Impact de la nature et de la taille des corpus d'apprentissage sur les performances dans la détection automatique des entités nommées

Ollagnier Anaïs[1, 2], Fournier Sébastien[1], Bellot Patrice[1,2], Béchet Frédéric[3]
(1) Aix-Marseille Université, CNRS, ENSAM, Université de Toulon LSIS UMR 7296, 13397, Marseille, France
(2) Aix-Marseille Université, CNRS, CLEO OpenEdition UMS 3287, 13451, Marseille, France
(3) Aix-Marseille Université, CNRS, LIF UMR 7279, 13288, Marseille, France
anais.ollagnier@openedition.org

Résumé. Nous présentons une étude comparative sur l'impact de la nature et de la taille des corpus d'apprentissage sur les performances dans la détection automatique des entités nommées. Cette évaluation se présente sous la forme de multiples modulations de trois corpus français. Deux des corpus sont issus du catalogue des ressources linguistiques de l'Association Européenne pour les Ressources Linguistiques (ELRA) et le troisième est composé de documents extraits de la plateforme OpenEdition.org.

Abstract. We present a comparative study on the impact of the nature and size of the training corpus on performance in automatic named entities recognition. This evaluation is in the form of multiple modulations on three French corpus. Two corpora are from the catalog of the European Language Resources Association (ELRA) and the third is composed of documents extract from the OpenEdition.org platform.

Mots-clés : Reconnaissance d'entités nommées, Adaptation au domaine, comparaison d'outils.

Keywords: Named entity recognition, Domain adptation, performance comparison.

1 Introduction

La reconnaissance des entités nommées (REN) est une sous tâche de l'activité d'extraction d'information. Elle consiste à rechercher des objets textuels (c'est à dire un mot ou un groupe de mots) catégorisables dans des classes telles que les noms de personnes, les noms d'organisations ou d'entreprises, les noms de lieux, etc. De nombreuses campagnes d'évaluation, telles que la Message Understanting Conference (Grishman & Sundheim, 1996), l'Automatic Content Extraction [1] ou encore la Document Understanding Conference [2] pour l'Europe, permettent d'évaluer et de comparer les différents systèmes de détection d'entités nommées. Les résultats présentés lors des grandes campagnes d'évaluation montrent régulièrement des F-mesures avoisinant les 90% (Marsh & Perzanowski, 1998). Cependant ces résultats sont à prendre avec du recul, en effet, les tâches d'évaluation proposées sont souvent spécialisées et orientées sur un seul type de données, et les capacités de généralisation des systèmes sont rarement évaluées. Cependant les besoins en annotation en entités sont très divers, que ce soit par rapport au jeu d'étiquettes considéré, au type de texte ou au domaine sémantique visés. Il est donc intéressant d'évaluer les capacités de généralisation des systèmes de REN ainsi que d'étudier des processus d'adaptation peu couteux à un nouveau cadre d'utilisation.

Dans cet article nous allons nous focaliser sur l'étude de cette capacité à généraliser pour un système donné à travers le traitement de corpus provenant de la plateforme OpenEdition [3]. Ces corpus sont particulièrement intéressants car ils présentent une diversité de formes (livres, revues, actualité et blog) ainsi que de contenu (thèmes très variés issus des domaines des SHS). En confrontant des outils génériques d'extraction d'entités nommées à ces corpus, nous allons pouvoir déterminer quelle est la meilleure stratégie pour adapter un système existant à un nouveau cadre d'utilisation, tout en minimisant l'effort d'adaptation.

1. http ://www.itl.nist.gov/iad/mig/tests/ace/
2. http ://duc.nist.gov/
3. http ://www.openedition.org/

