MajecSTIC 2009
Avignon, France, du 16 au 18 novembre 2009

Interrogations de moteurs de recherche par des requêtes formulées en langage naturel

Ludovic Bonnefoy, Romain Deveaud et Eric Charton[1]

1 : LIA / Université d'Avignon, 339 chemin des Meinajaries, 84911 Avignon
Contact : ludovic.bonnefoy@etd.univ-avignon.fr,
romain.deveaud@etd.univ-avignon.fr, eric.charton@univ-avignon.fr

Résumé
L'utilisation de requêtes écrites en langage naturel est un des enjeux important du secteur de la recherche d'information pour les prochaines années. Malgré le fait que ce domaine commence à être couvert par quelques grands noms de l'informatique, les réalisations disponibles à ce jour ne peuvent être considérées que comme des prototypes compte tenu des résultats actuellement visibles. Ceci peut-être expliqué par la grande diversité du langage naturel et par les nombreux sens qui peuvent être donnés à un même mot ou une même expression. Les différents critères sémantiques naturellement présents dans les phrases permettent notamment de combiner la recherche d'information classique - extrayant des documents comparés à des termes, ou mots-clés - avec des « entités » ou des « concepts » pouvant être apparentés à des catégories sémantiques (par exemple : « personne »,« organisation »,« date »...). Nous proposerons dans cet article un système de catégorisation et d'extraction de mots-clés à partir de phrases formulées en langage naturel.

Abstract
Nowadays, requesting a search engine with natural language requests is a significant issue in the information retrieval research field, and some of its biggest actors begin to take it seriously. Some prototypes are actually available, but the error rate, inferred by the huge diversity of natural language and the different semantics of words or expressions, is still too large. Sentences naturally contain semantic criteria such as "entities", "concepts" or "categories" which can be combined with standard information retrieval in order to filter the documents with these semantic categories (e.g. "person", "organisation", "date"...). In this article we propose a categorization and keywords extraction system for natural language sentences.

Mots-clés : Sémantique, catégorisation, langage naturel, recherche d'information.

Keywords: Information retrieval, semantic, categorization, natural language.

1. Introduction

De nos jours, les moteurs de recherche sont un outil pleinement utilisable par les personnes averties et habituées, malgré la volonté d'en améliorer l'accessibilité. En effet, le processus consistant à passer d'une interrogation à un enchainement de mots-clés pertinents retranscrivant correctement la pensée initiale n'est pas un exercice aisé, du moins pour les personnes qui ne sont pas familières avec l'informatique ou internet, parmi lesquelles nous pouvons par exemple compter les enfants ou certaines personnes âgées.
Il serait en effet idéal de pouvoir simplement formuler une question à un moteur de recherche et que celui-ci puisse donner la réponse ou du moins un ensemble de documents dans lesquels une ou des réponses se trouveraient. Des sociétés telles que Google[1], Powerset[2] (propriété de Microsoft) ou Hakia[3] sont actuellement fortement investies dans le développement de solution

[1] http ://www.google.com
[2] http ://www.powerset.com
[3] http ://www.hakia.com

sémantiques à l'interrogation des moteurs de recherche et l'enrichissement communautaire en est un élément central, au moins pour les deux premiers protagonistes. Ils ont en effet recours à de nombreux sites dont les informations sont éditées par des internautes contributeurs, Wikipédia[4] étant le plus célèbre d'entre eux.
C'est également notre cas, puisque le système que nous proposons s'interface NLGbAse[5], une base de données classifiées provenant de Wikipédia qui peut être interrogée par le biais de trois moteurs de recherche différents. Le premier d'entre eux met en $\frac{1}{2}$uvre un algorithme calculant la *similarité cosinus* entre l'ensemble des mots-clés entrés et les documents issus de Wikipédia ; le deuxième est semblable au premier en tous points, à l'exception que l'on peut affiner la recherche en précisant une catégorie, ainsi seuls les documents classifiés comme appartenant à la catégorie spécifiée seront relevés. Le troisième applique quant à lui un algorithme de compacité[6] permettant de trouver une entité précise étant d'une catégorie donnée, proche d'un ou plusieurs mots donnés, dans le document Wikipédia se rapportant à une entité nommée donnée, ce qui permet notamment de pouvoir proposer une réponse factuelle à une requête.
Ces outils constituent le système de recherche d'information sur lequel nous appliquons les sorties de notre propre système ; ce dernier peut, à partir d'une phrase en langage naturel - de préférence une question, donner les différents mots-clés et catégories attendus par les moteurs de recherche de NLGbAse, et ainsi obtenir une liste de résultats - et éventuellement des réponses factuelles - pertinents suite à une interrogation en langage naturel.

