RECITAL 2009, Senlis, 24–26 juin 2009

Combinaison de contenus encyclopédiques multilingues pour une reconnaissance d'entités nommées en contexte

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

Résumé. Dans cet article, nous présentons une méthode de transformation de Wikipédia en ressource d'information externe pour détecter et désambiguïser des entités nommées, en milieu ouvert et sans apprentissage spécifique. Nous expliquons comment nous construisons notre système, puis nous utilisons cinq éditions linguistiques de Wikipédia afin d'enrichir son lexique. Pour finir nous réalisons une évaluation et comparons les performances du système avec et sans compléments lexicaux issus des informations inter-linguistiques, sur une tâche d'extraction d'entités nommées appliquée à un corpus d'articles journalistiques.

Abstract. In this paper, we present a way to use of Wikipedia as an external resource to disambiguate and detect named entities, without learning step. We explain how we build our system and why we used five linguistic editions of the Wikipedia corpus to increase the volume of potentially matching candidates. We finally experiment our system on a news corpus.

Mots-clés : Etiquetage d'entités nommées, ressources sémantiques.

Keywords: Named entity labeling, semantic resources.

1 Introduction

L'extraction d'entités nommées (EEN) consiste à localiser précisément des assemblages de lettres ou de mots en leur attribuant une classe sémantique. Ces assemblages - que nous appellerons des *formes de surface* - peuvent être des acronymes, des mots ou des groupes de mots correspondant de manière générale à des concepts tels que des produits, des organisations, des personnes ou encore des lieux.

Attribuer une classe sémantique exacte à une expression écrite est une tâche difficile. Pour une même entité sémantique, il existe souvent plusieurs formes de surface. A titre d'exemple, considérons les noms d'entreprises qui sont aussi bien exprimés sous forme d'acronymes que de mots isolés ou groupés (*IBM* pour *International Business Machine*). Citons encore les formes de surface qui, pour décrire un même lieu, peuvent être composées d'un ou plusieurs mots (*Paris, Paris Intra Muros*), mais aussi prendre la forme de descriptions familières ou d'expressions (dans le cas de Paris, *Paname, la Ville Lumière*). Par ailleurs, à une forme de surface identifiée peuvent aussi parfois correspondre plusieurs entités sémantiques (Le paquebot *Ville d'Alger*, ou la ville du même nom), que seule l'analyse du contexte permet de départager.

Dans cet article, nous décrivons un système d'EEN dont l'originalité est de collecter un maximum de formes de surface disponibles pour une même entité sémantique en exploitant plusieurs versions linguistiques de l'encyclopédie Wikipédia [1]. Nous utilisons ensuite un algorithme d'EEN reposant sur le calcul de la similarité cosinus entre le contexte textuel d'une

1. www.wikipedia.org, promu par la fondation www.wikimedia.org.

forme candidate identifiée et les mots contenus dans l'article encyclopédique qui lui correspond. Pour évaluer l'influence de l'apport de formes de surface issues d'autres langues que celle de départ sur le processus d'EEN, nous mesurons les performances obtenues par notre système avec les seules formes de surface issues du corpus français de Wikipédia puis celles obtenues après introduction de formes de surface complémentaires extraites de corpus Wikipédia d'autres langues. Les résultats obtenus démontrent que l'introduction de formes de surface collectées dans plusieurs éditions linguistiques de Wikipédia peut améliorer les capacités de détection d'un système d'EEN.

Cet article est organisé comme suit. Dans la section 2, après une rapide description des systèmes existants et de leurs évolutions récentes vers des propositions reposant sur des ressources lexicales ou encyclopédiques, nous justifions nos propres choix. Dans la section 3 nous présentons notre lexique de formes de surface d'entités nommées (EN) et le système d'apprentissage conçu pour le construire. Dans la section 4 nous décrivons un système complet de détection et de désambiguïsation d'EN. Enfin, dans la section 5, nous présentons les résultats de nos expériences, puis nous concluons sur nos projets futurs.

