Adaptation d'un système de reconnaissance d'entités nommées pour le français à l'anglais à moindre coût

Mohamed Hatmi
LINA, UMR 6241,Université de Nantes
mohamed.hatmi@univ-nantes.fr

RÉSUMÉ

La portabilité entre les langues des systèmes de reconnaissance d'entités nommées est coûteuse en termes de temps et de connaissances linguistiques requises. L'adaptation des systèmes symboliques souffrent du coût de développement de nouveaux lexiques et de la mise à jour des règles contextuelles. D'un autre côté, l'adaptation des systèmes statistiques se heurtent au problème du coût de préparation d'un nouveau corpus d'apprentissage. Cet article étudie l'intérêt et le coût associé pour porter un système existant de reconnaissance d'entités nommées pour du texte bien formé vers une autre langue. Nous présentons une méthode peu coûteuse pour porter un système symbolique dédié au français vers l'anglais. Pour ce faire, nous avons d'une part traduit automatiquement l'ensemble des lexiques de mots déclencheurs au moyen d'un dictionnaire bilingue. D'autre part, nous avons manuellement modifié quelques règles de manière à respecter la syntaxe de la langue anglaise. Les résultats expérimentaux sont comparés à ceux obtenus avec un système de référence développé pour l'anglais.

ABSTRACT

Adapting a French Named Entity Recognition System to English with Minimal Costs

Cross-language portability of Named Entity Recognition systems requires linguistic expertise and needs human effort. Adapting symbolic systems suffers from the cost of developing new lexicons and updating grammar rules. Porting statistical systems on the other hand faces the problem of the high cost of annotation of new training corpus. This paper examines the cost of adapting a rule-based Named Entity Recognition system designed for well-formed text to another language. We present a low-cost method to adapt a French rule-based Named Entity Recognition system to English. We first solve the problem of lexicon adaptation to English by simply translating the French lexical resources. We then get to the task of grammar adaptation by slightly modifying the grammar rules. Experimental results are compared to a state-of-the-art English system.

MOTS-CLÉS : Reconnaissance d'entités nommées, approche symbolique, portabilité entre les langues.

KEYWORDS: Named entity recognition, symbolic approache, cross-language portability.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 3: RECITAL, pages 151–161,
Grenoble, 4 au 8 juin 2012.

1 Introduction

La reconnaissance des entités nommées (REN) est une sous-tâche de l'extraction d'information consistant à délimiter et à catégoriser certaines expressions linguistiques autonomes et mono-référentielles (Ehrmann, 2008). Ces dernières correspondent traditionnellement à l'ensemble des noms propres (noms de personnes, de lieu et d'organisation) ainsi que certaines expressions numériques et temporelles (expressions de dates, de temps, de pourcentages, etc.). La délimitation et la catégorisation des entités nommées sont connues sous le nom d'annotation qui consiste généralement à encadrer une entité nommée par le biais d'une balise de début et de fin mentionnant sa typologie.

Certaines langues ont suscité beaucoup d'intérêt, notamment via les campagnes d'évaluations telles que MUC (Grishman et Sundheim, 1996) pour l'anglais et le japonais, CONLL (Tjong Kim Sang, 2002) pour l'espagnol et l'allemand et ESTER (Galliano *et al.*, 2009) pour le français. Plusieurs systèmes ont été développés pour ces différentes langues. La plupart d'entre eux utilisent soit des méthodes symboliques soit des méthodes statistiques. Les systèmes symboliques sont basés sur des lexiques (listes de prénom, de pays, etc.) et sur un ensemble de règles de réécriture (Maurel *et al.*, 2011; Brun et Ehrmann, 2010; Stern et Sagot, 2010). D'un autre côté, les systèmes statistiques sont basés sur un modèle appris à partir d'un corpus préalablement annoté (Bikel *et al.*, 1997; Raymond et Fayolle, 2010; Béchet et Charton, 2010).

