Un système de détection d'entités nommées adapté pour la campagne d'évaluation ESTER 2

Caroline Brun[1] Maud Ehrmann[2]

(1) XRCE, 6, Chemins de Maupertuis, Meylan, France
(2) JRC – European Commission, Ispra, Italie
Caroline.Brun@xrce.xerox.com, maud.ehrmann@jrc.ec.europa.eu

Résumé Dans cet article nous relatons notre participation à la campagne d'évaluation ESTER 2 (Evaluation des Systèmes de Transcription Enrichie d'Emissions Radiophoniques). Après avoir décrit les objectifs de cette campagne ainsi que ses spécificités et difficultés, nous présentons notre système d'extraction d'entités nommées en nous focalisant sur les adaptations réalisées dans le cadre de cette campagne. Nous décrivons ensuite les résultats obtenus lors de la compétition, ainsi que des résultats originaux obtenus par la suite. Nous concluons sur les leçons tirées de cette expérience.

Abstract In this paper, we report our participation to the ESTER 2 (Evaluation des Systèmes de Transcription Enrichie d'Emissions Radiophoniques) evaluation campaign. After describing the goals, specificities and challenges of the campaign, we present our named entity detection system and focus on the adaptations made in the framework of the campaign. We present the results obtained during the competition and then new results obtained afterward. We then conclude by the lessons we learned from this experiment.

Mots-clés : entités nommées, évaluation, extraction d'information.
Keywords: named entities, evaluation, information extraction.

1 Introduction

La campagne d'évaluation ESTER 2 s'est déroulée de janvier 2008 à avril 2009, dans la continuité de la première campagne ESTER[1]. L'objectif principal était «*de promouvoir une dynamique de l'évaluation en France, 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*»[2]. Ces campagnes visaient à évaluer les performances des systèmes de transcription de la parole, les performances des systèmes de segmentation en tours de paroles, et la capacité à extraire automatiquement des informations, en particulier les entités nommées (EN). Cette troisième tâche, à laquelle se sont attelés 7 participants dans le cadre d'ESTER 2, est l'objet de cet article. Elle était divisée

[1] Ces campagnes furent organisées par la DGA et l'AFCP. Site internet : http://www.afcp-parole.org/ester/index.html

[2] http://www.afcp-parole.org/ester/present.html

en deux sous-tâches : la détection d'entités nommées sur transcriptions de référence (NE-ref) et sur transcriptions automatiques (NE-asr).

2 ESTER 2 en détails

2.1 Spécificités et difficultés de la tâche de détection d'EN

Dans le cadre d'ESTER2, il s'agissait d'extraire et de catégoriser des mentions directes d'EN, selon un guide d'annotation comprenant 7 catégories principales et 38 sous-catégories :

Personnes : pers.hum (réelles ou fictives), pers.anim (animaux réels ou fictifs)
Fonctions : fonc.pol (politique), fonc.mil (militaire), fonc.admi (administrative), fonc.rel (religieuse), fonc.ari (aristocratique)
Lieux : loc.geo (lieu géographique), loc.admi (administratif), loc.line (voies de circulation), loc.adr (adresses), loc.adr.post (adresses postales), loc.adr.tel (fax et téléphones), loc.adr.elec (adresse électroniques), loc.fac (bâtiments)
Organisations : org.pol (politique), org.mil (militaire), org.edu (éducation), org.com (commerciale), org.non-profit (sans but lucratif), org.div (divertissement), org.gsp (géopolitique)
Produits: prod.vehic (vehicules), prod.award (récompenses), prod.art (produits artistiques), prod.doc (documents)
Temps : time.date (date), time.date.abs (date absolue), time.date.rel (date relative), hour (heures)
Quantités : amount.age (âge), amount.dur (durée), amount.temp (température), amount.len (longueur), amount.area (surface), amount.phys.vol (volume), amount.weight (poids), amount.spd (speed), amount.phys.cur (monnaies), amount.phys.other (autres)