2 Détecteur d'entités nommées et corpus d'apprentissage

Les travaux en REN sont classés principalement selon trois types d'approches : les approches symboliques, les approches statistiques et les approches hybrides. Les approches symboliques (Poibeau, 2003) se basent sur des règles écrites à la main, ces approches ont l'avantage de privilégier la précision des détections mais présentent l'inconvénient d'être très dépendantes du domaine d'application et d'être coûteuses à mettre en place si l'on veut obtenir une couverture satisfaisante (Nadeau & Sekine, 2007). Les approches statistiques (Burger *et al.*, 2002) se basent sur un mécanisme d'apprentissage à partir d'un corpus pré-étiqueté. De manière générale, ce type d'approche privilégie le rappel plutôt que la précision. Cependant, comme précisé précédemment, ces approches nécessitent l'utilisation de données annotées comme base d'apprentissage. Les approches hybrides consistent en une combinaison des deux approches précédentes (Shaalan & Oudah, 2014). Les campagnes d'évaluation des systèmes de REN ont montrées que les différences de performance entre les systèmes dépendent plus du soin mis dans l'adaptation fine des modèles (règles ou traits) que des choix des modèles théoriques. Ainsi les systèmes à base de règles rédigées manuellement (soit purement symboliques, soit hybrides) sont souvent les plus performants (Sun, 2010) lors des conférences ACL, MUC-7 ou CoNLL03. Cependant ces résultats sont constatés sur des domaines précis, lors d'évaluations sur des domaines plus larges les résultats se dégradent (Mansouri *et al.*, 2008). Le but de cette étude étant d'adapter à moindre coût des systèmes de REN à de nouveaux types de corpus, nous avons opté pour l'utilisation d'outils basés sur des approches statistiques pour lesquelles nous avons pu trouver des corpus français déjà annotés. Afin d'établir nos corpus d'apprentissage nous avons utilisé deux corpus déjà annotés en EN en français extraits du catalogue ELRA [4] : le corpus Quaero d'émissions télé-radio-diffusées (ELRA-S0349) et le corpus Quaero de presse anciennes (ELRA-W0073). Ainsi qu'un troisième corpus constituant la *cible* de notre étude a été annoté manuellement à partir des différents types de documents présents sur la plateforme *OpenEdition.org*. Dans les sections suivantes nous allons détailler les différentes caractéristiques de ces corpus ainsi que les outils de détection des entités nommées choisis.

2.1 Corpus Open Edition (OE)

Un corpus issu de la plateforme *OpenEdition.org* a été extrait et entièrement annoté manuellement par un seul annotateur. Il se compose de documents extraits des quatre plateformes Open Edition (Revues, Books, Calenda, Hypothèses) qui est un vaste catalogue de publications scientifiques en sciences humaines et sociales (SHS). La première plateforme, *Revues.org* est composée de revues en SHS. La seconde, *Books* met à disposition en libre accès l'intégralité du contenu de plus de mille ouvrages en SHS. La troisième plateforme, *Calenda* répertorie des actualités dans le domaine des SHS. Enfin la quatrième plateforme, *Hypothèses* est constituée de blogs académiques. Ces différentes plateformes dédiées à la valorisation des publications dans les sciences humaines et sociales possèdent plusieurs spécificités que ce soit au niveau structurel que thématique. Parmi ces spécificités nous pouvons citer : des documents de longueur variable, des dimensions stylistiques variées, la présence de multilinguisme ainsi que l'utilisation pour certains documents (notamment au sein des plateformes Books et Revues) de guide de recommandation aux auteurs. La table 1 présente des exemples de phrases annotés en entités nommées du corpus OE.

2.2 Corpus Quaero

L'ELRA met à disposition deux corpus Quaero. Ces corpus ont été entièrement annotés manuellement selon la définition étendue et structurée d'entités nommées Quaero, qui distingue les "types" et les "composants" d'entités (Rosset *et al.*, 2011). Dans le cadre de notre évaluation, nous avons décidé de conserver les annotations les plus communes entre nos trois corpus, à savoir, les noms de personnes, les toponymes et les noms d'organisations. Les entités structurées de Quaero ont été "*aplaties*" en ne gardant que la structuration de plus haut niveau (Ex : <pers.ind> <name.first> Jacques </name.first> <name.last> Chirac </name.last> </pers.ind> : <pers> Jacques Chirac </pers>). Deux corpus ont été utilisés :

1. Corpus Quaero d'émissions télé-radio-diffusées annoté en entités nommées (QO)
   Le corpus QO consiste en l'annotation manuelle du corpus ESTER2 et du corpus d'évaluation de systèmes de reconnaissance de la parole Quaero (Galliano *et al.*, 2006). Ce corpus comporte la particularité de n'être

