

Classification d'un contenu encyclopédique en vue d'un étiquetage par entités nommées

Eric Charton¹ Juan-Manuel Torres-Moreno¹

(1) LIA / Université d'Avignon, 339 chemin des Meinajariès, 84911 Avignon
eric.charton@univ-avignon.fr, juan-manuel.torres@univ-avignon.fr

Résumé. L'étiquetage par entités nommées recourant à des ressources d'étiquetage externes, fait l'objet d'un regain d'intérêt avec l'apparition de Wikipédia ou Wordnet. La disponibilité de ces corpus encyclopédiques numériques de grande taille, multilingues, exhaustifs et ouverts, conduit à proposer des solutions de labellisation qui exploitent les connaissances qu'ils contiennent. La mise en correspondance d'une séquence de mot avec un document encyclopédique est possible, mais la classification des étiquettes et des entités encyclopédiques mises en relation n'est pas encore complètement résolue. Le corpus encyclopédique ouvert tel que Wikipédia est en effet classé de manière anarchique, et sa consistance rend parfois difficile la mise relation des ses entités avec un système taxonomique contraint. Dans cet article nous explorons cette difficulté et proposons un système complet pour répondre à ce besoin.

Abstract. The advent of Wikipedia and WordNet aroused new interest in labeling by named entity aided by external resources. The availability of these large, multilingual, comprehensive and open digital encyclopaedic corpora suggest the development of labeling solutions that exploit the knowledge contained in these corpora. The mapping of a sequence to a word document is encyclopedic, but the classification labels and entities being encyclopedic relationship is not yet fully resolved. An open encyclopaedic corpus such as Wikipedia is in fact classified in an anarchic way. Its inconsistency makes it sometimes difficult to establish a relationship between its entities and a restricted taxonomy. In this article we explore this problem and propose a complete system to meet this need.

Mots-clés : Etiquetage, Entités nommées, classification, taxonomie.

Keywords: Named entity recognition, classification, taxonomie.

1 Introduction

Les systèmes d'étiquetage par entités nommées (EEN) font le plus souvent appel à plusieurs familles de ressources : des systèmes tels que celui de (Brun & Hagege, 2008), sont associés à des analyseurs syntaxiques robustes. Ils recourent également à des ressources externes telles que WordNet (Fellbaum, 1998) ou l'encyclopédie en ligne Wikipédia¹. Ces systèmes, qu'ils exploitent des ressources de type automate à états finis (FSM), ou lexicales externes, doivent régulièrement être configurés pour répondre à un besoin d'extraction précis : la bio-technologie ou les textes médicaux (Sasaki *et al.*, 2008) par exemple. Si les systèmes à base d'automates peuvent être rapidement adaptés par une séquence d'apprentissage, le problème posé par les contenus lexicaux externes tels que Wordnet ou Wikipédia est différent. En effet, par nature, les entités encyclopédiques sont très nombreuses (plusieurs millions désormais pour Wikipédia)

1. <http://www.wikipedia.org>

et constituent des réservoirs d'entités pré-existant sur lesquels doit être *appliqué*, pour chaque besoin, une nouvelle cartographie. Or, les analyses que nous avons conduites sur quatre versions linguistiques du corpus de Wikipédia (anglaise, française, allemande et espagnole) font état non seulement de grandes divergences de classement, mais aussi d'une grande disparité pour ce qui est la taille des classes.

Cette question de l'application d'une cartographie de classes à un lexique d'étiquetage est à la fois importante et délicate. Délicate car elle fait intervenir deux disciplines du traitement et d'analyse de la langue, à savoir la classification, et l'extraction d'information. Importante car de la finesse ou de la souplesse d'une catégorisation d'un corpus de connaissance appliqué à la détection d'entité, dépendront les performances finales du système d'extraction et son adaptabilité à des tâches variées.

Cet article propose un système capable de générer un lexique issu de Wikipédia, utilisable pour l'étiquetage d'entités nommées. Ce lexique est constitué par un ensemble de métadonnées, composées de formes de surfaces des termes encyclopédiques, et de sacs de mots utilisable pour la désambiguïsation.

