Système hybride pour la reconnaissance des entités nommées arabes à base des CRF

Emna Hkiri, Souheyl Mallat and Mounir Zrigui
Laboratoire LaTICE, Faculté des Sciences de Monastir, Tunisie

RESUME

La reconnaissance d'entités nommées (REN) pour les langues naturelles telles que l'arabe est une tâche essentielle et difficile. Dans cet article, nous décrivons notre système hybride afin d'améliorer la performance du système de REN et de combler le manque de ressources pour le TAL arabe. Notre système applique un modèle CRF, un lexique bilingue d'ENs et des règles linguistiques spécifiques à la tâche de reconnaissance d'entités nommées dans les textes arabes. Les résultats empiriques indiquent que notre système surpasse l'état-de l'art de la REN arabe lorsqu'il est appliqué au corpus d'évaluation standard ANERcorp.

ABSTRACT

Hybrid arabic NER system using CRF Model

Named Entity Recognition (NER) in Natural Languages like Arabic is an essential and challenging task. In this paper we describe our hybrid system that takes the advantages of machine learning-based and rule-based approaches in order to improve the NER system performance and overcome the lack of resources for Arabic. Our system applies Conditional Random Fields Model (CRF),bilingual named entity lexicon and language specific rules to the task of Named Entity Recognition in arabic texts. The empirical results indicate that our system outperforms the state-of-the-art of Arabic NER in terms of precision when applied to ANER corpus standard dataset

MOTS-CLES : REN arabe, approche hybride, Modèle CRF, Linked Data datasets, lexique bilingue des ENs .

KEYWORDS: ARABIC NER, Hybrid Approach , CRF model, Linked data datasets, bilingual NE lexicon

1 Introduction

Les entités nommées (ENs) sont des entités du monde réel qui représentent des indices indispensables pour la compréhension du sens des données textuelles. La REN consiste à identifier et classer toutes les entités nommées dans un texte donné selon deux classes intuitives; noms propres (personnes, lieux, organisations), expressions numériques (heure, date, monnaie et pourcentage) (Daille et al., 2000) . La REN arabe se confronte des défis principaux telsque; l'absence de capitalisation, la complexité morphologique, l'absence des voyelles (ou signes diacritiques) et le manque de ressources.

Dans notre travail, nous traitons le problème de la reconnaissance des entités nommées arabes par couplage des approches à base de règles et les approches par apprentissage. Pour ce faire nous avons mis en place un système hybride dont l'objectif est d'améliorer la performance de la REN globale. Le reste du papier est organisé comme suit: nous décrivons brièvement les principales ressources

linguistiques utilisées dans la deuxième section. Nous détaillons l'architecture et les modules de base de notre système dans la troisième section. La quatrième est dédiée à l'évaluation et à la comparaison avec l'état de l'art. L'article est clos par une conclusion et les travaux futurs.

2 Collection des données

Pour notre travail, les ressources linguistiques utilisées sont une combinaison de lexique d'ENs et de corpus disponibles gratuitement. Les corpus ont été préparés, nettoyés et annotés au format XML (trois étiquettes d'entités nommées, personne, organisation et lieu). Nous avons pu acquérir trois corpus. Le corpus ANER, le corpus News Commentary et le corpus United Nations, décrits ci-dessous.

-United Nations corpus (UN): Pour obtenir notre corpus d'apprentissage pour le module à base d'apprentissage, nous avons utilisé environ 15000 phrases du corpus pour l'année 2005. Avant d'utiliser ces textes, nous avons appliqué le prétraitement linguistique pour obtenir des données dans le format approprié pour le traitement automatique de ces textes.

-ANERcorp est développé par Yassine Benajiba. Il est utilisé comme corpus de référence pour la comparaison de notre système avec les travaux existants dans l'état de l'art. Nous l'avons exploité aussi pour la phase d'apprentissage dans notre méthode de REN hybride. Le corpus est étiqueté au format CONLL (Conference on Natural Language Learning) pour cela nous avons transformé son annotation en format XML.

