Repérage des entités nommées pour l'arabe : adaptation non-supervisée et combinaison de systèmes

Souhir Gahbiche-Braham[1,2] Hélène Bonneau-Maynard[1,2] Thomas Lavergne[1]
François Yvon[1,2]
(1) LIMSI-CNRS, (2) Université Paris-Sud
{souhir.gahbiche,helene.maynard,thomas.lavergne,francois.yvon}@limsi.fr

RÉSUMÉ
La détection des Entités Nommées (EN) en langue arabe est un prétraitement potentiellement utile pour de nombreuses applications du traitement des langues, en particulier pour la traduction automatique. Cette tâche représente toutefois un sérieux défi, compte tenu des spécificités de l'arabe. Dans cet article, nous présentons un compte-rendu de nos efforts pour développer un système de repérage des EN s'appuyant sur des méthodes statistiques, en détaillant les aspects liés à la sélection des caractéristiques les plus utiles pour la tâche ; puis diverses tentatives pour adapter ce système d'une manière entièrement non supervisée.

ABSTRACT
Named Entity Recognition for Arabic : Unsupervised adaptation and Systems combination

The recognition of Arabic Named Entities (NE) is a potentially useful preprocessing step for many Natural Language Processing Applications, such as Machine Translation. This task is however made very complex by some peculiarities of the Arabic language. In this paper, we present a summary of our recent efforts aimed at developing a statistical NE recognition system, with a specific focus on feature engineering aspects. We also report several approaches for adapting this system in an entirely unsupervised manner to a new domain.

MOTS-CLÉS : Adaptation non supervisée, Repérage des entités nommées.

KEYWORDS: Unsupervised domain adaptation, named entity recognition.

1 Introduction

La détection des Entités Nommées (EN) est un élément essentiel à de nombreuses tâches de TAL, qu'elles soient mono ou multilingues, comme la recherche d'information ou la traduction automatique. En témoignent les nombreuses campagnes d'évaluation internationales (MUC, CoNLL, ACE) ou nationales (ESTER) organisées au cours des 15 dernières années.

Nous nous intéressons à cette question dans un contexte de traduction automatique statistique depuis l'arabe vers le français (ou l'anglais). Il est fréquent, en effet, qu'une EN à traduire ne soit pas présente dans les corpus parallèles qui servent à entraîner les systèmes de traduction. Dans (Gahbiche-Braham *et al.*, 2011), nous avons observé que parmi les 1% des formes inconnues, environ un quart correspondent à des EN. Dans ce cas, le système recopie, par défaut, le mot inconnu *verbatim* dans la sortie, alors qu'il serait préférable, pour une EN, de produire une forme

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 487–494,
Grenoble, 4 au 8 juin 2012.

translittérée en alphabet latin (Hermjakob *et al.*, 2008; Zhang *et al.*, 2011), ou encore de consulter des dictionnaires. (Daumé III et Jagarlamudi, 2011) adaptent leur système de traduction en créant des dictionnaires à partir de mots fréquents dans le domaine cible. L'étiquetage en EN apparait donc comme un prétraitement potentiellement utile à la traduction.

L'arabe est une langue morphologiquement riche et complexe. L'analyse automatique des mots arabes est compliquée par l'absence de voyellation dans les textes écrits d'une part (Habash, 2010), et d'autre part par l'existence de nombreuses variantes orthographiques, notamment sur les noms propres, ce qui multiplie les formes inconnues dans les textes. L'étiquetage en EN en langue arabe représente de nombreux défis intéressants : l'arabe se caractérise par le manque de ressources dictionnairiques et surtout par l'absence de distinction majuscule/minuscule qui est un indicateur très utile pour identifier les noms propres dans les langues utilisant l'alphabet latin.

