Interrogations en langue naturelle

Projet M1

Ludovic Bonnefoy Romain Deveaud

Tutoré par Marc El-Bèze et encadré par Eric Charton

Jeudi 18 juin 2009

Introduction

- Trululu

Moteurs de recherche intégrant la sémantique

- Google, Powerset, Hakia...

Moteurs de recherche intégrant la sémantique

- Google, Powerset, Hakia...
- Algorithmes ayant recours à des sources extérieures.

Moteurs de recherche intégrant la sémantique

- Google, Powerset, Hakia...
- Algorithmes ayant recours à des sources extérieures.
- Enrichissement et activité communautaire indispensables pour la validité et la récence des informations.

Moteurs de recherche intégrant la sémantique

- Google, Powerset, Hakia...
- Algorithmes ayant recours à des sources extérieures.
- Enrichissement et activité communautaire indispensables pour la validité et la récence des informations.
- NLGbAse : base de données classifiée (ontologie) issue de Wikipédia.

Présentation de NLGbAse

- Trois outils de recherche d'informations.

Présentation de NLGbAse

- Trois outils de recherche d'informations.
- Un moteur "classique", prennant en entrée des mots-clés utilisant la similarité cosine.

Présentation de NLGbAse

- Trois outils de recherche d'informations.
- Un moteur "classique", prennant en entrée des mots-clés utilisant la similarité cosine.
- Un moteur "sémantique", reprennant le même algorithme que le précédent, mais permettant de sélectionner les résultats appartenant à une catégorie sémantique précise.

Présentation de NLGbAse

- Trois outils de recherche d'informations.
- Un moteur "classique", prennant en entrée des mots-clés utilisant la similarité cosine.
- Un moteur "sémantique", reprennant le même algorithme que le précédent, mais permettant de sélectionner les résultats appartenant à une catégorie sémantique précise.
- Un moteur "extracteur d'informations", basé sur un algorithme de compacité, permettant d'obtenir une information précise éventuellement contenu dans un document.

Les règles

- Application de règles sur les pronoms interrogatifs.
  - "Who","Whom","Whose" =¿ pers
  - "How long","How much", "How many" =¿ amount
  - "What","Why",... =¿ ?

Catégorisation des noms propres

- Si insuffisant : extraction du nom propre de la question.
  - on le catégorise avec NLGbAse
  - si échec on vérifie l'orthographe sur google.com
  - si modification on interroge de nouveau NLGbAse
  - enfin si NLGbAse n'a rien retourné on interroge CCG

Catégorisation de l'objet de la question via Wordnet

- Si la phrase ne contient pas de noms propres ou que l'étape précédente n'a rien donné :
- On récupère l'objet de la question et on va essayer de le catégoriser en ayant recours à Wordnet
- Wordnet associe une catégorie à la majorité des mots, cependant elles ne correspondent pas à Ester
- Nous utilisons donc une liste de correspondances : mot -¿ catégorie
  - Mots de plus au niveau dans les arbres d'hyperonymes qui n'ont pour hyponymes que des mots de même classe.
- L'algorithme est le suivant :

Catégorisation de l'objet de la question via Wordnet(2)

- Le mot est-il dans la liste?
- Si oui Fin
- Sinon on réessaye avec son hypéronyme.
- Tant qu'un hypéronyme n'est pas dans la liste ou que l'on a pas atteint le concept de plus haut niveau on réitère.
- Finalement si aucune classe n'est trouvée, on prend celle que Wordnet propose.
- Si toutes les stratégies ont échoués on prend la classe unk.

Améliorations

- Accepter plusieurs classes
- Ajouter des options de ri classique
  - Permettre d'élargir la requete (mots-clés et catégories)
  - Opérateurs logiques

Conclusion

- Trululu