TALN 2011, Montpellier, 27 juin – 1er juillet 2011

Génération automatique de motifs de détection d'entités nommées en utilisant des contenus encyclopédiques.

Eric Charton[1] Michel Gagnon[1] Benoit Ozell[1]
(1) École Polytechnique, 2900 boul. Edouard Montpetit, Montréal, Canada
{eric.charton, michel.gagnon, benoit.ozell}@polymtl.ca

Résumé. Les encyclopédies numériques contiennent aujourd'hui de vastes inventaires de formes d'écritures pour des noms de personnes, de lieux, de produits ou d'organisation. Nous présentons un système hybride de détection d'entités nommées qui combine un classifieur à base de Champs Conditionnel Aléatoires avec un ensemble de motifs de détection extraits automatiquement d'un contenu encyclopédique. Nous proposons d'extraire depuis des éditions en plusieurs langues de l'encyclopédie Wikipédia de grandes quantités de formes d'écriture que nous utilisons en tant que motifs de détection des entités nommées. Nous décrivons une méthode qui nous assure de ne conserver dans cette ressources que des formes non ambiguës susceptibles de venir renforcer un système de détection d'entités nommées automatique. Nous procédons à un ensemble d'expériences qui nous permettent de comparer un système d'étiquetage à base de CRF avec un système utilisant exclusivement des motifs de détection. Puis nous fusionnons les résultats des deux systèmes et montrons qu'un gain de performances est obtenu grâce à cette proposition.

Abstract. Encyclopedic content can provide numerous samples of surface writing forms for persons, places, products or organisations names. In this paper we present an hybrid named entities recognition system based on a gazetteer automatically extracted. We propose to extract it from various language editions of Wikipedia encyclopedia. The wide amount of surface forms extracted from this encyclopedic content is then used as detection pattern of named entities. We build a labelling tool using those patterns. This labelling tool is used as simple pattern detection component, to combine with a Conditional Random Field tagger. We compare the performances of each component of our system with the results previously obtained by various systems in the French NER campaign ESTER 2. Finally, we show that the fusion of a CRF label tool with a pattern based ones, can improve the global performances of a named entity recognition system.

Mots-clés : Étiqueteur, Entités nommées, Lexiques.

Keywords: Tagger, Named entities, Gazetteer.

1 Introduction

La tâche d'*étiquetage par des entités nommées* (EEN) est un processus lors duquel chaque mot d'une phrase correspondant à une *entité nommée* (EN) (généralement un nom propre et par extension des dates ou des quantités) reçoit une étiquette de classe. Cette classe correspond à un

arbre taxonomique dans la complexité et la nature sémantique peuvent varier. La tâche d'EEN s'étend à la reconnaissance de locution nominales (au sens de suite de mots, figée par l'usage, pouvant être substituée à un nom) en regroupant plusieurs mots étiquetés (comme par exemple dans le cas de *Paris* qui est une entité de type localité tout comme *Ville Lumière*, ou encore l'acronyme *TGV* qui décrit le même produit de type véhicule que la locution nominale *Train à Grande Vitesse*). Les campagnes d'évaluation telles que MUC [1], ACE (Doddington *et al.*, 2004), CoNLL (Tjong & Meulder, 2003) et dans le contexte francophone, la tâche d'étiquetage de la campagne ESTER 2 (Galliano *et al.*, 2009), ont permis d'expérimenter des approches variées dans un contexte standardisé et de mesurer leurs performances avec des métriques communes. A la suite des ces campagnes, deux grandes familles de systèmes d'EEN ont fait la démonstration de leur potentiel : celles dérivées de la linguistique computationnelle, recourant à des règles de détection plus ou moins sophistiquées, et celles par apprentissage automatique qui consistent à entraîner un classifieur sur un corpus pré-étiqueté. Ces deux grandes familles d'approches exploitent à des degrés divers des ressources lexicales externes dont la finalité est de renforcer leur capacités de détection d'EN. L'une des caractéristiques récurrente des systèmes d'EEN à base de règles est qu'ils intègrent dans leur processus de détection d'EN des lexiques plus ou moins riches, dont la disponibilité de grand corpus numériques favorise aujourd'hui l'extraction automatisée. Dans ce contexte, nous avons souhaité chercher à évaluer dans quelle mesure un lexique de grande taille, tel que ceux implémentés dans les détecteurs à base de règles, pourrait être utilisé en tant que système rudimentaire de détection par motif pour améliorer un étiqueteur numérique. Cette communication, présente une ressource lexicale automatiquement extraite du corpus Wikipédia, que nous utilisons en tant que motifs rudimentaires de détection d'EN. Nous évaluons les capacités d'un système d'EEN reposant uniquement sur ces motifs de détection, puis nous l'hybridons avec un système d'EEN par apprentissage automatique à base de CRF.