La principale instruction d'annotation est de considérer les entités *en contexte*, avec la prise en compte des phénomènes d'ambiguïtés et de métonymie : par exemple, selon les contextes, « *Charles de Gaulle* » doit être annoté en tant que personne (le président), véhicule (le porte-avion) ou encore lieu (l'aéroport). L'annotation des noms de personnes inclut celle des fonctions et l'annotation des expressions temporelles couvre pour sa part un large éventail de possibilités, des classiques « *Lundi matin* » aux plus complexes « *Il y a un peu moins de trois jours environ* ». Par ailleurs, dans la mesure où l'extraction d'entités est réalisée sur des transcriptions de la parole, certains phénomènes propres à l'oral (hésitations ou répétitions) doivent être inclus dans les annotations (<pers.hum> *Jacques* ***heu*** *Chirac*</pers.hum>). Ces directives d'annotation spécifiques, combinées au nombre important de catégories à prendre en compte, complexifient la tâche d'annotation. En effet, les quantités de type âge et durée sont particulièrement difficiles à distinguer des expressions temporelles, tout comme les lieux administratifs des entités géopolitiques, puisqu'il s'agit de noms de villes ou de pays fréquemment employés en tant que l'un ou l'autre. On peut donc constater que cette tâche est plus ambitieuse que l'extraction d'EN « classique » (i.e. à la MUC).

2.2 Questions ouvertes concernant l'annotation

Se mettre d'accord sur la manière d'annoter des EN n'est pas chose facile. Ce problème bien connu n'a pas manqué d'apparaître durant ESTER2, avec de nombreuses discussions et remises en cause du guide d'annotation, modifié au fur et à mesure de la campagne jusqu'à une version définitive en janvier 2009. Les points délicats ont concerné, parmi d'autres, les expressions temporelles et les fonctions. Pour ce qui est des premières, il fut principalement question de l'extension des expressions temporelles (inclusion ou non des prépositions, déterminants et relatives), des difficiles distinctions entre dates et durées, et entre une expression temporelle et une autre qui ne l'est pas. Concernant les fonctions, deux points furent soulevés: le manque de critères pour définir la portée de cette catégorie d'une part (il est facile de lister des fonctions « standards » mais bien d'autres posent problèmes) et, d'autre part, la pertinence d'annoter conjointement, comme il était demandé dans certains cas, personnes et fonctions (n'est-t-il pas préférable, d'un point de vue sémantique, d'annoter des relations entre noms de personnes et noms de fonction ?).

3 XIP à ESTER 2

Nous avons participé à la campagne d'évaluation ESTER2 en adaptant l'analyseur syntaxique robuste « Xerox Incremental Parser » (XIP, (Ait-Mokthar et al., 2002)). XIP prend en entrée du texte tout venant, sous format texte ou XML, et produit en sortie de façon robuste une analyse syntaxique profonde. A partir d'un ensemble de règles, l'analyseur désambiguise les catégories, construit les syntagmes noyaux et extrait des relations de dépendances syntaxiques. En plus de l'analyse des relations syntaxiques de surface, XIP effectue également une analyse syntaxique dite « profonde » ou « normalisée » (prise en compte des sujets et objets de verbes non finis, normalisation de la forme passive en forme active, etc.). Cet analyseur intègre également un module de reconnaissance des entités nommées (Rebotier 2006), prenant en compte les types classiques d'entités nommées, à savoir les expressions numériques, les monnaies, les dates, ainsi que les noms de lieux, de personnes et d'organisations. Il s'agit d'un module à base de règles, consistant en un ensemble de règles locales ordonnées utilisant des informations lexicales et des informations contextuelles concernant les parties du discours, les formes lemmatisées et un ensemble de traits lexico-sémantiques.

Nous avons dû adapter le système développé pour le français aux consignes d'annotation ESTER 2, selon les axes suivants :

- *Adaptation aux spécificités de la transcription* : Les transcriptions de la parole, manuelles ou automatiques, ont des particularités que l'on ne retrouve pas dans les textes « standards » ; il s'agit de disfluences, de répétitions, ou encore de bruits :

« *Il y a encore* ***euh*** *quelques mois…* », « *Une forme* ***de de*** *journalisme …* », **« [rires-en-fond-]** *Voila !* **[-rire-en-fond]** »

Afin d'ignorer les disfluences et les bruits, nous avons converti les fichiers d'entrée originaux sous format XML, en marquant ces éléments comme des balises ouvrantes/fermantes (</heu>, </[rires]>) totalement transparentes pour les traitements linguistiques. Dans le cas des répétitions, nous avons développé des règles qui groupent ces éléments sous un nœud de même catégorie qui hérite des traits du premier élément.

