L'apport d'une approche hybride pour la reconnaissance des entités nommées en langue arabe

Inès Zribi, Souha Mezghani Hammami, Lamia Hadrich Belguith

ANLP Research Group – Laboratoire MIRACL
Faculté des Sciences Economiques et de Gestion de Sfax
B.P. 1088, 3018 - Sfax – TUNISIE
Téléphone (216) 74 278 777, Fax (216) 74 279 139
ineszribi@gmail.com, souha.mezghani@fsegs.rnu.tn, l.belguith@fsegs.rnu.tn

Résumé. Dans cet article, nous proposons une méthode hybride pour la reconnaissance des entités nommées pour la langue arabe. Cette méthode profite, d'une part, des avantages de l'utilisation d'une méthode d'apprentissage pour extraire des règles permettant l'identification et la classification des entités nommées. D'autre part, elle repose sur un ensemble de règles extraites manuellement pour corriger et améliorer le résultat de la méthode d'apprentissage. Les résultats de l'évaluation de la méthode proposée sont encourageants. Nous avons obtenu un taux global de *F-mesure* égal à 79.24%.

Abstract. In this paper, we propose a hybrid method for Arabic named entities recognition. This method takes advantage of the use of a learning method to extract rules for the identification and classification of named entities. Moreover, it is based on a set of rules extracted manually to correct and improve the outcome of the learning method. The evaluation results are encouraging as we get an overall F-measure equal to 79.24%.

Mots-clés : Traitement de la langue arabe, reconnaissance des entités nommées, méthode d'apprentissage.
Keywords: Arabic language processing, named entity recognition, learning method.

1 Introduction

La reconnaissance des entités nommées (REN) est devenue une tâche nécessaire pour plusieurs applications de TAL et plus particulièrement pour le processus de résolution automatique des anaphores. Un système de résolution automatique des anaphores nécessite, tout d'abord, l'identification des expressions anaphoriques et la liste des antécédents candidats afin de résoudre le lien entre ces deux éléments. Donc, leur efficacité dépend de la bonne identification des antécédents (Hammami et al., 2009). Et puisque la plupart des antécédents sont des groupes nominaux, alors l'identification des entités nommées (EN) pour un système de résolution des anaphores est fondamentale pour assurer une bonne résolution.

La reconnaissance des noms propres pour l'arabe diffère des autres langues telles que les langues indo-européennes. Elle se heurte à plusieurs problèmes dus aux particularités de cette langue. Dans ce travail, nous proposons une méthode hybride qui se base sur des règles générées automatiquement capables d'identifier les mots qui composent les EN et sur un ensemble de règles génériques extraites manuellement pour reconnaitre

les EN. Nous exposons, en premier lieu, un aperçu sur les travaux qui ont été réalisés. Ensuite, nous exposons les problèmes de REN liés aux spécificités de la langue arabe. Puis, nous décrivons la méthode proposée pour résoudre la tâche de REN pour l'arabe. Enfin, nous présentons une évaluation et nous discutons les résultats obtenus.