L'article est structuré ainsi : dans la section 2, nous passons en revue les différentes méthodes d'étiquetages d'EN proposées et leur caractéristiques. Nous présentons ensuite dans la section 3 notre proposition de système d'EEN par motifs. Nous décrivons une méthode d'extraction de motifs de détection non ambigus contenus dans un corpus encyclopédique, et la ressource que nous avons obtenue. Puis, dans la section 4, nous évaluons ce système de détection à base de motifs en l'appliquant au corpus de test ESTER 2. Nous fusionnons ses résultats avec ceux obtenus par un EEN à base de CRF et discutons du gain de performance obtenu. Nous concluons dans la section 5 qu'il est possible d'élaborer une méthode peu coûteuse d'introduction de connaissance lexicale en complément des méthodes statistiques et que cette méthode permet d'améliorer la robustesse des système d'EEN.

2 Méthodes d'étiquetage d'entités nommées

Pour extraire les EN d'un texte l'utilisation, en tant que motifs de détection, de lexiques d'entités issus de corpus tels que Wikipédia est une solution applicable (Bunescu & Pasca, 2006; Nothman *et al.*, 2009; Kazama & Torisawa, 2007) mais insuffisante pour plusieurs raisons. En premier lieu, de nombreuses entités à détecter sont absentes de ces corpus de ressources,

1. Voir `http://www-nlpir.nist.gov/related_projects/muc/proceedings/ne_task.html`.

fussent-ils aussi vaste que Wikipédia (principe des *mots hors vocabulaires* ou OOV) ; en second lieu, de nombreuses EN sont hautement ambiguës et ne permettent pas la détection directe sans recours à une analyse de leur contexte. Les lexiques apparaissent donc toujours en renfort d'un étiqueteur à règle ou en tant que ressource pour améliorer l'apprentissage d'un étiqueteur statistique.

2.1 Méthodes automatiques par apprentissage

De manière générale, la plupart des approches de reconnaissance automatique reposent sur la théorie des probabilités et peuvent être caractérisées par une méthode générative ou discriminante selon que la distribution de probabilités de la caractéristique à reconnaître est modélisée ou non. Cette différence joue un rôle important dans la tâche d'étiquetage d'EN. En effet, un classifieur discriminant est théoriquement plus précis mais moins capable d'inférer qu'un classifieur génératif et donc moins adaptable aux innombrables possibilités de représentations d'une EN. Ces deux voies possibles d'approches se retrouvent dans la tâche d'étiquetage d'EN : les méthodes génératives, comme par exemple celle reposant sur des modèles de Markov cachés (HMM) (Bikel *et al.*, 1999), et les méthodes discriminantes telles que SVM, Maximum Entropie (MaxEnt) (Borthwick *et al.*, 1998). Les Champs Conditionnels Aléatoires (CRF) (Lafferty *et al.*, 2001) occupent une place à part car ils combinent une nature générative et discriminante. Comme les modèles discriminants, ils tiennent compte, lors de la construction du modèle, des nombreuses observations issues du corpus d'apprentissage et les corrèlent entre elles lors de l'entrainement. Mais à l'instar des modèles génératifs, les CRF probabilisent les décisions en fonction de la position des séquences d'apprentissage. Ce mode de fonctionnement hybride qui favorise l'inférence, c'est-à-dire la reconnaissance par un classifieur CRF d'une entité qu'il n'a jamais observée dans le corpus d'apprentissage, mais aussi la précision en utilisant les données d'apprentissage pour discriminer, explique pourquoi des études (Raymond & Riccardi, 2007) montrent régulièrement que les systèmes à base de CRF sont plus performants que ceux à base de HMM, de SVM ou de MaxEnt pour résoudre la tâche d'étiquetage d'EN avec un système automatique. La performance d'un système CRF est aussi très largement dépendante de la formation des échantillons d'apprentissage qui vont lui être soumis (McCallum & Li, 2003). Pourtant, dans un contexte expérimental normalisé tel que celui de la campagne ESTER, quelque soit le soin apporté à la sélection des échantillons et à la formation du corpus d'apprentissage, il est régulièrement observé (voir par exemple (Raymond & Fayolle, 2010)) que les performances du CRF demeurent inférieures à celles d'un EEN à base de règles sur des données non bruitées.

