Classification d'entités nommées de type « film »

Olivier Collin Aleksandra Guerraz
Orange Labs
2, avenue Pierre Marzin, 22307 Lannion Cedex, France
{olivier.collin,aleksandra.guerraz}@orange.com

Résumé. Dans cet article, nous nous intéressons à la classification contextuelle d'entités nommées de type « film ». Notre travail s'inscrit dans un cadre applicatif dont le but est de repérer, dans un texte, un titre de film contenu dans un catalogue (par exemple catalogue de films disponibles en VoD). Pour ce faire, nous combinons deux approches : nous partons d'un système à base de règles, qui présente une bonne précision, que nous couplons avec un modèle de langage permettant d'augmenter le rappel. La génération peu coûteuse de données d'apprentissage pour le modèle de langage à partir de Wikipedia est au coeur de ce travail. Nous montrons, à travers l'évaluation de notre système, la difficulté de classification des entités nommées de type « film » ainsi que la complémentarité des approches que nous utilisons pour cette tâche.

Abstract.

Named Entity Classification for Movie Titles

In this article, we focus on contextual classification of named entities for « movie » type. Our work is part of an application framework which aims to identify, in a text, a movie title contained in a catalog (e.g. VoD catalog). To do this, we combine two approaches : we use a rule-based system, which has good accuracy. To increase recall we couple our system with a language model. The generation of training data for the language model from Wikipedia is a crucial part of this work. We show, through the evaluation of our system, the complementarity of approaches we use.

Mots-clés : reconnaissance d'entités nommées, films, classification, règles, modèle de langage, Wikipedia.

Keywords: named entity recognition, movies, classification, rules, language model, Wikipedia.

1 Introduction

La détection de « film » est un cas particulier de détection d'entités nommées. La segmentation de ce type d'entités nommées, comme d'ailleurs celle des titres d'oeuvre en général, semble plus difficile à réaliser que celle des entités de type personne, lieu ou organisation. Cette tâche qui consiste à déterminer le début et la fin du titre est rendue plus complexe car de nombreux titres sont aussi des groupes de mots (noms communs, verbes, locutions adverbiales ou adjectivales) du lexique général de la langue : *A fleur de peau*, *A bout de souffle*, *Minuit à Paris*, *Vivre vite*. Dans le contexte de la phrase, il est donc difficile de déterminer l'étendue de ce type d'entité : A bout de souffle a été tourné...

Dans un cadre opérationnel, nous avons été confrontés à un problème différent. Les bases de données de titres étant de plus en plus nombreuses et le but applicatif étant de repérer un titre de film dans un texte et proposer au client de le voir en VoD, l'ensemble des titres est connu à l'avance. La détection est donc simplifiée puisqu'elle se résume essentiellement à repérer dans les documents des groupes de mots correspondant aux titres de la base, ce qui constitue techniquement un accès à un lexique de titres. Ce repérage entraîne cependant deux types de problème :

1. La forme d'un titre dans le document peut être différente de la forme du titre de la base. Cette différence peut être faible car liée à des variantes morphologiques (accentuation, ponctuation, conjugaison) ou plus importante car liée à des versions de l'oeuvre (*Mission impossible*, *Mission impossible 2*, *Mission impossible : Protocole Fantôme...*) ou à des versions non traduites (*Mission : Impossible – Ghost Protocol*)
2. Cette détection entraîne des annotations non désirées des entités qui sont ambiguës avec des entités de type différent (lieu, personne...) ou des noms communs (*Elle*, *Vie privée...*).

Le travail présenté se focalise uniquement sur ce problème de classification (film/pas film). La technique est relativement simple : nous avons réalisé un modèle de classification statistique qui s'appuie (apprentissage et test) sur la représentation fournie par notre système initial de détection d'entités nommées (basé sur les règles). La décision issue du classifieur statistique qui complète celle du premier système n'est utilisée que pour les données non étiquetées « film » de manière sure par notre système initial. Il existe différentes manières de coupler un système symbolique à un système probabiliste telles que dans (Béchet *et al.*, 2011) et (Nouvel *et al.*, 2012). Nous présentons une hybridation très simple qui consiste à utiliser un classifieur basé sur des règles offrant une très bonne précision en amont d'un classifieur statistique qui ne remet pas en cause la décision prise par le système symbolique mais la complète sur les cas ambigus. Pour réaliser notre classifieur statistique, nous avons utilisé une technique standard de désambiguïsation statistique : les modèles de langage. Leur mise en oeuvre a été faite par l'un des outils les plus connus et techniquement aboutis dans ce domaine : la librairie SRILM [1].