À la suite de nombreux travaux, nous abordons cette tâche avec des outils d'apprentissage automatique et utilisons le modèle des champs markoviens conditionnels (ou CRF (Lafferty *et al.*, 2001)), avec l'implémentation présentée dans (Lavergne *et al.*, 2010), qui permet de construire des modèles intégrant un très grand nombre de descripteurs. Cette démarche pose la question de la pertinence des corpus d'apprentissage au regard des données de test. Nous traitons cette question en explorant les possibilités d'une adaptation non-supervisée. Nous proposons enfin une hybridation entre un système statistique et un système symbolique. Le reste de l'article est organisé comme suit. Dans la section 2, nous passons en revue des travaux sur le repérage des EN dans les textes arabes, et sur l'adaptation de modèles statistiques. Nous présentons dans la section 3 les expériences qui ont conduit au développement de notre système de base. L'adaptation de notre système est décrite à la section 4. La section 5 conclut ces travaux.

2 État de l'art

2.1 Étiquetage en EN pour l'arabe

Les premiers travaux sur la reconnaissance des EN pour l'arabe datent de 1998 et reposent sur des méthodes à base de règles (Maloney et Niv, 1998), voir également le travail plus récent de (Shaalan et Raza, 2009) ou de (Zaghouani *et al.*, 2010). (Samy *et al.*, 2005) utilisent un corpus parallèle pour extraire des EN en arabe. Ils utilisent un étiqueteur à base de règles enrichies avec un lexique monolingue espagnol pour extraire les EN en espagnol qui sont, par la suite, translittérées vers l'arabe. (Zitouni *et al.*, 2005) utilisent des techniques d'apprentissage automatique (des *Maximum Entropy Markov Models*) en considérant des jeux de descripteurs idoines, et parviennent à de très bons résultats.

Ces travaux ont été prolongés en particulier par Benajiba et ses co-auteurs, et ont donné lieu notamment à la construction du corpus ANER (voir section 3). Dans une première approche (Benajiba et Rosso, 2007), un étiquetage fondé sur le maximum d'entropie est exploré. Cette approche est étendue ensuite en décomposant la prédiction en deux temps : d'abord les frontières de l'EN en introduisant des catégories morpho-syntaxiques (POS), puis à la détermination de son type. Une seconde approche, fondée sur l'utilisation des CRF (Benajiba et Rosso, 2008) a permis d'explorer l'intégration de l'ensemble des traits dans un modèle unique, amenant à de meilleures performances. (Benajiba *et al.*, 2008) montrent également l'efficacité d'un prétraitement des textes pour séparer les différents constituants du mot (préfixes, lemme, et suffixes). (Abdul Hamid et Darwish, 2010) intègrent des traits intra-mot (n-grammes de caractères) dans

une modélisation CRF. Cette approche permet de capturer implicitement les caractéristiques morphosyntaxiques, introduites explicitement dans les expériences de (Benajiba et Rosso, 2008).

2.2 Adaptation et combinaison de systèmes

En apprentissage automatique, l'adaptation consiste à développer un système de traitement pour un domaine cible à partir de données et/ou d'un système de traitement développé pour un domaine source. D'un point de vue statistique, cela implique que les distributions des exemples observés sont différentes au moment de l'apprentissage et au moment du test.

Cette problématique a fait l'objet de multiples propositions en modélisation statistique des langues (par exemple l'étude de (Bellagarda, 2001) pour les modèles statistiques de langue), utilisation de pondérations différentielles pour les exemples de la source et de la cible (Jiang et Zhai, 2007), utilisation de descripteurs spécifiques pour les exemples source et cible (Daume III, 2007), etc. (Daume III *et al.*, 2010) présentent des travaux plus récents. Dans un cadre non supervisé, la stratégie la plus commune est l'auto-apprentissage (*self-training*) générant automatiquement des données d'apprentissage pour le domaine cible à partir du système source (Mihalcea, 2004).