2 Systèmes de détection d'entités nommées

De nombreuses propositions ont été faites pour résoudre le problème posé par la tâche d'extraction et de reconnaissance des EN (Nadeau & Sekine, 2007). A la fin des années 90, les systèmes de référence (Lafferty *et al.*, 2001; Favre *et al.*, 2005) réalisaient la tâche d'EEN en utilisant des réseaux de connaissances et d'identifications obtenus d'après des automates à états finis ou des modèles stochastiques appris sur un corpus pré-étiqueté. Ces méthodes qui reposent sur un entrainement, posent le problème de l'exhaustivité des corpus d'apprentissage. Pour modéliser finement les motifs et séquences à détecter, il faut disposer des corpus d'apprentissage de très grande taille, étiquetés et donc coûteux. On a donc souvent complété les systèmes automatiques par des éléments lexicaux (dictionnaires de noms propres, de villes, de lieux) afin d'accroître leur robustesse.
Malgré cela, les meilleurs systèmes hybrides (utilisant apprentissage et connaissances lexicales) ont parfois du mal à résoudre le problème des mots hors vocabulaires (Out Of Vocabulary abrégés par OOV) posé par l'apparition continuelle de nouvelles entités dans le langage. Pour répondre à ce problème, une nouvelle génération de systèmes d'EEN reposant sur des connaissances encyclopédiques (Bunescu & Pasca, 2006; Juníchi & Kentaro, 2007) a récemment vu le jour. La plupart de ces systèmes utilisent l'encyclopédie collaborative et multilingue Wikipédia. Cette ressource contient plus de 6 millions d'entités [2] couvrant un spectre très large de concepts (des objets, des individus, des lieux).
Un des avantages de cette ressource encyclopédique est que son contenu est très rapidement mis à jour et augmente continuellement. Dans Wikipédia, les articles concernant des événements ou des nouveautés sont parfois incorporés quotidiennement. Ces mises à jour sont diffusées de manière instantanée [3]. Cette possibilité est actuellement l'une des plus performantes qui soit donnée pour résoudre le problème des OOV.
L'aspect multilingue, interconnecté et encyclopédique de Wikipédia peut, lui aussi, se révéler essentiel dans le cadre applicatif d'un système d'EEN. Un nom étranger ou des formes de ce

2. Décompte réalisé sur les fichiers publics de Wikipédia au 16 Oct 2008 dans les 7 langues les plus utilisées à savoir l'Anglais, l'Allemand, le Français, l'Espagnol, l'Italien, le Portugais, le Polonais.

3. Sous forme de dump XML régulièrement diffusés, complétés par la fonction "Modifications Récentes" du logiciel Média Wiki.

nom (d'entreprise, d'individu) peuvent apparaitre dans un corpus spécifique à une langue et être absents de celui d'une autre : or, cette connaissance interlinguale peut s'avérer très utile dans le cadre de l'EEN. On peut illustrer cette affirmation par la forme *AMD* (le nom d'une entreprise fabricant des composants électroniques) qui est la seule connue dans le corpus francophone, alors que la forme *Advanced Micro Device* n'existe que dans le corpus anglophone [4]. Pourtant ces deux formes de surface d'un même concept peuvent être rencontrés dans un texte en français pour faire référence à la société désignée. Nous souhaitons exploiter cette possibilité de découvrir dans un corpus linguistique une représentation sémantique utilisable dans une autre langue pour améliorer l'étiquetage d'EN.
Le système que nous proposons construit une représentation lexicale et statistique d'un concept sémantique d'après un article encyclopédique, en utilisant plusieurs versions linguistiques de Wikipédia. Nous intitulerons *metadata* cette représentation. Chaque *metadata* contient un graphe des formes de surface correspondant à une entité nommée à détecter. Ce graphe contient des formes de surface extraites des cinq éditions linguistiques principales de Wikipédia (anglais, allemand, français, italien, espagnol). Ces graphes sont construits d'après des liens interlingues (correspondance d'un terme entre plusieurs corpus linguistiques), des redirections internes (plusieurs formes d'écritures d'un mot qui dirigent vers une unique page encyclopédique) et des pages descriptives d'homonymie. Le jeu complet de *metadata* en français couvre la description de 215.287 personnes, 172.340 lieux, 72.519 organisations et 90.917 produits [5]. Ces concepts sont représentés par un total de 901.592 formes de surface. Pour réaliser la tâche de reconnaissance des EN et leur extraction ou labellisation, un jeu d'algorithmes utilise les *metadata*. Pour la partie détection et désambiguïsation de notre travail, la proposition théorique que nous avons retenue est celle de (Bunescu & Pasca, 2006). Ces auteurs utilisent une mesure de similarité cosinus entre les mots issus d'un article de Wikipédia et le contexte d'une séquence de mots à identifier. Par ailleurs, pour que notre système soit en mesure d'étiqueter des EN avec un jeu de classes réduit tel que celui rencontré lors des campagnes d'évaluation Ester 2 [6], ACE (NIST, 2007) ou ConLL (Sang & Meulder, 2003), nous avons développé un système de classification des articles de Wikipédia en un arbre à 4 classes sémantiques, compatibles avec les normes d'étiquetage d'EN, selon une proposition à base de SVM appliqués à Wikipédia proche de celle de (Wisam & Silviu, 2008) mais légèrement améliorée.