2. Analyse morpho-syntaxique et couplage des mots

Pour travailler sur la sémantique, il est indispensable de posséder des outils permettant à la machine de décomposer et d'analyser la structure des phrases. C'est pourquoi nous avons utilisé un analyseur morpho-syntaxique réalisant des couplages de mots selon leur position grammaticale dans la phrase et les liant selon leurs interdépendances [**?**].

3. Catégorisation des phrases

3.1. Catégorisation à base de règles simples

Notre approche pour catégoriser les questions fonctionne avec un ensemble de règles. Cet ensemble est relativement restreint car nous avons pu remarquer qu'une dizaine de règles environ pouvaientt couvrir une majorité des cas, et qu'ensuite chaque petit gain se traduisait par la production d'un nombre croissant exponentiellement de nouvelles règles. Les règles que nous avons formulées se basent principalement sur les pronoms interrogatifs des questions. En voici la liste :

- Who, Whom, Whose : *pers* (Person).
- Where, Whence, Wither : si il s'agit trouver une catégorie pour le deuxième moteur de NLGbAse, et si un nom propre ou un objet est trouvé à la question alors la catégorie sera celle du *nom propre* ou de l'*objet* grammatical de la phrase (par exemple : « Où a étudié Patrick Sébastien ? », il y a peu de chance de trouver dans la fiche de ce lieu une mention de cette personnalité et il est à priori plus judicieux de proposer la fiche de *Patrick Sébastien*, dans laquelle l'utilisateur sera à même de touver l'information). Dans le cas contraire (ou si nous voulons une catégorie pour le troisième moteur), la catégorie *loc* (Location) est attribuée.
- How : nous appliquons la même procédure que précédemment, à l'exception près que si les premières conditions ne sont pas remplies, la catégorie *unk* (Unknown) est attribuée. Pour ce pronom là nous avons ajouté quelques précisions lorsque nous cherchons une catégorie pour le troisème moteur : si le mot suivant directement *how* fait partie de la liste suivante (far, few, great, little, many, much, tall, wide, high, big, old), la catégorie *amount* est attribuée.
- What, Why, Which : le principe est toujours le même, avec *unk* (Unknown) pour valeur par défaut. Nous avons également établi, comme précédemment, une liste de mots pouvant être acceptés comme suivant directement le pronom interrogatif (comme par exemple day : « What day is the Independance Day ? ») et qui vont impliquer automatiquement l'attribution d'une

[4] http ://www.wikipedia.org
[5] http ://www.nlgbase.org
[6] Référence nécessaire

catégorie (*date* dans l'exemple précédent).
Nous allons maintenant détailler les méthodes de catégorisation des noms propres et des noms communs que nous avons mises en place.

