Système de Question Réponse Pour la langue Arabe

Nom: faiza beirade **Encadreur**: Me Azzoune
Affiliation: Traitement Automatique de la Langue
Etablissement: Ecole Supérieure d'Informatique ESI

Résumé

La mise au point d'un système de question/réponse demande de disposer d'outils assez évolués et de ressources linguistiques et / ou pragmatiques très importantes.

La réalisation d'un tel système nécessite des techniques d'intelligence artificielle, de recherche d'information et de traitement automatique de la langue, pour la langue arabe peu de systèmes existent.

Pour cela nous avons contribués par notre travail modeste, notre système contient trois module le premier d'analyse de la question qui utilise la ressource lexicale Arabic Wordnet AWN qui contient plus de 23496 mots de la langue arabe et procède à une extension des mots clés originaux en utilisant la plate forme « AMINE » open-source implémentée en java .

Le deuxième module d'extraction de la réponse utilise le système de recherche d'information de java JIRS qui est une multi plateforme indépendante du langage. Enfin le module de sélection de la réponse.

Introduction

Aujourd'hui la surcharge d'information est devenue de plus un défi que les systèmes d'information doivent prendre en charge. En effet, nous remarquons l'expansion du contenu disponible sur différents médias. 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.

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 recherche, les systèmes de questions réponses, les systèmes d'extraction de l'information, les analyseurs morphologiques et syntaxiques, etc.

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.

Notre travail est une contribution dans le domaine des systèmes de questions réponses pour le cas de la langue arabe.

Cet article contient les modules conçus pour notre système afin d'élaborer des réponses intelligentes à des requêtes en langage naturel. En premier lieu nous présentons quelques systèmes de question réponse existants, où nous allons présenter le système Qarabe et le système arabiQA. Ensuite nous présentons les ambiguïtés syntaxiques de la langue arabe qui ralentissent la progression du traitement de cette dernière.

Enfin, nous présentons la conception et l'implémentation des trois modules de notre système, nous commençons par le module d'analyse de question qui suit le principe d'extension de mot clés en utilisant la plateforme open

source « AMINE» ensuite nous présentons le module d'extraction de la réponse, c'est le module noyau qui fonctionne avec un système de recherche d'information de java ; et enfin le module de sélection de la réponse.

1. . Historique des systèmes de question / réponse

La problématique des systèmes de questions réponses se situe à l'intersection de plusieurs domaines dont notamment la recherche d'information et le traitement de la langue naturelle. Il s'agit de poser des questions à une machine et attendre d'elle des réponses satisfaisantes. Le processus d'un système de questions-réponses est le suivant :

1. Analyse de la question.
2. Indexation de la question (transformer la question en requête).
3. Interroger un moteur de recherche pour chercher le document pertinent.
4. Extraction de la réponse.

Les tâches 2 et 3 se basent sur des techniques de recherche d'information, tandis que les processus 1 et 4 sont des tâches de traitement automatique de la langue naturelle (TALN).

Actuellement les trois premiers processus ne posent pas de grandes difficultés du point de vue des techniques disponibles en recherche. Cependant, le dernier processus qui s'occupe de l'extraction de la réponse, reste un vrai défi pour le développement de tels systèmes, où les chercheurs posent les questions suivantes :

1. Comment peut-on répondre en temps réel à la question, justifier la réponse et savoir évaluer l'adéquation de la réponse à la question ?
2. Comment peut-on déterminer les documents les plus pertinents à la question de l'usager et choisir celui qui contient la réponse ?
3. Comment choisir entre plusieurs réponses candidates (qui ont le même degré de similarité) celle qui sera la plus pertinente ?

Le second problème se pose dans le traitement de la langue arabe qui reste toujours à son étape initiale comparé au travail des autres langues, Il y a quelques aspects qui ralentissent la progression du traitement de la langue arabe.