3 Construire un système pour apprendre les informations sur les entités

Notre système de construction des *metadata* assure une transformation de la structure interne de Wikipédia en une représentation sémantiques et statistique. Dans la section 3.1, nous définissons la structure interne des corpus Wikipédia. Dans la section 3.2 nous décrivons de manière formelle une *metadata* : chaque enregistrement de *metadata* est composé du nom de l'article encyclopédique utilisé en tant que clé, d'un graphe représentant les formes de surface potentielles d'une EN, d'un label représentant une classe sémantique et d'un ensemble de mots et leurs poids $tf.idf$. Nous expliquons quels éléments d'un article encyclopédique sont utilisés pour construire une *metadata*. Puis nous décrivons dans la section 3.3, l'algorithme déployé pour procéder à la transformation de Wikipédia en *metadata*. Nous illustrons, pour finir, ces étapes par un exemple.

4. ce cas est vérifiable à cette adresse http ://www.nlgbase.org/perl/display.pl ?query=Advanced Micro Devices&search=FR.

5. Comptage par le système NLGbAse, voir statistiques à jour sur www.nlgbase.org.

6. Articles à paraître, voir http ://www.afcp-parole.org/ester/index.html.

3.1 Structure de la ressource encyclopédique Wikipédia

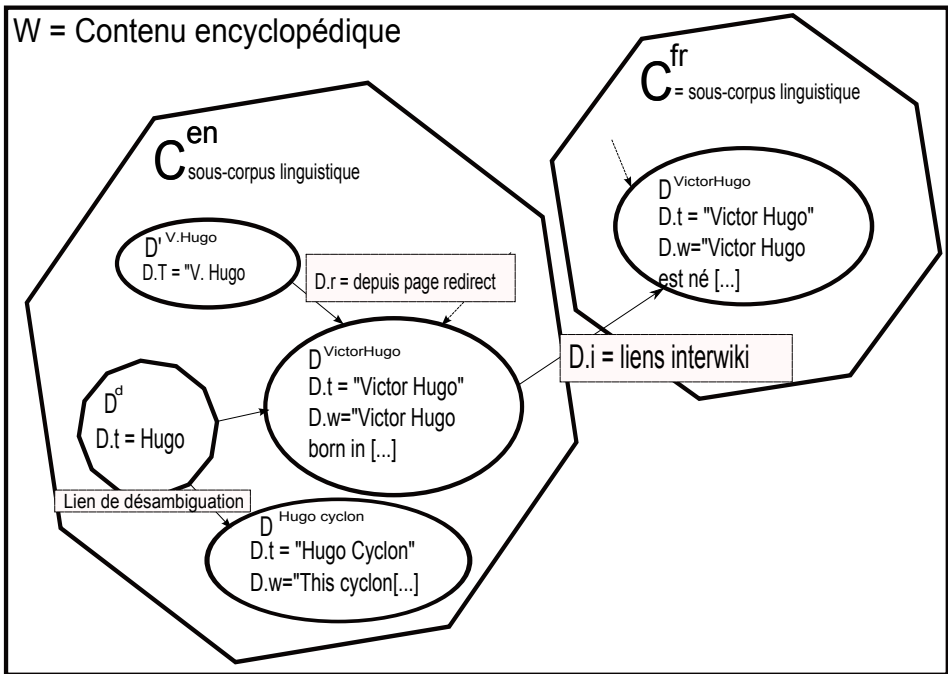

FIGURE 1 – Structure des données dans Wikipédia

Dans notre application, considérons M qui représente l'encyclopédie Wikipédia et C^l un corpus représentant une version linguistique. Cette version linguistique contient des **articles** structurés d'après une DTD [7]. Un ensemble d'*espaces de nom* [8] décrit par la DTD permet de reconnaitre les articles (contenus dans l'espace encyclopédique) des autres documents (Modèles, Utilisateurs, Images). Chaque article contient en ensemble de mots, en relation avec le concept encyclopédique qu'il décrit.
Considérons D un article du corpus Wikipédia C^l, définit par des propriétés :