2.2 Méthodes à règles et automates

Les systèmes d'EEN à base de règles utilisent des méthodes linguistiques ou des automates à états finis pour identifier les EN dans un texte. Certains de ces systèmes sont fortement inspirés par les méthodes linguistiques. Tel XIP (Brun *et al.*, 2010) qui en partant d'un ensemble de règles, identifie les syntagmes noyaux et extrait les relations de dépendance syntaxiques pour localiser des EN. L'analyse syntaxique peut d'ailleurs être plus ou moins profonde pour détecter des structures qui fiabiliseront le processus de détection des EN. D'autres systèmes

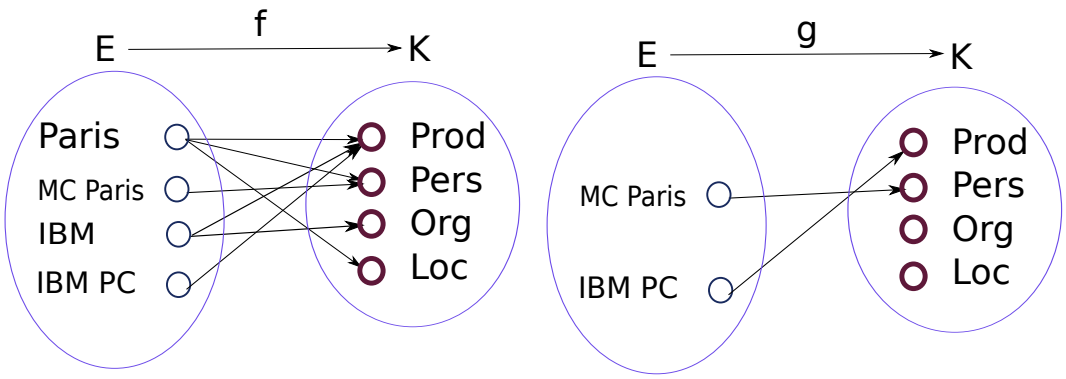

FIGURE 1 – Représentation du principe de réduction de l'ensemble de formes de surface disponibles par identifications des motifs non discriminants.

d'EEN à règles se contentent d'automates de détection plus ou moins sophistiqués. Ainsi, le système CasEN, déployé lors de la campagne ESTER 2 (Nouvel *et al.*, 2010) exploite exclusivement des transducteurs, environ 150, qui s'appliquent à reconnaitre des séquences de mots qui contiennent une EN.

La plupart des systèmes de cette famille des systèmes d'EEN non automatiques complètent le processus de détection par un ensemble de règles très spécialisées qui font appel à des ressources lexicales, des informations liées aux parties du discours, et parfois des traits lexico-sémantiques.

La littérature fait apparaitre que la famille des étiqueteurs à règles utilise quasi systématiquement des automates complémentaires pour identifier les expressions numériques, les quantités, les devises, les dates. Ces automates, lorsqu'ils sont appuyés par des ressources lexicales, peuvent être amenés à jouer un rôle plus ou moins important dans l'identification des EN de la famille des noms propres. Mais de manière générale, peu d'explications détaillées sont fournies sur le rôle et l'influence sur les performances globales des systèmes de ces ressources lexicales associées à des automates. L'un des objectifs du travail présenté dans cet article sera de contribuer à l'étude de l'influence des lexiques utilisés directement en tant qu'automates simples pour détecter des EN sans avoir à utiliser le système de règles ou le classifieur numérique.

3 Système proposé

Nous proposons d'enrichir un système d'EEN à base de CRF avec un ensemble de motifs de détection extraits automatiquement depuis l'encyclopédie Wikipédia. Notre idée est qu'il est possible d'améliorer les performances d'un système d'EEN par apprentissage automatique en lui associant un module qui détecterait les graphies non ambiguës des EN. Nous souhaitons ainsi évaluer à quel point les motifs issus de ressources lexicales qui renforcent les systèmes à base de règle influent sur les performances globales de cette famille d'étiqueteurs. Cette proposition permet d'envisager d'intégrer une connaissance lexicale rudimentaire dans un processus d'EEN par CRF pour le rapprocher des performances des systèmes linguistiques et à base de règles. Les motifs de détection que nous proposons d'extraire sont rudimentaires et ne concernent que des EN non ambiguës. Leur principe fonctionnel peut être illustré par ces exemples :

– Si nous prenons l'exemple du nom **Montréal**, celui-ci correspond à plusieurs entités distinctes (*Montréal (Québec)*, *Montréal (Ardèche)*) qui sont de classe identique, à savoir des localités (étiquette LOC). Considérons une graphie associée à une EN, que nous appelons forme de surface de cette EN. Il est possible d'exploiter une forme de surface *Montréal* en tant que motif pour détecter plusieurs entités nommées d'identités différentes mais qui sont toute de type LOC.
– La forme de surface *Paris*, en revanche, est associée à plusieurs entités de type localité telles que *Paris, France* ou *Paris, Texas* (LOC), mais aussi à des noms de personnes (PERS.HUM) *Antoine Paris*, *Paris Hilton*, de navires ou de produits (le Paquebot *Paris* (PROD.VEHICLE)) ou l'album musical *Paris* (PROD.DIV)). La forme de surface *Paris* est donc hautement ambiguë et ne peut être utilisée en tant que motif de détection susceptible d'identifier une entité et de lui attribuer une classe d'étiquetage valable.
– On pourra en revanche conserver les formes de surfaces intégrant un élément ambigu, mais plus longues - de type bi-grammes à n-grammes, si elles sont non ambiguës : ainsi les formes de surface *MC Paris* (nom de personne) ou *SS Paris* (nom de véhicule) peuvent être utilisées en tant que motifs de détection.