Ces aspects incluent :

- L'arabe est fortement flexionnel et dérivative, ce qui fait l'analyse morphologique une tâche très complexe.
- L'absence des diacritiques (qui représentent la plupart des voyelles) dans le texte écrit crée une ambiguïté et donc des règles morphologiques complexes sont exigées pour identifier la marque et analyser le texte.
- La direction de l'écriture est de droit à gauche et certains des caractères changent leurs formes basées sur leur endroit dans le mot.
- La capitalisation n'est pas employée dans l'arabe, ce qui le rend dur pour identifier des noms propres, anonymes, et abréviations.
- En plus des issues linguistiques ci-dessus, il y a également un manque de corpus arabes, de lexiques, et des dictionnaires lisibles, pour cela il est essentiel de faire des recherches dans différents secteurs.

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) [DPS].
Pour la langue Arabe, de nombreuses implémentations de systèmes de Q/R existent. Nous allons citer ci-dessous le système QARAB et le système Arabiqa.

2. .Le système QARAB

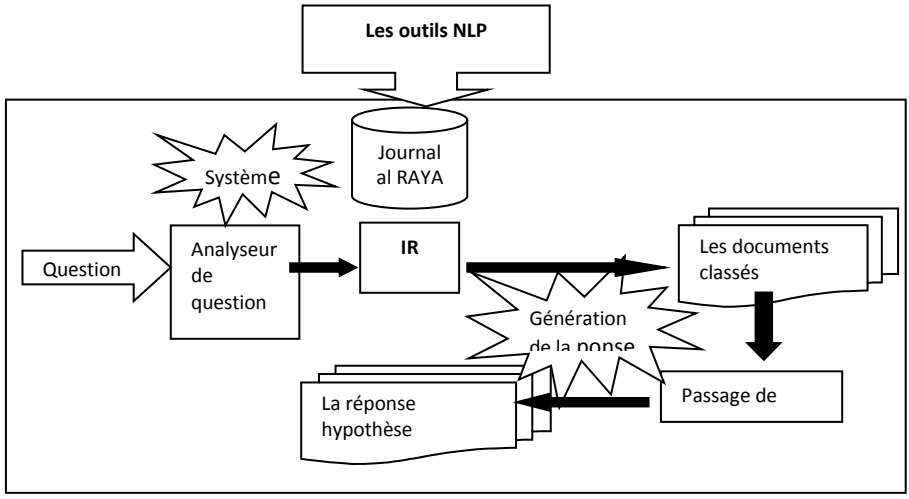
3.1 Présentation

Le système QARAB est un système qui traite des questions exprimées en langue arabe (langue naturelle) et essaie de fournir des réponses courtes. Le système a pour principale source de connaissance une collection de journaux arabes extraits d'Al-Raya, un journal publié au Qatar. QARAB ne procède pas à une analyse sémantique de la question **[HAM02]**.

Le système QARAB est le résultat de l'accouplement traditionnel Techniques (IRES) de recherche documentaire avec l'approche de traitement du langage naturel sophistiqué (NLP). L'approche peut être récapitulée comme suit :
Le système de recherche d'information IR (information retreival) traite la question afin d'identifier les documents candidats qui peuvent contenir la réponse ; puis les techniques de NLP sont employées pour analyser la question et analyser les documents rangés retournés par le système IR.

3.2 Architecture

Le système complet est présenté par La figure 1 ; il a la structure globale suivante :



Le but principal du système QARAB est l'identification des passages des textes qui répondent à une question de langue naturelle.

La tâche peut être récapitulée comme suit :

Pour un ensemble de questions exprimées en arabe, donner des réponses qui ont les caractéristiques suivantes :

- La réponse existe dans une collection de texte arabe de journal extrait à partir d'Al-Raya (journal édité au Qatar).

- La réponse ne franchit pas par des documents (c.-à-d. toute l'information de soutien pour la réponse mensonge dans un document)
- La réponse est un passage court.

Le traitement de base dans QARAB est composé de trois étapes principales :

- Traitement de la question d'entrée.
- Recherche des documents candidats (paragraphes) contenant des réponses avec le système IR.
- Traitement de chaque document candidat (paragraphes) et renvoi des phrases qui peuvent contenir la réponse.