- $D.t$ est un titre composé d'une séquence de mots.
- $D.w$ est la liste des mots contenus dans l'article.
- Si un article est susceptible d'ambigüité par homonymie de son titre descriptif, une page spéciale intitulée **Page d'homonymie** dans la version française ou **Disambiguation page** dans la version anglaise est créée. Nous intitulerons ces pages $D^d \in C^l$. Elles contiennent plusieurs références à des pages $D \in C^l$ à désambiguïser.
- Des articles uniques dits **Pages de redirection** existent pour répertorier les noms alternatifs susceptibles de correspondre à un article de Wikipédia. Nous utilisons ces pages pour générer les graphes de formes de surface. Nous appelons D^r une *page de redirection* et $D^r.R$ le lien de redirection unique qu'elle contient et qui correspond au nom original $D.t$ de l'article D vers lequel elle redirige.
- Chaque article contenu dans une version linguistique de Wikipédia peut inclure des liens vers les articles similaires contenus dans d'autres versions linguistiques de l'encyclopédie. Ces liens sont intitulés des **Interwiki**. Nous intitulons *relation interwiki* $D.i$, le lien de D vers son équivalent dans un corpus d'une autre langue $C^{l'}$ de M.

7. Consulter http ://meta.wikimedia.org/wiki/Wikipedia.DTD.
8. Lire fr.wikipedia.org/wiki/Aide :Espace_de_noms pour des précisions sur l'espace de nom.

3.2 Les *metadata* produites d'après la ressource encyclopédique

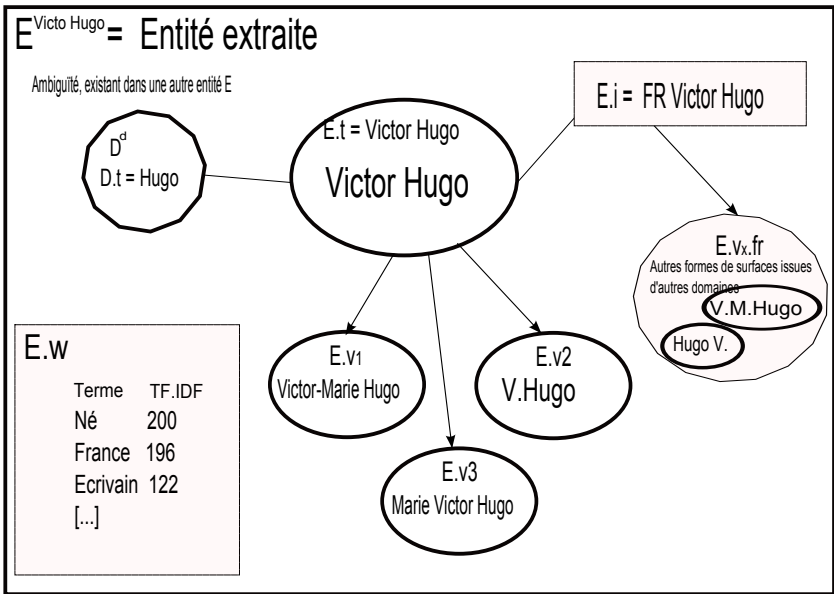

FIGURE 2 – Structure des données d'une *Metadata*

Les *metadata* de notre application sont composées d'un jeu d'entités E dérivées de $D \in C^l$. Chaque entité $E \in C^l$ est définie par un jeu de propriétés $(E.t, E.v, E.w, E.i, E.k)$.

- $E.t$ est le titre de l'entité, correspondant à $D.t$ titre unique d'une page Wikipedia.
- $E.v$ est l'ensemble de toutes les formes de surface qui peuvent écrire E. $E.v$ contient donc des formes de surface synonymes. Cet ensemble est construit d'après les *pages de redirection* D^r reliées à la page D utilisée pour construire E. On notera que lorsqu'une *page d'homonymie* D^d existe, contenant des liens vers la page D utilisée pour construire E, le titre $D^d.t$ de D^d est inclus dans $E.v$.
- $E.i$ est l'ensemble de *relations interwiki* contenues dans la propriété $D.i$ correspondante. $E.i$ représente la relation entre $E \in C^l$ et tout $E' \in C^k, k \neq l$. Ceci signifie que $E.i$ contient les références des entités correspondantes dans les autres corpus linguistiques que celui de départ.
- $E.w$ est un ensemble de mots avec leurs poids (exprimés sous la forme d'une valeur $tf.idf$) associés à l'entité. Cet ensemble de propriétés est construit d'après le texte $D.w$ contenu dans l'article D original de Wikipédia, utilisée pour construire E.