On peut formaliser d'après ces exemples que l'ensemble des motifs de détection non ambigus est le sous ensemble injectif constitué des relations entre l'ensemble des formes de surfaces et l'ensemble des classes qui leur sont reliées, si et seulement si tout élément de l'ensemble d'arrivée K possède au plus un antécédent par g de l'ensemble de départ E (voir figure 1).

3.1 Extraction automatique de motifs de détection

Nous souhaitons extraire les motifs de détection depuis l'encyclopédie Wikipédia. Nous avons présenté dans (Charton & Torres-Moreno, 2009) un système capable de produire, d'après un corpus encyclopédique tel que Wikipédia, une ressource multilingue de concepts que nous avons intitulé *métadonnées*. Ces *métadonnées* incluent des noms propres, des noms communs, des entités nommées, ainsi que des locutions rigoureusement classées selon la norme taxonomique de la campagne ESTER 2 et associées chacune à plusieurs formes de surface. La proportion des ensembles de fiches encyclopédiques transformées en *métadonnées* affectées à chaque classe est présentée dans la table 1. Pour chaque *métadonnée*, les formes de surfaces qui permettent d'écrire le concept encyclopédique sont collectées dans les éditions polonaise, italienne, française, anglaise, espagnole, allemande et italienne de Wikipédia. La quantité totale de formes de surfaces disponibles est indiquée dans la table 2. Un exemple d'ensemble de formes de surfaces contenu dans une *métadonnée* est montré dans la figure 2. Cet exemple [2] montre l'ambiguïté de certains motifs collectés dans le corpus encyclopédique. On peut observer dans cet exemple que la forme *Renault* est hautement ambiguë (puisqu'elle caractérise également un nom de personne dans l'encyclopédie), en revanche, des séquences telles que *Renault Nissan Group, Renault Motor* collectées depuis Wikipédia en Anglais ou encore le sigle *RNUR* collecté depuis Wikipédia en Polonais, sont des motifs de détection non ambigus. Nous obtenons ainsi un ensemble de paires composées de motifs de détections associés à une classe unique, que nous allons utiliser trivialement dans un étiqueteur d'EN à expression régulières.

2. Consultable en ligne sur `http://www.nlgbase.org/perl/display.pl?query=Renault&search=EN`

Qté	Contenu	Classe taxonomique
3515	Fonctions et titres	FONC
753629	Lieu	LOC
346218	Organisations	ORG
972663	Personne	PERS
411569	Produit	PROD
14294	Date	TIME
621082	Contenu encyclopédique	UNK
3122970	*métadonnées* disponibles	

TABLE 1 – Quantité de *métadonnées* disponibles pour chaque classe d'étiquetage.

3 122 970	métadonnées disponibles
8 142 183	formes de surface disponibles
5 832 730	formes de surface conservées

TABLE 2 – Formes de surfaces non ambiguës extraites depuis les *métadonnées* et utilisables en tant que motifs de détection.

3.2 Étiqueteur CRF

Nous utilisons en tant que *baseline* la première version de l'étiqueteur d'EN mis au point par le LIA pour la campagne ESTER 2 (Béchet & Charton, 2010). Nous l'intitulons CRF-V1. Cet étiqueteur a pour caractéristique d'être entraîné sur un corpus de grande taille préalablement étiqueté par un étiqueteur HMM, en utilisant une ressource lexicale issue des *métadonnées*. Des itérations successives permettent de diminuer le bruit qui subsiste sur le corpus d'entrainement.

La version que nous utilisons ici pour comparer notre système à CRF-V1 est intitulée CRF-V2 et décrite dans (Charton & Torres-Moreno, 2010). Elle complète la phase de préparation par HMM du corpus d'entrainement par un étiquetage supplémentaire utilisant les liens internes de Wikipédia. CRF-V2 est appris sur un ensemble de phrases issues du corpus d'entrainement d'ESTER 2, renforcé par 140 000 phrases étiquetées extraites depuis Wikipédia en français. CRF-V2 est légèrement plus performant que CRF-V1. Il a déjà été déployé dans le système Poly-Co du challenge GREC 2010 [3].

