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

- Les machines ne comprennent pas le langage naturel.

Introduction

- Les machines ne comprennent pas le langage naturel.
- Développement d'un système permettant d'interroger une ressource ontologique et sémantique avec des vraies questions.

Introduction

- Les machines ne comprennent pas le langage naturel.
- Développement d'un système permettant d'interroger une ressource ontologique et sémantique avec des vraies questions.
- Combinaison avec un extracteur d'informations afin d'afficher des résultats.

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 cosinus.

Présentation de NLGbAse

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

Utilisation de règles

- Analyse morpho-syntaxique de la question pour la décomposer en concepts grammaticaux compréhensibles par la machine.

Utilisation de règles

- Analyse morpho-syntaxique de la question pour la décomposer en concepts grammaticaux compréhensibles par la machine.
- Règles appliquées sur les pronoms interrogatifs (Who, Whom, Whose, How, What, Which...).

Utilisation de règles

- Analyse morpho-syntaxique de la question pour la décomposer en concepts grammaticaux compréhensibles par la machine.
- Règles appliquées sur les pronoms interrogatifs (Who, Whom, Whose, How, What, Which...).
- Règles appliquées sur les mots suivant ces pronoms (How many, What day...).

Catégorisation par les noms propres

- Détection d'un nom propre contenu dans la question

Catégorisation par les noms propres

- Détection d'un nom propre contenu dans la question
- Utilisation de NLGbAse pour récupérer la catégorie qui lui est associée (ex : Valentino Rossi $=>$ pers)

Catégorisation par les noms propres

- Détection d'un nom propre contenu dans la question
- Utilisation de NLGbAse pour récupérer la catégorie qui lui est associée (ex : Valentino Rossi $=>$ pers)
- Vérification de l'orthographe de l'entité nommée à l'aide de Google

Catégorisation par les noms propres

- Détection d'un nom propre contenu dans la question
- Utilisation de NLGbAse pour récupérer la catégorie qui lui est associée (ex : Valentino Rossi $=>$ pers)
- Vérification de l'orthographe de l'entité nommée à l'aide de Google
- Utilisation du module Named Entity Recognition de CCG (Cognitive Computation Group, University of Illinois)

Catégorisation utilisant Wordnet

- Wordnet : base de données lexicale, classifiant et mettant en relation le contenu sémantique et lexical de la langue anglaise.

Catégorisation utilisant Wordnet

- Wordnet : base de données lexicale, classifiant et mettant en relation le contenu sémantique et lexical de la langue anglaise.
- Catégorisation de l'objet de la phrase (mot fortement porteur de sens).

Catégorisation utilisant Wordnet

- Wordnet : base de données lexicale, classifiant et mettant en relation le contenu sémantique et lexical de la langue anglaise.
- Catégorisation de l'objet de la phrase (mot fortement porteur de sens).
- Algorithme récursif parcourant les hyperonymes de l'objet jusqu'à trouver un mot appartenant à une liste associative (mot $=>$ catégorie).

Extraction destinée aux moteurs utilisant la *similarité cosinus*

- Extraction automatique des entités nommées présentes dans la question.

Extraction destinée aux moteurs utilisant la *similarité cosinus*

- Extraction automatique des entités nommées présentes dans la question.
- Sinon l'analyse morpho-syntaxique nous permet de détecter les mots « grammaticalement importants »(groupes nominaux...).

Extraction destinée aux moteurs utilisant la *similarité cosinus*

- Extraction automatique des entités nommées présentes dans la question.
- Sinon l'analyse morpho-syntaxique nous permet de détecter les mots « grammaticalement importants »(groupes nominaux...).
- Utilisation d'un anti-dictionnaire pour éliminer les mots non-porteurs de sens.

Extraction destinée au moteur utilisant la *compacité*

- Deux champs doivent être remplis : l'entité nommée en rapport avec la question et une liste de mots représentant l'information recherchée.

Extraction destinée au moteur utilisant la *compacité*

- Deux champs doivent être remplis : l'entité nommée en rapport avec la question et une liste de mots représentant l'information recherchée.
- Si la question ne contient pas d'entité nommée, une requête est exécutée sur NLGbAse avec l'objet de la question pour récupérer l'entité nommée la plus pertinente.

Extraction destinée au moteur utilisant la *compacité*

- Deux champs doivent être remplis : l'entité nommée en rapport avec la question et une liste de mots représentant l'information recherchée.
- Si la question ne contient pas d'entité nommée, une requête est exécutée sur NLGbAse avec l'objet de la question pour récupérer l'entité nommée la plus pertinente.
- Récupération des mots porteurs de sens près desquels l'information cherchée devrait être trouvée.

Extraction destinée au moteur utilisant la *compacité*

- Deux champs doivent être remplis : l'entité nommée en rapport avec la question et une liste de mots représentant l'information recherchée.
- Si la question ne contient pas d'entité nommée, une requête est exécutée sur NLGbAse avec l'objet de la question pour récupérer l'entité nommée la plus pertinente.
- Récupération des mots porteurs de sens près desquels l'information cherchée devrait être trouvée.
- Recherche de synonymes aux mots porteurs de sens afin d'élargir les possiblités.

