

Systeme de Question/Réponse dans le cadre d'une plateforme intégrée : cas de l'Arabe

Lahsen Abouenour¹, Karim Bouzoubaa¹, Paolo Rosso²

Email: abouenour@yahoo.fr, karim.bouzoubaa@emi.ac.ma, proso@dsic.upv.es

1 Ecole Mohammadia d'Ingénieurs
Université Med V - Agdal,
Avenue Ibn Sina B.P. 765 Rabat Morocco

2 Natural Language Engineering
Lab., Dpto. Sistemas Informáticos
y Computación, Universidad
Politécnica Valencia, Spain

Résumé

Les outils de recherche jouent un rôle fondamental dans le cycle d'exploitation du contenu informationnel disponible sur différents médias électroniques notamment sur le Web. Cependant, pour des besoins précis et pointus le recours aux systèmes de Question/Réponse est indispensable. Le présent article présente un effort dans ce domaine consacré à la langue Arabe et mené dans le cadre d'une plateforme logicielle intégrée.

1. Introduction

Aujourd'hui, la surcharge d'information est devenue de plus en plus un défi que les systèmes d'information doivent prendre en charge. En effet, nous remarquons l'expansion du contenu disponible sur différents médias (en particulier Internet). Par conséquent, il serait intéressant de mettre en place des outils permettant d'automatiser les traitements liés à la recherche de l'information, de faciliter l'accès à celle-ci, de diminuer la surcharge d'information, etc.

Jusqu'à aujourd'hui le marché de l'informatique essaie de répondre à cette problématique en développant des outils spécifiques tels que : les moteurs de recherche¹, les systèmes de Question/Réponse [3], les systèmes d'extraction d'information [1], les analyseurs morphologiques et syntaxiques [5], etc. Notons que ces outils peuvent présenter une certaine forme de dépendance : par exemple, un système Q/R (ou d'extraction de l'information) peut faire appel à un analyseur morphologique. La maturité et l'efficacité de ces outils diffèrent selon le niveau de complexité du domaine traité et selon la langue cible. A ce titre, et malgré divers efforts, la maturité et l'efficacité de ce type d'outils pour le cas de la langue Arabe, reste relativement faible par rapport à d'autres langues.

Le présent article est une contribution dans le domaine des systèmes de Question/Réponse pour le cas de la langue Arabe. Ce travail est particulier dans la mesure où il n'est pas isolé. Il représente, en fait, une composante d'un projet de plus grande envergure dont la finalité est de mettre en œuvre une plateforme logicielle pour le développement des différentes applications susmentionnées autour de la langue Arabe. Cette plateforme est conçue en respectant les tendances des systèmes d'information urbanisés, à savoir : l'ouverture, la standardisation, la flexibilité, la réutilisation, etc.

¹ <http://ajeab.sakhr.com>

L'article est structuré comme suit. Dans la deuxième partie nous décrivons l'architecture et la conception de la plateforme logicielle proposée en soulignant les modules déjà réalisés notamment celui de l'analyse morphologique de la langue Arabe. La troisième partie est consacrée aux travaux connexes concernant les systèmes Question/Réponse de la langue Arabe. Dans la quatrième partie, nous présentons notre système de système Q/R avec un exemple spécifique qui illustre les avantages de notre approche. Enfin, nous dressons une conclusion en définissant les travaux futurs relatifs à la finalisation de ce projet.

2. Architecture de la plateforme logicielle

Notre plate-forme est conçue pour être une API Java open source multi-couche et pour offrir un environnement de développement modulaire, intégré, et dédié à la mise en œuvre d'applications liées à la recherche et au traitement du contenu en langue Arabe. Comme l'illustre la figure 1, l'architecture contient trois couches de base : morphologie, syntaxe et sémantique.

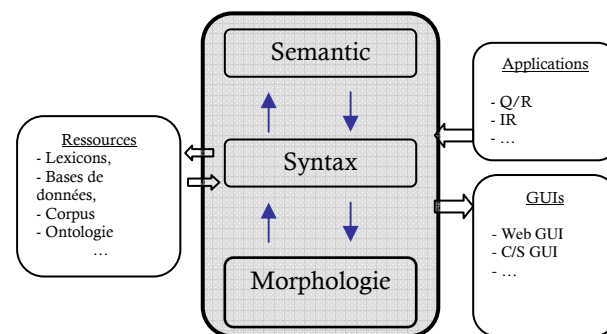


Figure 1: Architecture de la plateforme logicielle

Les trois couches forment une hiérarchie. Chaque couche utilise les couches basiques sur lesquelles elle est construite. Toutefois, une couche inférieure peut être utilisée par elle-même sans avoir recours à une couche plus haute : la couche morphologie (c'est-à-dire les API associés) peut être utilisée directement dans n'importe quelle application, sans faire appel aux autres couches, la couche syntaxique peut également être utilisée directement, etc. Atteindre un plus haut niveau de modularité et d'indépendance entre ses composantes a figuré parmi les objectifs qui ont influencé la conception de la plate-forme.