-Corpus News Commentary 2012: Ce corpus est constitué de commentaires politiques et économiques issus du site Web Project Syndicate. Ce corpus n'est pas annoté et il est conçu à l'origine pour évaluer la traduction automatique statistique en arabe. Par conséquent, dans cette recherche, nous l'avons utilisé comme corpus de référence pour l'évaluation de la REN arabe. Pour ce faire nous avons extrait 600 phrases dans lesquelles nous avons détecté et annoté manuellement 350 noms de personnes 410 noms de lieux, et 151 noms d'organisations.

Un autre type de ressources linguistiques est utilisé, qui consiste en un lexique bilingue des entités nommées. Ce lexique est construit à base des linked datasets de DBpedia. Il couvre les entités nommées personne, lieu et organisation pour le couple de langues arabe-anglais (voir le tableau 1).

Entités nommées	**Paires en Arabe-Anglais**
Personne	27480
Organisation	17237
Lieu	4036
Total	48753

Tableau1. Lexique Ar-EN des entités nommées

3 Architecture du système

L'architecture de notre système est basée sur une approche hybride combinant les méthodes à base de connaissances avec les méthodes par apprentissage. Le module à base de règles développé est une reproduction d'un système d'annotation existant. Le module par apprentissage est fondé sur les champs conditionnels aléatoires (CRF). Dans les sections suivantes, nous détaillons chacun de ces deux modules ainsi que les ressources utilisées.

3.1 Module à base de règles

Le module à base de règles est une reproduction du système ANNIE (A Nearly New Information Extraction system). C'est une chaine complète dédiée principalement à l'extraction d'ENs pour l'anglais. Cet annotateur intégré dans GATE comprend différents modules, il est basé principalement sur le langage JAPE (Java Annotation Patterns Engine) et des Gazetteers. Les Gazetteers cherchent les éléments dans les listes prédéfinies annotées en tant qu'entité « Lookup ». Le Transducteur JAPE permet de repérer les catégories (personne, organisation, lieu, date, URL, adresses, e-mail, etc.). Rappelons qu'ANNIE est développée essentiellement pour l'annotation des textes anglais. Par la suite les développeurs ont intégré un module pour la langue arabe. Néanmoins, le nombre et la qualité des qazetteers pour cette langue sont encore beaucoup inférieurs par rapport à ceux de la langue anglaise. Le coût et le temps élevés de construction de Gazetteers arabes nous amènent à nous interroger sur la manière d'en acquérir un nombre acceptable pour garantir une meilleure performance du système de REN. Pour remédier à ce problème, nous avons recours à notre lexique bilingue (décrit dans la section précédente). Dans cette étape, nous avons exploité la partie arabe de notre lexique, nous avons mappé les entités nommées vers les gazetteers prédéfinies de GATE comme présenté dans le tableau ci-dessous.

ANNIE/Gate	Entités prédéfinies	Entités Enrichies
Personne	1700	27480
Organisation	96	17237
Lieu	485	4036

Tableau 2 : Enrichissement des Gazetteers prédéfinies de Gate

3.2 Module par apprentissage

L'union des modules à base de règle et par apprentissage forme le système hybride, qui vise à améliorer la performance du système de REN de base. Le processus d'hybridation consiste à annoter automatiquement le corpus de test par l'annotateur à base de règles. Le corpus de test est annoté une seconde fois par CRF++, en considérant que les ENs annotées par notre module à base de règles sont correctes et n'utilisant CRF++ que pour prédire les zones qui n'ont pas été annotées.