Il est essentiel d'associer aux entités extraites depuis Wikipédia une classe pour rendre la détection exploitable mais aussi conforme aux spécifications des campagnes d'évaluation.

- Nous ajoutons donc aux *métadata* une propriété $E.k$ qui est un label d'étiquetage *personne, produit, lieu* ou *organisation* en accord avec le standard ESTER 2 [9]. Une classe spécifique intitulée *unknown* est par ailleurs introduite pour retirer de la liste de détection des EN les articles relatifs à des descriptions encyclopédiques inutiles pour la tâche d'EEN (un théorème mathématique, une mode).

La méthode de classification retenue est basée sur une combinaison de trois classifieurs : SVM-Lib, un classifier bayésien naïf, Icsiboost (AdaBoost). Plusieurs méthodes de normalisation sont appliquées aux textes utilisés pour construire les classes. Cette méthode est décrite en détails dans (Charton *et al.*, 2008) et a été déployée lors de la campagne d'évaluation DEFT'08 (Grouin *et al.*, 2008) sur un corpus incluant notamment des données de Wikipédia.

9. Voir la convention d'annotation sur http ://www.afcp-parole.org/ester/docs/Conventions_EN_ESTER2_v01.pdf.

3.3 Transformation d'un article encyclopédique en *metadata*

Une description formelle de l'algorithme de construction d'une *metadata* d'après un article encyclopédique peut être présentée comme suit :
Considérons que tout article Wikipedia $D \in C^l$ est défini par une ensemble de propriétés $(D.t, D.w, D.r, D.i)$ telles que $D.t$ est un titre représenté par une séquence de mots, $D.w$ un ensemble de mots, $D.r$ une relation unique entre D et n'importe quel élément de C^l, $D.i$ est un ensemble de *relations interwiki* entre des éléments de D et n'importe quel élément de $C^{l'} \in M \backslash C^l$. Pour générer une table T^l d'entités, nous explorons dans un premier temps C^l et conservons tout élément de D qui n'est pas une *page de redirection* (D^r) ou de *Page d'homonymie* (D^d). On considère que E et D sont en relation si et seulement si $E.t = D.t$. Ceci est formalisé par $E \rightarrow D$. Puis, pour tout E de la table T^l, nous cherchons tous les $D^d \in C^l$ en relation avec E et incluons le nom alternatif $D^d.t$ dans $E.v$. Nous cherchons aussi tous les $D^d \in C^l$ en relation avec E et définissons $E.v = D^d.t$.
Ces étapes sont répétées pour les corpus de toutes les versions linguistiques de Wikipédia retenues. Nous obtenons donc pour chaque corpus linguistique de Wikipédia C^l une entité E^l. Puis, pour procéder à l'agrégation des formes de surface localisées dans ces différentes versions, nous collectons les *relations interwiki* disponibles dans la section $D.i$ de l'article encyclopédique et les attribuons à la propriété $E.i$. Nous agrégeons toutes les formes de surface $E^{l.v}$ dans l'entité $E.v$.

3.4 Exemple de transformation d'un article encyclopédique en *metadata*

La structure d'une *metadata* E.*Victor Hugo* construite d'après l'article D.*Victor Hugo* est présentée dans la figure 2. Les informations originales D contenues dans Wikipédia ayant servi à construire E sont illustrées dans la figure 1. Pour construire E, considérons que le titre $D.t$ =*Victor Hugo* de la page Wikipedia est utilisé en tant que descripteur unique $E.t$ de la *metadata Victor Hugo*. Les formes de surface sont ensuite collectées pour construire le graphe $E.v$. D'abord, nous ajoutons le titre original $D.t$ à $E.v$. Ce titre encyclopédique original est toujours la forme minimale (et parfois unique) d'écriture possible pour $E.v$. Les *pages de redirection* de Wikipédia D^r contenant un lien $D^r.R$ vers D =*Victor Hugo* sont recherchées ; chaque titre $D^r.t$ d'une *page de redirection* est un synonyme de *Victor Hugo* et est inclus dans $E.v$. Nous cherchons les *Page d'homonymie* D^d qui contiennent des liens vers l'entité D. Ces liens sont contenus à l'intérieur du corps de texte de D^d en accord avec les spécifications du formalisme de Wikipédia. Dans l'exemple de *Victor Hugo*, la *page d'homonymie* D^d est "Hugo" et est reliée à 45 autres pages encyclopédiques qui décrivent des personnes (la page *Victor Hugo* elle-même), des lieux, des événements (ouragans) ou des organisations (un prix littéraire). Nous ajoutons le titre $D^d.t = Hugo$ à $E.v$.
L'agrégation finale de toutes les formes de surface est réalisée par l'exploration des *relations interwiki*. Les formes collectées sont indiquées dans la figure 2 par le lien $E.i$. On peut observer une introduction de bruit lors de la collecte des formes de surface. En effet le graphe de formes de surface de *Victor Hugo* contient 14 formes valides et une forme erronée, *Adèle Foucher*. Ce phénomène est lié au caractère manuel de la saisie des *pages de redirection* par les contributeurs de Wikipédia. Nous discutons de son influence dans nos expériences.
Tous les termes de l'article *Victor Hugo* de la version linguistique française de Wikipédia font l'objet d'un calcul de poids $tf.idf$ et sont intégrés dans $E.w$. On observe que les termes de poids le plus fort sont hautement contextuels (Ecrivain, France, etc) [10].