Concernant le repérage des EN, le problème de l'adaptation se pose avec une acuité particulière, dûe au fait que les EN (i) sont souvent associées avec un thème particulier et (ii) ont également des distributions d'occurences très variables dans le temps. Cette problématique est étudiée en particulier par (Béchet *et al.*, 2011) qui (i) combinent deux approches d'étiquetage en EN pour le français : une approche symbolique avec une approche probabiliste et (ii) adaptent le système probabiliste fondé sur un processus discriminant à base de CRF, au domaine des données de test.

3 Étiquetage en entités nommées : systèmes de base

Dans cette section, nous décrivons les expériences réalisées pour développer des systèmes de base et en particulier pour identifier les descripteurs linguistiques utilisés. Tous ces modèles sont entraînés avec l'implémentation des CRF réalisée dans l'outil Wapiti[1] (Lavergne *et al.*, 2010). Cette implémentation permet (i) d'utiliser de très gros modèles incluant nominalement des centaines de millions de descripteurs, et (ii) de sélectionner les descripteurs les plus utiles par le biais d'une pénalité L_1 (Sokolovska *et al.*, 2009).

3.1 Protocole expérimental

Les expériences ont été réalisées sur le corpus ANER[2] (Benajiba *et al.*, 2007) constitué à partir d'articles de presse, et composé de plus de 150 000 occurrences de mots (4 871 phrases). Le corpus distingue 4 types d'EN : localisation (LOC : 40% des EN observées), personne (PERS : 32%), organisation (ORG : 18%) et une classe « divers » regroupant tous les autres types (MISC : 10%)[3], et peut être considéré comme le corpus de référence pour la tâche. Il utilise le schéma d'annotation IOB-2 et distingue 9 étiquettes. Les expériences sont produites à partir de données translittérées[4], sans faire d'analyse morphologique. Les scores sont calculés en utilisant l'outil

[1] Wapiti est librement disponible à l'adresse `http://wapiti.limsi.fr`.
[2] `http://users.dsic.upv.es/~ybenajiba/downloads.html`
[3] Seuls les trois premiers types sont utilisés dans nos évaluations.
[4] `http://www.qamus.org/transliteration.htm`

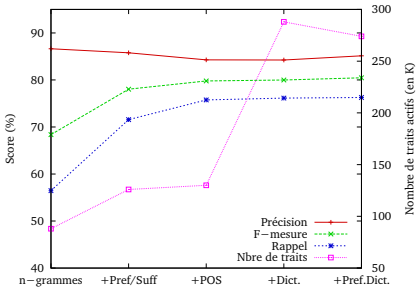

FIG. 1 – Précision (en %), rappel (en %), F-mesure et nombre de traits actifs pour des modèles de complexité croissante (à chaque nouveau modèle, de nouveaux traits sont ajoutés).

d'évaluation de CoNLL 2002[5]. Les modèles sont évalués par validation croisée à 10 partitions, sur des tests d'environ 25 000 mots chacun.

3.2 Sélection de caractéristiques et comparaison à l'état de l'art

Différentes versions du modèle de base ont été développées, qui incluent des jeux de descripteurs de richesse croissante. Nous décrivons ci-dessous les principales familles de descripteurs ; chaque réalisation x d'un élément d'une de ces familles donne lieu à un ensemble de fonctions booléennes testant x avec chaque étiquette et avec chaque bigramme d'étiquettes possibles.
N-grammes de mots : ces caractéristiques testent tous les unigrammes, les bigrammes, les trigrammes et les quadrigrammes dans, respectivement, des fenêtres de tailles 5, 3, 4 et 5.
Préfixes et suffixes : chaque séquence d'une, deux, ou trois lettres observée à l'initiale ou à la finale d'un mot du corpus d'apprentissage donne lieu à un nouveau descripteur. L'apparition de ces préfixes et suffixes est testée dans une fenêtre de taille 5 centrée sur le mot courant.
POS-tags : ce trait concerne les étiquettes morpho-syntaxiques prédites en utilisant un modèle entraîné par Wapiti sur l'*Arabic Tree Bank*[6] (Gahbiche-Braham *et al.*, 2012). Les tests évaluent les unigrammes et bigrammes d'étiquettes respectivement dans des fenêtres de tailles 5 et 3.
Ponctuation et nombres : ce trait teste la présence de caractères de ponctuations et de chiffres dans le mot courant ainsi que dans les deux mots voisins.
Dictionnaires : Ces dictionnaires proviennent de l'ANERGazet [2], d'extraits de Wikipedia et de la base de noms propres distribuée par JRC[7]. Pour chaque mot w, on teste s'il figure dans le dictionnaire (Dict sur la figure 1), ou s'il y figure précédé de préfixes (Pref.Dict). Ces dictionnaires contiennent 3 798 noms de lieux, 386 noms d'organisation et 13 648 noms de personnes.