Le corpus d'apprentissage pour le modèle de langage a été constitué à partir de Wikipedia. En effet, cette ressource a suscité un grand intérêt pour la tâche de Reconnaissance d'Entités Nommées et son utilisation a été largement approuvée. (Nothman *et al.*, 2008), (Charton & Torres-Moreno, 2009) montrent qu'il est possible d'obtenir à partir de Wikipedia une ressource annotée en entités nommées de large couverture et de très bonne qualité. Les systèmes de détection d'entités nommées entraînés avec des ressources issues de Wikipedia peuvent obtenir de très bons résultats sur d'autres types de corpus (Balasuriya *et al.*, 2009). Les ressources de Wikipedia sont utilisées pour compléter les ressources des systèmes de détection d'entités par (Stern & Sagot, 2010) et (Okinina *et al.*, 2013).

Dans cet article, nous présentons d'abord comment les données d'apprentissage pour le modèle de langage ont été produites (section 2). Le couplage de notre système de détection d'entités nommées à base de règles avec un modèle de langage est décrit dans la section 3. Enfin, nous présentons une évaluation de notre approche (section 4).

2 Production de données d'apprentissage pour le modèle de langage

La production de données d'apprentissage ou l'annotation consiste à ajouter des métadonnées (étiquettes sémantiques) telles que *<film> Hollywood </film>*, en regard des entités à classer, de manière à lever contextuellement l'ambiguïté portant sur l'entité. Dans le cas présent *Hollywood* est potentiellement un lieu ou un film mais ne possède qu'une catégorie dans un contexte donné. Ce travail est généralement réalisé manuellement en parcourant visuellement les textes à annoter, ce qui est long et fastidieux et peut nécessiter des ressources humaines importantes pour obtenir des données annotées en quantité suffisante.

Les données textuelles des pages de Wikipedia possèdent des liens internes qui sont un type d'annotation sémantique « gratuite ». Ce sont les rédacteurs des pages qui ajoutent manuellement ces annotations. Nous avons réalisé un travail de récupération et de filtrage des pages de Wikipedia (français). Le résultat de ce travail est un gros fichier texte qui associe à chaque page de Wikipedia une partie de son texte (hors tableaux, citations ...) en conservant les liens internes. Ces liens sont des références à des pages de Wikipedia qui associent à la forme du texte le nom de la page correspondante. Par exemple :

```
[[Jean Renoir]] : texte = Jean Renoir, page = Jean Renoir
[[Jacques Martin (auteur)|Jacques Martin]] : texte = Jacques Martin, page = Jacques Martin (auteur)
[[Jacques Martin (animateur)|Jacques Martin]] : texte = Jacques Martin, page = Jacques Martin (animateur)
```

Généralement, ces liens internes sont désambiguïsés par les rédacteurs des pages. Cela signifie que dans le cas ou plusieurs formes existent (comme dans le cas de *Jacques Martin*), le nom de la page associée spécifie quel est le « bon » *Jacques Martin* dans le contexte de la phrase. Sans précision particulière (cas de *Jean Renoir*), c'est la page unique de *Jean Renoir* qui est référencée. Ceci étant, les références ne possèdent pas souvent un type sémantique tel que « auteur » ou « animateur ». Dans le cas de *Jean Renoir*, il faut donc trouver un moyen de connaître quel est son type.

Pour récupérer une étiquette sémantique nous avons utilisé une stratégie automatique qui est très simple mais qui ne permet pas d'étiqueter tous les liens de Wikipedia. En ce qui concerne les films et un certain nombre d'autres types d'entités, cela permet toutefois de constituer un corpus partiellement annoté. L'hypothèse est donc que ce corpus partiellement annoté, et probablement un peu bruité, nous permette de réaliser un modèle de langage utile. Nous avons en quelque sorte inversé la problématique. Plutôt que de compléter l'annotation de tous les liens internes de Wikipedia, nous avons