2 Les travaux antérieurs

Les travaux de REN sont classés principalement selon trois types d'approches : *(i)* l'approche linguistique qui se base sur des règles génériques écrites à la main, *(ii)* l'approche numérique qui se base sur un mécanisme d'apprentissage à partir d'un grand corpus de textes pré-étiquetés et *(iii)* l'approche hybride qui présente une combinaison de ces deux approches (Mansouri et al., 2008). Dans la littérature, plusieurs travaux de REN ont été effectués pour les langues indo-européennes. Dans le cadre de l'approche linguistique, nous signalons le système de REN proposé par (Wacholder et al., 1997) qui ont définit un ensemble d'heuristiques basées sur des indicateurs sans l'utilisation ni de listes de noms propres ni d'informations syntaxiques pour la détection des EN. Dans le cadre de l'approche numérique (Malouf, 2003) a utilisé les modèles de Markov caché pour résoudre la tâche de REN pour la langue anglaise. Par contre, peu de travaux ont étudié la tâche de REN pour la langue arabe. Parmi ces travaux, nous notons celui Shaalan et Raza (Shaalan, Raza, 2008) ont conçu leur système NERA qui permet de reconnaitre dix types des EN. Ce système se base sur un lexique formé de listes de dictionnaires des EN et sur une grammaire sous forme d'expressions régulières pour l'identification des EN. Les valeurs F-mesure obtenues par ce système varient entre 83.15% et 98.6%. De même, (Zaghouani et al., 2010) ont présenté un module de repérage des EN à base de règles pour la langue arabe. Ce module est constitué de trois étapes : une étape de prétraitement lexicale qui prépare le texte pour son analyse linguistique, une étape de repérage des EN reconnues grâce à un dictionnaire et une étape de repérage des EN basée sur des règles écrites à la main sous forme d'expressions régulières. Ce module a été évalué sur un corpus composé d'articles de presse. Les valeurs de F-mesure apportées par ce système varient respectivement entre 47.35% pour le type organisation et 95.10% pour le type date. D'autre part, et dans le cadre de l'approche numérique, Benajiba et al (Benajiba et al., 2009) ont proposé le modèle d'apprentissage SVM pour la réalisation de leur système de REN en exploitant un ensemble de caractéristiques. Certaines de ces caractéristiques sont spécifiques à la langue arabe et d'autres non. Ce système a rapporté une F-mesure globale qui atteint 82.71%. De même, Farber et al. (Farber et al., 2008) ont exploré des caractéristiques morphologiques de l'arabe dans un modèle d'apprentissage basé sur le perceptron structuré. L'exploitation de cette information a donné une valeur de F-mesure équivalente à 71.5%. En tirant profit de l'approche linguistique et statistique, Abuleil (Abuleil, 2006) a adopté une approche hybride. En effet, il a exploité des règles pour le marquage des syntagmes nominaux contenant des noms propres, des graphiques pour l'extraction des noms propres candidates et, enfin, des statistiques pour la classification des noms propres et ceci en exploitant des listes de mots déclencheurs.

3 Difficultés de la reconnaissance des entités nommées en arabe

La REN est une tâche complexe mais son niveau de complexité ne se limite pas seulement aux problèmes théoriques de REN tels que l'imbrication ou la coordination des EN si la langue étudiée est l'arabe, mais aussi à certaines spécificités de cette langue à savoir :

- **L'absence des lettres majuscules** : c'est un obstacle majeur pour la langue arabe. En fait, la REN pour certaines langues comme les langues indo-européennes se base principalement sur la présence des lettres majuscules.
- **L'absence des voyelles** : un mot non voyellé présente de nombreuses ambiguïtés au niveau du sens ou de la fonction syntaxique. Par exemple, le mot « حافظ » /HAfZ/ peut être un verbe (*sauvegarder*) pour une voyellation ou un nom propre (*Hafez*) pour une autre. Donc, l'absence des voyelles est considérée comme un autre obstacle pour la tâche de REN.

- **Morphologie complexe** : l'arabe est une langue fortement flexionnelle. Elle utilise une stratégie agglutinative pour former un mot. Si une EN apparait avec une forme agglutinative alors ceci pose une difficulté pour l'identification de cette entité.

4 Méthode proposée

L'idée principale de notre méthode consiste à utiliser un ensemble de règles pour extraire et classer les EN. En effet, nous utilisons, d'une part, un ensemble de règles générées automatiquement à l'aide d'un algorithme d'apprentissage. Ces règles doivent apprendre la structure et le contexte dans lequel apparaissent les EN. D'autre part, nous proposons un autre ensemble de règles extraites manuellement pour corriger et améliorer le résultat de l'application des règles générées automatiquement. Notre méthode est composée de trois principales étapes (voir Figure 1) : une étape d'extraction de règles, une étape de validation de ces règles et, enfin, une étape d'extraction des EN. Au niveau de ces trois étapes, nous choisissons d'appliquer une analyse morphologique afin de remédier aux erreurs engendrées par la nature agglutinative de la langue arabe, et pour déterminer pour chaque mot du corpus les différentes caractéristiques morphosyntaxiques (genre, nombre, type, etc.). Par ailleurs, le module d'extraction des attributs se base sur des listes de lexique de noms propres et sur l'ensemble des caractéristiques morphologiques pour déterminer les valeurs des attributs dans les différents contextes pour chaque mot du corpus. Les deux modules (analyse morphologique et extraction d'attributs) constituent une étape de préparation pour les trois étapes de notre méthode. Pour standardiser notre travail de REN, nous avons adopté la définition proposée par la conférence CoNLL (Erik et al., 2003) pour les EN. Par conséquent, les EN appartiennent à quatre classes : *personne*, *lieu*, *organisation* et *divers*. Cette dernière classe regroupe toutes les autres EN. De même, nous avons utilisé un corpus d'apprentissage annoté selon le format proposé par cette conférence et le schème d'annotation BIO[1]. Nous avons exploité le corpus ANERcorp[2]. Ce dernier est formé de 150.000 mots extraits d'un ensemble d'articles de presse de types variés. Il contient 10.519 EN de type lieu, organisation et personne. L'apprentissage a été réalisé sur 76.608 mots contenant 5.192 EN. Pour la validation des règles, nous avons utilisé 38.488 mots contenant 2.778 EN et le reste du corpus a été utilisé pour tester la performance du notre système. Actuellement, notre système se limite uniquement aux types *personne*, *lieu* et *organisation*.