Le module par apprentissage nécessite un grand volume de données annotées pour ce faire nous ajoutons à notre corpus des UN de base (environ15000 phrases annotées par le module à base de règles), le corpus ANER. Ce dernier est composé de 4 871 phrases (plus de 150 mille occurrences de mots). Notre module par apprentissage supervisé utilise les champs conditionnels aléatoires (Conditional Random Fields CRF), qui sont une généralisation des réseaux bayésiens. Les CRF sont des modèles statistiques d'annotation qui obtiennent d'excellentes performances pour la reconnaissance d'ENs (Lafferty et al., 2001). Pour ce faire le graphe utilisé est présenté par une "chaîne linéaire du premier ordre" où chaque variable Yv est reliée uniquement à sa voisine gauche et à sa voisine droite. Dans notre application nous avons utilisé CRF ++ [1]pour annoter des séquences des entités nommées (personne, lieu et organisation).

Le module d'apprentissage repose sur la sortie du système à base de règles, les descripteurs et l'algorithme de classification. La sortie du modèle de classification est utilisée dans la phase de prédiction pour générer l'annotation finale des ENs. La sortie du système hybride de REN, à base de CRF, est analysée et exploitée afin d'améliorer le module à base de règles

La sélection des descripteurs (features) implique la sélection d'une combinaison de fonctions de classification de l'espace de caractéristiques global. Les descripteurs étudiés dans notre application sont divisés en: caractéristiques dérivées du module à base de règles, les caractéristiques morphologiques, étiquetage morphosyntaxiques (POS-tag), les caractéristiques des gazetteers.

[1] http://crfpp.sourceforge.net/

Chaque réalisation x d'un élément d'une de ces catégories donne lieu à des fonctions booléennes testant x avec chaque étiquette et avec chaque n-gramme d'étiquettes possibles.
Le jeu de descripteurs utilisé pour l'extraction des entités nommées comprend les caractéristiques suivantes:
-Caractéristiques à base de règles: Ces éléments contextuels sont la principale contribution du module à base de règles au système hybride. Ils sont issus des décisions fondées sur des règles et définis en termes de fenêtre de taille 5 mots autour du mot courant.
-Caractéristiques linguistiques: sont dérivées de l'analyse morphologique. Ces caractéristiques aident à distinguer l'entité nommée du texte régulier en se basant sur son état morphologique. Ces caractéristiques sont respectivement: l'aspect, le mode et l'état du verbe, le nombre le genre, la personne, la voix, l'existence ou non des proclitiques (tels que les conjonctions proclitique (Fa), conjonction de subordination (Wa), les particules, les prépositions (Fi, Bi), la jussive (Li), marqueur du futur (Sa) les particules négatives, les pronoms relatifs, etc.
-Etiquetage morphosyntaxique: La caractéristique du POS est la catégorie morphosyntaxique du mot cible estimée par l'outil SAPA[2]. Cette caractéristique permet au classificateur d'apprendre les étiquettes morphosyntaxiques que les entités nommées se produisent avec. Ces étiquettes sont: nom, nombre, nom propre, adjectif, adverbe, pronom, verbe, particule, préposition, conjonctions et ponctuation.
-Caractéristique du Gazetteer : vérifie la classe de l'entité nommée (personne, lieu et organisation): une fonction binaire pour vérifier si le mot (voisin gauche / voisin de droite du mot courant) appartient aux Gazetteers (personne, localisation, organisation). Cette caractéristique aide à révéler le contexte des entités nommées.
-Ponctuation : cette caractéristique indique si le mot a un point adjacent, par exemple, au début ou à la fin de la phrase ou il fait partie d'une abréviation. Cette fonction permet d'exploiter la position du texte dans le cadre des modèles de classification.

4 Protocole expérimental

Le processus d'évaluation est divisé sur deux expérimentations principales. La première est dédiée pour la comparaison de notre système hybride par rapport à l'état de l'art. Dans la deuxième nous comparons les performances des différentes adaptations de notre système. Les expériences sont réalisées sur les données décrites ci-dessous (section 4.1). Les scores sont calculés en termes de rappel(R), précision (P) F-mesure(F). Nous avons évalué notre système sur le corpus ANER et le Corpus News Commentary 2012en utilisant Corpus Quality Assurance. Ce dernier est un plugin dans l'outil GATE qui permet de comparer entre plusieurs annotations sur un même corpus.