1. http://www.speech.sri.com/projects/srilm/manpages/

essayé de qualifier les pages dont nous sommes surs (ou presque !) de la catégorie sémantique. Une fois ces pages qualifiées, nous avons complété les liens où ces pages apparaissent avec la catégorie récupérée. Par exemple : *Jean Renoir* sera complété avec le type « réalisateur » et son lien interne pourra être transformé en une annotation de type XML <*realisateur*>*Jean Renoir*</*realisateur*>. La récupération des types sémantiques a été réalisée à partir de deux sources d'information suivantes :
– le type fourni dans le nom de la page Wikipedia, lorsqu'il existe, tel que dans l'exemple précédent de *Jacques Martin*
– la relation « est un » entre le nom de la page et le premier lien à droite qui apparaît généralement dans le premier paragraphe de la description de la page.
Pour la page de *Jean Renoir* [2] on a :

```
Jean Renoir est un [[réalisateur]] et [[scénariste]] [[français]], né à [[Paris]]...
```

Le type associé directement à la page n'est pas très fréquent et n'est pas toujours un type sémantique mais peut être un type thématique ou une date. La relation « est un » peut induire des erreurs mais est généralement précise. En fusionnant ces deux types d'informations, nous avons pu étiqueter automatiquement près de 200 000 pages de Wikipedia français et notamment près de 6 000 titres de film. Chaque titre apparaissant en moyenne plusieurs fois dans Wikipedia, nous avons obtenu environ 26 000 phrases contenant au moins un titre de film étiqueté par phrase. Ce type de corpus contenant un étiquetage contextuel de films, avec une telle quantité d'annotation est unique. D'autre part, cette stratégie n'est pas limitée aux films et peut être appliquée sur tous les types d'entités rencontrés lors de l'étiquetage initial des pages de Wikipedia. Elle n'est pas non plus limitée au français, elle devrait être reproductible à faible coût pour toutes les autres langues traitées par Wikipedia. Nous avons ensuite projeté les étiquettes extraites dans une taxonomie simple permettant de catégoriser les principaux types d'entités : film, evt (évènement), jeu, time, livre, loc (lieu), org (organisation), album, pers(personne), amount. Cette projection a été réalisée automatiquement. Nous avons utilisé ce corpus annoté pour réaliser une désambiguïsation statistique avec un modèle de langage que nous couplons avec notre système de détection d'entités nommées à base de règles.

3 Couplage du système à base de règles avec un modèle de langage

Nous avons choisi un couplage relativement simple qui consiste à utiliser notre système à base de règles pour générer une représentation qui est ensuite utilisée par le modèle de langage statistique, aussi bien en apprentissage qu'en test. L'apprentissage se fait « à la suite » du système à base de règles ce qui permet d'utiliser la segmentation produite par notre système (notamment multi-mots et entités nommées) ainsi que les étiquettes sémantiques des entités nommées, notamment des personnes, des lieux et des organisations. Le modèle de langage est ensuite généré et utilisé en reprenant cette segmentation des données. Il ne sert donc pas à détecter l'étendue des entités nommées (réalisée par notre système) mais uniquement à classer le type des entités en fonction du contexte. Le modèle généré est donc dépendant de la segmentation et de la précision de notre système en ce qui concerne le typage des entités nommées.

3.1 Configuration initiale du système de détection d'entités nommées à base de règles

Notre système de détection d'entités nommées est basé sur la plate-forme de traitement linguistique de textes TiLT décrite dans (Heinecke *et al.*, 2008). Ce système permet de manière générale de :
– repérer des entités nommées déjà connues grâce à des informations lexicales,
– découvrir des entités nommées sur la base de déclencheurs lexicaux et de contraintes de bonne formation.
Le repérage d'entités nommées s'appuie sur des indices lexicaux, typographiques et contextuels. En l'absence d'indices contextuels, la confiance que l'on peut avoir dans le lexique varie en fonction des caractéristiques de l'entité :
– une entité nommée doit comporter une majuscule,
– une entité nommée mono-mot ne sera pas repérée en cas d'ambiguïté avec un autre mot du lexique et en l'absence d'un contexte syntaxique fiable (*Puma* lieu ne sera pas étiqueté lieu en l'absence de contexte fiable, par exemple *la ville de Puma*),
– une entité nommée multi-mots ne sera pas repérée si elle est de type mot outil + mot du lexique (*La Flèche* ne sera pas étiqueté lieu en l'absence de contexte fiable, par exemple *la piscine de La Flèche*).

