L'intégration d'un outil de repérage d'entités nommées pour la langue arabe dans un système de veille

Wajdi Zaghouani
Linguistic Data Consortium, 3600 Market street suite 810, Philadelphia, USA
wajdiz@ldc.upenn.edu

Résumé. Dans cette démonstration, nous présentons l'implémentation d'un outil de repérage d'entités nommées à base de règle pour la langue arabe dans le système de veille médiatique EMM (Europe Media Monitor).

Abstract. We will present in this demo an Arabic rule-based named entity recogntion tool which is integrated within the news monitoring system EMM (Europe Media Montior).

Mots-clés : Étiquetage des entités nommées, langue arabe, système de veille médiatique.
Keywords: Named entity recognition, Arabic language, news monitoring system.

1 Présentation du système de veille médiatique EMM

Le système de veille médiatique EMM parcourt quotidiennement une moyenne de 100 000 articles en 50 langues différentes provenant de plus de 900 sites Web (Steinberger *et al*., 2009). Il permet de regrouper les articles couvrant le même sujet, fusionnant ces articles automatiquement en un seul groupe afin d'éviter la redondance des nouvelles en n'affichant qu'un seul événement pour plusieurs articles traitant le même sujet. Le système fonctionne comme un système d'information et d'alerte médiatique en temps réel permettant d'avoir un aperçu quotidien des nouvelles à travers le monde, tout en affichant les informations sous un format lisible et permettant l'accès rapide au contenu pertinent d'un article donné (cf. figure 1). Derrière le système EMM il existe une panoplie d'outils et de modules multilingues, comme les modules de représentation graphique, les outils d'analyse statistique, le générateur du réseau social des personnes ou le module de repérage des entités nommées (EN).

Figure 1 : L'environnement EMM

2 Présentation de RENAR

RENAR est un outil de repérage des EN à base de règles que nous avons créé spécifiquement pour la langue arabe (Zaghouani *et al.*, 2010). Il est totalement intégré dans l'environnement EMM à l'instar des autres langues (Pouliquen *et al.*, 2005). Le processus de repérage (cf. figure 2) commence par une étape de prétraitement lexical qui permet de préparer le texte brut à l'analyse linguistique en segmentant tout d'abord le texte en phrases, puis en normalisant son orthographe. Par la suite, l'opération d'extraction des EN se fait sur deux étapes : la première étape est basée sur la consultation directe du lexique qui se compose de plusieurs dictionnaires. Lors de cette étape, le système commence par la comparaison de chaque entrée dans le texte brut avec chacune des entrées des différents dictionnaires que nous avons construits. Une fois une EN reconnue grâce à un dictionnaire, elle sera automatiquement retenue sans passer par la deuxième étape, qui est réservée exclusivement à la détection des EN ne figurant pas dans le lexique. La deuxième étape repose sur des fichiers de règles écrites à la main sous forme d'expressions régulières qui permettent de détecter les EN grâce aux dictionnaires et à la liste des marqueurs lexicaux qui sont des indices textuels permettant de localiser les EN dans un texte.

3 Disponibilité

L'accès à l'environnement EMM y compris l'outil RENAR est proposé gratuitement à la communauté sous forme de page Web[1]. Par contre, EMM et RENAR ne sont pas disponibles sous forme de source pour le grand public.

[1] L'environnement EMM et RENAR sont accessible sur : <http://emm.newsexplorer.eu/NewsExplorer/home/ar/latest.html>

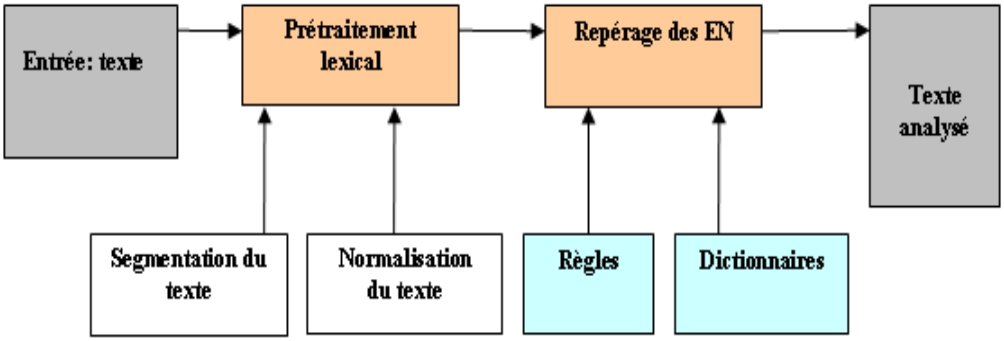

Figure 2 : Architecture de RENAR

4 Aperçu de la démonstration et équipement nécessaire

La démonstration ne demandera pas d'équipement particulier mis à part un ordinateur et une connexion Internet. Les participants auront le matériel nécessaire pour naviguer sur le portail EMM et découvrir les différentes fonctionnalités du module RENAR.

La figure 3 illustre l'environnement EMM avec la page en langue arabe qui intègre l'outil RENAR, les EN de type pays et personnes sont affichées dans des sections réservées à la droite de l'écran. La figure 4 illustre des détails sur un exemple d'une entité nommée trouvée.

Figure 3 : L'environnement EMM avec RENAR

Figure 4 : Illustration d'un exemple d'entité nommée

Références

POULIQUEN B., STEINBERGER R., IGNAT C., TEMNIKOVA I., WIDIGER A., ZAGHOUANI W. & ŽIŽKA J. (2005). Multilingual person name recognition and transliteration. CORELA, Poitiers, France, CERLICO. ISSN 1638-5748. Volume 3/3, numéro 2, pp. 115-123.

STEINBERGER R., POULIQUEN B. & VAN DER GOOT E. (2009). An Introduction to the Europe Media Monitor Family of Applications. In Gey F., Kando N. & Karlgren J. (eds.): Information Access in a Multilingual World - dans *Proceedings of* the SIGIR 2009 Workshop (SIGIR-CLIR'2009). Boston, USA.

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