4. http ://www.elra.info/ELRA.html

Books Plus tard, <pers> R. Janko </pers> s'est inspiré des remarques de <pers> T. Weischadle </pers>. À propos de l'introduction, il relève plusieurs exceptions aux remarques formulées par <pers> T. Weischadle </pers>, parmi lesquelles la présence du nom du dieu au vocatif.
Revues <pers> Michael Spens </pers>, (<pers> M. Spens </pers>, Paysages contemporains, <date> 2005 </date>) examine les créations des paysagistes de différentes générations <pers> B. Lassus </pers>, <pers> L. Halprin </pers>
Blog C'est <pers> Jean-Marc Berlière </pers> qui l'annonce: les <org> Archives de la Préfecture de Police </org> vont fermer pour déménager au <loc> Pré Saint-Gervais </loc>, dans la banlieue est de la capitale.
Calenda <date> 12 février </date> (9h30-12h) - Collecter - « Les éboueurs entre collecte des ordures ménagères et récupération », <pers> Christophe Clerfeuille </pers> (éboueur, président de l'<org> association Collectif Ripeurs </org>) et <pers> Stéphane Le Lay </pers> (sociologue, <org> CNAM </org>)

Table 1 – Exemples de phrases annotées en EN du corpus OE provenant du site *OpenEdition.org*

constitué que de données non structurées. En outre, les textes comportent de nombreuses disfluences de surface (hésitations, répétition, etc.) qui peuvent générer des difficultés dans la généralisation des modèles. Ce corpus a été utilisé lors de la campagne ESTER2 [5] durant laquelle les meilleurs résultats oscillent entre 78,4% et 92,5% de F-mesure (Nouvel *et al.*, 2010). Dans le cadre de notre évaluation, nous avons utilisé la partie dédiée à l'apprentissage constituée d'actualités télé-radio-diffusées, soit 188 émissions pour 1 291 225 mots afin de constituer nos corpus d'apprentissage.

2. Corpus Quaero de presse ancienne étendue en entités nommées (QP)
   Le corpus QP consiste en l'océrisation de 76 numéros de journaux, publiés entre 1890-1891, fournis par la Bibliothèque Nationale de France. Trois publications sont utilisées (Le Temps, La Croix et Le Figaro) pour un total de 295 pages. Ce corpus présente plusieurs caractéristiques. Premièrement, il est entièrement constitué de documents OCR-isés dont le taux de qualité a été estimé bon par rapport à l'état de l'art dans le domaine (Character Error Rate de 5,09 et Word Error Rate de 36,59) (Rosset *et al.*, 2012). Quelques erreurs résiduelles persistent notamment dans la reconnaissance de certain caractère comme le « e » souvent écrit « o ». Deuxièmement, ce corpus réfère à une période assez ancienne dont les informations diffèrent des actuelles. Et troisièmement, la particularité des journaux OCR-isés basés sur des éditions papiers se retrouve également dans leurs structures en colonnes. De ce fait, les textes conservent de nombreux sauts de ligne ainsi que de nombreuses césures. Dans l'article de Rosset (Rosset *et al.*, 2012), trois systèmes de REN à base de méthodes statistiques sont évalués sur ce corpus. Les résultats présentent un taux d'erreur oscillant entre 44,2 et 60,3 (Slot Error Rate). Dans le cadre de notre évaluation, nous avons utilisé la partie dédiée à l'entraînement, constituée de 231 pages pour 1 297 742 mots afin de constituer nos corpus d'apprentissage.

2.3 Les outils

Nous avons choisi plusieurs outils basés sur des approches d'apprentissage statistiques, à savoir, l'entropie maximale (Chieu & Ng, 2002) ainsi que les champs conditionnels aléatoires (CRFs) (Sobhana *et al.*, 2010), avec et sans *boosting* (Favre *et al.*, 2005). Le premier outil que nous avons testé est intégré dans l'environnement OpenNLP [6]. Plusieurs outils sont disponibles au sein de cette plateforme qui s'intéresse à la tokénisation, la détection de syntagmes, la détection des entités nommées (NameFinder), tous basés sur le *framework* Maxent implémentant des modèles à maximum d'entropie. Le second est Stanford NER [7], il fournit un détecteur d'entités nommées basé sur des CRFs. Une évaluation de ces deux outils présentée par (Rodriquez *et al.*, 2012) sur des documents OCR-isés anglais extraits du Wiener Library dataset et du King College London dataset montre une meilleure performance globale pour Stanford NER (F-mesure : 54%) et une performance beaucoup plus faible pour OpenNLP surtout au niveau de la précision (F-mesure : 19%). Le troisième outil que nous avons choisi d'utiliser est LiaNE [8], basé sur une combinaison d'un modèle HMM pour la prédiction d'étiquettes pour chaque

