EVALDA : constitution d'une infrastructure d'évaluation pérenne

Kevin McTait, Maria Nava, Khalid Choukri

ELRA/ELDA
55-57 rue Brillat-Savarin – 75013 Paris, France
Tél.: ++33 (0)1 43 13 33 33 - Fax: ++33 (0)1 43 13 33 30
Mél: {mctait,nava,choukri}@elda.fr, - http://www.elda.fr

ABSTRACT

The aim of the EVALDA project is to establish a permanent evaluation infrastructure for both text and speech processing systems for the French language. This infrastructure is designed to assemble several components involved within the context of evaluation in language engineering i.e. an organisational model, logistics, language resources, metrics, methodologies, evaluation protocols, scoring software as well assembling major players within the field (scientific advisory committees, experts, project partners etc). In order to capitalise on the outcome of the project, the resources produced within the campaigns are to be made available to external players, in the form of evaluation packages available via the ELRA/ELDA catalogue, enabling them to reproduce the evaluation campaigns and thus benchmark their systems.

1. INTRODUCTION

Le projet EVALDA a pour objectif la constitution d'une infrastructure d'évaluation des systèmes d'ingénierie linguistique du français, pérenne et permanente, au sein d'ELDA (Evaluation and Language Resource Distribution Agency), ainsi que son exploitation par la mise en oeuvre de plusieurs expérimentations.

L'initiative s'inspire de précédents efforts en évaluation déjà réalisés au niveau national et international, par exemple les campagnes GRACE (Adda et coll. [1]), AUPELF/AUF (Mariani [2]), ARCADE (Véronis & Langlais [3]), Amaryllis (Landi et coll. [4]), TREC/NIST (cf. Voorhees & Buckland [5] pour la plus récente publication de synthèse) et CLEF (cf. Peters [6], pour la dernière campagne).

L'infrastructure d'évaluation doit assembler plusieurs composants réutilisables : organisation, logistique, ressources linguistiques, métriques et outils, expertise technique (comités scientifiques, spécialistes en évaluation, etc.). Elle vise, d'une part, à garantir la possibilité de capitaliser les résultats des expérimentations déjà réalisées et, d'autre part, favoriser la recherche de collaborations et la mise en place de nouvelles actions d'évaluation. Les résultats capitalisés doivent permettre de reproduire à tout moment une expérimentation déjà terminée. A titre d'exemple, ELDA distribue déjà le cédérom avec les données issues de la campagne Amaryllis, qui permet de reproduire une évaluation de moteur de recherche d'informations francophone.

2. MOTIVATION ET OBJECTIFS

Une service d'évaluation permanent peut se justifier à différents niveaux :

- scientifique : l'évaluation permet des avancées technologiques et soutient des technologies innovantes ; des expérimentations permettent de rassembler la communauté scientifique et/ou économique et de favoriser les échanges, elles diminuent les barrières culturelles entre les différents domaines d'application, etc.
- économique : des expérimentations d'évaluation pourraient aider à voir plus clair, en terme d'avantages et d'inconvénients respectifs des outils par rapport à leurs domaines d'application privilégiés, elles favorisent le transfert de technologie entre, d'une part, la recherche et l'industrie et, d'autre part, l'industrie et le marché
- institutionnel : l'activité d'évaluation en France et en Europe est moins homogène qu'aux Etats-Unis et au Japon qui jouissent d'un soutien institutionnel et financier.

C'est pour répondre à ce manque, aussi bien en France qu'en Europe, que le projet d'une infrastructure permanente a été conçu.

S'inspirant des précédents efforts pour offrir un cadre générique à l'évaluation, le projet EVALDA se propose :

- de favoriser la réutilisation des données produites ;
- de travailler sur la langue française et aborder des pistes multilingues ;
- d'initier des nouvelles campagnes d'évaluation qui permettent de développer soit des alternatives à des méthodes/métriques existantes, soit des propositions pour des domaines qui demeurent inexplorés (identification des langues, etc.).

3. STRUCTURE ACTUELLE D'EVALDA

Le projet EVALDA est composée de plusieurs lots, conçus pour être autonomes : d'une part, un projet de mise en place d'une infrastructure générale illustrée en Figure 1, d'autre part, la mise en œuvre de campagnes d'évaluation, qui constituent des lots indépendants.

Dans un souci d'efficacité, d'économie d'échelle et de capitalisation, les tâches d'organisation et de gestion ont été regroupées au maximum.

Les campagnes d'évaluation ont été conçues en toute indépendance. Cependant, la production de ressources communes a été prévue. Actuellement, plusieurs campagnes permettent de mettre en oeuvre l'infrastructure d'évaluation. Ces campagnes portent sur des thèmes variés, constituant un ensemble cohérent, couvrant le domaine de l'oral et de l'écrit.