2. http://fr.wikipedia.org/wiki/Jean_Renoir

La découverte d'entités nommées se fait au moyen de déclencheurs typographiques et lexicaux. La présence de majuscule est un élément majeur permettant de soupçonner la présence d'une entité nommée et de la segmenter. Les déclencheurs lexicaux sont internes ou externes selon qu'ils font (ou non) partie de l'entité nommée elle-même. La chaîne de caractères constituée par le(s) déclencheur(s) interne(s) et les noms propres est identifiée via les règles de grammaire en dépendance et est visible sous forme d'une locution : *Jean-Marcel Dupont* devient une locution de type nom de personne. Les déclencheurs sont principalement des noms qui introduisent de façon directe ou indirecte (c'est-à-dire à l'intermédiaire d'une préposition) les entités nommées.

Le repérage des titres de film consiste alors à valider un titre de film (existant dans le lexique) dans un contexte donné. La fonctionnalité de découverte n'est pas utilisée pour la détection de films. Le repérage des titres de film exploitent des informations lexicales, typographiques (guillemets) et stylistiques (énumération), contextuelles modélisées par une grammaire locale sous forme de règles de « chunking » et des informations sur la longueur des titres de film (on part de l'hypothèse que plus le titre est long, moins il peut être confondu avec un autre élément de la phrase). Le lexique contient 70 054 entités nommées de type film. 41% des titres de film sont ambigus avec un autre mot du lexique (par exemple, *Vampires*, *Pirates*, *A fleur de peau*). Un contexte déclencheur est nécessaire pour leur identification dans les phrases. 2% de titres de film sont ambigus avec une autre entité nommée (parmi les types personne, lieu, organisation). Il est à noter que le lexique ne contient pas de titres d'autres oeuvres artistiques tels que les titres de livre, les titres de musique ou les titres de spectacle, etc.

3.2 Apprentissage du modèle de langage

Afin de coupler le système décrit dans la section précédente avec un modèle de langage, une mesure de confiance binaire est attribuée à chaque titre de film pour distinguer les films en contexte fiable. Les films en contexte fiable sont étiquetés 1, les autres titres de film potentiels sont étiquetés 0.

Le modèle de langage va attribuer une probabilité à tous les titres de film étiquetés 0 (indice de confiance). Le but est d'augmenter le rappel sans trop diminuer la précision. Nous considérons comme résultat positif soit les films initialement étiquetés 1 par notre système soit les films initialement étiquetés 0 mais dont le résultat du modèle de langage donne pour l'hypothèse du film (pour l'étiquette NAM:film) une probabilité forte. Nous pouvons donc définir une valeur de seuil pour cette probabilité. Dans un premier temps, les résultats ont été calculés avec une valeur de seuil de 0,8. Si $P(NAM{:}film) \geq 0,8$ alors le résultat est positif (on considère que l'entité est un film). Cette valeur de seuil permet d'effectuer un réglage précision/rappel, ce qui peut être utile dans un cadre applicatif de manière à favoriser le rappel ou la précision. Le seuil est relatif aux probabilités conditionnelles de l'étiquette NAM:film versus toutes autres catégories (personne, lieu, organisation...). Nous considérons donc implicitement que notre modèle de classification des « non-films » est constitué de la somme des probabilités des étiquettes autres que NAM:film.

Lors de l'apprentissage du modèle de langage nous avons utilisé des données de supervision (étiquetage des entités à classer) issues des étiquettes supposées fiables par notre système initial. Mais en nous limitant à cet étiquetage, le modèle ne pourrait qu'apprendre implicitement des règles statistiques conduisant au fonctionnement actuel de notre système à base de règles. Nous avons donc remplacé ou complété les étiquettes générées par des étiquettes apportant une information complémentaire. Cette information est essentiellement contenue dans de nouveaux contextes. De cette manière, nous espérons obtenir un système conservant la précision de notre système initial mais qui étend ses performances à d'autres contextes non gérés par notre système, et par conséquent augmente le rappel. L'exemple suivant illustre ce processus :

```
En 1928, pour le bimillénaire de la cité de Carcassonne, Jean Renoir réalise
Le Tournoi dans la cité.
```

1. Segmentation et étiquetage produit par notre système à base de règles (NAM: indique la catégorie de l'entité nommée, les « _ » relient les mots d'un même segment) :

   ```
   En 1928/NAM:time , pour le bimillénaire de la cité de Carcassonne/NAM:loc ,
   Jean_Renoir/NAM:pers réalise Le Tournoi dans la cité .
   ```

   Notre système étiquette *1928*, *Carcassonne*, *Jean_Renoir* mais pas *Le_Tournoi_dans_la_cité*