4.1 Extraction de règles

La première étape consiste à extraire automatiquement des règles capables de détecter les mots qui composent les EN selon leur type. Ces règles permettent la classification des mots selon trois types. Elles déterminent si un mot constitue le premier syntagme d'une EN (B-TYPE), ou s'il appartient à une EN (I-TYPE) ou s'il n'appartient pas à une EN (O). Pour ce faire, nous avons utilisé l'algorithme d'apprentissage des règles RIPPER (Cohen, 1995). Cet algorithme s'avère efficace pour des données bruitées et il a une évolutivité quasi linéaire avec le nombre d'exemples dans un corpus d'apprentissage. RIPPER (Cohen, 1995) a utilisé un ensemble d'attributs qui présentent les éléments les plus influents sur le résultat de l'apprentissage. Nous avons choisi d'utiliser deux types d'attributs pour l'extraction des règles : attributs morphologiques et attributs à base de lexique de noms propres. Pour ces attributs, nous avons exploité un contexte de +/- 5 mots ; c'est-à-dire, on étudie les cinq mots qui sont situés avant et après le mot à classer. Nous avons sélectionné quatre caractéristiques morphologiques pour les inclure dans cette étude : la catégorie et le type du mot, le proclitique et le type du proclitique qui est rattaché à ce mot. Pour les attributs à base de lexique de noms propres, nous avons utilisé des listes contenant des EN connues auparavant (2.391, 615 et 683 entrées respectivement pour le type personne, lieu et organisation) ainsi que des listes de mots déclencheurs (189, 122 et 50 entrées respectivement pour le type personne, lieu et organisation). De même, nous avons employé des listes pour les nationalités (250 nationalités). Pour former ces listes, nous nous sommes basés sur ANERgazet[2] utilisés par (Benajiba, Rosso, 2007) et les listes de nom propre et de mots déclencheurs définis dans le cadre de logiciel GATE (Hepple et al., 2000). Nous les avons enrichies par

[1] Le schème d'annotation BIO ou IOB : on étiquette en tant que «B» le premier mot d'une EN, «I» pour le mot à l'intérieur d'une EN et les mots qui ne sont pas des EN sont étiquetés comme «O».

[2] http://www1.ccls.columbia.edu/~ybenajiba/downloads.html

d'autres entrées à partir du web. L'algorithme RIPPER a généré 46 règles pour le type *personne*, 48 pour le type *lieu* et 40 règles pour le type *organisation*. Considérons l'EN «شرق غزة» (*Est de Gaza*) /$rq gzp/. La détection de cette EN se base sur les deux règles suivantes :

- *(Direction = y) and (POS+1 = اسم_علم) => Decision=B-LOC* : si le mot à classer désigne une direction et il est suivi par un nom propre alors ce mot présente le premier mot d'une EN de lieu.
- *(Direction-1 = y) and (POS = اسم_علم) => Decision=I-LOC*: si le mot à classer est un nom propre et il est précédé par une direction alors ce mot appartient à une EN de type lieu.