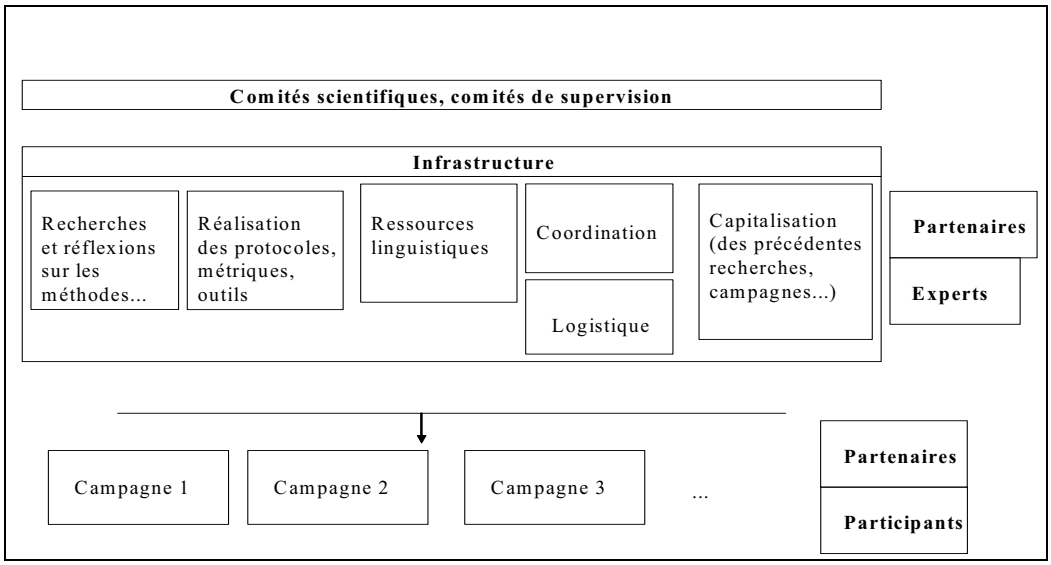

Figure 1 : Organisation de l'infrastructure pour l'évaluation en ingénierie linguistique

Huit campagnes sont actives aujourd'hui, animée par des consortium d'une quinzaine d'acteurs, en moyenne[1] :

ARCADE II : Action de Recherche Concertée sur l'Alignement de Documents et son Evaluation. Le travail se trouve dans la phase d'élaboration d'importants corpus parallèles multilingues : une collection trilingue (français, anglais, allemand ou espagnol), une collection bilingue français-arabe et, pour une deuxième phase prévue en 2005, plusieurs triplets de langues comprenant le français, l'anglais (langue pivot), plus le russe, le grec, le persan, le chinois ou le japonais.

CESART : Campagne d'Evaluation de Systèmes d'Acquisition de Ressources Terminologiques. Les spécifications de l'évaluation – tâches, protocole, métriques et corpus – sont en cours d'élaboration et les ressources linguistiques en cours d'acquisition (corpus de spécialité dans le domaine médical). La campagne propose des tâches orientées vers des applications spécifiques, comme l'acquisition de connaissances ou la construction de thésaurus.

CESTA : Campagne d'Evaluation de Systèmes de Traduction Automatique. De nouveaux protocoles sont en cours d'élaboration (évaluation manuelle et automatique). Les profils des corpus d'apprentissage et de test sont à l'étude, pour des combinaisons de langues comme français-anglais et français-(anglais)-arabe.

Une activité de méta-évaluation est également prévue, avec l'objectif notamment de comparer les résultats de l'évaluation manuelle et automatique.

EASy : Evaluation d'Analyseurs Syntaxiques. Un formalisme d'annotation syntaxique a été élaboré par les partenaires du projet et fait aujourd'hui l'objet d'un consensus. Un corpus de référence, constitué à partir de cinq sources de contenu et de style rédactionnel très variés, est en train d'être annoté syntaxiquement.

EQueR : Evaluation en Questions-Réponses. Deux importantes collections textuelles sont en cours de préparation. Elles constituent deux corpus distincts : un corpus généraliste comprenant des articles et des dépêches de presse et des rapports d'information d'origine institutionnelle sur des sujets très varies ; un deuxième corpus de texte du domaine médical, construit avec la contribution de médecins documentalistes du CISMeF (Catalogue et Index des Sites Médicaux Français, Centre Hospitalier Universitaire de Rouen). Les spécifications de l'évaluation ont également été élaborées dans une première phase du projet.

ESTER : Evaluation des Systèmes de Transcription enrichie d'Emissions Radiophoniques. Une première phase de ce projet vient de se clôturer avec la distribution du premier corpus d'apprentissage et la collecte des résultats d'un test à blanc.

[1] Des pages dédiées à la description générale des campagnes et des participants peuvent être consultées sur le site http://www.elda.fr.

La production d'un corpus supplémentaire d'apprentissage (90 heures d'enregistrement annoté et 2000 heures d'enregistrement brut) vient de se terminer. Le corpus de test (10 heures d'enregistrement) est en cours de préparation.