4.1 Comparaison avec l'état de l'art

Dans cette section, nous comparons les performances de notre système hybride de REN par rapport aux travaux existants dans le domaine de REN arabe. Notre système est entraîné sur des corpus des UN et ANER. La comparaison est réalisée avec les systèmes NERar de (Gahbiche et al,2013) et le système de base de (Benajiba et al, 2008). Ces deux sont entrainés et évalués sur le corpus ANER. Le système de (Gahbiche et al,2013) est basé sur les CRF pour l'apprentissage comme dans notre système. Le tableau montre les performances de notre système sur le corpus ANER en comparaison avec ces deux systèmes sur les EN (personne, organisation et lieu).

[2] https://github.com/SouhirG/SAPA

Système de REN utilisé	Personne			Organisation			Lieu			F-mesure globale
	P	R	F	P	R	F	P	R	F	
Benajiba et Rosso	**80.41**	**67.42**	**73.35**	**84.23**	**53.94**	**65.76**	**93.03**	**86.67**	**89.74**	**76.28**
NERAr (Gahbiche)	**82.49**	**78.53**	**80.46**	**83.37**	**66.01**	**73.68**	**89.75**	**89.39**	**89.57**	**81.23**
Notre système Hybride	**84**	**82.74**	**83.36**	**85.64**	**62.50**	**72.26**	**89.81**	**89.37**	**89.58**	**81.73**

Tableau 3 : Comparaison de notre système hybride de REN avec le système de Benajiba sur le corpus ANER

D'après le tableau ci-dessus nous remarquons que notre système hybride présente de bons scores même si la répartition des données dans la phase d'apprentissage n'est pas équitable entre les deux corpus ANER et UN. En effet, nous avons utilisé 4871 phrases du corpus ANER et 15000 phrases du corpus UN. Notons bien que notre système donne des performances comparables à l'état de l'art pour les noms de personnes et organisations, et améliore les performances pour les lieux. Pour montrer l'impact de l'hybridation et l'ajout du lexique comme étant une ressource de renforcement de la REN, nous procédons à une deuxième évaluation sur le corpus News Commentary.

4.2 R EN pour le corpus News Commentary

Le corpus News Commentary est différent du corpus ANER que nous avons utilisé dans l'évaluation précédente. Le corpus News Commentary est extrait des sites politiques et économiques dont les thèmes abordés sont proches de ceux de notre corpus d'apprentissage de base, qui est extrait des travaux de l'organisation des nations unies (UN). Le tableau ci-dessous montre les performances du système de REN de base avec le système de REN optimisé et hybride sur le corpus News Commentary.

Entités nommées		REN à base de règles	REN à base de règles + lexique	REN Hybride
Personne	Précision	48.3	80.6	84.3
	Rappel	45.7	79.8	83.34
	f-mesure	46.96	80.19	83.81
Organisation	Précision	52.12	71.54	86.24
	Rappel	33.4	59.12	62.5
	f-mesure	40.71	64.73	72.47
Lieu	Précision	59.5	86.7	89.86
	Rappel	44.6	80.35	89.5
	f-mesure	50.98	83.40	89.67
F-mesure globale		46.21	76.10	81.9

Tableau 4: Comparaison de système de REN de base avec le système de REN optimisé et hybride sur le corpus News Commentary.