10. Voir http ://www.nlgbase.org/perl/display.pl ?query=Victor Hugo&search=FR pour un descriptif complet.

Gao (Région)	Gao Xingjian	Gao (Ville)
mali :44.92	nobel :31.3	mali :63.78
bozo :35.36	litteracy :25.49	songhay :51.27
region :19.44	tien :25.22	city :12.45
kidal :18.67	chinese :24.23	river :9.81
...	...	...

TABLE 1 – Trois représentations différentes de *metadata* partageant le même nom (Gao) et la table des mots contextuels, accompagnés de leurs poids $tf.idf$, pour chacun de ces noms

4 Algorithme du système d'extraction, d'identification et de désambiguïsation des entités nommées

Considérons une phrase S contenant une séquence de mots s ; pour tout s, nous recherchons dans les *metadata* extraites depuis Wikipédia une forme de surface ou plus, candidate pour étiqueter l'EN éventuellement contenue dans s. Nous intitulons Rm l'ensemble des formes de surface candidates. Pour détecter les EN, et éventuellement les désambiguïser, le système d'EEN utilise les mots contextuels contenus dans $E.w$ avec leurs poids $tf.idf$ afin de calculer le degré de similarité entre une forme de surface candidate et son contexte textuel dans s.
Le système de détection est divisé en 2 algorithmes intitulés $A1$ et $A2$ qui exploitent deux fonctions :

1. La fonction $Rm = f_{synsets}(s)$ recherche dans les *metadata* les éléments des $E.v$ qui correspondent à un sous ensemble de s.
2. La fonction $f_{simcos}(Rm, S)$ calcule la similarité cosinus de S comparée à tout élément E de Rm d'après $E.w$.

Dans $A1$, nous considérons que si $|Rm| = 1$, Rm est l'unique EN candidate valable. Nous mesurons alors la similarité cosinus entre le contexte de l'EN et les poids de mots contenus dans $E.w$. Si $|Rm| > 1$, nous en concluons que l'algorithme a trouvé un ensemble d'entités ambigües : $A2$ recherche alors la meilleure entité contenue dans Rm avec les scores de similarité cosinus. On considère que $A1$ ou $A2$ ne proposent d'entité que si le score de similarité cosinus entre des entités candidates et le contexte d'étiquetage est supérieur à un seuil c_s.
Finalement, si $A1$ or $A2$ ne proposent pas d'entité candidate pour s, nous décidons que s n'est pas une entité. Conformément à cette description, notre système fonctionne de la manière suivante :

- $Rm = f_{synset}(s)$
- A1 : Une seule proposition dans Rm (*métadonnée candidate unique*) ▷ si $f_{simcos} > c_s$ ▷ **extraction d'entité**
- A2 : Plusieurs propositions dans Rm (*Ambiguité de détection entre plusieurs EN*)
    1. Utilise un score de similarité cosinus pour classer les propositions
        (a) A2.1 : Le meilleur du classement de $f_{simcos} > c_s$ ▷ **extraction d'EN**
        (b) A2.2 : $f_{simcos} < c_s$ pas d'EN candidate acceptable ▷ ***pas d'EN***

Le processus de détection débute avec $f_{synset}(s)$ qui doit identifier les séquences de mots s de la phrase à étiqueter S identiques à des formes de surface contenues dans les *metadata*.