EvaSy : Evaluation des Synthétiseurs de parole en français. Quatre volets d'évaluation, correspondant à des « maillons » de synthétiseur, sont prévus : traduction graphème-phonème, prosodie et expressivité de la synthèse ; traitement du signal acoustique ; qualité globales des systèmes.
La production de ressources linguistiques annotées est en cours pour le volet graphème-phonème, qui se propose d'analyser en priorité des textes de courriers électroniques ainsi que les entités nommées.

MEDIA : Méthodologie d'Evaluation de la compréhension du DIAlogue hors et en contexte. Un protocole de production d'un corpus de dialogues de référence a été élaboré. Après un important effort de recrutement de locuteurs, un corpus de 1250 dialogues a été enregistré à l'aide d'un système de type « magicien d'Oz ». Parallèlement, les enregistrements sont en cours de transcription et annotation.

Les campagnes suivent majoritairement un protocole de type « boîte noire » et des méthodes d'évaluation quantitatives, en s'inspirant de ce qui a été fait dans des campagnes d'évaluation précédentes.

Néanmoins, des tâches exploratoires originales ont été prévues dans certaines campagnes, par exemple : la détection de ruptures de parallélisme entre corpus de langue éloignées du français (arabe, persan, chinois), dans ARCADE II ; les aspects « utilisateur » de l'indexation dans une tâche de détection d'entités nommées et d'indexation thématique dans ESTER ; l'utilisation d'un corpus spécialisé du domaine médical dans EQueR ; ou encore, un nouveau protocole d'enregistrement d'un corpus de dialogue de référence un utilisant le système « magicien d'Oz » pour MEDIA.

4. ORGANISATION DES CAMPAGNES

La mise en œuvre des campagnes comprend trois phases. La première phase est une phase d'exploration et de mise au point du cadre général d'évaluation. Cette phase prévoit les activités suivantes :

- la réalisation d'études sur l'état de l'art ;
- la définition et la diffusion des tâches d'évaluation, ;
- la mise au point de métriques et d'outils ;
- la spécification ou la validation des ressources adéquates pour les expérimentations envisagées.

La seconde phase est une phase de définition et de mise en place effective de campagnes. Chaque campagne comprend :

- un appel à participation ;
- une phase d'entraînement ;
- une phase de test : résultats accessibles aux seuls participants, publications non réservées ;
- un atelier final.

La troisième phase consiste à faire un bilan de l'expérience et à capitaliser les résultats obtenus. Elle implique les activités suivantes :

- la réalisation de publications ;
- la distribution des paquets d'évaluation comprenant toutes les ressources, protocoles, outils de *scoring* etc. utilisés ou produits pendant la campagne.

La capitalisation des connaissances au sein d'ELDA, en terme d'organisation de campagne doit tout d'abord permettre de ré-initier une campagne d'évaluation à la demande. De plus, ELDA prévoit de distribuer les ressources créées ou enrichies lors des campagnes, sous forme d'un paquet d'évaluation, pour garantir la pérennité des ressources et leur accès à toute la communauté, permettant à tous les acteurs du domaine de reproduire à n'importe quel moment l'évaluation sur une thématique donnée.

5. CONCLUSION

La structuration de l'ensemble des services d'évaluation en une infrastructure permanente, d'un côté, et des campagnes spécifique, de l'autre côté, devrait permettre de tirer parti de la synergie produite par la mise en œuvre parallèle.

Les questionnements, les problèmes rencontrés ou, au contraire, les solutions trouvées au sein des campagnes spécifiques représentent autant de suggestions pour l'ajustement du modèle d'activité propre au service d'évaluation permanent.

RÉFÉRENCES

[1] G. Adda, J. Mariani, J. Lecomte, P. Paroubek, and M. Rajman. The GRACE French part-of-speech tagging evaluation task. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC)*, volume I, pages 433-441, Granada, Spain, 1998.

[2]J. Mariani. The Aupelf-Uref evaluation-based language engineering actions and related projects. In *Proceedings of the First International Conference on Language Resources and Evaluation*, volume I, pages 123-128, Granada, Spain, 1998.

[3] J. Véronis and P. Langlais. Evaluation of parallel text alignment systems. The ARCADE project. In Véronis J. (Editor), *Parallel text processing*, pages 369-388, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000.

[4] B. Landi, P. Kremer, D. Schibler, L. Schmitt. Amaryllis: an evaluation experiment on search engines in a French-speaking context. In *Proceeding of the First International Conference on Language Resourcesand Evaluation (LREC),* pages 1211–1214, Granada, Spain, 1998.

[5] E.M. Voorhees and L.P. Buckland (Editors). *NIST Special Publication 500-251: The eleventh Text Retrieval Conference (TREC 2002)*. Department of Commerce, National Institute of Standards and Technology, Gaithesburg, Maryland, 2002.

[6] C. Peters (Editor). Working papers of CLEF 2003, http://clef.iei.pi.cnr.it.