Le système ArabiQA

4.1 Présentation

ArabiQA [BEN 07] est un système de Question/Réponse pour la langue arabe. Il est basé sur un module d'extraction de texte et sur un système de reconnaissance d'entités nommées (NER). Il intègre un module d'extraction de réponses dédié plus particulièrement aux types de questions. 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).

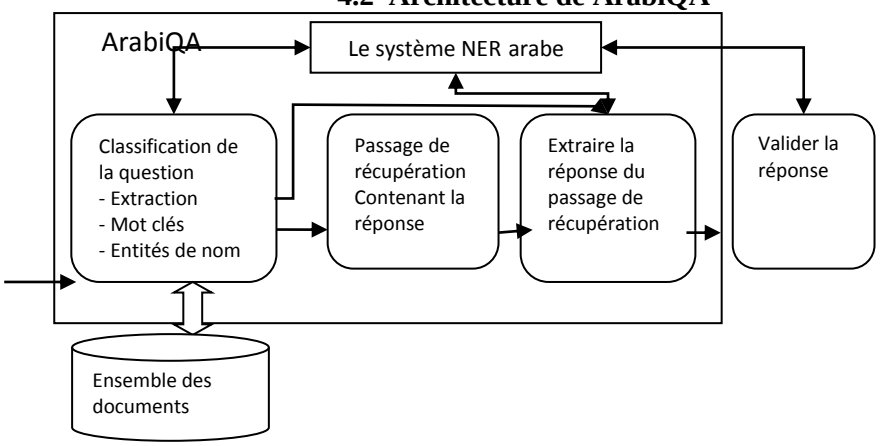
Ce travail a été réalisé par l'équipe suivante :

Y. Benajiba qui est un étudiant PHD à l'université polytechnique de Valence en Espagne.

P. Rosso qui est un docteur, un membre dans le laboratoire de la technologie de langage naturel et un conférencier permanent dans le service des systèmes informatiques de l'université polytechnique de Valence en Espagne

A. Lyhyaoui qui est un docteur, un membre du laboratoire des systèmes de technologie et un conférencier permanent dans l'Abdelmalek Essaadi Université de Maroc.

4.2 Architecture de ArabiQA



D'un point de vue général, le système se compose de composants suivants :

(i**) Le Module *d'analyse de question*** : il détermine le type de la question donnée (afin d'informer le module d'AE au sujet du type de réponse prévu), les mots-clés de la question (utilisé par le module de passage de récupération) et

les entités de noms apparaissant dans la question (qui sont très essentiel pour valider les réponses candidates).

(ii) **Le module de *récupération* de passage** : c'est le module de noyau du système. Il recherche les passages estimés appropriés à contenir la réponse.

(iii) **Le Module d'extraction *de la réponse*** : il extrait une liste de réponses candidates des passages appropriés.

(iv) **Le Module de *validation de réponses*** : il évalue pour chaque réponse candidate la probabilité d'exactitude et ils les rangent par leurs probabilité d'exactitude.

Les premiers, troisièmes et quatrièmes modules ont besoin d'un système fiable appelé Système d'identification d'entités (NER). Le système arabiQA, utilise son propre système de NER [BEN 07].