Nous décrivons dans un premier temps les propositions récentes de dictionnaires et de lexiques appliqués à la détection d'entités et reposant sur des contenus encyclopédiques. Nous mettons en perspective ces ressources avec les besoins taxonomiques d'une tâche d'étiquetage par entités nommées. Nous utilisons à titre d'exemple, les besoins rencontrés lors des campagnes d'évaluation Ace², CoNLL³ ou Ester⁴. Nous présentons ensuite le système de classification que nous avons appliqué au corpus Wikipédia pour produire un lexique aisément adaptable aux besoins de l'étiquetage d'entités nommées. Puis nous décrivons les résultats de nos expériences. Nous avons mesuré la qualité des classes d'étiquetage affectées aux entités encyclopédiques par notre système, puis nous avons déployé un système d'étiquetage complet sur des données de la campagne Ester. Nous présentons pour finir nos conclusions et perspectives de développement sur ce projet naissant, et indiquons comment nous mettons à la disposition de la communauté scientifique notre système et les données lexicales qu'il contient.

2 Utilisation de Wikipédia comme lexique d'entités nommées

L'un des premiers systèmes proposés pour exploiter un contenu encyclopédique tel que Wikipédia à des fins de désambiguïsation d'entités nommées est celui de (Bunescu & Pasca, 2006). L'idée est d'extraire de la structure particulière de Wikipédia (pages de redirection, de synonymie, d'entités) des dictionnaires d'entités nommées, et d'associer à chaque entité un sac de mot extrait de sa description encyclopédique. Chaque terme d'un sac de mots est ensuite affecté d'un poids *TF.Idf*. Le principe du système de détection et de désambiguïsation consiste à calculer la *similarité cosinus* entre le contexte d'un mot à étiqueter, et les sacs de mots reliées à des entités candidates issues de Wikipédia.

La finalité de ce système est d'associer, avec précision, un mot détecté avec son entité encyclopédique. Dans cette démonstration, il n'existe pas à proprement parler d'étiquetage selon les standards d'évaluation. Un noyau taxonomique reposant sur les catégories de Wikipédia est entraîné avec des classifieurs du type Machine à Vecteur Support (SVM), mais à fin exclusive de restreindre le champs des recherches de similarités entre les mots d'une requête, et une entité encyclopédique. Il est donc relativement difficile d'appliquer tel quel ce système dans un

2. <http://www.nist.gov/speech/tests/ace/>

3. <http://www.cnts.ua.ac.be/conll2003/ner/>

4. <http://www.afcp-parole.org/ester/index.html>

contexte d'extraction d'entités nommées.

Un système dérivé de celui de Bunescu et Pasca est présenté dans (Jun'ichi & Kentaro, 2007). Il adjoint à l'index d'entités extraites, en exploitant ses liens internes du corpus de Wikipédia, un ensemble de relations complémentaires entre termes. Dans Wikipédia, une relation entre deux documents peut être représentée par un lien `[[BobDylan(Singer)|Dylan]]`. C'est ce type de lien qui est ici repris pour étendre le dictionnaire de formes de surfaces reliées à une entité. Le système de détection d'entités nommées n'est pas directement appliqué à une tâche de reconnaissance d'entités. Ce système – après entraînement d'un classifieur à base de CRF – associe à des étiquettes du corpus de CoNLL (PER, LOC, ORG, MISC) des entités de Wikipédia, mais ne procède pas directement à un étiquetage des entités.

C'est dans (Wisam & Silviu, 2008), que l'on trouve le système le plus proche de la tâche d'étiquetage. Dans ce travail, qui reprend le système de (Bunescu & Pasca, 2006), les pages d'une version linguistique Wikipédia sont classifiées en utilisant le système taxonomique des entités nommées de (Sang & Meulder, 2003). C'est un classifieur SVM qui est déployé pour attribuer une des cinq étiquettes de classes à des pages d'un corpus linguistique de Wikipédia.

3 Normes taxonomiques des entités nommées

Il n'existe pas à proprement parler de standard taxonomiques pour les étiquettes. Néanmoins, on observe que les propositions de classes retenues pour les systèmes d'étiquetage génériques sont généralement issues d'un tronc commun de définitions conceptuelles, proposées lors des campagnes d'évaluation. Dans les systèmes anglophones, ce sont les règles de la campagne MUC⁵ ACE⁶ ou CoNLL qui sont mises en œuvre. Dans le monde francophone, l'unique disponibilité des corpus de la tâche d'étiquetage de la campagne Ester⁷ a fait des règles taxonomiques proposées par ce groupe, un standard de-facto.