La portabilité entre les langues des systèmes de REN est coûteuse en termes de temps et de connaissances linguistiques requises. Par exemple, les systèmes symboliques nécessitent un traitement lourd incluant, notamment, la modification des règles contextuelles et le développement de nouvelles listes de mots déclencheurs et de noms propres. La rareté et le coût de construction de ces ressources représentent un problème majeur pour certaines langues (Poibeau, 2003; Gamon *et al.*, 1997). D'un autre côté, les systèmes probabilistes se heurtent au problème de disponibilité des corpus d'apprentissages. Ces derniers ne sont pas faciles à constituer et ne sont pas disponibles pour toutes les langues. Plusieurs travaux visent à automatiser le processus d'annotation en exploitant l'encyclopédie multilingue Wikipédia (Nothman *et al.*, 2009) et les corpus multilingues comparables (Klementiev et Roth, 2006).

Dans cet article, nous examinons l'intérêt et le coût associé pour porter un système existant de REN pour du texte bien formé vers une autre langue. Nous nous sommes appuyés pour cela sur le système symbolique français Nemesis (Fourour, 2002). Nous présentons une méthode permettant de porter Nemesis vers l'anglais à moindre coût. Nous décrivons le système Nemesis dans la section 2 et les corpus d'évaluation dans la section 3. Nous présentons ensuite le détail du processus d'adaptation dans la section 4. La section 5 présente les résultats sur les corpus ayant servi aux évaluations et les compare aux résultats obtenus par Stanford Named Entity Recognizer [1] (Finkel *et al.*, 2005), un système de REN développé pour l'anglais. Pour terminer, dans la section 6, nous discutons les apports et les limites de cette approche.

1. Ce système est disponible gratuitement à : http ://nlp.stanford.edu/software/CRF-NER.shtml

2 Nemesis : un système symbolique de REN pour le français

Nemesis (Fourour, 2002) est un système qui permet la délimitation et la catégorisation des entités nommées développé pour le français et pour du texte bien formé (c'est-à-dire qui respecte les règles du français écrit). Il se base essentiellement sur les indices internes et externes définis par McDonald (1996). L'architecture de Nemesis se compose principalement de trois modules qui s'exécutent séquentiellement : prétraitement lexical, projection des lexiques et application des règles.

Prétraitement lexical : segmentation du texte en occurrences de formes et de phrases, puis association des sigles à leur forme étendue.
Projection des lexiques : les lexiques ont été construits soit manuellement, soit automatiquement à partir du Web. Les éléments composant ces lexiques (79 476 éléments) sont répartis en 45 listes selon les catégories dans lesquelles ils sont utilisés : prénom connu, mot déclencheur d'un nom d'organisation (l'élément fait partie de l'entité nommée : « Fédération française de handball »), contexte d'un nom de personne (l'élément appartient au contexte gauche immédiat de l'entité nommée, mais ne fait pas partie de celle-ci : « philosophe Emmanuel Kant »), fin d'un nom d'organisation (l'élément est la dernière forme composant l'entité nommée : « Conseil régional », « Coupe du monde de football »), etc.
La projection des lexiques consiste à associer les étiquettes liées aux lexiques aux différentes formes du texte. Une forme peut avoir plusieurs étiquettes (Washington|prénom-connu|lieu-connu).
Application des règles : les règles de réécriture permettent l'annotation du texte par des balises identifiant les entités nommées (délimitation et catégorisation). Elles sont basées sur des étiquettes sémantiques référant à une forme capitalisée ou à une forme appartenant à un lexique. En tout, Nemesis utilise 93 règles qui s'exécutent dans un ordre prédéfini. Lorsque plusieurs règles s'appliquent, Nemesis opte pour la règle ayant la priorité la plus élevée. Voici un exemple de règles de réécriture :

```
$Clé-oronyme $Article-min [ $Forme-capitalisée+ ]→ ORONYME
```

et le résultat de son application :

"montagne du <ORONYME> Mont-Blanc </ORONYME>"

L'évaluation de Nemesis a été réalisée sur un corpus composé de textes issus du journal Le Monde et du Web (31 000 mots). Les performances sur l'ensemble des entités nommées montrent un rappel de 79 % et une précision de 91 % (Fourour, 2003).