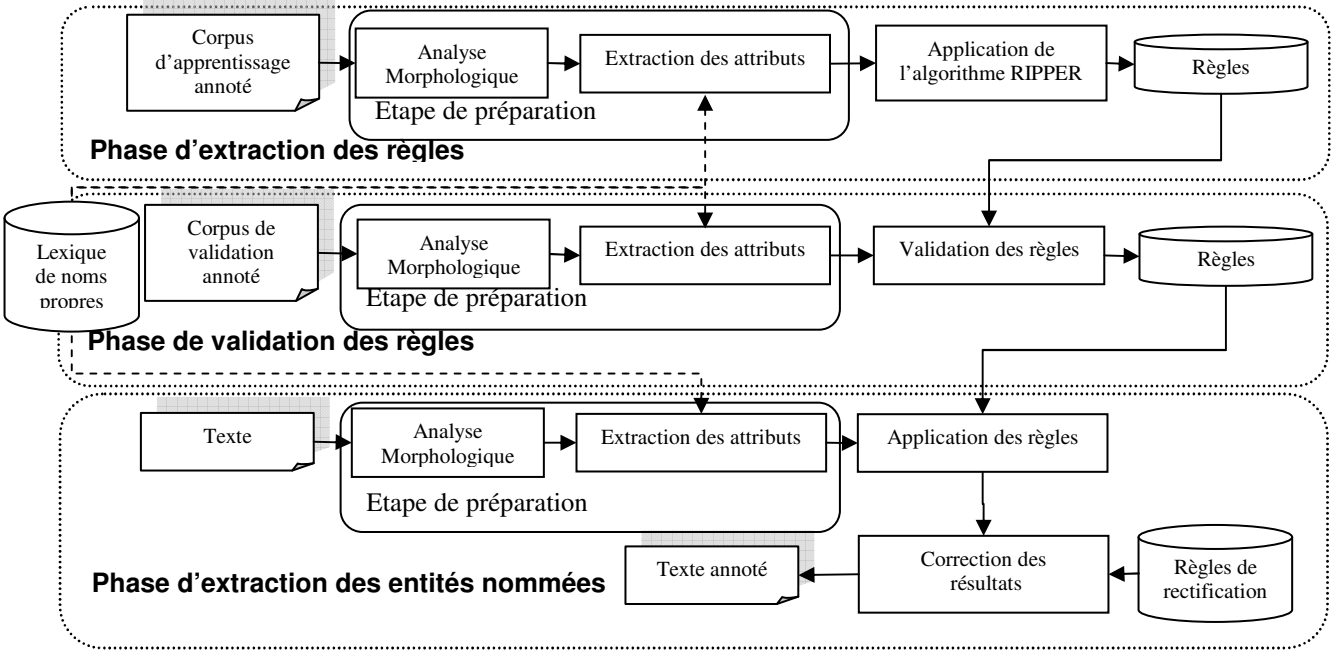

Figure 1 : Les étapes de la méthode proposée.

4.2 Validation des règles

Après l'extraction des règles, l'étape suivante a pour but de les raffiner afin d'assurer leurs validités. Le processus de validation calcule pour chaque règle le nombre de mots bien classés et le nombre de mots mal classés. Si le nombre de mots correctement classés (*NBbc*) par une règle est supérieur au nombre de mots mal classés (*NBmc*), alors cette règle est dite valide et elle sera ajoutée à l'ensemble des règles valides avec un score *S= NBbc* / (*NBbc+NBmc*). Dans le cas où une règle n'est pas activée dans le corpus de validation, elle est automatiquement ajoutée à l'ensemble des règles validées, mais avec un score égal à 0. Selon les scores trouvés, nous trions ces règles en ordre décroissant. Au niveau de cette étape, nous avons éliminé 11 règles parmi 37 pour le type personne, 18 règles parmi 48 pour le type lieu et 8 règles parmi 43 pour le type organisation.

4.3 Extraction des entités nommées

L'étape d'extraction des EN se déroule en deux phases. La première vise à identifier les mots qui composent les EN en appliquant les règles générées automatiquement, issues de l'étape de validation. Une EN est bien classée si tous les mots qui la composent prennent les étiquettes correctes. Mais, dans certain cas, l'application de ces règles ne suffisse pas pour reconnaitre les différents syntagmes qui composent une EN. Par exemple, l'application des règles n'a pas pu reconnaitre le mot « العمال » (*ouvriers*) /AlEmAl/ comme une partie de l'EN «حزب العمال الكردستاني » (*Parti des ouvriers Kurdistan*) /Hzb AlEmAl AlkrdstAny/ et donc on obtient l'annotation (B-ORG O I-ORG) au lieu de (B-ORG I-ORG I-ORG). La seconde phase de l'extraction des EN permet la correction des erreurs engendrées par la première phase. Par exemple, si on veut corriger l'erreur dans l'annotation (B-ORG O I-ORG) pour l'EN «حزب العمال الكردستاني », alors, on doit appliquer la règle suivante : « *classe_mot= O and classe_mot-1=B-ORG and classe_mot+1= I-ORG => classe_mot=I-*