L'architecture complète du système est la suivante. Dans un premier temps, les deux étiqueteurs d'EN, celui à motifs et celui à CRF, sont appliqués sur le document à étiqueter. On obtient par ce moyen deux documents étiquetés que nous nommerons $doc.crf$ et $doc.rule$. L'étiquetage des EN de ces documents est soit un nom de classe k (issu de la taxonomie ESTER) soit le label indéfini UNK appliqué lorsqu'aucune étiquette n'est attribuée. Dans un second temps une fusion de $doc.crf$ et $doc.rule$ est réalisée. Le processus de fusion est trivial et consiste à comparer les étiquettes appliquées à $doc.crf$ et $doc.rule$ en donnant priorité à l'un des documents. L'algorithme de fusion donne ici priorité aux EN contenues dans $doc.crf$.

3. Voir `http://www.itri.brighton.ac.uk/research/genchal10/grec/`

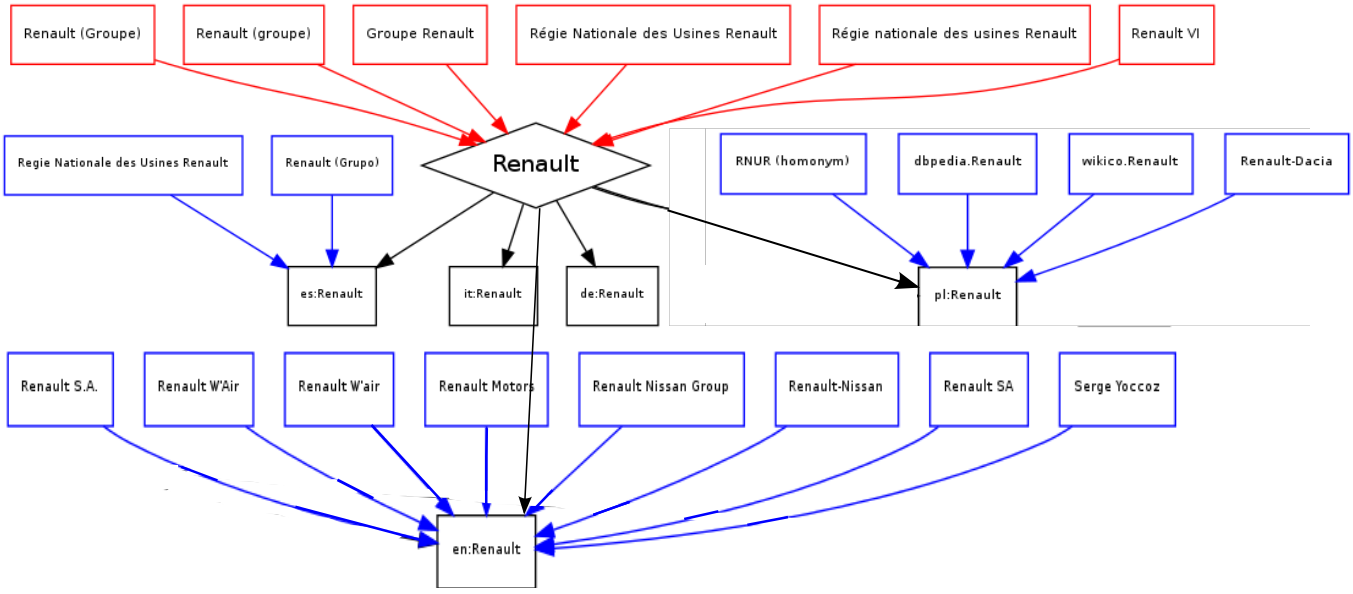

FIGURE 2 – Un exemple de formes de surface collectées pour la *métadonnées* de classe ORG correspondant à la fiche encyclopédique du constructeur automobile *Renault* dans plusieurs éditions linguistiques de Wikipédia.

4 Évaluation et résultats

Nous évaluons les capacités d'un EEN à base de motifs en le comparant aux autres méthodes d'étiquetage dont les résultats sont connus pour un corpus de référence. Puis nous évaluons les performances d'un EEN à CRF renforcé par l'EEN à motifs. Notre expérience vise à mesurer jusqu'à quel point l'introduction de motifs de détection non ambigus et collectés automatiquement peuvent améliorer les performances du CRF, et le cas échéant jusqu'à quel point il permet de rapprocher les performances des CRF de ceux à base de règles, sur des corpus non bruités. Nous utiliserons le corpus de test de la campagne ESTER 2.