- *Adaptation pour les catégories «standards»* : Nous avons tout d'abord utilisé les corpus d'entraînement et de développement pour collecter semi-automatiquement le vocabulaire inconnu (noms de lieux, d'organisations, etc.) et l'intégrer à nos lexiques. Nous avons ensuite adapté le système pour prendre en compte de nouvelles catégories, telles que les fonctions, les âges, les productions humaines, et la plupart des quantités, qui n'étaient pas préalablement couvertes par notre système. Nous avons également adapté les règles existantes selon le guide d'annotation, en particulier pour couvrir la portée des entités ; par exemple, les déterminants et prépositions sont inclus dans les quantités et les noms de fonction en apposition d'un nom de personne sont inclus dans ce dernier :

« *Il est âgé* <amount.age> *de 18 ans* </amount.age>
« <pers.hum> *Nicolas Sarkozy, président de la république* </pers.hum> … »
Il s'est principalement agit ici de développer et de modifier des règles locales de regroupement des noms propres, en amont de l'analyse en syntagmes noyaux (chunks).

- *Traitement des expressions temporelles* : Le vocabulaire relatif aux expressions temporelles étant une liste fermée, le cœur du travail fut l'écriture de règles locales et de chunking. L'attention fut portée sur les prépositions et adverbes principalement, ces derniers affectant radicalement le sens de telle ou telle expression (*Il est parti* <amount.phys.dur> *pendant 10 mois* </amount.phys.dur> vs. *Il est parti*

<time.date.rel>*10 mois après* </time.date.rel>). Nous avons également dû prendre en compte certaines incidences de la transcription de parole, comme par exemple avec l'expression « *19 cent 97* ».

- *Ambiguïtés et métonymies :* Une des spécificités les plus intéressantes d'ESTER 2 est la prise en compte des ambiguités et des phénomènes de métonymies. Afin d'être à même de traiter ces cas, nous avons utilisé les résultats de l'analyse syntaxique profonde fournis par XIP. En nous référant au guide d'annotation, nous avons réalisé une étude de corpus pour détecter les régularités syntaxiques et lexicales déclenchant un glissement métonymique ou permettant de résoudre une ambiguïté, selon la méthodologie appliquée dans (Brun et al 2007). Cette étude a conduit à des hypothèses telles que « Si un nom de lieu de type administratif est sujet d'un verbe de communication, il est employé comme nom d'organisation géopolitique». L'analyseur fut alors enrichi par des lexiques sémantiques dédiés et par des règles de dépendance modifiant l'interprétation des entités, appliquées en aval de l'analyse syntaxique, par exemple :

If (^LIEU[ADMI](#1) & SUBJ(#2[v_communication],#1)) → ORG[GSP=+](#1)[3]

Cette règle peut s'appliquer sur une phrase comme « *Dakar parle de 28 millions d'euros* », alors annotée « <org.gsp> *Dakar* </org.gsp> *parle de 28 millions d'euros* ». Notre étude s'est concentrée sur les relations de type sujet, objet, modifieur (nominal et propositionnel) et attribut, et nous a conduites à développer environ 150 règles de dépendances supplémentaires.