-Le système de base présente notre système sans ajout du lexique. Il s'agit de l'annotateur à base de règles intégré dans l'outil GATE. Ce mode présente des scores modestes au niveau de précision, rappel et F-mesure pour toutes les classes des ENs. Cela s'explique par la carence des gazetteers arabe dans cet annotateur. On rappelle qu'il est développé principalement pour l'anglais et par la suite on lui a intégré un plugin pour le traitement arabe. Les règles de la grammaire sont satisfaisantes mais les gazetteers sont encore incomparables par rapport à celles de l'anglais.
-La version système de base + lexicon est la version optimisée. Nous avons enrichi le système de base par notre lexique bilingue. En fait nous aons mappé la partie arabe du lexique bilingue vers les gazetteers de GATE. Comme résultat, nous notons une amélioration de la précision pour les organisations et les noms de personnes et lieu. La force de notre système d'annotation à base de règles réside dans la reconnaissance des entités lieu, qui est attribuée à la haute couverture de notre lexique d'ENs contenant les DBpedia datasets. Il est important de constater que notre système présente un bon rappel pour les noms de personne, qui étaient plus abondants dans le corpus des Nations Unies et dans notre lexique (27480 EN Personne). Outre cela, le corpus avait un mélange hétérogène de noms propres de personne non seulement des pays arabes mais aussi des continents d'Afrique, d'Asie et d'Amérique. Un bon pourcentage de rappel pour les noms de l'entité personne est encourageant parce que les entités nommées de l'Asie du Sud et l'Amérique n'ont pas de similitude phonétique avec les noms de personnes des pays arabes. Un examen détaillé des résultats des entités organisations révèle que notre système ne gère pas efficacement les acronymes et les abréviations.
-Le système hybride est la version finale de notre système. Les résultats prouvent une amélioration par rapport aux résultats antérieurs. On note que le système hybride améliore pratiquement la reconnaissance de toutes les ENs. .Nous constatons de plus que le système hybride donne de meilleures performances sur le corpus News Commentary (81.2%) par rapport au corpus ANER(80.9%). Ceci, comme déjà expliqué, revient que les thèmes traités dans le corpus News Commentary sont proches de ceux traités dans le corpus UN d'apprentissage.

5 Conclusion

D'après la littérature dans le domaine de REN arabe, l'utilisation des systèmes à base de règles ou par apprentissage est considérée comme des tentatives réussites dus aux spécificités et à la complexité de la langue arabe. Notre méthode s'inscrit dans les méthodes de REN hybride. En outre, notre système se diffère des autres approches sur plusieurs niveaux. La différence majeure est que notre système se base sur un module à base de règles renforcé par un lexique des entités nommées extraites des linked data datasets. Alors que le module par apprentissage est utilisé pour supporter le module à base de règles et améliorer la qualité de la REN globale. Les résultats expérimentaux ont prouvé que le système hybride surpasse l'état de l'art ainsi que les modules à base de règles et par apprentissage.
Dans les travaux futurs, nous avons l'intention d'améliorer les règles développées pour le module à base de règles. Cela pourrait se faire par l'intermédiaire d'une analyse de la sortie du système hybride afin d'automatiser la tâche d'amélioration. Une autre piste intéressante consiste à combiner différentes techniques d'apprentissage supervisé, autre que CRF, et évaluer son impact sur la performance du système REN

Références

ABDELRAHMAN S., ELARNAOTY M., MAGDY M., FAHMY A. (2010). Integrated machine learning techniques for Arabic named entity recognition. *International Journal of Computer Science Issues (IJCSI)*, 7(4):27–368.

ABDUL-HAMID A, DARWISH K. (2010). Simplified feature set for Arabic named entity recognition. In *Proceedings of the 2010 Named Entities Workshop (NEWS 2010)*, pages 110–115, Stroudsburg, PA.

AL-JUMAILY H., PALOMA M., ERIK G. (2012). A real time named entity recognition system for Arabic text mining. *Language Resources and Evaluation Journal*, 46(4):543–563.

AL-SHALABI R., GHASSAN K., BASHAR A., KAIGHONMEIN K., SALEM A. (2009). Proper noun extracting algorithm for the Arabic language. In *International Conference on IT to Celebrate S. Charmonman's 72nd Birthday*, pages 28.1–28.9, Bangkok.