Les résultats de ces expériences sont reportés sur la figure 1, qui représente la variation de la précision, du rappel et de la F-mesure, ainsi que le nombre de traits actifs. On constate qu'au fur et à mesure que de nouveaux traits sont ajoutés au modèle précédent, le rappel et la F-mesure augmentent, parfois au prix d'une légère dégradation de la précision.

[5] `http://bredt.uib.no/download/conlleval.txt`
[6] `http://www.ircs.upenn.edu/arabic/`
[7] *Joint Research Center* de la Communauté européenne : `http://langtech.jrc.it/JRC-Names.html`
[8] Le total reporté dans (Benajiba et Rosso, 2008) inclue l'EN MISC. Le total ici a été fait en calculant la moyenne.

Résultats du système de base				Résultats de (Benajiba et Rosso, 2008)[8]			
	Précision	Rappel	$F_{\beta=1}$		Précision	Rappel	$F_{\beta=1}$
LOC	90,59%	85,42%	87,83	LOC	93,03%	86,67%	89,74
ORG	78,67%	61,05%	68,75	ORG	84,23%	53,94%	65,76
PERS	81,31%	73,61%	77,27	PERS	80,41%	67,42%	73,35
Total	85,14%	76,27%	80,46	Total	85,89%	69,34%	76,28

TAB. 1 – Précision, rappel et F-mesure du modèle de base (qui combine tous les traits) sur le corpus ANER en comparaison avec les résultats de (Benajiba et Rosso, 2008)

Le tableau 1 donne une autre vue des performances du modèle le plus complet qui comprend environ 275 000 traits finalement sélectionnés par Wapiti sur un potentiel d'environ 80 millions. Afin de comparer notre système à l'état de l'art, le tableau 1 présente également les résultats obtenus par (Benajiba et Rosso, 2008) sur le corpus ANER avec un système utilisant également les CRF. Notre modèle de base semble donc cohérent avec les performances décrites dans l'état de l'art et atteint des performances globales semblables à celles décrites dans (Abdul Hamid et Darwish, 2010).

4 Adaptation du système d'étiquetage des entités nommées

Les applications étudiées dans ce travail s'inscrivent dans le cadre du projet SAMAR[9], qui vise à développer une plateforme de traitement de dépêches en langue arabe. Les données sont principalement produites par l'Agence France Presse (AFP). L'étiquetage en EN est envisagé ici comme un pré-traitement pour la traduction des données de l'arabe vers le français et l'anglais.

Nous disposons dans ce cadre de ressources supplémentaires pour adapter la détection des EN :
– de données du domaine (AFP), non-annotées (130 000 phrases, 3 500K mots) ;
– d'un étiquetage automatique d'une partie des données réalisé par un système symbolique développé par un des partenaires du projet, TEMIS (Guillemin-Lanne *et al.*, 2007).
– d'un corpus de test annoté manuellement et constitué de 900 phrases issues de l'AFP.

Les dépêches traitées dans notre application diffèrent substantiellement des données du corpus ANER, qui contient à la fois des articles de presse, des données collectées en ligne, en particulier des extraits de Wikipedia. Il existe également un décalage temporel entre la constitution de ce corpus (2007) et les données que nous devons traiter, qui sont postérieures à 2009.