De manière générale, on peut observer que les principales entités nommées recherchées sont regroupées dans des classes racines de type Organisationnelles (ORG), Individuelles (PERS), Géographiques (LOC), des descriptions de produits ou d'objets, et plus généralement de toute conception humaine (PROD). À ces définitions d'entités sont adjointes parfois des propriétés (les titres ou fonctions), des entités temporelles telles que les dates (DATE), les horaires (TIME), ou numériques (AMOUNT) telles que les poids ou les mesures.

La granularité des classes se développe régulièrement. Les classes racines se voient étendre lors des versions les plus récentes des évaluations (ACE 2007 et Ester 2), par des sous-classes filles plus précises : ORG.COM ou ORG.NON-PROFIT, etc.

3.1 Autres besoins taxonomiques exprimés

Ces grandes familles de classes sont inadaptées pour répondre à certains besoins émergents ou particuliers : c'est le cas des classes d'étiquetages pour des entités de type biologiques ou chimiques (application à la recherche médicale (Kulick *et al.*, 2004)). Par ailleurs, de nouvelles tâches d'extraction d'information ouvertes au cours des deux dernières années peuvent imposer l'usage d'un étiquetage adapté à une thématique particulière : la construction automatique d'ontologies sur des individus d'après des corpus de pages web ou de texte telles que proposée lors de campagnes telles que *Web People Search* (Artiles *et al.*, 2007) en 2008 et une tâche spéciali-

5. http://www.itl.nist.gov/iaui/894.02/related_projects/muc/

6. <http://nist.gov/speech/tests/ace/2008/doc/ace08-evalplan.v1.2d.pdf>

7. http://www.afcp-parole.org/ester/docs/Conventions_EN_ESTER2_v01.pdf

sée, héritant de ACE, intitulée *Knowledge Base Population*), entamée par le Nist en 2009⁸. Dans ces expériences, les extractions d'information proposées peuvent amener à rechercher des entités de type propriétés telles que les titres, les diplômes, des adresses postales, etc.

3.2 Spécificités de la taxonomie appliquée au corpus Wikipédia

Les versions récentes Wikipédia contiennent de nombreuses entités nommées. Mais pour exploiter cette connaissance, il est important de la cartographier en l'adaptant à des exigences taxonomiques qui varient selon le cadre applicatif. Les particularités taxonomiques d'une campagne d'évaluation par exemple. Or, l'application à Wikipédia d'une cartographie de classes adaptées à l'EEN est une tâche qui se heurte à plusieurs difficultés :

- Il a été souligné lors de précédentes campagnes d'évaluation, et notamment lors du Défi Fouille de texte 2008 (Hurault-Plantet *et al.*, 2008) qui consistait à segmenter un corpus contenant notamment des fiches Wikipédia, qu'il est difficile de dépasser une précision de classification supérieure à 90%, même avec un nombre de classe restreint. Cette précision est insuffisante puisqu'elle implique qu'une tâche d'EEN reposant sur les entités dérivées de Wikipédia pourrait produire 10% d'erreurs.
- La mise au point d'un classifieur non équiprobable (*i.e.* modélisant à la fois des classes de très grande taille et d'autres de petite taille sur un même corpus) est particulièrement délicate, et donc peu adaptée à la modélisation de certaines classes peu représentées dans Wikipédia et pourtant recherchées en EEN (c'est le cas des *fonctions* et des *diplômes* par exemple).
- La réutilisation de la taxonomie interne de Wikipédia pour en dériver un nouveau système de classes adaptées à l'EEN, comme elle a été récemment proposée par (Suchanek *et al.*, 2007) pour le domaine ontologique, est délicate dans celui de l'EEN, car elle implique de transformer un graphe de classes volumineux, granulaire, et dont les nœuds sont très imbriqués en un arbre taxonomique très compact et précis.