3 Description des corpus et mesure des performances

Trois corpus dont deux en langue française et un autre en langue anglaise ont été utilisés dans nos expérimentations.

Le corpus de référence anglais, *BBN Pronoun Coreference and Entity Type Corpus* [2], est composé d'articles provenant de *Wall Street Journal* manuellement annoté en entités nommées. Le guide d'annotation comporte 12 catégories principales (*Person*, *Facility*, *Organization*, *GPE*, *Location*, *Nationality*, *Product*, *Event*, *Work of Art*, *Law*, *Language*, *Contact-Info*) et plusieurs sous-catégories. Seules les entités sont prises en compte dans l'étiquette. Les formes qui ne font pas partie de l'entité elle-même sont exclues de l'annotation (Mr. <PERSON> Spoon </PERSON>). Ce corpus a été divisé en deux parties, développement (3/4 du corpus) et test (1/4 du corpus).

Le corpus de référence français (Stern et Sagot, 2010) [3] est constitué de dépêches provenant de l'Agence France-Presse (AFP) et contient des annotations manuelles des entités de type Personne, Lieu et Organisation (comprenant les noms d'entreprises). Les formes non constitutives du nom de l'entité lui-même sont exclues de l'annotation (M. <PERSON> Spoon </PERSON>). La taille de ce corpus est bien plus faible que celle du corpus anglais. Pour des raisons de comparaison, nous avons également eu recours à un corpus non annoté constitué des articles du journal Le Monde (2007).

Dans ce travail, seules les entités communes entre les deux corpus annotés ont été retenues pour les expériences (Personne, Organisation, Lieu et Entreprise). La table 1 décrit les différents corpus utilisés dans ce travail. Les performances sont mesurées en termes de rappel et

Corpus	**Langue**	**Nb de mots**	**Nb d'entités**
Corpus BBN (développement)	anglais	938 330	46 478
Corpus BBN (test)	anglais	235 274	11 930
Corpus AFP	français	38 831	1 497
Corpus Le Monde	français	1 010 000	-

TABLE 1: Description des corpus

de précision. Le rappel est défini par le nombre d'entités correctement étiquetées au regard du nombre d'entités étiquetées dans la référence. La précision est le nombre d'entités correctement étiquetées au regard du nombre d'entités correctement et incorrectement étiquetées. La F-mesure combine ces deux mesures.

$$Rappel = \frac{Nombre\ d'entités\ correctement\ étiquetées}{Nombre\ d'entités\ étiquetées\ dans\ la\ référence} \quad (1)$$

$$Précision = \frac{Nombre\ d'entités\ correctement\ étiquetées}{Nombre\ d'entités\ étiquetées\ fournis} \quad (2)$$

$$F - mesure = \frac{2 * rappel * précision}{rappel + précision} \quad (3)$$

2. Catalogue LDC n° LDC2005T33
3. Ce corpus est disponible librement dans le cadre de la distribution de SXPipe

4 Adaptation de Nemesis à l'anglais

La mesure des performances des systèmes de REN dépend directement de la cohérence entre les annotations manuelles et les annotations automatiques. Nous avons donc commencé par ajuster les règles de délimitation et de catégorisation de Nemesis aux normes d'annotation des corpus d'évaluation.

Notre objectif est de porter Nemesis vers l'anglais d'une façon simple et peu coûteuse. La méthode proposée consiste à adapter séparément et séquentiellement les deux principaux éléments constitutifs de Nemesis : les lexiques et les règles de réécriture.