4.1 Corpus et mesures de référence

Le corpus complet de la tâche de détection d'EN d'ESTER 2 se compose de 72 heures d'émissions radiophoniques francophones (France-Inter, France Info, RFI, RTM, France Culture, Radio Classique) manuellement transcrites et annotées en EN suivant les conventions des deux campagnes ESTER. La première campagne comportait un jeu de 30 types d'EN réparties en 9 catégories racines, alors que la seconde possède un jeu de 37 types d'entités nommées réparties en 7 catégories racines (personne, fonction, organisation, lieu, fabrication humaine, date et heure, quantités). Seules les catégories racines sont mesurées dans les résultats de référence. La campagne ESTER 2 prévoit plusieurs tâches de reconnaissance d'EN : la première consiste à reconnaître les EN dans la transcription manuelle du corpus de test (NE-Ref). La seconde s'applique à trois transcriptions automatiques dites ASR et dont les taux d'erreurs de reconnaissance de mots vont croissants : 12.11%, 17.83% et 26.09%. La volonté de l'organisateur est ici de tester la précision des systèmes sur NE-Ref qui est non bruité, mais aussi leur ro-

EN	AMOUNT	FONC	LOC	ORG	PERS	PROD	TIME	tous
Qté	239	196	1215	1267	1108	58	1025	5123
précision	0,85	0,61	0,77	0,79	0,93	0,53	0,91	0,86
rappel	0,56	0,559	0,81	0,63	0,75	0,12	0,60	0,718
F-Score	0,68	0,58	0,79	0,70	0,84	0,20	0,73	0,78

TABLE 3 – Résultats par entité à étiqueter du système LIA dit CRF-V1 appliqué au corpus NE-Ref lors de la campagne ESTER 2.

EN	AMOUNT	FONC	LOC	ORG	PERS	PROD	TIME	tous
Qté	239	196	1215	1267	1108	58	1025	5123
précision	0,93	0,818	0,897	0,89	0,97	100	0,95	0,93
rappel	0,86	0,899	0,88	0,83	0,95	0,42	0,95	0,91
F-Score	0,90	0,85	0,89	0,87	0,97	0,59	0,96	0,93

TABLE 4 – Résultats par entité à étiqueter du système XIP de Xerox à base de règles appliqué au corpus NE-Ref lors de la campagne ESTER 2.

bustesse dans le contexte plus difficile des corpus de test ASR bruités de manière croissante.

Les résultats de la campagne ESTER 2 (Galliano *et al.*, 2009) soulignent l'efficacité d'un système EEN à base de règle linguistique sur la transcription de référence (NE-Ref). Sur ce corpus, les deux meilleurs systèmes sont à base de règle, et le troisième est de type automatique à base de CRF. Le tableau 4 présente les résultats obtenus par le meilleur système sur transcriptions de référence (NE-Ref) en termes de Précision, Rappel, F-Score. Le tableau 3 présente les résultats du meilleur système automatique à CRF sur ce même corpus de référence. Nous considérerons les résultats du meilleur système linguistique (XIP) et du meilleur système automatique (CRF-V1) sur le corpus NE-Ref pour situer les performances obtenues par l'hybridation de l'étiqueteur à motif, que nous appellerons ici EEN-M, avec CRF-V2.

Ce plan d'expérience vise à évaluer dans quelle mesure un ensemble de motifs appris automatiquement sur un corpus encyclopédique peut améliorer les performances du système CRF et jusqu'à quel point les performances de ce système CRF amélioré peuvent se rapprocher d'un système d'EEN de nature linguistique à l'état de l'art (en l'occurrence le système XIP). Notre expérience consistera à appliquer au corpus NE-Ref d'ESTER 2 les détecteurs EEN-M et CRF-V2 et à fusionner leurs résultats puis à mesurer les performances de chaque élément de notre système.

4.2 Résultats

Le tableau 5 expose les résultats des différents composants de notre système d'EEN. Il indique pour chaque jeu d'étiquettes du corpus de test NE-Ref ESTER 2 les performances individuelles de chaque composant. Dans la section motif du tableau, qui présente les résultats de l'étiqueteur par détection de motif EEN-M, on remarque que la classe AMOUNT n'est pas traitée par ce composant d'étiquetage car non observée dans les *métadonnées* utilisées pour collecter les motifs. La classe TIME qui correspond aux dates dans le corpus ESTER 2 est pour ce qui la

	EN	AMOUNT	FONC	LOC	ORG	PERS	PROD	TIME	tous
	Qté	239	196	1215	1267	1108	58	1025	5123
EEN-M	précision	x	0,85	0,73	0,94	0,98	0.11	0,96	0,88
/ Motifs	rappel	x	0,30	0,32	0,27	0,50	0,07	0,36	0,34
	F-Score	x	0,43	0,44	0,42	0,66	0,08	0,53	0,48
CRF-V2	précision	0,90	0,99	0,77	0,92	0,94	0.38	0,97	0,88
	rappel	0,70	0,46	0,90	0,61	0,93	0,25	0,69	0,74
	F-Score	0,79	0,63	0,83	**0,73**	0,93	**0,30**	0,89	0,80
Hybride	précision	0.90	0.91	0.76	0,91	0,96	0.27	0,96	0,88
	rappel	0,70	0,55	0,92	0,60	0,93	0,25	0,83	0,78
	F-Score	**0,79**	**0,69**	**0,83**	0,72	**0,94**	0,26	**0,90**	**0,83**
CRF-V1	F-Score	0,68	0,58	0,79	0,70	0,84	0,20	0,73	0,78