On peut résumer la difficulté de la tâche par l'exigence de réduction de classes d'un facteur conséquent : au maximum, quelques centaines de classes sont exigées en EEN. Or, nos mesures statistiques⁹ indiquent que la version française¹⁰ de Wikipédia contient 100 731 catégories et la version anglaise, 256 000. Par ailleurs la distribution de ces catégories peut être représentée par une loi de Zipf-Mandelbrot : dans la version française, la classe la mieux représentée est la catégorie des dates de naissances (regroupants 144 000 entités de types personnes, soit 7,68% du corpus), suivie des décès (69 970 entités soit 3,7% du corpus). A la suite de ces classes volumineuses, on trouve rapidement des milliers de catégories qui caractérisent moins de 0,1% des entités. Et ces micro-classes sont parfois exigées en EEN. C'est le cas des fonctions militaires par exemple (FONC.MIL de la campagne ESTER 2), représentées par 151 entités dans Wikipédia¹¹, ventilées sur trois classes.

On ajoutera qu'à cette difficulté de transformation du graphe de catégories de Wikipédia en arbre taxonomique adapté à l'EEN, s'ajoute celle de la sélection des entités pertinentes pour l'EEN. Wikipédia intègre en effet, 140 000 descriptions encyclopédiques non utiles pour les tâches d'EEN (des noms communs par exemple tel que la fiche *Voiture* ou celle des *dirigeables*) qui sont elles aussi référencées par le système taxonomique de l'encyclopédie, parfois mélan-

8. Voir <http://apl.jhu.edu/paulmac/kbp/090220-KBPTaskGuidelines.pdf>

9. voir notamment www.nlgbase.org, et www.nlgbase.org/fr/stat.html

10. Version du dump XML

11. voir www.nlgbase.org/fr/stat/stat_clust.html

Classification encyclopédique en vue d'un étiquetage par entités nommées

Victor Hugo

Pour les articles homonymes, voir Hugo et Victor Hugo (homonymie).

Victor-Marie Hugo, né le 26 février 1802 à Besançon, mort le 22 mai 1885 à Paris, est un écrivain, dramaturge, poète, homme politique, académicien et intellectuel engagé français considéré comme le plus important des écrivains romantiques de langue française et un des plus importants écrivains de la littérature française.

Son œuvre est très diverse : romans, poésie lyrique, drames en vers et en prose, discours politiques à la Chambre des Pairs, correspondance abondante.

Il a contribué, tout comme Baudelaire, au renouvellement de la poésie et de la littérature.

Sommaire [masquer]

- 1 Biographie
 - 1.1 Enfance et jeunesse
 - 1.2 Le jeune écrivain
 - 1.3 L'exil
 - 1.4 Le retour en France et la mort
- 2 Une œuvre monumentale
 - 2.1 Le romancier
 - 2.1.1 Un romancier inclassable
 - 2.1.2 Une œuvre de combat
 - 2.2 Le dramaturge
 - 2.3 Le poète
 - 2.3.1 Vers de jeunesse
 - 2.3.2 La première maturité
 - 2.3.3 L'exil
 - 2.3.4 Une place à part dans son siècle
 - 2.4 Le témoin voyageur
- 3 Sa pensée politique
 - 3.1 Politique intérieure
 - 3.2 La Commune
 - 3.3 Combats sociaux

Victor Hugo

Victor Hugo en 1852

Activité(s)	Écrivain Poète Dramaturge Personnalité politique Dessinateur
Naissance	26 février 1802 Besançon
Décès	22 mai 1885 Paris
Mouvement(s)	Romantisme

FIGURE 1 – La fiche de Victor Hugo dans Wikipédia contient un texte descriptif et une boîte d'information (sur la droite). Dans le cas de cette fiche, la boîte d'information est celle des écrivains.

Catégories : Bon article | Victor Hugo | Besançon | Personnalité du XIXe siècle | Écrivain français du XIXe siècle | Écrivain romantique | Poète français du XIXe siècle | Poète romantique français | Dramaturge français du XIXe siècle | Opposant au Second Empire | Philhellène | Personne inhumée au Panthéon de Paris | Membre de l'Académie française | Naissance en 1802 | Naissance à Besançon | Décès en 1885 | Décès à Paris | Naissance dans le Doubs | [+]

FIGURE 2 – Les catégories associées à la fiche Victor Hugo de Wikipédia.

gées avec des entités elles mêmes pertinentes¹². Ces entités non désirées doivent être isolées pour ne pas introduire de bruit dans la tâche de détection. Elles sont référencées dans notre système par le label UNK.