4.1 Adaptation des lexiques

Nous avons construit l'ensemble des lexiques pour l'anglais en traduisant tout simplement les lexiques existants pour le français. La traduction est faite automatiquement en utilisant un dictionnaire bilingue [4] sans aucune information contextuelle. Cela concerne principalement l'ensemble des lexiques de mots déclencheurs. Les lexiques des noms de personne et d'entreprise sont conservés. Cette tâche n'est pas coûteuse en termes de temps et ne demande pas une expertise linguistique (des outils en ligne comme *Google Translate* peuvent être utilisés pour les langues pour lesquelles les dictionnaires électroniques ne sont pas disponibles). Lorsque le dictionnaire comporte plusieurs traductions pour un mot, l'ensemble des traductions sont conservées (nous ne traitons pas à ce niveau les problèmes de polysémie). Une fois la phase de traduction terminée, nous avons utilisé une liste des mots outils en anglais pour écarter certaines entrées pouvant produire de bruit, par exemple le mot *even* qui se présente comme un nom de personne dans le lexique français. En définitive, l'ensemble des lexiques pour l'anglais compte 83 305 entrées.

4.2 Adaptation des règles

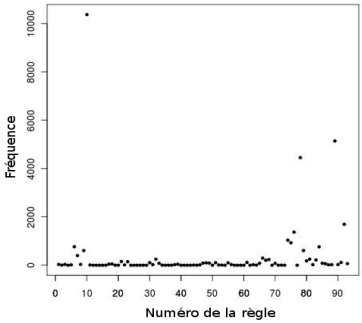
4.2.1 Classification des règles

Avant d'adapter les règles, nous avons commencé par mesurer la fréquence et la précision relatives de chacune des règles utilisées par Nemesis (93 règles). La fréquence représente le nombre de fois où chaque règle est déclenchée pour reconnaître une entité nommée. La figure 1 présente les résultats obtenus sur le corpus Le Monde pour le français (1a) et sur le corpus de développement BBN pour l'anglais (après adaptation des lexiques) (1b). Les observations montrent que la loi de Zipf est respectée pour les deux langues : un nombre limité de règles est à l'origine de la plupart des entités extraites, les autres règles sont rarement déclenchées. Par exemple pour l'anglais, 11 règles couvrent 86% des entités extraites. On remarque aussi que de nombreuses règles ne sont pas déclenchées (38 pour l'anglais contre 17 pour le français).

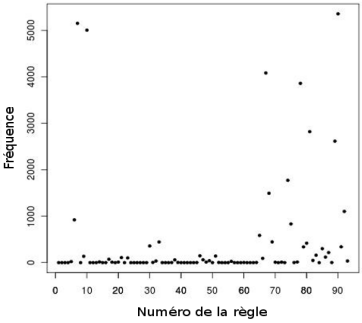
La précision mesure la quantité d'entités correctement reconnues parmi les réponses retournées. En effet, ce n'est pas parce qu'une règle est fréquente qu'elle est pour autant précise.

4. Catalogue ELRA-M0033 (http ://catalog.elra.info/product_info.phpproducts_id=666&language=fr)

Nous avons calculé la précision de chacune des règles déclenchées pour le français sur le corpus AFP et pour l'anglais sur le corpus de développement BBN. La figure 2 présente les résultats obtenus pour l'anglais. Plusieurs règles déclenchées obtiennent une précision relativement faible (32 règles ont une précision inférieure à 50%, ce qui représente environ 60% des règles déclenchées). En se basant sur les critères de fréquence et de précision, nous avons ensuite classé



(a) Français (corpus Le Monde)



(b) Anglais (corpus de développement BBN)

FIGURE 1: Nombre de déclenchements des règles pour le français et pour l'anglais

les règles en différentes catégories, par exemple : règles fréquentes [5] ayant une bonne précision [6] dans les deux langues (7 règles), règles fréquentes pour l'anglais avec une faible précision [7] (3 règles), règles déclenchées seulement pour le français (19 règles), règles non déclenchées dans les deux langues (15 règles), etc. Ces différentes catégories vont nous permettre de déterminer quelles sont les règles à modifier pour la langue anglaise.