5. http ://www.afcp-parole.org/camp_eval_systemes _transcription/index.html
6. http ://opennlp.apache.org
7. http ://nlp.stanford.edu/software/CRF-NER.shtml
8. http ://pageperso.lif.univ-mrs.fr/ frederic.bechet/download.html

mot et d'un modèle CRF pour l'étiquetage de séquences constituant les entités nommées. Sa particularité est d'être orienté sur le traitement de données orales. LiaNE a obtenu lors de l'évaluation ESTER2 une F-mesure de 78,4% (Nouvel *et al.*, 2010).

3 Expérimentations

Notre évaluation s'est déroulée en deux parties. La première partie s'est établie en tenant compte de la nature des corpus, c'est à dire que nous avons voulu au vu des spécificités des deux corpus Quaero, évaluer l'impact de leurs caractéristiques lors de l'apprentissage des outils ainsi que leur portabilité. La seconde partie de notre évaluation s'est focalisée sur l'adaptation des modèles en utilisant une partie du corpus OE pour apprendre ou adapter les modèles d'EN de nos systèmes.

Evaluation de la nature et de la taille des corpus d'apprentissage Dans le cadre de nos évaluations de l'impact de la nature et de la taille des corpus d'apprentissage sur les performances en NER, nous avons utilisé les données du corpus OE afin de constituer notre corpus de test et les corpus QO et QP comme corpus d'apprentissage. Les trois systèmes LiaNE, OpenNLP et StandfordNER sont comparés dans la table 2 et la table 3.

	Précision (%)		Rappel (%)		F-mesure (%)	
Corpus	QO	QP	QO	QP	QO	QP
LiaNE	**55,7**	15,5	48,3	21,9	**51,8**	18,1
Open NLP	30,4	27,3	29,0	23,7	29,7	25,4
Stanford NER	44,0	**36,4**	**52,3**	**50,7**	47,8	**42,3**

Table 2 – Performance en NER sur le corpus de test OE avec 3 systèmes et 2 corpus d'apprentissage différents

De manière générale, nous pouvons constater que les performances sont beaucoup plus faibles que celles obtenues par ces mêmes systèmes dans la littérature. Cette dégradation a deux explications : d'une part à cause des spécificités du corpus OE, assez différent des corpus Quaero utilisés pour l'apprentissage ; d'autre part à cause de l'utilisation directe des corpus Quaero pour l'apprentissage des systèmes, sans adaptation des traits utilisés pour décrire les données d'apprentissage. Cette table nous permet de constater un certain écart sur les performances globales des outils sur les différents corpus. Nous pouvons noter une constante assez faible de la part de OpenNLP. A l'inverse, nous pouvons constater que LiaNE obtient une performance bien meilleure après son entraînement sur le corpus QO, ce qui s'explique par la participation de LiaNE à la campagne ESTER2, ce qui a favorisé son adaptation aux données non structurées présentes dans le corpus QO. Les résultats obtenus par Stanford NER sont les plus constants. Au vu de cette table nous pouvons tirer deux enseignements de ces résultats : premièrement, malgré le fait que le corpus QP contienne du texte écrit *a priori* plus proche des données OE que le corpus oral QO, les résultats obtenus en entraînant les systèmes sur QP sont globalement plus faible que ceux obtenus avec QO. Sans doute la distance temporelle entre le données QP et OE est-elle trop forte. Deuxièmement, en dehors de la méthode et du choix du corpus d'apprentissage, les performances des systèmes sont fortement affectées par l'adaptation des traits utilisés pour décrire les données. Ainsi, bien que les deux systèmes LiaNE et Standford NER partagent tous deux une approche à base de CRF, le système LiaNE donne les meilleurs résultats sur le corpus QO alors qu'il se dégrade sur le corpus QP ; tandis que le système Standford NER affiche plus de stabilité sur les deux corpus, bien qu'atteignant des performances plus faibles. On peut donc voir que l'adaptation spécifique du système LiaNE sur les données QO lui permet de tirer un meilleur parti de ces données lors de l'apprentissage.