3.2. Catégorisation par les noms propres utilisant NLGbAse

Comme nous l'avons vu la majorité des règles ne nous permettent pas de trancher directement, nous devons donc compléter notre analyse par un autre moyen et cela passe notamment par la catégorisation des noms propres. Nous avons remarqué que dans la majorité des cas, si un nom propre est présent dans une quest, il en est l'objet ou du moins l'objet est l'une de ses caractéristiques. Prennons par exemple ces deux questions « What is the date of birth of Bruce Dickinson ? », « Who is Batman's team-mate ? » ; nous voyons bien que les informations désirées sont *forcément* en relation avec nos noms propres.
Nous avons donc prit le parti de prendre comme catégorie la catégorie du nom propre se trouvant dans la question - s'il y en a un. Pour cela nous adressons une requête à un script issu de NLGbAse, qui comme nous l'avons vu associe une catégorie à chaque entité de Wikipédia, qui récupère la catégorie de l'entité correspondant à ce nom propre. Si une entité porte exactement le même nom alors la catégorie sera celle de cette entité ; si ce n'est pas le cas mais que des entités ont un nom similaire, alors la catégorie sera celle de la plus pertinente d'entres elles. Enfin si ce n'est pas le cas nous effectuons une recherche par *TF.Idf* en prenant la catégorie qui a le plus fort score, pour cela chaque catégorie se voit attribué comme score la somme des scores des documents ayant cette catégorie. De ce fait la catégorie qui rassemble le plus de pertinence sera sélectionnée.
Cependant nous sommes conscients qu'il arrive parfois que cette stratégie ne soit pas idéale, comme pour : « What is the name of Batman's car ? ». Il y a des chances que cette méthode donne *pers* (Person) comme catégorie attendue alors que la solution idéale aurait probablement été *prod* (Product). Cependant elle n'est pas totalement inapproprié car nous devrions trouver l'information désirée dans la fiche Wikipédia de Batman, néanmoins l'accès à la réponse est moins direct.

3.3. Catégorisation par les noms communs utilisant WordNet

Cependant toutes les questions ne comportent évidemment pas de noms propres mais généralement des noms communs, c'est pourquoi nous avons du trouver un moyen de traiter ces questions par une approche assez simple. La décomposition morpho-syntaxique nous permet généralement de trouver l'objet de la question, étant donné qu'il est généralement fortement porteur de sens dans une phrase, c'est celui-ci que nous allons étudier. En effet, pour la question « What are the generals ? », l'objet de la question est « generals » ; l'enjeu est d'arriver à associer « generals » à l'étiquette *fonc.mil* (fonction militaire).
Pour arriver à cela nous utilisons WordNet et ses hyperonymes ainsi que sa capacité à fournir une classe pour chaque mot : en effet WordNet associe déjà à la totalité des termes une étiquette, par exemple à « general » WordNet associe *noun.person*. L'étiquette que fournit WordNet est bien souvent satisfaisante, cependant son jeu d'étiquette ne correspond pas aux exigences d'Ester auquel notre projet doit se plier.
La première étape pour arranger ça fut d'associer « à la main » des étiquettes à des mots qui prendront le dessus sur celles de WordNet ; pour reprendre notre exemple, nous ne voulons pas l'étiquette *pers* (Person) pour « general » mais bien *fonc.mil*. Cependant faire ce travail sur tous les mots demanderait un investissement titanesque, et nous avons pu palier à ce problème en réfléchissant aux mots les plus généraux possibles pour chaque catégorie dont nous avions besoin. Ensuite nous vérifions que les hyponymes de ces mots sur WordNet correspondent bien à la même catégorie, et si ce n'était pas le cas nous sélectionnions tous les hyponymes pour lesquels c'est le cas et répétions cette opération. Nous sommes donc arrivés à une liste de mots caractérisant parfaitement chaque catégorie - et étant compatibles avec WordNet.
L'algorithme de catégorisation en lui-même consiste en une fonction récursive qui va vérifier si le nom commun ne fait pas partie des mots étiquetés. Si ce n'est pas le cas cette vérification est faite pour son hyperonyme, et ainsi de suite. La récursivité s'arrête si un mot associé à une étiquette est trouvé ou si on arrive sur l'hyperonyme de plus haut niveau. Dans le premier cas le mot de départ se voit associé cette étiquette et donc la catégorie recherchée aussi, dans le deuxième cas c'est l'étiquette associée au mot de départ qui prévaut et qui est donc définie comme la catégorie recherchée.

Cette méthode utilisée seule ne peut bien évidemment pas couvrir tous les cas, néanmoins c'est l'association des différentes - mais surtout complémentaires - méthodes de classification qui permet d'obtenir des résultats satisfaisants.

4. Extraction des mots-clés

Comme nous l'avons expliqué, les différents moteurs de recherche de NLGbAse n'attendent pas les mêmes entrées, nous allons donc détailler ici les deux types d'extraction de mots-clés - ou mots pertinents.