5. Fréquence > 1 000
6. Précision > 70 %
7. Précision < 50 %

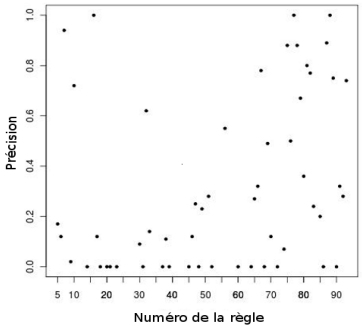


FIGURE 2: Précision des règles déclenchées pour l'anglais (corpus BBN)

4.2.2 Adaptation des règles

Pour adapter les règles, nous avons sélectionné celles n'appartenant pas aux catégories de bonne précision pour l'anglais. Nous avons également éliminé certaines règles déclenchées seulement pour le français car elles sont très spécifiques à cette langue, par exemple :

`[ $Clé-organisation $Article-min $Item $Article-min $Forme-capitalisée+ ] → ORGANISATION`

et un exemple d'application :

"<ORGANISATION> Centre de recherche de Solaize </ORGANISATION>"

Nous avons ensuite modifié les règles sélectionnées de manière à respecter la syntaxe de la langue anglaise (29 règles), comme la règle suivante pour le français :

`$Fonction $Adjectif-de-nationalité [ $Forme-capitalisée+ ] → PERSON`

et le résultat de son application :

"Le président français <PERSON> Jacques Chirac </PERSON>"

et la règle pour l'anglais après adaptation :

`$Adjectif-de-nationalité $Fonction [ $Forme-capitalisée+ ] → PERSON`

et le résultat de son application :

"The French President <PERSON> Jacques Chirac </PERSON>"

5 Résultats

Pour des raisons de comparaison, nous avons dans un premier temps appliqué Nemesis au corpus de test BBN et au corpus AFP. Aucune adaptation n'a été réalisée. La table 2 présente les performances réalisées par catégorie. Nous pouvons remarquer que la reconnaissance des noms de personne demeure satisfaisante (perte d'environ 5 points de F-mesure). Cela s'explique par le fait que les lexiques de Nemesis contiennent des prénoms anglais et qu'il y a des règles de réécriture communes. La reconnaissance se voit fortement dégradée pour les autres catégories. Nous avons ensuite mesuré l'apport de l'adaptation des lexiques et des règles. La table 3 affiche

	Nemesis Français	**Nemesis Anglais (sans aucune adaptation)**
	Corpus AFP (français)	**Corpus de test BBN (anglais)**
	F1 (P/R)	**F1 (P/R)**
Personne	77,33 (85,83/71,03)	71,82 (66,65/77,85)
Lieu	88,82 (90,24/87,44)	37,46 (45,61/31,79)
Organisation	54,24 (65,04/46,51)	1,5 (1,1/2)
Entreprise	37,73 (54,05/30)	10 (36,83/5,84)

TABLE 2: Performances de Nemesis pour le français et l'anglais sans adaptation

les gains obtenus pour chaque adaptation. Ces résultats sont comparés à ceux obtenus avec un système natif (Stanford NER).