	[...] cinq équipe de F1 : Toyota, [**Williams**], Sauber, Red-Bull, et Minardi, sur dix engagées dans le [...]	
	Entité correspondante pour for *Williams*	Score Cosinus
1	Williams (F1 Team)	0.91
2	Franck Williams	0.45
3	John Williams	0
4	Robby Williams	0

TABLE 2 – Exemple de localisation d'une entité dans son contexte, en utilisant le classement de la mesure de similarité cosinus pour identifier le terme le plus approprié.

Considérons M le corpus Wikipédia et E les entités extraites représentées par des formes de surface contenues dans $E.v$:

– $f_{synset}(s)$

1. $\mathcal{E} = \{E_0, ..., E_n\}$ ensemble d'entités extraites de M.
2. Chaque E_n contient un ensemble de $E_n.v$, $\{v_0, ...v_m\}$ formes de surface utilisées pour la détection
3. $s = \{t_0, ..., t_u\}$ ensemble de séquences de mots (n-gram à 1-gram) $t \in S$
4. **Pour tout** $s^m \in S$,
   (a) $Rm = \emptyset$
   (b) **Pour tout** $E_n \in \mathcal{E}$, $v_m \in E_n.v$
   (c) SI ($v_m = s^m$) alors add E_n to Sd
5. Retourner Rm

Si Rm retourné par $f_{synset}(s)$ ne contient qu'une proposition, $A1$ s'applique, sinon $A2$ est invoqué. Ce serait le cas avec l'entité *Gao* (voir tableau 1), qui exprime trois concepts différents (personne, région, ville).
Chaque E inclut un ensemble de mots avec leurs $tf.idf$ contenus dans $E.w$. Considérons pour la phrase S un groupe de mots X avec ses poids $X.tf.idf$. Ce groupe est intitulé *contexte*. Le *contexte* est composé de q mots à droite et à gauche de l'entité E identique à $s \in S$.
Considérant que la fonction de similarité cosinus $cos(E.tf.idf, X.tf.idf)$ existe, nous pouvons déterminer quelle entité candidate E_c de Rm obtient le plus haut score en mesurant le cosinus de l'angle entre les vecteurs de poids des mots correspondant à $E_c.w$ de Rm et les $X.tf.idf$. Nous obtenons alors une liste de scores indiquant la meilleure entité candidate $\hat{E_c}$ dans son contexte. Nous utilisons dans $A2.1$ la formule :

$$\hat{E} = argmax_{E_n}\ score(cos(E_n.tf.idf, X.tf.idf))$$

Un exemple de liste de scores produite par $A2$ pour un contexte donné, est présenté dans le tableau 2.

5 Expériences et résultats

Pour nos expériences, nous avons utilisé 100 articles de presse du quotidien Français *Le Monde* issus du corpus fourni durant la campagne d'évaluation DEFT'07. Nous avons étiqueté ces articles de manière semi-automatique, par une première application de notre système, suivie d'une correction manuelle. Les 100 articles représentaient 1.996 phrases, 61.000 mots et 1.307 entités à étiqueter. Nous avons appliqué un système de détection d'EN conçu d'après les algorithmes

décrits précédemment. Nous avons évalué les résultats de notre système par précision, rappel et F-Score. Afin de déterminer quelle était l'influence de l'incorporation de formes de surface issues de corpus d'une autre langue sur l'étiquetage de documents en français, nous avons testé notre système avec les deux configurations suivantes :

1. Etiquetage d'EN en utilisant des *metadata* incluant uniquement des formes de surface issues de la version linguistique française de Wikipédia.
2. Etiquetage d'EN en utilisant des *metadata* incluant des formes de surface issues de cinq versions linguistiques de Wikipédia (français, anglais, espagnol, italien, allemand).

Les résultats obtenus sont indiqués dans le tableau 3.

Système	($\bar{p}$)	($\bar{r}$)	($\bar{F}$-m)
Métadata françaises	0.89	0.87	**0.88**
Métadata françaises et cross-linguistiques	0.90	0.91	**0.91**

TABLE 3 – Evaluation et comparaison des performances avec des *métadata* et des formes de surface issues du corpus linguistique français de Wikipédia, et de celles issues de cinq corpus linguistiques

5.1 Discussions

Le tableau 3 montre que l'introduction de ressources interlinguales dans le système de détection d'entités augmente ses performances. On peut en déduire que si les performances de l'algorithme de mesure par similarité cosinus restent constantes, l'ajout de nouvelles formes de surface pour les entités, collectées depuis les versions anglaise, espagnole, italienne et allemande du corpus Wikipédia a amélioré la couverture du système et limité les OOV. Nous notons que le bruit introduit par la présence des formes de surface sémantiquement non correctes (voir la section 3.3) est sans influence sur les performances du système d'EEN : le calcul de similarité cosinus entre la *metadata* et le contexte textuel d'une EN candidate les écarte automatiquement et efficacement du processus de détection.