TABLE 5 – Résultats détaillés du système EEN à motifs de détection, CRF (dit CRF-V2) et hybride, comparé au système CRF du LIA (dit CRF-V1) ayant obtenu les meilleures perfomances sur le corpus de test NE-Ref de la campagne ESTER 2.

concerne traitée car cette classe de contenu est représentée dans Wikipédia et donc modélisée dans les métadonnées [4]. On note que EEN-M offre une couverture de détection des EN relativement faible (rappel de 0,34) mais une précision supérieure à celle de CRF-V1. Cette précision est également supérieure à celle de l'étiqueteur à règles linguistique présenté dans le tableau 4 pour les classes FONC, ORG et TIME. Il est important de remarquer les performances inférieures de EEN-M sur la détection de la classe LOC qui sont attribuables à l'impossibilité pour EEN-M de traiter la différence entre les notions LOC.ADMI et ORG.GSP (un nom toponymique peut désigner une localité ou une organisation géo-politique dans le corpus ESTER 2) par un système à motif.

La section CRF-V2 présente les résultats de l'étiqueteur CRF amélioré tel que décrit dans la section 3. On observe que les performances de cet étiqueteur sont légèrement meilleures que celles de l'étiqueteur déployé par le LIA lors de la campagne ESTER 2, et dont les résultats sont indiqués dans la section CRF-V1 du tableau. Une comparaison plus détaillée avec le tableau 3 montre que les performances de CRF-V2 sont améliorées globalement tant pour la précision que le rappel, avec les mêmes difficultés de modélisation des séquences d'EN de type PROD. L'amélioration des performances du système CRF-V2 appliqué sur NE-Ref, par rapport à CRF-V1, n'est pas l'objet de cette communication, mais doivent être commentées ici car l'amélioration de la précision joue un rôle sur le processus de fusion. Les expériences de fusions que nous avons menées entre les résultats produits par CRF-V1 et ceux de EEN-M nous ont montré une très légère minoration des performances globales du système (les moindres performances de CRF-V1 étant compensées par l'introduction des EN détectées par EEN-M).

L'hybridation de EEN-M et CRF-V2 indiquée dans la ligne *Hybride* du tableau 5, est le résultat de la fusion entre les deux sorties de ces systèmes. Elle montre un gain de performance de 3% sur CRF-V2 seul, et de plus de 5% par rapport à CRF-V1 déployé lors de ESTER 2.

4. Voir par exemple la catégorie `http://fr.wikipedia.org/wiki/Categorie:Jour_de_septembre` et une *métadonnée* telle que `http://www.nlgbase.org/perl/display.pl?query=2septembre&search=FR`

4.3 Discussion

Il apparaît que le système de détection hybride à base de motifs et de CRF proposé améliore substantiellement les performances du système CRF-V1 déployé lors de la campagne ESTER 2 sur le corpus de référence NE-Ref.

On notera que les expériences d'hybridation de *métadonnées* et du système CRF qui avaient été employées lors de cette campagne, qui reposaient sur une détection de motifs non désambiguïsés associée à un calcul de similarité cosinus entre le contexte du motif et les métadonnées, n'avaient produit qu'un gain de 1% sur un F-Score du CRF de 0,77 (voir sur ce point (Béchet & Charton, 2010)). Le module de détection de motifs expérimenté dans cet article introduit un gain de 3% sur un F-Score de CRF de 0,80. Ce gain souligne le potentiel de la méthode. On observe par ailleurs que notre proposition réduit globalement l'écart de performance entre un système à règles et un système statistique complété par des motifs, sur une transcription de référence corrigée telle que NE-Ref de ESTER2.

En terme de précisions, les performances de notre système s'approchent pour plusieurs classes de celles obtenues par le meilleur système à règles linguistiques appliqué sur NE-Ref, lors de la campagne ESTER 2. Ces résultats permettent d'envisager qu'une augmentation de la couverture des formes de surfaces extraites des *métadonnées* (par exemple à la suite d'une augmentation de la quantité de formes de surfaces disponibles dans Wikipédia) puisse fournir d'autres gains de performance.