La plateforme présente des caractéristiques comme la portabilité, la réutilisabilité, l'ouverture à la communauté des utilisateurs et chercheurs (source libre, interfaces web, etc.), la qualité de la documentation, l'utilisation de standards (ex. XML), etc.

Dans le cadre de cette plateforme, les travaux qui ont été achevés sont [2] :

- Développement d'un dictionnaire lexical des noms de l'Arabe (comprenant plus de 40 000 entrées) ;
- Développement de la couche 'Morphologie' et l'API associé. Rappelons que l'analyse morphologique s'intéresse aux structures que les mots prennent dans les différentes utilisations [7]. Pour analyser un texte arabe, notre analyseur le découpe en unités lexicales et retourne l'analyse morphologique des tous les noms. Le résultat final est structuré sous format XML permettant d'être exploité soit directement, soit par une autre couche de la plateforme ou par une application

particulière. A titre d'exemple, l'analyse morphologique du mot اللغة retournera sous format XML les solutions possibles à savoir (ل-ال-لغة) et (ل-لغة);

- Une interface classique (par utilisation de swing de Java) et une interface web pour permettre à la communauté d'exploiter la couche morphologique et le dictionnaire associé.

L'étape sur laquelle nous travaillons actuellement consiste à mettre en place une application de Q/R dans le cadre de notre plateforme. Avant de décrire notre approche, nous allons consacrer la partie suivante aux travaux déjà réalisés dans ce domaine.

3. Travaux connexes des systèmes de Q/R

Contrairement aux moteurs de recherche, les systèmes de Question/Réponses ne se contentent pas de retrouver les documents contenant une certaine combinaison de chaîne de caractères mais essaient plutôt d'obtenir une réponse exacte à une question spécifique (la question et la réponse sont formulées toutes les deux en langage naturel). Pour la langue Arabe, de nombreuses implémentations de systèmes de Q/R existent :

- QARAB [8] est un système qui traite des questions exprimées en langue arabe et essaie de fournir des réponses courtes. Le système a pour principale source de connaissance une collection de journaux arabes extraits d'Al-Raya², un journal publié au Qatar. QARAB ne procède pas à une analyse sémantique de la question;
- AQAS [9] est basé sur la connaissance et, par conséquent, extrait les réponses seulement des données structurées et non pas de textes bruts (textes non structurés écrits en langage naturel);
- ArabiQA [3] est un système de Q/R pour l'arabe. Il est basé sur un module d'extraction de texte et sur un système de reconnaissance d'entités nommées (NER) [4]. Il intègre un module d'extraction de réponses dédié plus particulièrement aux questions types. Afin de mettre en place ce module les auteurs ont élaboré un système de NER arabe et un ensemble de modèles pour chaque type de question (élaboré à la main).

4. Application Question/Réponse

4.1 Architecture du système Q/R

Un système de Q/R est composé en général de trois principaux modules (figure 2) [8] :

1. Module d'analyse de la question; généralement ce module requiert un classificateur de questions, un module d'extension de requête (par mots clés) et un module de reconnaissance des noms d'entités NER (personnes, pays, organismes, etc.) ;
2. Module de recherche des documents candidats contenant des réponses ;
3. Module de traitement de chaque document candidat de la même manière que la question et extraction des phrases qui contiennent la réponse.