	Nemesis Anglais (adaptation lexiques)	**Nemesis Anglais (adaptation lexiques et règles)**	**Stanford NER**
	Corpus BBN (anglais)	**Corpus BBN (anglais)**	**Corpus BBN (anglais)**
	F1 (P/R)	**F1 (P/R)**	**F1 (P/R)**
Personne	77,3 (74,48/80,34)	79,51 (81,08/78,01)	90,9 (88,07/93,9)
Lieu	74,38 (72,5/76,35)	79,17 (78,59/79,76)	90,8 (87,6/94,2)
Organisation	21,02 (24,2/18,58)	38,15 (41,9/35,02)	84,6 (89,2/80,04)
Entreprise	27,52 (50,72/18,88)	30,15 (68,34/19,34)	

TABLE 3: Performances de Nemesis pour le français et l'anglais

L'adaptation des lexiques a permis un apport significatif concernant la reconnaissance des noms de lieu (environ 37 points de F-mesure). La reconnaissance des noms d'organisation et d'entreprise se voit améliorée mais elle reste toutefois faible. En effet, ce problème semble lié à une couverture insuffisante des lexiques de Nemesis et à une ambiguité liée à la catégorisation des organisations et des entreprises (entreprise catégorisée en tant qu'organisation et vice-versa).

L'adaptation des règles montre un gain relativement bon pour la catégorie organisation

(environ 17 points de F-mesure) et une légère amélioration concernant les autres catégories. Les résultats globaux sont bien en dessous de ceux obtenus avec Stanford NER, surtout pour la catégorie organisation. Stanford NER est un système à base d'apprentissage développé pour l'anglais. Pour ce dernier, les noms d'entreprises sont catégorisés comme étant organisation.

6 Discussion

Cette approche peu coûteuse pour porter Nemesis vers l'anglais montre des résultats satisfaisants pour la reconnaissance des noms de personne et de lieu. Elle montre ses limites pour la reconnaissance des noms d'organisation et d'entreprise. L'analyse des résultats obtenus nous a permis de souligner deux problèmes récurrents de la reconnaissance des entités nommées : la délimitation des entités nommées et la délimitation des catégories.

Les règles d'annotation sont loin de faire consensus. Elles suscitent toujours des vives discussions et plusieurs remises en cause du guide d'annotation, par exemple dans le cadre de la campagne ETAPE (Évaluations en Traitement Automatique de la Parole). Le premier problème rencontré concerne la délimitation des entités nommées. En effet, les règles de délimitation de Nemesis ne sont pas les mêmes que celles utilisées pour annoter le corpus AFP et le corpus BBN. Par exemple, Nemesis inclut les titres dans les noms de personne (<PERSON> M. Dorgan </PERSON>) alors que ces derniers ne sont pas inclus dans le corpus AFP (M. <PERSON> Dorgan </PERSON>). Nous avons dû adapter les règles de délimitation de Nemesis aux spécificités du corpus traité. Cependant, plusieurs erreurs de délimitation ont été relevées. Le deuxième problème concerne la délimitation des catégories. Nemesis adopte une catégorisation fine (5 catégories et 30 sous-catégories). Cette typologie a l'avantage de pouvoir s'adapter aux typologies moins fines en regroupant des sous-catégories. Nous avons donc adapté la typologie de Nemesis en fonction des catégories du corpus AFP et du corpus BBN. Toutefois, beaucoup d'erreurs sont dues à une incohérence de catégorisation. Par exemple, la notion d'organisation et d'entreprise n'est pas identique entre Nemesis et le corpus BBN (l'entité « Federal Reserve Board » est annotée comme étant une entreprise par le système Nemesis alors qu'elle est considérée comme étant une organisation dans le corpus BBN).

7 Conclusion et perspectives

Cet article présente une méthode peu coûteuse pour porter un système symbolique de REN dédié pour le français vers l'anglais. L'adaptation est basée principalement sur les ressources développées pour le français. Elle consiste à traduire les lexiques français et à adapter légèrement quelques règles de la grammaire. L'évaluation du système adapté montre des résultats satisfaisants pour la reconnaissance des noms de personne et de lieu. En revanche, les résultats restent insuffisants pour les noms d'organisation et d'entreprise. Un traitement plus approfondi est nécessaire pour ces deux catégories. Pour cela, nous comptons mesurer l'impact d'un enrichissement des lexiques de Nemesis (notamment les listes d'organisation et d'entreprise). D'un autre côté, nous envisageons d'adapter les règles de réécriture automatiquement et de tester cette méthode sur d'autres langues qui sont moins proches du français que l'anglais.