5 Conclusion et perspectives

Nous avons décrit une méthode d'introduction dans un système d'étiquetage d'entités nommées à base de CRF d'un composant d'identification d'EN exploitant des motifs de détection collectés automatiquement dans un corpus encyclopédique. Il était apparu lors de la campagne ESTER 2 qu'un système CRF correctement entrainé pouvait obtenir les meilleurs résultats sur un corpus bruité, mais que les systèmes d'EEN à règles linguistiques étaient plus performants sur des corpus non bruités. Nous avons donc cherché à évaluer dans quelle mesure le renforcement d'un CRF par des motifs de détection simples pouvait réduire l'écart de performances entre un étiqueteur CRF et un étiqueteur à base de règle sur un document textuel non bruité. Nous avons montré que notre proposition pouvait amener un gain de performances global important sur un système CRF, et améliorer de manière conséquente sa précision. La solution que nous proposons améliore la robustesse d'un système CRF sur des corpus non bruités et réduit l'écart avec un système d'EEN linguistique tout en conservant au CRF son faible coût de développement, l'intégralité du processus d'entrainement de notre système demeurant automatique. La précision du système que nous avons élaboré et sa facilité de déploiement nous ont permis de l'entrainer dans plusieurs versions linguistiques (Français, Espagnol et Anglais) et de l'exploiter en tant que module dans des applications qui prolongent la tâche d'étiquetage d'entités nommées. Nous travaillons en particulier sur la détection de co-références et avons à ce titre déployé cet étiqueteur dans sa version anglaise en tant que composant de l'architecture de détection de co-référence du challenge Grec 2010 ou il a obtenu des résultats satisfaisants [5].

5. Les ressources décrites sont disponibles sous forme d'API et en téléchargement sur `www.nlgbase.org`.

Références

BÉCHET F. & CHARTON E. (2010). Unsupervised knowledge acquisition for extracting named entities from speech. In *ICASSP 2010*, Dallas : ICASSP.

BIKEL D., SCHWARTZ R. & WEISCHEDEL R. (1999). An algorithm that learns whats in a name. *Machine learning*, **7**.

BORTHWICK A., STERLING J., AGICHTEIN E. & R (1998). Exploiting diverse knowledge sources via maximum entropy in named entity. *Proc. of the Sixth*, p. 152–160.

BRUN C., EHRMANN M. & MAUPERTUIS C. D. (2010). Un système de détection d'entités nommées adapté pour la campagne d'évaluation ESTER 2. In *TALN 2010*, volume 2.

BUNESCU R. & PASCA M. (2006). Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of EACL*, volume 6.

CHARTON E. & TORRES-MORENO J. (2009). Classification d'un contenu encyclopédique en vue d'un étiquetage par entités nommées. In *Taln 2009*, volume 1, p. 24–26 : TALN.

CHARTON E. & TORRES-MORENO J. (2010). NLGbAse : a free linguistic resource for Natural Language Processing systems. In LREC, Ed., *LREC 2010*, number 1, Matla : Proceedings of LREC 2010.

DODDINGTON G., MITCHELL A., PRZYBOCKI M., RAMSHAW L., STRASSEL S. & WEISCHEDEL R. (2004). The automatic content extraction (ACE) program–tasks, data, and evaluation. In *Proceedings of LREC*, volume 4, p. 837–840 : Citeseer.

GALLIANO S., GRAVIER G. & CHAUBARD L. (2009). The ESTER 2 Evaluation Campaign for the Rich Transcription of French Radio Broadcasts. In *International Speech Communication Association conference 2009*, p. 2583–2586 : Interspeech 2010.

KAZAMA J. & TORISAWA K. (2007). Exploiting Wikipedia as external knowledge for named entity recognition. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, p. 698–707.

LAFFERTY J., MCCALLUM A. & PEREIRA F. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, p. 282–289 : Citeseer.

MCCALLUM A. & LI W. (2003). Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003* -, p. 188–191, Morristown, NJ, USA : Association for Computational Linguistics.

NOTHMAN J., MURPHY T. & CURRAN J. (2009). Analysing Wikipedia and gold-standard corpora for NER training. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, number April, p. 612–620 : Association for Computational Linguistics.

NOUVEL D., ANTOINE J., FRIBURGER N. & MAUREL D. (2010). An analysis of the performances of the casen named entities recognition system in the ester2 evaluation campaign. *LREC 2010*.

{ERIC.CHARTON, MICHEL.GAGNON, BENOIT.OZELL}@POLYMTL.CA

RAYMOND C. & FAYOLLE J. (2010). Reconnaissance robuste d'entités nommées sur de la parole transcrite automatiquement. In *Traitement Automatique des Langues Naturelles*, volume 1, p. 19–23.

RAYMOND C. & RICCARDI G. (2007). Generative and discriminative algorithms for spoken language understanding. In *Proceedings of Interspeech2007, Antwerp, Belgium*, p.$\tilde{2}$: Citeseer.

TJONG E. & MEULDER F. D. (2003). Introduction to the conll-2003 shared task : Language-independent named entity recognition. In *In CoNLL*.