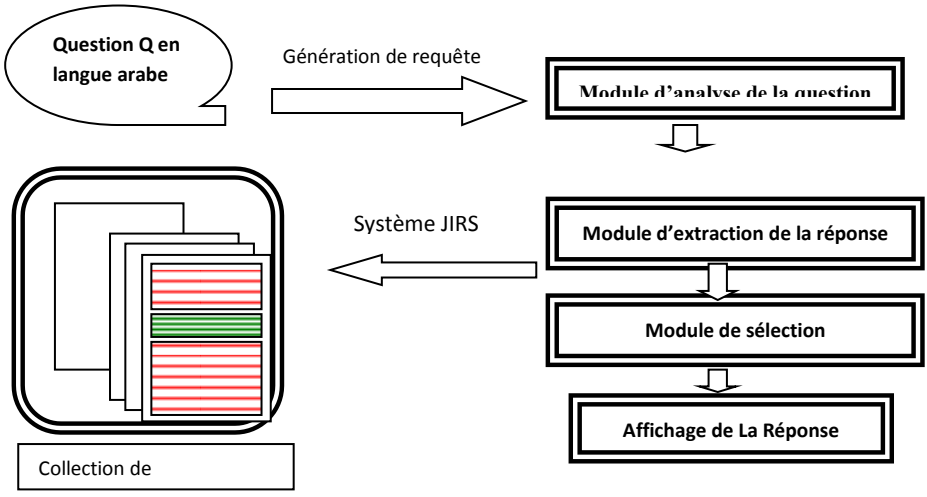
Conception et implémentation de système

5.1 Présentation :

Nous présentons ci-dessous la conception de notre système, un système de question réponse pour la langue arabe. Cependant, nous allons décrire brièvement son architecture modulaire, nous commençons par le module d'analyse de la question qui utilise la ressource lexicale AWN Arabic Word Net et la plate forme « amine », puis le module d'extraction de La réponse qui utilise le système de recherche d'information de java JIRS et nous terminerons par le module de sélection de la réponse.

5.2 Architecture de système

L'architecture de notre système de question / réponse est la suivante :



Architecture de système Q/R

5.3 Module d'analyse de la question

Ce module analyse les mots de la question et détermine leurs parties du discours (verbes, noms, particules)

- il détermine le type des noms (nom, personnels, endroit, extc).
- Extraction et extension des mots clés pertinents.

- Détermine la classe de la question pour déterminer le type de traitement.

Les types de questions possibles sont :

Qui, Dont : من ⟶ type : personne.

Quand : متى ⟶ **type :** Date, Temps.

Ce qui, Qui ماءماذا ⟶ **type :** Organisation, Produit, Événement.

Où أين ⟶ type : endroit (normal, politique).

Combien كم ⟶ **type :** Nombre, Quantité.

Pour cela nous allons utiliser une plate forme open-source implémentée en Java, la plate forme « AMINE ». Cette dernière nous a aidés aux différentes extensions des mots clés par synonymes et par définition, par types et par sous types.

On va se baser sur la ressource lexicale Arabic Wordnet (AWN) qui contient plus de 23496 mots de la langue arabe [MAL 06].

AWN est une ressource lexicale de la langue Arabe disponible gratuitement [MAL 06]. Elle est basée sur la conception et le contenu de Princeton WordNet et peut être liée à d'autres WN pour d'autres langues, ce qui permet une traduction à des dizaines de langues.

La base de données AWN est spécifiée dans le format XML et implémentée en SQL SERVER hébergée par l'un des sites du projet. Cette dernière à quatre types d'entités principales : item, mot, forme et lien.

La plateforme Amine est une plateforme open-source et multicouche, implémentée en Java et dédiée au développement de systèmes intelligents [KAR 07].

Elle permet le traitement des divers aspects du langage naturel. Amine n'est pas dédiée spécifiquement au TLN. Cependant, elle permet d'explorer les niveaux du langage (morphologie

, syntaxe, sémantique, pragmatique) et de développer des applications orientées langage naturel telles que l'analyse de phrases, la traduction ou encore les systèmes de questions /réponses.

Notre but est de faire l'extension des mots clefs de la question pour améliorer la précision de la réponse, pour cela on a utilisé amine pour définir les différentes extensions des mots clés par synonymes et par définition, par types et par sous types.

Grâce à Amine, nous avons développé une application traitant l'aspect morphologique de la langue arabe par exploration d'une base de règles. Il est possible de traiter en analyse aussi bien les noms que les verbes. La base de règles est exprimée dans le langage Prolog+C G. L'interface d'utilisation de l'application est développée dans le langage Java.

