

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

Ludovic Bonnefoy¹, Romain Deveaud¹ et Eric Charton²

1 : Centre d'Enseignement et de Recherche en Informatique / Université d'Avignon et des Pays de Vaucluse ludovic.bonnefoy, romain.deveaud {@etd.univ-avignon.fr}

2 : LIA / Université d'Avignon et des Pays de Vaucluse, 339 chemin des Meinajaries, 84911 Avignon eric.charton@univ-avignon.fr

Résumé

La difficulté de la tâche d'interprétation de requêtes en langue naturelle réside dans la transformation d'une phrase grammaticale en une requête pertinente pour interroger un système de RI ou d'extraction d'information. Dans cet article, nous présentons un système d'extraction d'information associé à un analyseur de requêtes, dédié à l'analyse et à l'interprétation de questions factuelles formulées en langue naturelle issues de la campagne TREC. Notre système fournit des réponses en exploitant un système composé d'une ressource ontologique et sémantique issue d'un corpus encyclopédique, dont on aura extrait les réponses candidates à l'aide de méthodes statistiques.

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

L'interprétation de requêtes écrites en langage naturel par un système de recherche d'information (RI) ou de dialogue est un enjeu important de l'ingénierie documentaire. Outre le fait que la requête en langue naturelle rend possible la formulation d'une requête de recherche selon une forme bien plus précise que la simple saisie de mots clés, elle permet, dans le cadre de la compréhension des langages parlés et des systèmes de dialogue, de déployer des composants de compréhension (SLU) conviviaux pour les utilisateurs.

Dans le cadre des systèmes de Question et Réponses (SQR), l'analyse d'une requête en langue naturelle doit permettre de transformer en processus d'extraction d'information, une question factuelle ou complexe. La difficulté de la tâche d'interprétation de requêtes en langue naturelle réside dans la transformation d'une phrase grammaticale en une requête pertinente pour interroger un système de RI ou d'extraction d'information.

Dans cet article, nous présentons le système *NLGbAse QR*. Ce système est un extracteur d'information contenues dans une ontologie, associé à un analyseur de requêtes formulées en langue naturelle. Ce système est testé à l'analyse et à l'interprétation de questions factuelles formulées en langue naturelle issues de la campagne TREC.

La structure de cet article est la suivante :

Nous décrivons dans notre introduction les grands principes de la littérature sur l'analyse de requêtes en langue naturelle, et son déploiement dans le cadre de systèmes de Question Réponses (SQR). Puis nous présentons les principaux systèmes de QR existants, et les méthodes de transformation de requêtes sémantiques et d'interrogation qu'ils mettent en oeuvre. Nous expliquons également dans cette section la relation existant entre ontologies issues de corpus encyclopédiques, et le rôle que ces ontologies peuvent jouer dans des SQR.

Dans la section suivante, nous présentons l'architecture du SQR *NLGbAse.QR*. Ce dernier est divisé en trois parties : analyse des requêtes en langue naturelle, transformation des requêtes en paramètres d'extraction, extracteur d'informations appliqué sur une ontologie. Puis nous détaillons dans la section 3 ces trois modules et les algorithmes qu'ils sous-tendent.

Dans la section 4, nous présentons une évaluation de notre système avec le corpus de question de TREC. Nous concluons par une analyse critique des résultats obtenus, qui sont du niveau de l'état de l'art, et introduisons nos perspectives futures de recherche.

2. Principe de l'analyse de requêtes en langue naturelle et des SQR

Des sociétés telles que Google¹, Powerset² (propriété de Microsoft) ou Hakia³ sont actuellement fortement investies dans le développement de solutions 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, Wikipédia⁴ étant le plus célèbre d'entre eux.

C'est également notre cas, puisque le système que nous proposons s'interface avec NLGbAse⁵, 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 œuvre 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 compacité⁶ permettant de trouver une entité précise appartenant à 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.

2.1. Encyclopédies, ontologies et SQR

3. Algorithmes déployés

3.1. 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 [1].