4 Approches de classification adaptées à Wikipédia

Nous avons développé un ensemble d'algorithmes de classification qui permettent d'extraire et de structurer les contenus encyclopédique, mais aussi de transformer la classification interne très complexe et ouverte de Wikipédia.

Considérons pour la mise au point de notre système une entité encyclopédique et ses éléments constitutifs :

- Elle contient un texte qui peut être analysé par un classifieur numérique (voir figure 1).
- Des "Catégories" lui sont attribuées, qui font partie d'un graphe qui peut être exploré et partiellement décrit par un arbre taxonomique (voir figure 2).
- Elle contient des "Boîtes d'information" ou "Infobox" qui décrivent des propriétés standardisés de l'objet décrit par l'entité encyclopédique. Par exemple pour un écrivain, sa date de naissance, son lieu de naissance, etc. (voir figure 1)

Notre idée est de tirer partie de l'existence de ces trois éléments susceptibles de participer d'une classification. Dans un tel cadre applicatif, nous nous trouvons face à un jeu d'ensembles, avec

12. Voir par exemple la classe *Catégorie :Université* de la version française de Wikipédia qui inclut aussi bien des établissements, que des personnes, ou des concepts

des intersections plus ou moins prononcées suivant des règles à finalité divergente, mais qu'il est possible de modéliser par une approche statistique.

Notre système fonctionne en deux phases que nous intitulerons W_k1 et W_k2 :

Phase W_k1 : première classification

La **première phase** permet de créer une classification générale, du niveau de l'état de l'art pour ce qui est de la classe racine (par exemple la classe descriptive des *Organisation*, ORG), et moins exhaustive et précise pour ce qui est du second niveau d'étiquetage de l'arbre taxonomique (ORG.DIV ou LOC.GEO). Ceci est la phase W_k1 :

- Une étape de classification numérique état de l'art appliqué à la racine des classes (par exemple ORG, PERS, LOC ou PROD)
- Une étape de classification par règles établies d'après des catégories sélectionnées au sein de la taxonomie de Wikipédia, pour le second niveau de classe (par exemple ORG.DIV, LOC.ADMI ou PERS.HUM).
- Une étape de classification par règles établies d'après des catégories pour les micro-classes (tel que les titres nobiliaires ou les fonctions).

À l'issue de cette première phase nous obtenons une première classification qui décrit le second niveau taxonomique très partiellement. Par exemple 85% à 89% des lieux sont correctement étiquetés LOC, mais moins de la moitié des éléments de sous classe LOC.GEO ou LOC.ADMIN ne sont détectés. Il faut donc affiner le processus.

Phase W_k2 : perfectionnement de la classification

Nous utilisons les informations déjà disponibles pour procéder à un nouvel apprentissage suivi d'une **nouvelle phase** de classification, la phase W_k2 :

- L'apprentissage consiste à inventorier pour chaque document Wikipédia contenant une *boîte d'information* ou *Infobox*, la classe attribuée lors de la phase précédente.
- Cette information disponible, à évaluer la probabilité qu'une classe soit associée à une *Infobox* puis à élaborer des règles d'association entre *Infobox* et classes (par exemple *Infobox Communes de France* sera associée à LOC.ADMI).
- A reclasser le corpus Wikipédia en utilisant comme règle de détection de classe la présence d'une *Infobox*

En théorie on peut considérer que si la seconde phase de détection dite W_k2 s'applique à la totalité des pages contenant une *Infobox*¹³, et que les classes associées aux *Infobox* sont fiables à 99%, il est possible d'envisager que l'augmentation de qualité finale de la classification – par rapport à W_k1 – soit de l'ordre de 8% à 15%.

4.1 Classification numérique

Le système numérique utilisé pour W_k1 est une fusion ternaire des résultats produits par un classifieur SVM, un classifieur Bayésien Naïf et BoosTexter, tel que présenté dans (Charton *et al.*,). Les classes de textes sont construites après pré-traitement : à savoir, linéarisation des classes par comparaison des distributions de Zipf, utilisation de trigrammes et applications d'un anti-dictionnaires.

