la campagne d'évaluation de systèmes de transcription enrichie d'émissions radio ESTER, dont la première phase est achevée et dont la deuxième commence. Cette première phase a été un succès, avec la participation d'une très large majorité des acteurs académiques du domaine et la participation ou l'intérêt croissant des industriels du secteur.

Cette première phase a également montré la justesse des choix de départ. D'une part, l'association dans le comité d'organisation d'un centre indépendant (DGA, assurant la mesure des performances), d'une société savante représentant la communauté (AFCP, en charge des aspects scientifiques) et d'un spécialiste des corpus (ELDA, garantissant la pérennité des efforts, notamment en terme de corpus) a permis de tenir les délais et les objectifs de départ. D'autre part, malgré ou grâce au bénévolat des participants, les laboratoires sont présents en nombre et participent très dynamiquement à ESTER.

La deuxième phase, avec un planning un peu plus large et la mise à disposition de ressources plus conséquentes, permettra de mesurer les gains, en terme scientifiques, qu'apporte la logique et la logistique d'une campagne d'évaluation telle qu'ESTER.

Enfin, si les objectifs scientifiques de la campagne ESTER sont centrés sur la transcription automatique et enrichie d'émissions radiophoniques, les corpus, les outils et le dynamisme des participants autorisent des collaborations dans d'autres spécialités de la parole et les organisateurs restent ouverts à toute demande allant dans cette direction.

RÉFÉRENCES

[1] DARPA. *Broadcast News Transcription and Understanding Workshop*, 1998. http://www.nist.gov/speech/publications/darpa98.

[2] DARPA. *Broadcast News Workshop*, 1999. http://www.nist.gov/speech/publications/darpa99.

[3] Alvin Martin and Mark Przybocki. The NIST Speaker Recognition Evaluations: 1996-2001. In *2001, A speaker Odyssey*, 2001.

[4] NIST. *Spring 2003 Rich Transcription Workshop*, 2003.

[5] C. Wayne. Multilingual topic detection and tracking: Successful research enabled by corpora and evaluation. In *Language Resources and Evaluation Conference*, pages 1487–1494, 2000.

[6] J.M. Dolmazon, F. Bimbot, G. Adda, M. El-Bèze, J.C. Caërou, J. Zeiliger, M. Adda-Decker, "Organisation de la première campagne AUPELF pour l'évaluation des systèmes de dictée vocale," in *Actes des 1ère Journée Scientifique et Techniques Francil*, pp. 13–18, 1997.