4.1. Moteurs de recherche d'information par similarité cosinus

Dans un premier temps, les mots-outils de la phrase sont automatiquement supprimés à l'aide d'un anti-dictionnaire. Nous utilisons ensuite l'analyse morpho-syntaxique de la phrase, et notamment l'arbre constitutif, pour récupérer les mots - qui ne sont pas des mots-outils - qui font partie des groupes nominaux ; malgrès son aspect simple, voire simpliste, nous avons pu prouver empyriquement son efficacité.

4.2. Moteur de recherche d'information par algorithme de compacité (question-réponse)

Nous l'avons déjà précisé plus haut, ce troisième moteur accepte plusieurs entrées différentes, et notamment deux champs de mots-clés. Le premier champ est une entité nommée qui va déterminer dans quel document sera appliqué l'algorithme de compacité, tandis que le deuxième champ consiste en une liste de mots qui représentent l'information cherchée ; par exemple pour la question « When was Albert Einstein born ? », le mot « born » devrait être sélectionné car l'information recherchée - une date de naissance en l'occurence - se trouvera certainement très proche de ce mot.

Dans un premier temps nous devons donc trouver l'entité nommée ; si un nom propre est présent dans la question, c'est lui qui sera directement utilisé comme entité nommée. Dans le cas contraire, nous éxécutons une requête sur NLGbAse avec l'objet de la question afin de récupérer le nom de l'entité nommée la plus pertinente.

Dans un second temps vient l'extraction des mots porteurs de sens ; il s'agit tout d'abord de supprimer tout les mots-outils, les noms propres et le verbe présents dans la phrase. Il s'agit ensuite de récupérer les synonymes des mots restant avec WordNet ; pour reprendre notre exemple précédent, nous ne savons pas si le mot »born« sera effectivement employé dans le document dans lequel l'information sera cherchée, c'est pourquoi nous cherchons des dérivations afin de les rajouter à notre liste et ainsi améliorer nos chances de trouver l'information. Toujours pour notre exemple, cette recherche de synonymes pourrait nous mener au mot »birth« , qui serait en effet intéressant à garder dans l'optique de la recherche d'une date de naissance.

5. Expériences et résultats

Nous avons procédé à un certain nombre de tests et d'expériences afin d'évaluer les performances de la catégorisation comme de l'extraction des mots-clés. Nous n'avons retenu que les catégories *pers*, *org*, *loc*, *date*, *amount* et *unk*, les standards d'Ester 2 n'étant pas encore complètement implémentés sur la version anglaise de NLGbAse. Nous avons récupéré un corpus de questions formulées par des utilisateurs et nous les avons étiquetées en faisant mentalement le même travail que notre système, afin de pouvoir comparer ses résultats aux notres.

5.1. Mesures de la catégorisation

Les résultats de l'attributions de catégories aux 107 phrases du corpus étiqueté sont présentés dans le tableau 1. Nous avons calculé pour chacune d'elles la précision et le rappel, puis le F-Score[7] en découlant.

Si les résultats ne semblent pas très satisfaisants, ils peuvent malgré tout mettre en perspective le fait que des règles simples couplées à une catégorisation par recherche dans une base de données sémantique peuvent être viables. Les résultats de la catégorie *org* peuvent être expliqués par le fait

[7] Mesure harmonique combinant la précision et le rappel

que nous n'avons pas pu définir de règle spécifique à cette catégorie sans impacter les résultats des autres catégories, notamment *pers* et *loc*.

Catégorie	($\bar{p}$)	($\bar{r}$)	($\bar{F}$-s)
Pers	0.74	0.86	**0.80**
Org	0.17	0.07	**0.10**
Loc	0.60	0.53	**0.56**
Date	0.89	0.89	**0.89**
Amount	1	0.58	**0.73**
Unk	0.48	0.59	**0.53**
Total			**0.60**

TAB. 1 – Précision ($\bar{p}$), Rappel ($\bar{r}$), F-Score ($\bar{F}$-s) obtenus sur le corpus de test

5.2. Mesures de l'extraction de mots-clés

6. Conclusion