13. Mesure réalisée sur le dump XML Wikipédia référencé frwiki-20081201-pages-articles.xml disponible sur download.wikipedia.org

La fusion par vote majoritaire que nous avons déployée est triviale. Elle consiste à confronter les propositions des trois meilleurs classifieurs (SVM, Naïve Bayes, Icsiboost) pour chaque document. Si une majorité l'emporte (2/3 ou 3/3), la classe majoritaire est choisie, dans le cas contraire, la stratégie de fusion se replie sur le système le plus performant (en l'occurrence SVM sur tâche1-catégorie et tâche2-catégorie et icsiboost sur tâche1-genre).

Ce système n'est efficace que sur des classes de poids équivalents et n'est donc déployé que sur les classes PERS, PROD, ORG, LOC et UNK. Les classes FONC et DATE (représentant moins de 1% du corpus) ne sont pas modélisables avec ce système de classification, tous comme les sous-classes.

4.2 Classification d'après des infobox

Pour exécuter la phase W_k2 nous utilisons les *infobox* de Wikipédia afin d'introduire un second niveau de classement des entités encyclopédiques.

Soient un ensemble D de documents issus de Wikipédia, un ensemble C de catégories de Wikipédia et un ensemble I d'*Infobox* de Wikipédia. Dénommons *étiquette de second niveau* (ESN) la sous classe d'une classe racine. Considérons un ensemble de documents $E \in D$ appartenant à une catégorie $c \in C$, un ensemble $F \in D$ munis d'une d'*infobox* $i \in I$, et un ensemble de documents $G \in D$ étiqueté par un label $l \in ESN$.

On peut déduire que l'ensemble U composé de g document constituant l'intersection $U = E \cap F \cap G$ possède pour tout U_g un ensemble de propriétés identiques pour $U.c$, $U.i$ et $U.l$.

On peut donc déduire que l'*Infobox* i correspond à l' ESN_l .

Ceci nous permet d'élaborer automatiquement en partant d'un petit groupe de catégories représentatives de chaque ESN la table des associations entre les *Infobox* et les étiquettes ESN . Nous utilisons cette table d'associations pour reclasser toutes les entités de Wikipédia en détectant la présence éventuelle d'une *Infobox* dans le document qui les décrit.

4.3 Classification d'après des catégories

Un résidu de documents du corpus de Wikipédia ne peut être classé par les deux méthodes précédentes : leur contenu est soit trop faiblement informatif pour ce qui est de la classification numérique, soit ils ne contiennent pas d'*Infobox*. Pour ces documents, nous prévoyons dans la phase W_k2 une dernière étape qui consiste à associer à une catégorie de Wikipédia une étiquette. Cette méthode peut être très performante si une classe est fortement représentative, mais très coûteuse lorsque les catégories sont mal définies ou très granulaires.

On notera, par exemple, que plus de 144 000 entités encyclopédiques sont associées à la catégorie *Naissance en* : ce type de catégorie utilisé comme règle de détection augmente considérablement la précision de la classe PERS.HUM. En revanche, la catégorie *Locution ou expression latine* qui devrait catégoriser des entités de type UNK n'est que d'une faible utilité avec les 124 éléments encyclopédiques qu'elle contient¹⁴. Les règles catégorielles, du fait de leur variabilité, ne peuvent donc être employées que pour affiner ou renforcer l'étiquetage des entités, ou pour traiter des micro-classes non modélisables par méthodes numériques.

14. Voir les analyses de catégories sur www.nlgbase.org/fr/stat/stat_cat.html

5 Expériences et résultats

Pour évaluer l'intérêt de notre dictionnaire d'entités nous avons procédé à un ensemble d'expériences et de mesures. Le standard retenu pour nos expériences est celui établi par la campagne Ester 2 pour ce qui est de la taxonomie des entités nommées, et de la vérification de couverture du dictionnaire. Seules les entités racine et de second niveau des familles PERS, ORG, LOC, PROD et FONC ont été retenues pour ces expériences. Les étiquettes de type AMOUNT, TIME, ne sont pas concernées par le dictionnaire extrait de Wikipédia et ne sont donc pas mesurées ici. Les étiquettes de type DATE qui sont présentes en tant que descriptifs de dates historiques dans Wikipédia, sont mesurées à titre indicatif mais ne sont pas exploitées dans le cadre applicatif de la détection.