ORG ». Cette règle signifie que si un mot porte l'étiquette O et s'il est précédé par un mot qui porte l'étiquette B-ORG et suivi par un mot qui porte l'étiquette I-ORG, alors, ce mot appartient à une EN de type *organisation* et l'étiquette O sera remplacée par l'étiquette I-ORG.

5 Evaluation

Nous avons mené deux évaluations. Au niveau de la première évaluation, nous avons utilisé seulement les règles générées automatiquement. A la seconde, nous avons ajouté la phase de correction des résultats qui utilise des règles extraites manuellement. En examinant les mesures de rappel, précision et F-mesure calculées sur le corpus d'évaluation, nous remarquons que les résultats sont très encourageants. En effet, l'utilisation d'un module de rectification a amélioré les résultats d'une manière significative (à peu prés 9%). Cela justifie notre choix pour une méthode hybride.

	Evaluation avec les règles générées automatiquement			**Evaluation avec les deux phases**		
	Rappel	**Précision**	**F-mesure**	**Rappel**	**Précision**	**F-mesure**
Personne	63.69%	76.94%	69.69%	74.86%	86.54%	80.28%
Lieu	81.95%	81.59%	81.77%	8494%	87.96%	86.42%
Organisation	45.361%	59.81%	51.61%	55.85%	75.36%	64.15%
Global	66.12%	75.54%	70.51%	74.22%	84.99%	79.24

Tableau 1 : Résultats de l'évaluation.

Notons qu'à la différence de la méthode proposée par Abuleil (Abuleil, 2006), notre méthode utilise un ensemble des règles générées automatiquement et d'autres extraites manuellement pour la classification et l'identification simultanée des EN, en profitant d'un ensemble de caractéristiques morphologiques et d'une liste de lexique de noms propres. De plus, nous avons comparé notre système avec le système ANERsys2.0 (Benajiba, Rosso, 2007) qui modélise le problème de REN avec l'entropie maximale comme une approche en deux étapes de classification séparant la détection des frontières de la classification des EN sans l'exploitation de l'information morphologique (ce système a obtenu une F-mesure égale à 86.71%, 45.02% et 52.13% respectivement pour les types lieu, organisation et personne). La comparaison de ces résultats montre l'apport de l'utilisation des informations morphologiques pour la tâche de REN pour l'arabe. Nous avons, aussi, comparé notre système avec ceux développés par (Benajiba et al., 2009) et (Zaghouani et al., 2010). Les résultats de la comparaison sont utilisés à des fins d'illustration uniquement et ils sont résumés dans le tableau 2. Bien que les trois systèmes aient utilisés des corpus différents pour leurs évaluations, nous pouvons affirmer que les résultats obtenus par notre système sont, dans certains cas, meilleurs que celles obtenus par les deux autres systèmes.

	Notre système	**Système de (Benajiba et al., 2009)**	**Système de (Zaghouani et al., 2010)**
Personne	80.28%	81.55%	75.40 %
Lieu	86.42%	87.04%	82.33 %
Organisation	64.15%	54.36%	47.35 %

Tableau 2: Tableau comparatif des valeurs des F-mesures par les trois systèmes.

L'échec d'identification et de la classification de quelques EN peut s'expliquer par la présence de quelques erreurs. Parmi ces erreurs, nous notons : *(i)* des erreurs dues à la segmentation des mots avec l'analyse morphologique (certains noms propres étrangers ressemblent à des mots agglutinées lorsqu'ils sont translittérés en arabe). Par exemple, le nom propre « Virginie » sera translittéré en « فيرجيني» (et il espère de moi) qui désigne un verbe auquel s'attache une conjonction de coordination « ف» (et). Donc avec cette

segmentation, on ne peut pas appliquer certaines règles et l'EN ne sera pas reconnue) *(ii)* des erreurs dues à des insuffisances dans le corpus d'apprentissage (la génération des règles de REN se base sur la structure et le contexte de l'EN. Si une EN apparait dans un contexte différent à celui du corpus d'apprentissage, alors elle ne sera pas reconnue).