4.1 Adaptation non-supervisée par auto-apprentissage

Le système de base, constitué à partir du corpus ANER, est utilisé pour annoter automatiquement le corpus AFP. Deux systèmes adaptés sont alors obtenus en utilisant comme corpus d'entraînement soit (i) le corpus étiqueté automatiquement seul, soit (ii) les deux corpus. Le tableau 2 donne les résultats des trois systèmes sur les données de test AFP. On constate une baisse sensible des performances du système de base (la F-mesure passe de 80,46 sur les données de test ANER à 72,64 sur les données de test AFP). Après adaptation (AFP et ANER+AFP), on constate une amélioration de la F-mesure pour les noms de lieux et d'organisations.

[9] http://samar.fr

	Modèle de base ANER			Modèle AFP			Modèle ANER+AFP		
	Précision	Rappel	$F_{\beta=1}$	Précision	Rappel	$F_{\beta=1}$	Précision	Rappel	$F_{\beta=1}$
LOC	89,30%	78,85%	83,75	90,81%	77,84%	83,83	91,45%	78,18%	84,29
ORG	50,24%	37,72%	43,09	51,01%	35,94%	42,17	51,76%	36,65%	42,92
PERS	68,07%	66,03%	67,03	70,83%	69,29%	70,05	70,87%	68,75%	69,79
Total	77,61%	68,27%	72,64	79,39%	68,14%	73,34	79,86%	68,34%	73,65

TAB. 2 – Comparaison et Adaptation de système de reconnaissance d'entités nommées

En moyenne, nous gagnons 1 point en F-mesure pour le système adapté. La bonne qualité des performances obtenues avec les annotations automatiques est principalement dûe à une augmentation très sensible de la couverture. Alors que seules 11% environ des EN de type personne du test sont dans le corpus ANER, on en retrouve plus de 60% quand on utilise le corpus automatique AFP pour l'apprentissage. Des écarts similaires, quoique moins importants, sont obtenus pour les organisations, et dans une moindre mesure, pour les lieux. Ceci illustre bien le caractère très localisé des occurrences des EN dont la distribution fluctue en fonction de l'actualité. D'une manière générale, ces améliorations restent toutefois limitées. Il est possible que la sélection des données de test (par l'AFP) conduise à sous-estimer l'apport de l'adaptation : le jeu de test contient majoritairement des dépêches ressortissant aux thèmes « guerre » et « politique », mais aucune de la catégorie « sport », pourtant très présente dans le corpus d'entraînement.

4.2 Un système hybride

Nous présentons ici les performances des trois modèles d'étiquetage décrits à la section 4.1 dans un cadre de combinaison de systèmes. La démarche suivie consiste à étiqueter automatiquement le corpus de test par l'annotateur de Temis, qui atteint une précision de 81% et une F-measure de 74% sur le corpus de test de l'AFP.

Le corpus de test est ensuite étiqueté une seconde fois par Wapiti, en considérant que les EN annotées par Temis sont correctes et en n'utilisant Wapiti que pour prédire les zones qui n'ont pas été détectées comme EN par l'étiqueteur symbolique. Les résultats sont donnés dans le tableau 3.

	Modèle de base ANER			Modèle AFP			Modèle ANER+AFP		
	Précision	Rappel	$F_{\beta=1}$	Précision	Rappel	$F_{\beta=1}$	Précision	Rappel	$F_{\beta=1}$
LOC	94,01%	85,12%	89,35	92,52%	81,31%	86,56	92,76%	81,31%	86,66
ORG	86,26%	66,18%	74,90	84,85%	61,09%	71,04	85,93%	62,18%	72,15
PERS	84,31%	76,01%	79,95	79,44%	72,22%	75,66	80,67%	72,73%	76,49
Total	90,24%	79,39%	84,47	87,80%	75,36%	81,11	88,45%	75,68%	81,57