Nous avons tout d'abord évalué la qualité de la classification des entités de Wikipédia obtenues avec les algorithmes W_k1 et W_k2 en mesurant la précision et le rappel d'après un échantillon de référence. Puis nous avons introduit le dictionnaire produit d'après Wikipédia dans un système d'étiquetage inspiré de la proposition de (Bunescu & Pasca, 2006), et évalué ses performances sur une partie des transcriptions étiquetées de la campagne ESTER 2.

5.1 Mesure de la classification

Le résultat de l'attribution de classes d'étiquetages aux entités nommées représentées par les entités encyclopédiques de Wikipédia est présenté dans le tableau 1. Les données de références utilisées pour mesurer la précision et le rappel sont fournies par un corpus de 5 500 entités extraites au hasard de Wikipédia. 4 800 entités de références sont étiquetées de manière semi-automatique, et corrigées à la main. 700 entités de références sont entièrement étiquetées à la main.

Class	W_k1 Classification numérique			W_k2 : Classification complète		
	(\bar{p})	(\bar{r})	(\bar{F} -m)	(\bar{p})	(\bar{r})	(\bar{F} -m)
Pers	0.92	0.91	0.92	0.95	0.97	0.96
Org	0.74	0.83	0.79	0.82	0.89	0.87
Loc	0.85	0.89	0.87	0.93	0.96	0.945
Prod	0.80	0.94	0.86	0.90	0.95	0.925
Unk	0.95	0.78	0.85	0.96	0.92	0.939
Date	-	-	-	1	1	1
Fonc	-	-	-	1	1	1
Total			0.86			0.95

TABLE 1 – Precision (\bar{p}), Rappel (\bar{r}) Et FScore (\bar{F} -m) obtenus sur le jeux de test

Ce tableau est divisé en deux sections : la première fait état du F-Score¹⁵ obtenu pour chaque classe avec la première classification numérique par l'algorithme W_k1 , la seconde l'amélioration obtenue après utilisations des modes de détection de classes complémentaires offerts par les classifieur de l'algorithme W_k2 . On notera que les scores détaillés des sous classes introduites dans l'algorithme W_k2 ¹⁶, ne sont pas présentés ici, seule la progression des scores globaux de classification sont comparables donc décrits.

15. Mesure harmonique combinant la précision et le rappel.

16. Voir le détail des sous classes sur www.nlgbase.org/fr/stat/stat_clust.html

À l'exception de la classe ORG qui pose des problèmes de modélisation spécifiques¹⁷, le niveau de précision et rappel obtenus sur les entités classées semble suffisamment élevé (de l'ordre de 95% de précision) pour envisager une utilisation dans une tâche de NER.

5.2 Mesure de l'étiquetage

L'étiquetage est réalisé sur les transcriptions pré-étiquetées et manuellement corrigées de la campagne ESTER 2. Nous avons sélectionné un ensemble de 3 transcriptions représentatives de ce corpus¹⁸, et avons mesuré le Slot Error Rate (SER)¹⁹. Le SER est mesuré en distinguant trois types d'erreurs : *I* les insertions, qui sont des entités détectées et qui n'ont aucun mot commun avec une entité de référence, *D* les délétions ou entités manquées par le système, et *S* les entités substituées, c'est à dire correspondant de manière incorrecte à des entités de référence. Cette dernière mesure est particulièrement intéressante dans notre cadre applicatif puisqu'elle correspond aux erreurs de classification des entités. En notant *R* l'ensemble des entités de références on obtient la formule de calcul de SER comme suit : $SER = (I + D + S)/R$.

Entités	R	I	D	S	SER
ORG	240	23	5	10	0.15
PERS	1071	15	60	6	0.07
LOC	190	19	6	2	0.14
PROD	110	21	0	2	0.20
FONC	26	10	0	7	0.65
Total	1637	88	71	27	0.12

Ces transcriptions représentent 35 000 mots, et 1637 entités à étiqueter. Nous ne mesurons pas dans cette expérience les performances de détection sur les classes d'EN temporelles (DATE, AMOUNT). Nous n'avons pas mesuré dans cette expérience la granularité des entités, laissant ces travaux pour une publication future plus spécifique à l'activité d'étiquetage. On observe dans ces résultats un bon niveau de performance sur les entités de type PERS (SER < 7%) qui sont par ailleurs particulièrement bien retrouvées et étiquetées par notre système (F-Score de 0.96), dans le corpus de Wikipédia.