6 Conclusion et perspectives

Dans cet article, nous avons proposé une méthode hybride pour la reconnaissance des entités en arabe qui se base sur une méthode d'apprentissage automatique des règles et sur un ensemble de règles extraites manuellement. La méthode d'apprentissage permet la génération d'un ensemble de règles permettant l'identification des entités nommées. Le résultat de l'application de ces règles est amélioré par un ensemble de règles extraites manuellement. La valeur de F-mesure obtenue par la première étape de notre méthode a atteint une valeur *F-mesure* globale égale à 70.51%. Cette valeur a été améliorée d'environ 9% par l'application de la deuxième étape. Comme perspective, nous envisageons d'abord d'étudier d'autres attributs pour améliorer la qualité des règles générées avec RIPPER. Ensuite, nous visons à étendre les corpus d'apprentissage et de validation afin de couvrir le maximum d'EN.

Références

ABULEIL S. (2006). Hybrid System for Extracting and Classifying Arabic Proper Names. *Proceedings of the WSEAS Int.Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases,* Madrid-Spain*,* 205-210*.*

BENAJIBA Y, DIAB M, ROSSO P. (2009). Using Language Independent and Language Specific Features to Enhance Arabic Named Entity Recognition. *The International Arab Journal of Information Technology, Vol. 6, No. 5, November* 2009, 464-472.

BENAJIBA Y, ROSSO P. (2007). ANERsys2.0: Conquering the NER Task for the Arabic Language by Combining the Maximum Entropy with POS-tag Information. *The Proceedings of Workshop on Natural Language Independent Engineering IICAI, India*, 36-40.

COHEN W. (1995). Fast Effective Rule Induction. *In Proceedings of the 12th International Conference on Machine Learning (ICML'95), Lake Tahoe, USA, Morgan Kaufmann,* 115-123*.*

ERIK F, TJONG KIM S, FIEN DE M. (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. *In Proc. of CoNLL-2003.* 142-147.

FARBER B, FREITAG D, HABASH N, RAMBOW O. (2008). *Improving NER in Arabic using a morphological tagger. In Proceedings of LREC 2008, Marrakech, Marroc*.

HAMMAMI S, BELGUITH L, BEN HAMADOU A. (2009). Arabic Anaphora Resolution: Corpora Annotation with Coreferential links. *The international Arab Journal of Information Technology, vol. 6, No.5, November* 2009, pp481-489.

HEPPLE M, HERRING P, MITCHELL B, OAKES M, PETERS W, SETZER A, STEVENSON M, TABLAN V, URSU C, WILKS Y.(2000). A Survey of Uses of GATE, *Technical Report CS-00-06*, Department of Computer Science, University of Sheffield.

MALOUF R. (2003). Markov Models for Language-Independent Named Entity Recognition. *In Proceedings of CoNLL-2003,* Edmonton, Canada.

MANSOURI A, SURIANI AFFENDEY L, MAMAT A. (2008). Named Entity Recognition Approaches. *IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.2, February 2008.* 339-344.

SHAALAN K, RAZA A. (2008). Arabic Named Entity Recognition from Diverse Text Types. *In Eds. Nordström, B., and Ranta, A., (Eds.) GoTAL 2008: 6th International Conference on Natural Language Processing, Gothenburg, Sweden, August 25-27*, *Lecture Notes in Computer Science/Lecture Notes in Artificial Intelligence (LNCS/LNAI): Advances in Natural Language Proceedings*, *Vol.* 5221, 440-451, Springer-Verlag, Berlin, Germany, 2008.

WACHOLDER N, RAVIN Y, CHOI M. (1997). Disambiguation of Proper Names in Text. *In Proceedings of the 5th Conference on Applied Natural Language Processing*, Washington, D.C, 202-208.

ZAGHOUANI W, POULIQUEN B, EBRAHIM M, STEINBERGER R. (2010). Adapting a resource-light highly multilingual Named Entity Recognition system to Arabic. *In Proceedings of LREC 2010*, Valetta-Malta.