TAB. 3 – Adaptation et test sur un corpus pré-étiqueté par un analyseur symbolique

Ces résultats montrent dans tous les cas une amélioration très sensible (+8 points) par rapport aux résultats antérieurs, en particulier quand on utilise le modèle non-adapté, qui a de meilleures performances que les modèles adaptés. Ceci est dû au fait que le système ANER a été entraîné sur un corpus annoté manuellement quand les systèmes adaptés utilisent des annotations automatiques potentiellement bruitées. Par comparaison avec le tableau 2, l'hybridation améliore les performances de chacun des systèmes pris séparément. Ceci ouvre des perspectives, en particulier pour mettre en place l'hybridation dès la construction du corpus d'apprentissage.

5 Conclusion

Dans cet article, nous avons présenté un système d'étiquetage en Entités Nommées construit par des méthodes d'apprentissage supervisé. Ce système, qui embarque des centaines de milliers de descripteurs, obtient des performances comparables aux meilleurs systèmes de l'état de l'art. Nous avons ensuite exploré diverses manières de réaliser une adaptation non-supervisée, par auto-apprentissage, de ce système conduisant à une légère amélioration des performances. Nous avons enfin montré qu'une hybridation du système statistique avec un système symbolique pouvait donner lieu à des gains bien supérieurs.

Ce travail ouvre de multiples perspectives portant sur les aspects liés à l'adaptation comme sur les aspects relatifs à la traduction. Concernant l'adaptation, il reste à reprendre les expériences précédentes avec un corpus produit par combinaison d'annotations ; de manière plus fondamentale, il reste également à voir comment *entraîner* Wapiti avec ces pré-étiquetages partiels, en utilisant par exemple des modèles à données latentes. Du point de vue de l'application finale, deux questions restent posées. L'une concerne l'ordre dans lequel effectuer les traitements préalables à la traduction : trois étapes s'enchaînent dans notre pipeline actuel : analyse morpho-syntaxique, détection des EN, puis segmentation des formes complexes. Il n'est pas dit que cet ordre soit optimal, et d'autres architectures devront être explorées. Ensuite, l'impact de la détection des EN sur la qualité de la traduction doit être évalué. Un travail préalable consistera à étudier comment les EN sont transférées d'une langue à l'autre, à partir de corpus parallèles annotés en EN.

Remerciements

Ces travaux ont été partiellement financés par le projet Cap-Digital SAMAR et par le programme Quaero. Merci à TEMIS pour l'annotation des corpus et à l'AFP pour le corpus de référence.

Références

Abdul Hamid, A. et Darwish, K. (2010). Simplified feature set for Arabic named entity recognition. *In Proc. of the 2010 Named Entities Workshop*, pages 110–115, Uppsala.

Béchet, F., Sagot, B. et 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 la conférence TALN*, Montpellier, France.

Bellagarda, J. R. (2001). An overview of statistical language model adaptation. *In Proc. of the ISCA Workshop on Adaptation Methods for Speech Recognition*, pages 165–174, Sophia Antipolis.

Benajiba, Y., Diab, M. et Rosso, P. (2008). Arabic named entity recognition using optimized feature sets. *In Proc. of EMNLP*, EMNLP, pages 284–293.

Benajiba, Y. et Rosso, P. (2007). Conquering the NER task for the Arabic language by combining the maximum entropy with POS-tag information. *In Proceedings of Workshop on Natural Language-Independant Engineering*, IJCAI.

Benajiba, Y. et Rosso, P. (2008). Arabic named entity recognition using conditional random fields. *In Proceedings of the Conference on Language Resources and Evaluation*.

Benajiba, Y., Rosso, P. et Benedí, J.-M. (2007). Anersys : An arabic named entity recognition system based on maximum entropy. *In CICLing*, pages 143–153.

DAUME III, H. (2007). Frustratingly easy domain adaptation. *In Proc. of the 45th Annual Meeting of the ACL*, pages 256–263, Prague, Czech Republic.