² <http://www.raya.com>

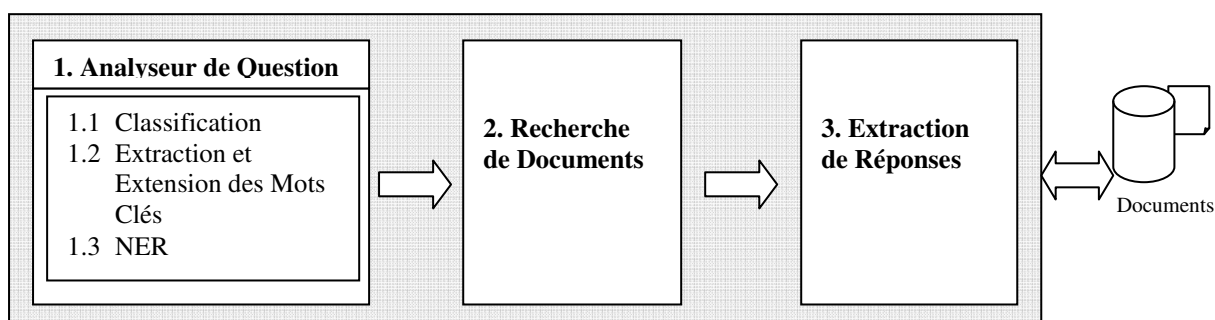


Figure 2 : Schéma type d'un système de Q/R

Il est connu que l'étape d'extension des mots clés est l'une des étapes cruciales d'un système Q/R. En effet, en se basant seulement sur les mots clés apparaissant dans la question, le nombre de documents contenant ces mots clés peut être faible et par conséquent diminue les chances de construire une réponse viable. C'est pour cela que nous procédons à une extension des mots clés originaux. Cette étape consiste à retourner d'autres mots clés proches (ayant une certaine relation avec les mots clés d'origine) qui permettraient d'augmenter les chances de trouver des documents contenant une ou plusieurs parties de la réponse.

Cette extension est généralement réalisée par utilisation d'un outil morphologique. Par exemple, la question *ماهي الشركات التي وظفت السيد جودت الحلبي ؟* comporte le mot clé: *وظفت*. Un outil morphologique générerait d'autres formes du même mot tel que : *وظيفة، موظف، توظيف، وظيفة*. Ces nouvelles formes pourraient également jouer le rôle de mots clés.

Notre approche ne consiste pas à étendre les mots clés seulement sur le plan morphologique mais également sur le plan sémantique. Pour ce faire, nous avons développé et utilisé une ontologie lexicale et sémantique construite à partir de :

- L'ontologie SUMO (Suggested Upper Merged Ontology) qui a été créée au Teknowledge Corporation [10, 11] et qui est destinée à être unifiée, cohérente et générale. Elle contient près de 2000 concepts indépendants de tout langage et ceci avec leurs définitions sémantiques. Par ailleurs ces concepts sont liés par relations de hiérarchie (spécialisation et généralisation), i.e. un concept peut avoir plusieurs sous types et/ou plusieurs super types.
- Arabic WordNet (AWN) [6] qui est une ressource lexicale contenant plus de 23496 mots de la langue Arabe. Ces entrées sont disposées de manière à ce que chacune d'entre elles peut avoir plusieurs synonymes et peut être sémantiquement connectée aux concepts SUMO.

Ainsi, notre approche d'extension des mots clés ne s'opère pas seulement sur le plan morphologique mais également sur le plan sémantique et s'exécute en quatre phases (figure 3) : (i) nous cherchons les synonymes (issus de AWN) d'un mot clé d'origine ; (ii) nous cherchons les concepts formant la définition du concept lié au mot clé d'origine ; (iii) nous cherchons les sous types du concept lié au mot clé d'origine ; (iv) nous cherchons les super types du concept lié au mot clé d'origine.

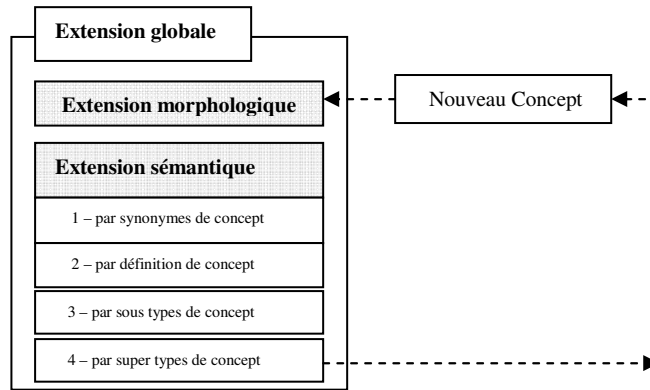


Figure 3 : Phases de l'extension des mots clés proposée

L'exemple de la partie suivante permet l'illustration des avantages de cette approche sémantique concernant l'extension des mots clés.

4.2 Exemple d'application

A partir du site web³ du groupe Al Arabi Investment, considérons l'exemple suivant :

السيد الحلبي هو مدير الدائرة البنكية الخاصة وإدارة الثروات والاستثمار في البنك العربي، والمدير العام لمجموعة العربي للاستثمار. قبل التحاقه بالبنك، كان يشغل منصب رئيس المنطقة الغربية لقسم تمويل الشركات في البنك الأهلي السعودي (NCB) في جدة، وانتقل بعدها ليصبح رئيساً للدائرة البنكية الخاصة عام 1999 والتي تشمل خدمات الوساطة المالية.