BENAJIBA Y., DIAB M., ROSSO P. (2008). Arabic named entity recognition: An SVM-based approach. In *Proceedings of Arab International Conference on Information Technology (ACIT 2008)*, pages 16–18, Hammamet

BENAJIBA Y., ROSSO P., BENEDÍ J.-M. (2007). ANERsys : An Arabic Named Entity Recognition System Based on Maximum Entropy , CICLing, p. 143-153,.

DAILLE B., FOUROUR N., MORIN E. (2000). Catégorisation des noms propres : une étude en corpus. Cahiers de Grammaire, 25, 115–12

ELSEBAI A., FARID M. (2011). Extracting persons names from Arabic newspapers. In *Proceedings of the International Conference on Innovations in Information Technology*, pages 87–89, Dubai.

GAHBICHE S., MAYNARD H., YVON F. (2013). traitement automatique des entités nommées en arabe : détection et traduction. TAL 54(2): 101-132

KOULALI M., ABDELOUAFI M. (2012). A contribution to Arabic named entity recognition. In *Proceedings of the 10th International Conference on ICT and Knowledge Engineering*, pages 46–52, Morocco.

KÜÇÜK, D., YAZICI A. (2012). A hybrid named entity recognizer for Turkish. Expert Systems with Applications. 39, 2733-2742.

LAFFERTY J.D., MCCALLUM A., PEREIRA F. (2001). Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data. In ICML '01 : Proceedings of the Eighteenth International Conference on Machine Learning, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

MESFAR S. (2007). Named entity recognition for Arabic using syntactic grammars. In *Proceedings of the 12th International Conference on Applications of Natural Language to Information Systems (NLDB'07)*, pages 305–316, Berlin.

OUDAH M., SHAALAN K. (2012). A pipeline Arabic named entity recognition using a hybrid approach. In *Proceedings of the International Conference on Computational Linguistics*, pages 2,159–2,176, Mumbai.

PETASIS G., FRANTZ VT., FRANCIS W., GEORGIOS PS., VANGELIS K., CONSTANTINE D.(2001). Using machine learning to maintain rule-based named-entity recognition and classification systems. In Proceedings of the 39th AnnualMeeting of the Association for Computational Linguistics, pages 426–433, Stroudsburg, PA.

SEON C., KO Y., KIM J., SEO J. (2001). Named Entity Recognition using Machine Learning Methods and Pattern-Selection Rules. In Proceedings of the 6th Natural Language Processing Pacific Rim Symposium. 229-236.

SHAALAN K., HAFSA R. (2008). Arabic named entity recognition from diverse text types. In Bengt Nordstr ¨om and Aarne Ranta, editors, *Advances in Natural Language Processing*, volume 5221 of *Lecture Notes in Computer Science*. Springer, Berlin Heidelberg, pages 440–451.

TRAN T., PHAM HQ., NGO D., COLLIER N. (2007). Named entity recognition in vietnamese documents. Progress in Informatics, 4:5-13.

TSAI T., WU S., LEE C., SHIH C., HSU W. (2004). Mencius: A Chinese Named Entity recognizer Using the Maximum Entropy-based Hybrid Model. Computational Linguistics and Chinese Language Processing. 9, 65-82.

TUERLINCKX T. (2004). Lemmatisation de l'arabe non classique ", JADT 2004 : 7es Journées internationales d'Analyse statistique des Données Textuelles, 2004.

W LI. MCCALLUM A. (2003). Rapid development of hindi named entity recognition using conditional random fields and feature induction. Special Issue of ACM Transactions on Asian Language Information Processing: Rapid Development of Language Capabilities: The Surprise Languages,.

ZAGHOUANI W. (2012). RENAR: A rule-based Arabic named entity recognition system. *ACM Transactions on Asian Language Information Processing (TALIP)*, 11(1):2:1–2:13.

ZHOU J., LIANG D., JIAJUN C. (2006). Chinese named entity recognition with a multi-phase model. In Proc. of the Fifth SIGHAN Workshop on Chinese Language Processing, pages 213-216, Sydney, Australia, 2006. Association for Computational Linguistics.