3.2. Catégorisation des phrases

Par définition, la sémantique crée naturellement différentes classes de sens plus ou moins générales, au niveau du mot comme de la phrase. Nous adhérons à l'idée selon laquelle la sémantique

¹ <http://www.google.com>

² <http://www.powerset.com>

³ <http://www.hakia.com>

⁴ <http://www.wikipedia.org>

⁵ <http://www.nlgbase.org>

⁶ Référence nécessaire

globale d'une phrase est déterminée par un nombre réduit de mots appartenant à une catégorie grammaticale précise, et c'est selon ce point de vue que nous allons présenter la catégorisation de phrases en catégories sémantiques larges que nous avons implémentée.

3.2.1. 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 pouvaient 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 de trouver une catégorie pour le deuxième moteur de NLGbAse, et si un nom propre ou un objet est trouvé, la catégorie sera celle du *nom propre* ou de l'*objet* grammatical de la phrase (par exemple : « Where did Patrick Sébastien study ? », 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 trouver 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 troisième 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 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.

3.2.2. 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 question, il en est l'objet ou du moins l'objet est l'une de ses caractéristiques. Prenons 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'entre 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 attribuée 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). Elle n'est néanmoins pas totalement inappropriée car nous devrions trouver l'information désirée dans la fiche Wikipédia de Batman, néanmoins l'accès à la réponse est moins direct.

3.2.3. Catégorisation par les noms communs utilisant WordNet

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, qui possède généralement une forte valeur sémantique ; c'est donc 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 WordNet⁷ 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 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 avons besoin. Ensuite nous vérifions que les hyponymes de ces mots sur WordNet correspondaient 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 répétons 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 par WordNet 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 catégorisation qui permet d'obtenir des résultats satisfaisants.

3.3. 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.

3.3.1. Extraction destinées aux 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 constituent groupes nominaux ; malgré son aspect simple, voire simpliste, nous avons pu prouver empiriquement son efficacité.

3.3.2. Extraction destinées au 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'occurrence - se trouvera certainement très proche de ce mot.

⁷ <http://wordnet.princeton.edu>

Dans un premier temps nous devons donc trouver l'entité nommée ; si un nom propre est présent dans la question, il sera directement utilisé comme entité nommée. Dans le cas contraire, nous exécutons 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 tous 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.

4. Expériences et résultats

La vocation de notre algorithme est de transformer une question en langue naturelle en une requête compatible avec un système de RI. Pour mesurer les performances de notre système, nous évaluons dans un premier temps sa capacité à étiqueter une question en vue de l'extraction des informations requise pour construire une requête. Dans un second temps nous construisons une requête d'après les informations extraites et mesurons la pertinence des résultats retournés par l'un des trois moteurs de NLGbAse.

4.1. Corpus de référence et standard de mesure

Nous avons utilisé pour ces mesures des corpus de questions reliées à une réponse factuelles ou de type liste. Des corpus de références ont été mis au points dans le cadre de campagnes d'évaluation telles que Trec 8

Le corpus *Question Answering Collections* de Trec est composé de XXX questions factuelles et des réponses standard qui doivent être retournées⁸. Ces paires de questions et réponses sont organisées comme suit. Un fichier XML décrit les question factuelles :

```
<target id="2" text="Fred Durst">
<q><q id="2.1" type="FACTOID">
What is the name of Durst's group?
</q></q>
<q>
<q id="2.2" type="FACTOID">
What record company is he with?
</q>
</q>
[...]
```

Des réponses « standard » sont livrées sous forme de liste :

```
2.5 1 vital tattoo artist
2.5 2 vital operated skate park
```

Ce corpus n'étant pas initialement prévu pour vérifier l'expérience d'étiquetage de questions tel que prévu dans le cadre de notre algorithme, nous avons procédé à son enrichissement.