وقد تبوّ السيد الحلبي مراكز عديدة في البنك السعودي الأمريكي في جدة (بنك تابع لـ - سيتي بنك) وذلك ابتداء من العام 1984 كمدير لوحدة التمويل للشركات البتروكيمياوية. وفي العام 1990 التحق السيد الحلبي في البنك السعودي الهولندي (بنك تابع لمجموعة ABN AMRO) في جدة كعضو من الفريق الذي تم استقطابه لإعادة هيكلة البنك حيث شغل منصب مدير قسم تمويل الشركات المنطقة الغربية. وفي العام 1995 عاد إلى البنك السعودي الأمريكي كمدير لوحدة القروض والتمويل حيث أسس وحدة العمل المصرفي الإسلامي .

Ce document est un extrait du CV de Mr. Jawdat Halabi. Considérons maintenant la question : **ماهي الشركات التي وظفت السيد جودت الحلبي ؟** (càd quelles sont les entreprises pour lesquelles a travaillé M. Jawdat HALABI ?). L'analyse de la question donne lieu aux mots clés : **الحلبي - جودت - السيد- وظفت -الشركات**. Prenons le mot clé : **وظفت** (employé par) qui joue un rôle important dans le sens de la question et par conséquent dans celui de la réponse attendue. En procédant à une reformulation morphologique nous obtenons les nouveaux mots clés : **موظف - وظيفة - توظيف**. Cependant, aucun de ces mots n'existe dans le document ci-dessus. Une nouvelle façon de procéder est de chercher des mots sémantiquement liés aux mots clés d'origine en appliquant le cycle de traitement décrit dans la figure 3 :

1. Le mot **توظيف** est un synonyme du concept « Hiring » qui existe comme concept dans notre ontologie. La figure 4 est un extrait du contexte de ce concept dans notre ontologie.

³ http://www.ab-invest.net/ar/index.php?option=com_content&task=view&id=53&Itemid=98

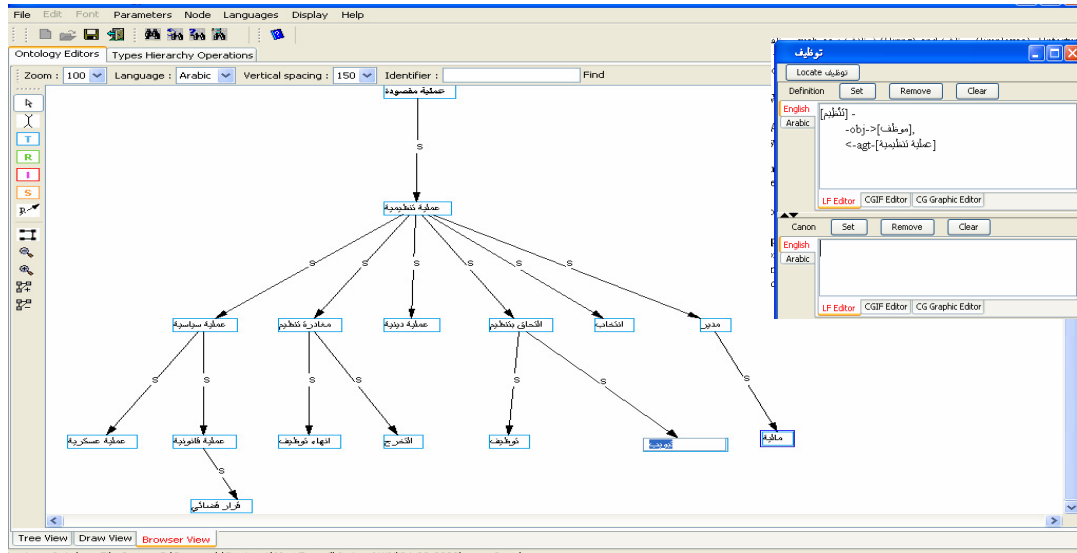


Figure 4: Extrait de L'ontologie et définition du concept "Organization"

Le fait que **توظيف** n'existe pas dans le document ci-dessus nous impose de passer à l'étape suivante du cycle d'extension.