6 Conclusion et travaux futurs

Dans cet article nous avons présenté une nouvelle ressource sémantique qui peut être utilisée pour détecter les variations d'écriture et les ambiguïtés d'une EN dans son contexte.
L'originalité de ce système est qu'il utilise des formes de surface collectées via une exploration des relations interlinguales de corpus encyclopédiques. Nos expériences ont montré que cette démarche pouvait améliorer les performances du système d'EEN et sa capacité à détecter des variations d'écritures.
Nous envisageons maintenant de mener des expériences sur des combinaisons linguistiques en utilisant le potentiel offert par les 253 langues disponibles dans Wikipédia. Notre idée serait de déterminer s'il existe des combinaisons interlinguales de formes de surface plus performantes pour la tâche d'EEN que celle expérimentée ici.
La ressource mise au point, intitulée **NLG**b**A**se peut être consultée et téléchargée librement. Le système d'EEN peut être expérimenté en ligne et sera prochainement diffusé sous une forme libre [11]. Dans une perspective plus large d'amélioration d'un système de reconnaissance existant, nous avons également prévu d'intégrer notre système en tant que ressource lexicale et

11. Consulter le site www.nlgbase.org.

fonction complémentaire d'EEN à postériori à un système d'EEN stochastique. Un prototype basé sur le programme LIA_NE [12] a été déployé lors de la campagne ESTER 2 et fera l'objet de communication ultérieure.
Notre objectif principal lorsque nous avons entamé ce travail était de concevoir un ensemble de *metadata* issu d'un contenu encyclopédique et de l'associer à des algorithmes d'étiquetage et d'extraction d'information en vue de participer aux campagnes d'évaluation telles que KBP [13]. Nous travaillons actuellement à cette évolution.

Remerciements

Je remercie vivement les relecteurs pour leurs commentaires et propositions particulièrement détaillés et utiles.

Références

BUNESCU R. & PASCA M. (2006). Using encyclopedic knowledge for named entity disambiguation. In *EACL Proceedings of the Third International Joint Conference on Natural Language Processing, April 3-7, 2006, Trento, Italy* : EACL.

CHARTON E., CAMELIN N., ACUNA-AGOST R., GOTAB P., KESSLER R., LAVALLEY R. & FERNANDEZ S. (2008). Pré-traitements classiques ou par analyse distributionnelle :application aux méthodes de classification. In *Atelier Défi Fouille de Texte, Actes de TALN 2008, Avignon* : DEFT.

FAVRE B., BECHET F. & NOCERA P. (2005). Robust named entity extraction from large spoken archives. In *Proceedings of HLT-EMNLP'05, Vancouver (Canada)* : HLT-EMNLP.

GROUIN C., BERTHELIN J.-B., AYARI S. E., HURAULT-PLANTET M. & LOISEAU S. (2008). Présentation de deft08 (defi fouille de textes). In *Atelier Défi Fouille de Texte, Actes de TALN 2008, Avignon* : DEFT.

JUNÍCHI 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.

LAFFERTY J., MCCALLUM A. & PEREIRA F. (2001). Crf : Probalistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning(ICML-2001)* : IMCL.

NADEAU D. & SEKINE S. (2007). A survey of named entity recognition and classification. In *Lingvisticae Investigationes, Vol 30, number 1, September 2007*.

NIST (2007). : NIST.

SANG E. F. T. K. & MEULDER F. D. (2003). Introduction to the conll-2003 shared task : Language-independent named entity recognition. In *Seventh Conference on Natural Language Learning, May 31 and June 1, 2003, HLT-NAACL 2003, Edmonton, Canada* : CoNLL.

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

12. Disponible sur http ://lia.univ-avignon.fr/fileadmin/documents/Users/Intranet/chercheurs/bechet/download_fred.html.
13. Knowledge Base Population, Tâche de TAC 2009, en cours. Voir http ://apl.jhu.edu/p̃aulmac/kbp.html.