Cet enrichissement consiste à attribuer à chaque question la catégorie sémantique de la réponse attendue (nous n'avons retenu que les classes racines à savoir *pers*, *org*, *loc*, *date*, *amount* et *unk*), ainsi que les différents mots-clés qui peuvent permettre d'obtenir une réponse. Nous disposons ainsi d'un corpus de questions et de réponses de références complétées par des informations sémantiques compatibles avec notre système⁹.

⁸ En téléchargement sur <http://trecnistgovdataqat2004.qadata.html>

⁹ Ce corpus complété est disponible sur www.nlgbase.org

Ainsi, dans l'exemple précité, nous complétons la description de la question par plusieurs étiquettes de classes, à savoir que l'entité source est une personne, et que la réponse cible est un nom d'organisation (celui du groupe musical), et d'éditeur de disque :

```
<target id="2" text="Fred Durst">
<k>PERS</k>
<qa><q id="2.1" type="FACTOID">
What is the name of Durst's group?
</q><kt>ORG</kt>
</qa>
<qa>
<q id="2.2" type="FACTOID">
What record company is he with?
</q><kt>ORG</kt>
</qa>
[...]
```

Le corpus est également composé de 108 questions prélevées aléatoirement sur le site Answer-Bus¹⁰, qui est également un système de recherche d'information par question-réponse. La particularité de ces questions est qu'elles ont été formulées par des utilisateurs, une majorité d'entre elles n'est donc pas grammaticalement parfaite ; nous avons trouvé ce corpus intéressant pour tester notre système car cela nous permet de nous rapprocher de la réalité par rapport aux questions classiques de campagnes d'évaluation.

Nous avons lancé notre système et mesuré pour chaque question la validité de son étiquetage sémantique et de l'identification des mots clés pertinents pour une extraction de réponse, puis la pertinence des réponses retournées par le moteur de recherche par rapport à la réponse de référence. Ces mesures sont réalisées par un calcul de score de précision et le rappel, complété par le F-Score¹¹.

4.2. Mesures de la catégorisation sémantique

Les résultats de l'attribution de catégories aux XXX phrases du corpus étiqueté sont présentés dans le tableau 1.

Ces résultats mettent 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 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	0.81	0.81	0.81
Org	0.64	0.61	0.63
Loc	0.76	0.77	0.76
Date	0.91	0.98	0.95
Amount	0.99	0.92	0.92
Unk	0.69	0.64	0.66
Total	0.80	0.78	0.79

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

¹⁰ <http://www.answerbus.com/>

¹¹ Mesure harmonique combinant la précision et le rappel

4.3. Mesures de l'extraction de mots-clés

Nous avons également mesuré la pertinence des mots-clés extraits par notre système. Comme précédemment nous avons tout d'abord effectué « à la main » le travail d'extraction, puis nous avons comparé ceci avec les résultats du système afin d'obtenir un pourcentage de satisfaction calculé selon la formule suivante :

$$S(C) = \frac{\sum_{i=1}^N P(Q_i)}{N}$$

Ils sont présentés dans les tableaux ci-dessous.

Type de mots-clés	(S(C))
Mots-clés extraits pour une recherche par similarité cosinus	54.03%
Entités nommées extraites pour une recherche de type question-réponse (compacité)	57.76%
Mots-clés extraits pour une recherche de type question-réponse (compacité)	73.28%

TAB. 2 – Satisfaction (S(C)) obtenue sur le corpus de test

La grande majorité des erreurs dans la détection de mots-clés est due à une formulation grammaticale approximative de la question, ce qui induit en erreur l'analyseur morpho-syntaxique qui ne va pas correctement étiqueter certains mots.

5. Conclusion

Nous avons présenté un système analysant et décomposant des requêtes formulées en langage naturel dans le but d'en extraire leur catégorie sémantique ainsi que les mots porteurs de sens et d'information, destiné à être couplé avec un système de recherche d'information que nous avons présenté.

Bibliographie

1. Daniel Sleator et Davy Temperley. Parsing english with a link grammar. 1993.