Outils	F-mesure (%)				
Partitions	100%	80%	60%	40%	20%
LiaNE	51,8	52,1	**52,5**	39,3	37,0
Open NLP	29,7	29,8	**30,1**	27,1	26,9
Stanford NER	47,8	48,9	**50,6**	45,8	44,7

Table 3 – Les valeurs de F-mesure obtenues sur le corpus OE en faisant varier la taille du corpus d'apprentissage QO pour chacun des outils

La deuxième expérience effectuée sur les mêmes données a consisté à faire varier la taille du corpus d'apprentissage QO. Plusieurs corpus ont été constitués en découpant proportionnellement le corpus QO, soit 1 291 225 mots. Une scission de 20% composée de 433 951 mots, une scission de 40% de 657 071 mots, une scission de 60% de 888 280 mots et une scission de 80% de 1 000 009 mots. Nous avons ensuite évalué ces données d'apprentissage en les testant sur le corpus OE soit 102 744 mots. Les résultats présentés dans la table 3 nous permettent de constater que, de manière générale, la quantité de données d'apprentissage n'a pas un impact linéaire sur les performances des systèmes. Nous pouvons constater que pour les trois outils les meilleurs résultats sont obtenus sur la scission constituée de 60% du corpus d'apprentissage. Plus précisément, nous pouvons noter une différence moyenne de 1,3% entre la scission de 60% et 100% et une différence moyenne de 8,4% entre la scission de 60% et 20%. Ces constations nous permettent de corroborer plusieurs études qui démontrent, que pour certaine tâche, un trop grand nombre de données d'apprentissage ne correspondant pas exactement aux données de tests ne permet pas d'augmenter la couverture mais qu'au contraire ceci peut engendrer du bruit et de la confusion (Biber, 1993; Gildea, 2001).

Evaluation de l'adaptation des modèles au corpus OE Cette série d'évaluation a pour but de mesurer l'impact sur les performances de nos modèles de l'ajout de données correspondant à la tâche cible. Pour cela nous présentons deux séries d'expériences effectuées en validation croisée sur le corpus OE et OE+QO. Chaque corpus d'apprentissage de OE (5 validations croisées) est composé d'environ 79 000 mots pour environ 1200 noms de personnes, 495 toponymes et 678 noms d'organisations. Les corpus de test sont composés d'environ 22 700 mots pour environ 560 noms de personnes, 220 toponymes et 205 noms d'organisations. Les évaluations OE+QO ajoutent le corpus QO à chaque partition d'apprentissage du corpus OE.

	Précision (%)		Rappel (%)		F-mesure (%)	
Corpus	OE	OE+QO	OE	OE+QO	OE	OE+QO
LiaNE	18,1	**60,0**	**66,7**	**56,3**	28,6	**57,9**
Open NLP	20,0	28,7	52,8	42,9	28,3	34,3
Stanford NER	**44,3**	56,3	56,3	48,3	**49,1**	51,1

Table 4 – Performances obtenues en évaluation croisée sur le corpus OE avec ou sans le corpus QO lors de l'apprentissage

Nous pouvons voir que malgré la petite taille du corpus OE, comparé au corpus QO, le système Standford NER permet d'obtenir des performances comparables (F-mesure : 49,1%) aux meilleures performances obtenues dans les tables 2 et 3. Le système LiaNE, lorsqu'il est appris uniquement sur le corpus OE présente de bons résultats de rappel, mais la présence de trop nombreux faux positifs conduit à un effondrement de la précision. L'ajout du corpus QO améliore significativement les résultats pour les systèmes LiaNE et OpenNLP, par contre il est surprenant de constater que le système Standford NER ne tire que peu de bénéfice de cet ajout. Nous pouvons également noter, si nous reprenons les résultats de la table 3, que l'ajout du corpus d'apprentissage OE au corpus d'apprentissage QO augmente dans le cas de LiaNE et Stanford NER la valeur de la précision, du rappel ainsi que de la F mesure par rapport à l'apprentissage sur QO seul. La combinaison des deux corpus a permis d'améliorer la couverture. En conclusion nous pouvons dire que l'ajout de corpus annoté spécifique permet d'améliorer globalement les résultats de tous les outils (F-mesure moyenne : 47,8%) comparativement aux résultats obtenus pour le seul corpus QO (F-mesure moyenne : 43,1%.