L'étiquette de fonction FONC pose des difficultés particulières : cette classe est non seulement difficile à construire d'après les informations trop fragmentaire fournies par Wikipédia, et la détection de ses entités par des systèmes à base de règles ou d'automates semble plus appropriée pour son étiquetage. On doit probablement voir dans cet particularité que le contexte d'une entité de type attribut (comme une fonction) est moins bien mesurée par similarité cosinus avec l'algorithme de (Bunescu & Pasca, 2006) qu'une entité de type nom propre.

6 Conclusions

Nous avons présenté un système capable de produire d'après un corpus encyclopédique tel que Wikipédia un dictionnaire d'entités nommées étiquetées et prêtes à l'emploi dans une tâche de

17. La classe ORG prévue par la campagne ESTER 2 fait cohabiter des partis politiques, des organisations commerciales, des entités géopolitiques et des émissions de divertissements. Ces entités sont suffisamment différentes pour rendre délicate la mise au point de leur classe de détection

18. L'expérience est reproductible sur www.nlgbase.org/perl/nlgetiq.pl

19. J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel, Performance measures for information extraction, in Proceedings of DARPA Broadcast News Workshop, February 1999

NER et utilisant des mesures de similarité pour désambigüiser des entités en contexte. Les performances obtenues nous permettent d'envisager que ce type de ressources peut faire progresser significativement les applications de NER.

Le corpus d'entités classées que nous proposons est disponible en téléchargement²⁰. Avec cette mise à disposition nous espérons que notre contribution permettra à des chercheurs désireux de développer leur propre système d'étiquetage, de gagner du temps.

D'ores et déjà, notre corpus d'entités labellisées permet d'extraire facilement des sous-lexiques spécialisés de noms propres, de noms de personnes, ou de lieux. À terme, nous envisageons de faire évoluer nos algorithmes vers un système d'étiquetage et d'extraction d'information applicables dans des campagnes tels que Weps ou KBP.

Références

ARTILES J., GONZALO J. & SEKINE S. (2007). The semeval-2007 weps evaluation : Establishing a benchmark for the web people search task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, p. 64–69, Prague, Czech Republic : Association for Computational Linguistics.

BRUN C. & HAGEGE C. (2008). Vérification sémantique pour l'annotation d'entités nommées. In *Actes TALN08 Avignon*.

BUNESCU & PASCA (2006). Exploiting wikipedia as external knowledge for named entity recognition. In *ACLWEB Antology 2006*.

CHARTON E., CAMELIN N., ACUNA-AGOST R., GOTAB P., LAVALLEY R., KESSLER R. & FERNANDEZ S. Pré-traitements classiques ou par analyse distributionnelle : application aux méthodes de classification automatique déployée pour deft08. p. 101–110.

C. FELLBAUM, Ed. (1998). *WordNet An Electronic Lexical Database*. Cambridge, MA ; London : The MIT Press.

HURAUULT-PLANTET M., BERTHELIN J.-B., AYARI S. E., GROUIN C., LOISEAU S. & PAROUBEK P. (2008). : limsi.fr.

JUN'ICHI K. & KENTARO T. (2007). Exploiting wikipedia as external knowledge for named entity recognition. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, p. 698–707.

KULICK S., BIES A., LIBERMAN M., MANDEL M., MCDONALD R., PALMER M., SCHEIN A. & UNGAR L. (2004). Integrated annotation for biomedical information extraction. In *HLT/NAACL 2004 Workshop : Biolink 2004*, pp. 61-68.

SANG E. T. K. & MEULDER F. D. (2003). Introduction to the conll-2003 shared task : Language-independent named entity recognition. In *Proceedings of CoNLL-2003, Edmonton, Canada, 2003*, pp. 142-147.

SASAKI Y., TSURUOKA Y., MCNAUGHT J. & ANANIADOU S. (2008). How to make the most of ne dictionaries in statistical ner. *BMC bioinformatics*, **9 Suppl 11**.

SUCHANEK F. M., KASNECI G. & WEIKUM G. (2007). Yago, a core of semantic knowledge. In *Proceedings of WWW 2007*.

WISAM D. & SILVIU C. (2008). Augmenting wikipedia with named entity tags. In *ACL Proceedings of the Third International Joint Conference on Natural Language Processing*.

20. Consulter www.nlgbase.org