2. Etiquetage externe à partir de Wikipedia (cf. section 2) :

   ```
   En 1928, pour le bimillénaire de la cité de Carcassonne,
   <pers>Jean Renoir</pers> réalise <film>Le Tournoi dans la cité</film>.
   ```

Les données produites étiquettent *Jean Renoir* et *Le Tournoi dans la cité*

3. Etiquetage cumulé et projeté dans le même formalisme d'annotation :

```
En 1928/NAM:time , pour le bimillénaire de la cité de Carcassonne/NAM:loc ,
Jean_Renoir/NAM:pers réalise Le_Tournoi_dans_la_cité/NAM:film .
```

4. Etiquetage utilisé pour l'apprentissage du modèle de langage :

```
En NAM:time , pour le bimillénaire de la cité de NAM:loc ,
NAM:pers réalise NAM:film .
```

Chaque segment annoté est remplacé par sa catégorie sémantique.

En cas de double étiquetage effectué par notre système et par l'étiquetage externe à partir de Wikipedia (cf. section 2), ce sont ces dernières étiquettes qui sont utilisées. Les étiquettes externes peuvent donc soit remplacer celles de notre système soit les compléter.

Pour réaliser le calcul des probabilités du modèle de langage, nous avons constitué un ensemble d'exemples représentatifs des contextes d'apparition des films, mais aussi des contextes d'apparition d'entités qui ne sont pas des films, de manière à en « peser » les probabilités respectives. En phase d'utilisation du modèle de langage, ces probabilités calculées à partir des données annotées vont être réutilisées (trigrammes, bigrammes, unigrammes) pour calculer la probabilité maximale d'émission de chaque catégorie sémantique pour l'ensemble de la phrase. Le modèle statistique ne s'applique que sur les décisions portant sur les titres de film jugés « non fiables » (étiquetés 0) par le système initial.

Au final, nous avons donc un système qui réalise un apprentissage statistique à partir d'une représentation des données filtrées par notre système et annotées automatiquement par Wikipedia.

4 Evaluation

4.1 Corpus d'évaluation

Nous avons constitué un corpus d'évaluation. L'annotation de ce corpus en entités nommées a été effectuée manuellement par un seul annotateur, selon les conventions d'annotations internes largement inspirées de celles de la campagne ESTER2 et de celles de la campagne QUAERO [3]. Aucun accord inter-annotateurs n'a pu être mesuré. Le corpus se compose de 287 documents textuels répartis comme suit :

– 257 textes courts de type dépêche AFP,
– 7 textes issus de Wikipedia décrivant des personnalités du monde cinématographique et musical,
– 5 dépêches AFP issues du portail Orange,
– 18 textes issus des Inrockuptibles, du portail Orange portant sur des chanteurs ou des groupes de musique.

4.2 Résultats

Le tableau 1 donne les résultats obtenus pour la détection de film avec le système initial et avec le même système auquel un modèle de langage a été ajouté a posteriori.

	Précision	Rappel	F-mesure
Système initial	0,84	0,33	0,48
Système initial+ML (Seuil $\geq$ 0,8)	0,76	0,53	0,63

TABLE 1 – Résultats

La configuration hybride (système initial + modèle de langage) apporte une amélioration du système. La F-mesure augmente de 31% (elle passe de 0,48 à 0,63). Ceci se traduit par une augmentation du rappel (de 60%) et une baisse de la précision (de 9%) qui reste, toutefois, acceptable.

3. https://perso.limsi.fr/rosset/quaero-guide-annotation-2011.pdf

4.2.1 Bruit

42 % du bruit est lié à la confusion de catégorie. On y trouve principalement (18% du bruit) des confusions avec d'autres sous-types de la catégorie oeuvre artistique, par exemple avec :
– un titre de musique (*Carla Bruni enregistre **Douce France***)
– un titre d'album (*Le premier single extrait de l'album, **Falling Down**, s'accompagne d'un clip*)
– un titre de livre (*tout en reprenant le thème du **Désert des Tartares** de Dino Buzzati*)
Une sous-catégorisation de la classe personne en acteurs, écrivains, réalisateurs, chanteurs pourrait limiter ce genre de bruit. Les confusions avec le sous-type personnage de la catégorie personne sont également fréquentes et représentent 17% du bruit. Par exemple, une confusion avec un personnage éponyme du film dans « 1993 Fanfan de Alexandre Jardin : ***Fanfan*** ». Le reste de bruit est dû notamment à des erreurs de segmentation. Le titre de film détecté n'est pas complet (par exemple dans le lexique on a *Die Hard* alors que le film à détecter est *Die Hard 4*). Ce type d'erreur est fréquent pour les différents volets de films (par exemple *Mission Impossible 2*) et pourrait être limité par un repérage par *approximate string matching* (technique fréquemment utilisée en correction). On trouve également ce type d'erreur pour des films avec des titres contenant un sous-titre (par exemple dans le lexique on a *Belphégor* pour le titre de film *Belphégor, le fantôme du Louvre*). Le titre de film détecté fait partie d'un titre d'une autre oeuvre artistique (titre de livre, titre d'album ou titre de chanson) Ce type d'erreurs (de segmentation et de catégorisation) est fréquent dans les textes qui comportent beaucoup de mots en anglais, par exemple : sur l'album *This Is The Sea* où *The Sea* est détecté comme titre de film.