Références

Béchet, F. et Charton, E. (2010). Unsupervised knowledge acquisition for extracting named entities from speech. *In Proceedings of the 35th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'10)*, pages 5338–5341, Dallas, Texas, USA.

Bikel, D. M., Miller, S., Schwartz, R. et Weischedel, R. (1997). Nymble : a high-performance learning name-finder. *In Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP'97)*, pages 194–201, Washington, DC, USA.

Brun, C. et Ehrmann, M. (2010). Un système de détection d'entités nommées adapté pour la campagne d'évaluation ester 2. *In Actes de la 17ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'10)*, Montréal, Canada.

Ehrmann, M. (2008). *Les Entités Nommées, de la Linguistique au TAL - Statut Théorique et Méthodes de Désambiguïsation*. Thèse de doctorat, Université Paris 7, France.

Finkel, J. R., Grenager, T. et Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. *In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL'05)*, pages 363–370, Ann Arbor, Michigan, USA.

Fourour, N. (2002). Nemesis, un système de reconnaissance incrémentielle des entités nommées pour le français. *In Actes de la 9ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'02)*, pages 265–274, Nancy, France.

Fourour, N. (2003). Apport du web dans la reconnaissance des entités nommées. *In Revue Québécoise de Linguistique (RQL)*, pages 41–60.

Galliano, S., Gravier, G. et Chaubard, L. (2009). The ester 2 evaluation campaign for the rich transcription of french radio broadcasts. *In Proceedings of the 10th conference Interspeech*, pages 2583–2586, Brighton, UK.

Gamon, M., Lozano, C., Pinkham, J. et Reutter, T. (1997). Practical experience with grammar sharing in multilingual nlp. *In Workshop From research to commercial applications : making NLP work in practice*, pages 49–56, Madrid, Spain.

Grishman, R. et Sundheim, B. (1996). Message Understanding Conference-6 : a brief history. *In Proceedings of the 16th conference on Computational linguistics (COLING'06)*, pages 466–471, Copenhagen, Denmark.

Klementiev, A. et Roth, D. (2006). Named entity transliteration and discovery from multilingual comparable corpora. *In Proceedings of Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL'06)*, pages 82–88, New York, USA.

Maurel, D., Friburger, N., Antoine, J.-Y., Eshkol-Taravella, I. et Nouvel, D. (2011). Cascades de transducteurs autour de la reconnaissance des entités nommées. *In Traitement Automatique des Langues (TAL)*, pages 69–96.

McDonald, D. D. (1996). Internal and external evidence in the identification and semantic categorization of proper names. *In Corpus processing for lexical acquisition*, pages 21–39. MIT Press, Cambridge, MA, USA.

Nothman, J., Murphy, T. et Curran, J. R. (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 (EACL'09)*, pages 612–620, Athens, Greece.

POIBEAU, T. (2003). The multilingual named entity recognition framework. *In Proceedings of the 10th Conference on European chapter of the Association for Computational Linguistics (EACL'03)*, pages 155–158, Budapest, Hungary.

RAYMOND, C. et FAYOLLE, J. (2010). Reconnaissance robuste d'entités nommés sur de la parole transcrite automatiquement. *In Actes de la 17ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'10)*, pages 19–23, Montréal, Canada.

STERN, R. et SAGOT, B. (2010). Détection et résolution d'entités nommées dans des dépêches d'agence. *In Actes de la 17ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'10)*, Montréal, Canada.

TJONG KIM SANG, E. F. (2002). Introduction to the conll-2002 shared task : language-independent named entity recognition. *In Proceedings of the 6th Workshop on Computational Language Learning (CoNLL'02)*, pages 155–158, Taipei, Taiwan.