Etant donné que la plateforme Amine est développée en Java, l'interface graphique permet de saisir les informations et d'envoyer la requête à partir de Java vers Prolog+CG. Ce dernier répond à la requête grâce à son moteur d'inférence et à la base de règles. Pour le traitement des chaînes de caractères, Prolog+CG fait appel directement aux méthodes Java (e.g._L : size () est un appel à la méthode Java size() qui retourne la taille d'une liste)

Selon la structure de l'ontologie Amine AWN, le passage d'un concept à l'autre se fait en utilisant les relations sémantiques suivantes :

(i) Synonymes du concept.
(ii) Définition de la structure conceptuelle.
(iii) Sous types du concept.
(iv) Super types du concept.

5.4 Module d'extraction de la réponse

Ce module réalise les étapes suivantes :

- récupère les termes de la question (verbes, noms) générer par le module précédant.
- Recherche les documents appropriés.

Dans ce module nous utiliserons le système de recherche d'information de java JIRS qui est une multi plateforme et un système indépendant du langage [NAE 09]. Il est basé sur une approche de densité parce qu'il s'est avéré être la technique la plus réussie pour la tâche de récupération du passage (PR).

5.5 Module de sélection de la réponse

Sélection de la réponse correcte

C'est un ordonnancement des réponses candidates, le composant de sélection de la réponse peut être considéré comme un classifieur binaire de réponses candidates : la classe « 1 » est associée aux bonnes réponses, et « -1 » aux mauvaises. Il est ensuite possible d'apprendre un classifieur (par exemple une MVS) sur une base d'apprentissage, à partir d'une représentation vectorielle des réponses candidatent.

Lorsque les paramètres de la MVS sont appris, la sélection de la réponse à renvoyer par le système parmi les candidates se fait comme suit : pour chaque réponse candidate, la sortie du classifieur est considérée comme un score de confiance sur la pertinence de la réponse par rapport à la question. La réponse candidate obtenant le meilleur sc ore est alors renvoyée par le système.

En utilisant toujours le système de recherche d'information de JAVA et en considérant l'ensemble des réponses candidates, ce composant détermine la réponse finale que le système doit renvoyer.

Par rapport au composant d'extraction, cette étape se focalise sur l'extraction de certains types de chaînes de caractères, elle utilise d'une part plus d'information sur la proximité lexicale ou sémantique entre le passage dans lequel les chaînes de caractères sont extraites et la question et d'autre part, utilise des informations externes, comme le nombre de fois que la même chaîne de caractères a été extraite.

Implémentation

Nous avons récupérer les termes de la question générer par le traitement précédant :

الشبكة , الشراك , الشص , المشتركة et مشاركة.

تستعمل , مستعمل , استعمال , تستخدم , تعمل , تؤدي et العمل.

Le module d'extraction par son système de recherche génère l'extrait suivant :

الشبكة هي مجموعة من أجهزة كمبيوتر متصلة ببعضها البعض. تسمح بانتقال المعلومات فيما بينها.

تستعمل الشبكة لعدة أسباب، من بينها **المشاركة** في الملفات والبرامج. **المشاركة** في الطباعة والاتصال بين المستعملين

لكل مستعمل لجهاز من **الشبكة**، الحق في **مشاركة** أو عدم **مشاركة** المستعملين الآخرين في ملحقات جهازه، حيث تستطيع مثلا، أن تسمح **بمشاركتهم** في القرص الصلب وعدم **مشاركتهم** في القرص المضغوط. هناك ثلاثة أنواع من **المشاركة** **مشاركة** كاملة: يستطيع مستعمل الجهاز الآخر أن يقوم بكل العمليات على الملحقة **المشتركة** (قراءة، تغيير، حذف ...). مشاركة في القراءة فقط: أي أن المستعملين الآخرين لا يستطيعون التغيير في محتوى أقراصك **المشتركة. مشاركة** حسب كلمة مرور: في هذه الحالة، يجب إعطاء كلمة مرور لاستعمال الملحقة **المشتركة**. هناك نوعان من هذه **المشاركة**.كلمة مرور للقراءة فقط كلمة مرور للتشغيل الكامل