4.2.2 Silence

41% du silence est dû à l'incomplétude du lexique : le titre de film est inconnu de la base de données, source du lexique de notre système. En se ramenant aux données présentes uniquement dans le lexique, le rappel serait donc nettement meilleur et par conséquent donc la F-mesure. Voici un exemple de non-détection (exemple en gras) : *Si **Manon** n'est pas une grande réussite, c'est sans doute parce que le metteur en scène n'était pas en forme.*

Les variations relatives des résultats sont toutefois indicatives des propriétés du système. Un lexique exhaustif devrait produire un accroissement relatif des résultats analogues à celui que nous constatons. En terme opérationnel, cela veut dire que les films qui ne sont pas dans la base ne seront pas détectés.

5 Conclusions

Ce travail nous a permis d'améliorer les résultats initiaux en classification des titres de film obtenus par notre système basé sur des règles. Cette amélioration globale de la F-mesure porte surtout sur le rappel, au prix d'une légère dégradation de la précision. Les avantages du processus sont :
– l'intégration simple du modèle de langage au système initial basé sur des règles par un couplage a posteriori,
– l'utilisation d'une mesure de probabilité permettant de régler les performances suivant les axes précision/rappel,
– la possibilité d'ordonner les solutions par probabilités décroissantes pour un filtrage utilisateur,
– une génération automatique du corpus d'apprentissage annoté, donc aucun coût humain d'annotation,
– la possibilité de reproduire cette technique sur d'autres langues.

Références

BALASURIYA D., RIGLAND N., NOTHMAN J., MURPHY T. & CURRAN J.-R. (2009). Named entity recognition in wikipedia. In *People's Web 2009*, p. 10–18, Morristown, NJ, USA : Association for Computational Linguistics.

BÉCHET F., SAGOT B. & STERN R. (2011). Coopération de méthodes statistiques et symboliques pour l'adaptation non-supervisée d'un système d'étiquetage en entités nommées. In *Actes de TALN 2011 (Traitement automatique des langues naturelles)*, Montpellier : ATALA LIRMM.

CHARTON E. & TORRES-MORENO J.-M. (2009). Classification d'un contenu encyclopédique en vue d'un étiquetage par entités nommées. In *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis : ATALA LIPN.

HEINECKE J., SMITS G., CHARDENON C., GUIMIER DE NEEF E., MAILLEBUAU E. & BOUALEM M. (2008). Tilt : plate-forme pour le traitement automatique des langues naturelles. In *TAL 2008 (Traitement automatique des langues)*, p. 17–41 : ATALA.

NOTHMAN J., CURRAN J.-R. & MURPHY T. (2008). Transforming wikipedia into named entity trainig data. In *Actes de ALTA 2008*, p. 124–132, Tasmania : ACL Australian Language Technology Workshop.

NOUVEL D., ANTOINE J.-Y., FRIBURGER N. & SOULET A. (2012). Coupling knowledge-based and data-driven systems for named entity recognition. In *Actes de EACL 2012*, Avignon : ACL.

OKININA N., NOUVEL D., FRIBURGER N. & ANTOINE J.-Y. (2013). Apprentissage supervisé sur ressources encyclopédiques pour l'enrichissement d'un lexique de noms propres destiné à la reconnaissance des entités nommées. In *Actes de TALN 2013 (Traitement automatique des langues naturelles)*, Sables d'Olonne : ATALA LINA.

STERN R. & SAGOT B. (2010). Détection et résolution d'entités nommées dans des dépêches d'agence. In *Actes de TALN 2010 (Traitement automatique des langues naturelles)*, Montréal : ATALA RALI.