4 Evaluation(s)

4.1 Corpus et calcul des scores

Comme dit précédemment, nous avons utilisé les corpus d'entraînement (100 heures d'émissions de radio transcrites et annotées manuellement) et de développement (6 heures de journaux radiophoniques transcrits et annotés manuellement) pour la mise au point du système. Le corpus de test était constitué de 7 heures de journaux radiophoniques datant de 2008. L'ensemble des corpus provenait de différentes sources : France Culture, France Inter, Radio France International, Radio Classique, Africa 1, Radio Congo et Radio Télévision Marocaine. Du point de vue de l'évaluation quantitative, même si les mesures classiques de précision et rappel étaient calculées, la mesure « officielle » était le « Slot Error Rate » (SER, voir (MAKHOUL et al. 1999)), qui combine et pondère les différents type d'erreurs (insertion, effacement, erreur de type) : SER = (Insertions+Effacements+Substitutions) / nb entités ref. C'est une mesure analogue au « Word Error rate » (WER) utilisé pour mesurer les performances des systèmes de transcriptions de la parole. D'autre part, si au début de la campagne il était prévu d'évaluer sur l'ensemble des sous-types, c'est seulement sur les 7 catégories principales que les résultats ont été calculés. Enfin, les résultats définitifs ont été obtenus après une phase d'adjudication qui permettait aux participants de contester les annotations du corpus de test (sans bien évidemment changer les résultats de leur système).

4.2 Résultats obtenus dans le cadre d'ESTER 2

Le tableau I présente les résultats obtenus par notre système sur transcriptions de référence (NE-Ref) en termes de précision, rappel, f-mesure et *slot error rate*. Les résultats de notre système sur les transcriptions automatiques (NE-Asr) sont publiés dans (Brun et Ehrmann, 2009), les résultats complets de la campagne dans (Galliano et al, 2009). Avec un SER de 9.80 (et une f-mesure de 0,93), ces résultats s'avèrent très

[3] Cette règle se lit de la manière suivante : si le parseur a détecté un nom de LIEU avec l'attribut « admi » (#1), et que ce nom est le sujet d'un verbe de communication (#2), alors une relation unaire ORG avec l'attribut « gsp » est crée pour ce nom (#1).

satisfaisants. On remarque que les scores pour les catégories *org* et *loc* sont quelque peu inférieurs aux résultats « standards» dans ce genre de compétition, ce qui montre l'impact (et la difficulté) du traitement de la métonymie, beaucoup d'erreurs venant de la confusion *loc.admi* et *org.gsp*. Un constat équivalent pour la catégorie *amount*, habituellement assez simple, peut être fait, dû aux ambiguïtés entre durées et âges d'un coté et expressions temporelles de l'autre. Une dernière remarque concerne les noms de productions humaines, dont le score est faible, en raison de leur faible représentation dans le corpus et de la diversité des éléments que cette catégorie est censée couvrir : des véhicules aux titre d'œuvres d'art en passant par les documents légaux.

Pers	3110	97.76	95.57	0.97	3.63
Fonc	754	81.81	89.46	0.85	24.90
Org	2663	89.24	83.97	0.87	16.08
Loc	1875	89.01	88.73	0.89	7.09
Prod	191	100	42.11	0.59	46.03
Time	3235	95.63	95.69	0.96	5.85
Amount	939	93.76	86.57	0.90	15.27
TOUT	**12767**	**93.61**	**91.50**	**0.93**	**9.80**

Tableau 1: résultats sur transcriptions de référence(NE-Ref)

4.3 Expérience post-ESTER 2

Nous avons trouvé intéressant de poursuivre de notre coté la campagne ESTER 2 en calculant les scores (initialement planifiés) pour l'ensemble des sous-catégories.

Pers.hum	97.8	95.6	0.97	3.63	**Prod.art**	100	8.6	0.16	78.3
Fonc.admi	51.2	71.8	0.60	48.7	**Prod.award**	100	65.5	0.79	28.3
Fonc.mil	0	0	0	200	**Prod.doc**	100	31.6	0.48	56.4
Fonc.pol	78.5	72.0	0.75	25.9	**Prod.vehic**	100	87.5	0.93	10
Fonc.reli	65.4	77.3	0.71	95.7	**Time.date.abs**	94.3	90.2	0.92	9.72
Org.com	81.4	57.4	0.67	15.2	**Time.date.rel**	94.2	87.9	0.91	9.15
Org.edu	100	32.7	0.49	25	**Time.hour**	86.6	95.5	0.91	4.12
Org.gsp	60	68.1	0.64	15.4	**Amount.cur**	96.9	92.5	0.95	14.5
Org.div	96.3	62.12	0.76	26.1	**Amount.age**	92	56.1	0.70	26.9
Org.non-profit	57.8	50.7	0.54	36.3	**Amount.len**	100	87.5	0.93	12.5
Loc.admi	83	86.4	0.85	6.3	**Amount.area**	100	95.2	0.98	7.7
Loc.fac	89.7	64.6	0.75	33.6	**Amount.vol**	100	100	1	0
Loc.geo	48.4	16.4	0.25	46.6	**Amount.wei**	100	91.1	0.95	9.1
Loc.line	82.5	60	0.69	27.2	**Amount.temp**	71.4	53.6	0.61	66.7
Loc.addr.elec	100	72.27	0.84	20	**Amount.dur**	83.3	81.8	0.83	18.9
Loc.addr.tel	100	100	1	0	**TOUT**	89.6	82.8	0.86	14.22