Enfin le module de sélection de la réponse utilise les deux résultats pour donner au premier paragraphe plus de poids parce qu'il contient tous les mots de la question, d'où la réponse générer sera :

تستعمل الشبكة لعدة أسباب، من بينها المشاركة في الملفات والبرامج. المشاركة في الطباعة والاتصال بين المستعملين

3. Conclusion

Dans toute la panoplie de moteurs de recherche disponible sur le marché, la qualité du résultat présenté à l'utilisateur est cruciale. L'handicap majeur de ces applications est le fait que l'internaute doit initier explicitement toutes les tâches suivantes : consulter les documents présentés, reformuler la question si nécessaire, bien choisir les mots clés (maîtriser le langage requête).

Cette métaphore d'interaction doit évoluer pour permettre aux usagers non spécialisés d'exploiter efficacement la masse d'information disponible. Grâce à un système de questions réponses, cet usager peut obtenir une réponse précise à une question posée en langue naturelle.

Un système de question-réponse recherche généralement dans un corpus de documents une réponse précise. Par exemple, si un usager pose la question "*Les étoiles se déplacent-elles dans le ciel ?*" Le système répondra par une réponse bien spécifique (Oui, Non) et non par une liste de liens hypertextes.

La mise au point d'un système de question/réponse demande de disposer d'outils assez évolués et de ressources linguistiques et/ou pragmatiques très importantes.

La réalisation d'un tel système nécessite des techniques d'intelligence artificielle, de recherche d'information et de traitement automatique de la langue arabe.

Pour cela nous avons contribués par notre travail modeste en utilisant la ressource lexicale Arabic Wordnet AWN qui contient plus de 23496 mots de la langue arabe.

nous avons procéder à une extension des mots clés originaux en utilisant la plate forme « AMINE » open-source implémentée en java .

Enfin nous avons utilisé pour l'extraction de la réponse le système de recherche d'information de java JIRS qui est une multi plateforme indépendante du langage.

7. **Bibliographie**

[BEN 07] BENAJIBA Y.ROSSO P .LYHYAOUI A. implémentation of the ArabiQA Question Answering Systèm's components. Morroco, April 3-5, 2007.

[CHA] Chafik aloulou, lamia hadrich belguith, ahmed hadj kacem, abdelmajid ben hamadou TUNISIE, conception et développement du système MASPAR d'analyse de l'arabe selon une approche agent.

[DPS] DOMINIQUE LAURENT, PATRIK SEGUELA QRISTAL,système de questions réponses.

[ERW] Langage XML Elliotte Rusty Harold et w.scott Means O'Reilly

[HAM 02] HAMMOU B.ABU-SALAM H, LYTINEN S .EVENS M. QARAB : A Question answering system to support the ARABic language. In :Proc. Of the workshop on computational approches to semitic language, ACL, Philadelphia, 2002.

[IOM] Traitement de requêtes XML et applications distribuées
Ioana Manolescu INRIA Futurs- LRI, projet Gemo Active XML www.purl.org/axml

[KAR 07] An integrate development plateform for arabic language processing
Karim bouzouba ,adilkabbaj
The 1er international symposium on computers and Arabic language 2007

[LKP] Three-level approach for Passage Retrieval in Arabic Question/Answering Systems Lahsen Abouenour, Karim Bouzoubaa and Paolo Rosso

[LKR] structure-based evaluation of an Arabic semantic query expansion Using the JIRS *Lahsen abouenour, karim bouzouba, paolo rosso*

[MAL 06] WordNet et arabe PNL arabe Musa Alkhalifa , Sciences cognitives et langagières, University of Pompeu Fabra, Bracelona,Spain.
JETALA 5-7 June 2006, Rabat.

[NAE 09] design and implementation of an information retrieval system by integrating semantique knowledge in the indexing phase
N.Tazzite,A.Yousfi, E.Bouyakhf MOROCCO February 2009

[OLB 01] Introduction aux Support Vector Machines (SVM)
Olivier Bousquet Centre de Mathématiques Appliquées
Ecole Polytechnique, Palaiseau Orsay, 15 Novembre 2001