DAUME III, H., DEOSKAR, T., MCCLOSKY, D., PLANK, B. et TIEDEMANN, J., éditeurs (2010). *Proc. of the 2010 Workshop on Domain Adaptation for NLP*. Uppsala, Sweden.

DAUMÉ III, H. et JAGARLAMUDI, J. (2011). Domain adaptation for machine translation by mining unseen words. *In ACL*, Portland, OR.

GAHBICHE-BRAHAM, S., BONNEAU-MAYNARD, H., LAVERGNE, T. et YVON, F. (2012). Joint segmentation and POS tagging for arabic using a CRF-based classifier. *In Proc. of LREC'12*.

GAHBICHE-BRAHAM, S., BONNEAU-MAYNARD, H. et YVON, F. (2011). Two ways to use a noisy parallel news corpus for improving statistical machine translation. *In Proc. of Workshop on Building and Using Comparable Corpora*, pages 44–51, Portland, OR.

GUILLEMIN-LANNE, S., DEBILI, F., TAHAR, Z. B. et GACI, C. (2007). Reconnaissance des entités nommées en arabe. *In Colloque VSST, Veille Stratégique Scientifique et Technologique*.

HABASH, N. (2010). *Introduction to Arabic Natural Language Processing*. Morgan Claypool.

HERMJAKOB, U., KNIGHT, K. et DAUMÉ III, H. (2008). Name translation in statistical machine translation - learning when to transliterate. *In Proc. of ACL-08 : HLT*, pages 389–397, Ohio.

JIANG, J. et ZHAI, C. (2007). Instance weighting for domain adaptation in nlp. *In Proc. of the 45th Annual Meeting of the ACL*, pages 264–271, Prague, Czech Republic.

LAFFERTY, J., MCCALLUM, A. et PEREIRA, F. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. *In Proc. ICML*, pages 282–289, San Francisco.

LAVERGNE, T., CAPPÉ, O. et YVON, F. (2010). Practical very large scale CRFs. *In Proceedings the 48th Annual Meeting of the Association for Computational Linguistics*, pages 504–513.

MALONEY, J. et NIV, M. (1998). TAGARAB : a fast, accurate Arabic name recognizer using high-precision morphological analysis. *In Proc. of the Workshop on Computational Approaches to Semitic Languages*, Semitic '98, pages 8–15, Stroudsburg, PA, USA.

MIHALCEA, R. (2004). Co-training and self-training for word sense disambiguation. *In* NG, H. T. et RILOFF, E., éditeurs : *HLT-NAACL Workshop : CoNLL-2004*, pages 33–40, Boston.

SAMY, D., MORENO, A. et MA GUIRAO, J. (2005). A proposal for an Arabic named entity tagger leveraging a parallel corpus. RANLP '05.

SHAALAN, K. et RAZA, H. (2009). NERA : Named entity recognition for arabic. *Journal of the American Society for Information Science and Technology*, 60(9):1652–1663.

SOKOLOVSKA, N., CAPPÉ, O. et YVON, F. (2009). Sélection de caractéristiques pour les champs aléatoires conditionnels par pénalisation l_1. *TAL*, 50(3):139–171.

ZAGHOUANI, W., POULIQUEN, B., EBRAHIM, M. et STEINBERGER, R. (2010). Adapting a resource-light highly multilingual named entity recognition system to arabic. Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), pages 563–567.

ZHANG, M., LI, H., KUMARAN, A. et LIU, M. (2011). Report of news2011 machin transliteration shared task. *In Proceedings of the 2011 Named Entities Workshop*.

ZITOUNI, I., SORENSEN, J., LUO, X. et FLORIAN, R. (2005). The impact of morphological stemming on Arabic mention detection and coreference resolution. *In Proc. of Workshop on Computational Approaches to Semitic Languages*, pages 63–70, Ann Arbor, Michigan.