2. La définition du concept Hiring ou **توظيف** dans notre ontologie est le suivant : [تنظيم] -obj->[موظف]<-agt-[عملية تنظيمية]. Cette définition comporte de nouveaux concepts tels que تنظيم. Ce dernier mot clé ne figure pas non plus dans le document ci-dessus. Passons alors à la troisième étape.
3. Le concept Hiring ou **توظيف** n'a pas de sous types donc nous passons à l'étape suivante.
4. JoiningAnOrganization (التحاق بتنظيم) est un super type direct de notre concept. Ce mot clé existe dans le document ce qui nous permet de le juger candidat pour abriter la réponse attendue.

Nous appliquons récursivement le cycle de traitement ainsi décrit. Le résultat final fait ressortir les nouveaux mots clés suivants : الرئيس - مدير - عضو - مجموعة - التحاق بتنظيم - موارد مالية - مالية.

Comme nous pouvons le constater, une grande partie de ces nouveaux mots clés apparaissent dans le document ci-dessus (mots soulignés dans le texte). Cela permettra au module de recherche associé au système de Question/Réponse de trouver un document candidat pertinent. Ces mots clés seront utilisés également pour extraire la réponse attendue.

5. Conclusion et Travaux Futurs

Dans cet article, nous avons présenté le projet d'application Q/R intégré dans notre plateforme de développement autour de la langue Arabe. Cette approche essaie de surmonter les limites de l'extension morphologique des mots clés en procédant à une extension sémantique basée sur les éléments d'une ontologie que nous avons construite à partir de Arabic WordNet et de l'ontologie SUMO.

Les prochaines étapes consistent à raffiner notre modèle d'extension sémantique en affectant des poids aux différents mots clés issus de cette extension selon la nature de la relation avec le mot d'origine (synonyme, définition, etc.) et de développer les

autres couches de notre plate-forme ainsi que d'autres applications de la langue Arabe permettant de contribuer à mieux gérer automatiquement la surcharge d'informations.

6. Remerciements

Pour leur contribution pour cet article, nos remerciements sont adressés aux projets de recherche 'AECI-PCI A01031707' et 'CICYT TIN2006-15265-C06'.

7. Références

- [1] ABDELALI A., COWIE J., SOLIMAN S. H. Arabic Information Retrieval Perspectives. JEP-TALN 2004, Arabic Language Processing, FEZ, 19-22 April, 2004.
- [2] ABOUENOUR L., EL HASSANI S., YAZIDY T., BOUZOUBA K., HAMDANI A. Building an Arabic Morphological Analyzer as part of an Open Arabic NLP Platform. In the Language Resources and Evaluation Conference (LREC), Marrakech, Morocco, 31st May, 2008.
- [3] BENAJIBA Y., ROSSO P., LYHYAOUI A. Implementation of the ArabiQA Question Answering System's components. In: Proc. Workshop on Arabic Natural Language Processing, 2nd Information Communication Technologies Int. Symposium, ICTIS-2007, Fez, Morocco, April 3-5, 2007.
- [4] BENAJIBA Y., ROSSO P. ANERsys 2.0 : Conquering the NER task for the Arabic language by combining the Maximum Entropy with POS-tag information. In: Proc. Workshop on Natural Language-Independent Engineering, 3rd Indian Int. Conf. on Artificial Intelligence, IICAI-2007, Pune, India, December 17-19, 2007.
- [5] BUCKWALTER T. Issues in Arabic Orthography and Morphology Analysis. In Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages, COLING 2004, Geneva, 2004.
- [6] ELKATEB S., BLACK W., VOSSEN P., FARWELL D., RODRIGUEZ H., PEASE A., ALKHALIFA M. Arabic WordNet and the Challenges of Arabic. In proceedings of Arabic NLP/MT Conference, London, U.K, 2006.
- [7] EL-SADANY T. A., HASHISH M. A. An Arabic Morphological System. IBM System Journal vol 28-no 4, 1989.
- [8] HAMMOU B., ABU-SALEM H., LYTINEN S., EVENS M. QARAB: A Question answering system to support the ARABic language. In: Proc. of the workshop on Computational approaches to Semitic languages, ACL, pages 55-65, Philadelphia, 2002.
- [9] MOHAMMED F. A., NASSER K., HARB H. M. A knowledge-based Arabic Question Answering System (AQAS). In: ACM SIGART Bulletin, pp. 21-33, 1993.
- [10] NILES I., PEASE A. Towards a Standard Upper Ontology. In: Proceedings of FOIS 2001, Ogunquit, Maine, pp. 2-9, 2001.
- [11] NILES I., PEASE A. Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. In Proceedings of the International Conference on Information and Knowledge Engineering., Las Vegas, Nevada, 2003.