Conclusion

Les expériences effectuées dans cette étude sur plusieurs outils de REN et plusieurs corpus d'apprentissage ont permis de mettre en évidence les points suivants : tout d'abord une perte importante de performance est constatée lorsque l'on change le contexte applicatif sur lesquels les modèles ont été appris ; malgré cette baisse de performance, des corpus telles que Quaero sont utiles pour amorcer l'apprentissage de premiers modèles en obtenant des résultats acceptables. Entre le choix du type de document (texte écrit, transcription de l'oral) et la période temporelle visée, les expériences comparatives ont clairement montré l'impact de cette distance temporelle, au regard des plus faibles performances obtenues avec QP plutôt que QO. Enfin, l'ajout d'un corpus d'adaptation, même de taille réduite, permet d'augmenter significativement les performances, même sans autre adaptation au domaine.

Références

Biber D. (1993). Representativeness in corpus design. *Literary and linguistic computing*, **8**(4), 243–257.

Burger J. D., Henderson J. C. & Morgan W. T. (2002). Statistical named entity recognizer adaptation. In *proceedings of the 6th conference on Natural language learning-Volume 20*, p. 1–4 : Association for Computational Linguistics.

Chieu H. L. & Ng H. T. (2002). Named entity recognition : a maximum entropy approach using global information. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, p. 1–7 : ACL.

Favre B., Béchet F. & Noera P. (2005). Robust named entity extraction from large spoken archives. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, p. 491–498, Stroudsburg, PA, USA : Association for Computational Linguistics.

Galliano S., Geoffrois E., Gravier G., Bonastre J.-F., Mostefa D. & Choukri K. (2006). Corpus description of the ester evaluation campaign for the rich transcription of french broadcast news. In *Proceedings of LREC*, volume 6, p. 315–320.

Gildea D. (2001). Corpus variation and parser performance. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, p. 167–202.

Grishman R. & Sundheim B. (1996). Message understanding conference-6 : A brief history. In *COLING*, volume 96, p. 466–471.

Mansouri A., Suriani L. A. & Mamat A. (2008). Named entity recognition approaches. *International Journal of Computer Science and Network Security*, **8**(2), 339–344.

Marsh E. & Perzanowski D. (1998). Muc-7 evaluation of ie technology : Overview of results.

Nadeau D. & Sekine S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, **30**(1), 3–26.

Nouvel D., Soulet A., Antoine J.-Y., Maurel D. & Friburger N. (2010). Reconnaissance d'entités nommées : enrichissement d'un système à base de connaissances à partir de techniques de fouille de textes. In *Traitement Automatique des Langues Naturelles*.

Poibeau T. (2003). The multilingual named entity recognition framework. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 2*, p. 155–158 : Association for Computational Linguistics.

Rodriquez K. J., Bryant M., Blanke T. & Luszczynska M. (2012). Comparison of named entity recognition tools for raw ocr text.

Rosset S., Grouin C., Fort K., Galibert O., Kahn J. & Zweigenbaum P. (2012). Structured named entities in two distinct press corpora : Contemporary broadcast news and old newspapers. In *Proceedings of the Sixth Linguistic Annotation Workshop*, p. 40–48 : Association for Computational Linguistics.

Rosset S., Grouin C. & Zweigenbaum P. (2011). *Entités nommées structurées : guide d'annotation Quaero*. LIMSI-Centre national de la recherche scientifique.

Sekine S. (1997). The domain dependence of parsing. In *Proceedings of the fifth conference on Applied natural language processing*, p. 96–102 : ACL.

Shaalan K. & Oudah M. (2014). A hybrid approach to arabic named entity recognition. *Journal of Information Science*, **40**(1), 67–87.

Sobhana N., Mitra P. & Ghosh S. (2010). Conditional random field based named entity recognition in geological text. *International Journal of Computer Applications*.

Sun B. (2010). Named entity recognition : Evaluation of existing systems. *Thèse*.

Zweigenbaum P. & Ben Abacha A. (2012). Une étude comparative empirique sur la reconnaissance des entités médicales. *TAL*, **53**.