Tableau II: résultats par catégorie fine

Nous avons donc appliqué le script d'évaluation sur les mêmes corpus, l'évaluation étant stricte car une erreur est comptée si les catégories hypothèse et référence ne sont pas exactement les mêmes, même si la catégorie générale est commune. Le tableau II montre que les résultats restent très satisfaisants globalement, mais on constate cependant une chute importante du rappel. Cette chute est particulièrement marquée pour les noms d'organisations, ce qui indique que leurs sous-types sont encore mal distingués par notre système.

5 Bilan et conclusions

Cet article décrit notre participation à la tâche de reconnaissance des entités nommées de la campagne d'évaluation ESTER 2, qui s'est terminée en juin 2009. Nous avons adapté un système d'extraction d'entités nommées préexistant au sein d'un analyseur robuste, XIP. La finesse d'annotation requise lors de cette campagne nous a ainsi poussées à utiliser les résultats de l'analyse syntaxique profonde, en particulier pour le traitement des problèmes d'ambigüités sémantiques et de métonymies. Les résultats obtenus sur transcriptions manuelles, pour l'annotation en catégories générales, étaient très satisfaisants. L'expérience que nous avons menée *a posteriori* sur l'annotation en catégories fines a permis de mettre en évidence certains éléments à améliorer dans notre système.

D'une façon générale, la participation à cette campagne s'est avérée extrêmement bénéfique pour notre système de repérage des entités nommées. Mais peut-être encore plus crucialement, cette évaluation a permis aux participants de mener une réflexion approfondie sur les problèmes d'annotations des entités nommées : quels sont les critères pour décider qu'une unité linguistique est une entité nommée, quel est l'étiquette à donner dans un contexte donné, quels sont leurs frontières, etc. Cette réflexion a permis d'aboutir à une première version d'un guide d'annotation qui vise à devenir un standard pour le français.

Références

AIT-MOKTHAR, S. CHANOD J.P, ROUX C. (2002). Robustness beyond Shallowness: Incremental Dependency Parsing. *Special Issue of NLE Journal.*

BRUN C., EHRMANN. (2009). Adaptation of a Named Entity Recognition System for the ESTER 2 Evaluation Campaign. *IEEE NLP-KE 2009 (IEEE International Conference on Natural Language Processing and Knowledge Engineering)*, Dalian, China, Sep 24-27.

BRUN C., EHRMANN M. , JACQUET G. (2007) , XRCE-M : A hybrid system for named entity metonymy resolution, Actes de 4th *International Workshop on Semantic Evaluations, ACL-SemEval 200*, Prague.

GALLIANO S., GRAVIER G. AND CHAUBARD L. (2009). The ESTER 2 Evaluation Campaign for the Rich Transcription of French Radio Broadcasts", *10th Annual Conference of the International Speech Communication Association , InterSpeech 2009, Brighton UK.*

MAKHOUL J., KUBALA F., SCHWARTZ R., WEISCHEDEL R. *(1999).* Performance Measures For Information Extraction, dans les actes du *DARPA Broadcast News Workshop*, 249—252.

REBOTIER A. (2006). Développement d'un module d'extraction d'Entités Nommées pour le français, Mémoire de DEA, Université Stendhal Grenoble III.