Mesures de la catégorisation sémantique

- Nécessité de se comparer à l'état de l'art pour évaluer les performances du système.

Mesures de la catégorisation sémantique

- Nécessité de se comparer à l'état de l'art pour évaluer les performances du système.
- Utilisation d'un corpus de traitement automatique de la langue naturelle : Question-Answer de TREC12 (500 questions formulées en langage naturel).

Mesures de la catégorisation sémantique

- Nécessité de se comparer à l'état de l'art pour évaluer les performances du système.
- Utilisation d'un corpus de traitement automatique de la langue naturelle : Question-Answer de TREC12 (500 questions formulées en langage naturel).
- Création d'un nouveau formalisme d'étiquetage pour pouvoir comparer QA-TREC12 avec les sorties de notre système.

Mesures de la catégorisation sémantique

- Nécessité de se comparer à l'état de l'art pour évaluer les performances du système.
- Utilisation d'un corpus de traitement automatique de la langue naturelle : Question-Answer de TREC12 (500 questions formulées en langage naturel).
- Création d'un nouveau formalisme d'étiquetage pour pouvoir comparer QA-TREC12 avec les sorties de notre système.
- Etiquetage des questions « à la main ».

Mesures de la catégorisation sémantique

- Nécessité de se comparer à l'état de l'art pour évaluer les performances du système.
- Utilisation d'un corpus de traitement automatique de la langue naturelle : Question-Answer de TREC12 (500 questions formulées en langage naturel).
- Création d'un nouveau formalisme d'étiquetage pour pouvoir comparer QA-TREC12 avec les sorties de notre système.
- Etiquetage des questions « à la main ».
- `How big is Mars?#Mars#loc#Mars#big#amount#`

Résultats de la catégorisation

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.: Précision ($\bar{p}$), Rappel ($\bar{r}$), F-Score ($\bar{F}$-s) obtenus sur le corpus QA de TREC 12

Mesures de l'extraction des mots-clés

- Difficultés pour extraire « à la main »les mots-clés pertinents d'une question.

Mesures de l'extraction des mots-clés

- Difficultés pour extraire « à la main »les mots-clés pertinents d'une question.
- Résultats à relativiser.

Résultats de l'extraction des mots-clés

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)	**66.23%**
Mots-clés extraits pour une recherche de type question-réponse (compacité)	**73.28%**

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

Apports du projet

- Acquisition de connaissances en TALN (analyse morpho-syntaxique, ontologies, extraction d'entités nommées, analyse sémantique, hyper et hyponimie. . .).

Apports du projet

- Acquisition de connaissances en TALN (analyse morpho-syntaxique, ontologies, extraction d'entités nommées, analyse sémantique, hyper et hyponimie. . .).
- Rédaction d'un article sur notre système pour la convention des jeunes chercheurs Majecstic.

Apports du projet

- Acquisition de connaissances en TALN (analyse morpho-syntaxique, ontologies, extraction d'entités nommées, analyse sémantique, hyper et hyponimie. . .).
- Rédaction d'un article sur notre système pour la convention des jeunes chercheurs Majecstic.
- Découverte de nombreux outils (Wordnet, LinkParser, Xip. . .).

Evolutions possibles

- Evaluation de l'ensemble des catégories.

Evolutions possibles

- Evaluation de l'ensemble des catégories.
- Certaines options de RI manquantes : relâchement des contraintes, opérateurs logiques.

Evolutions possibles

- Evaluation de l'ensemble des catégories.
- Certaines options de RI manquantes : relâchement des contraintes, opérateurs logiques.
- Pouvoir sélectionner plusieurs catégories ayant différents poids.

Evolutions possibles

- Evaluation de l'ensemble des catégories.
- Certaines options de RI manquantes : relâchement des contraintes, opérateurs logiques.
- Pouvoir sélectionner plusieurs catégories ayant différents poids.
- Nouvelle approche basée sur de l'apprentissage automatique.

? ?

- Produit fini facilement déployable : script d'installation, utilisation en ligne grâce à un CGI.

? ?

- Produit fini facilement déployable : script d'installation, utilisation en ligne grâce à un CGI.
- Résultats proches de ceux de l'état de l'art.

? ?

- Produit fini facilement déployable : script d'installation, utilisation en ligne grâce à un CGI.
- Résultats proches de ceux de l'état de l'art.
- Facilement adaptable à différentes langues.

? ?

- Produit fini facilement déployable : script d'installation, utilisation en ligne grâce à un CGI.
- Résultats proches de ceux de l'état de l'art.
- Facilement adaptable à différentes langues.
- Un corpus annoté de 490 questions à disposition libre de la communauté scientifique.

Conclusion

- Système expérimental mais fonctionnel.

Conclusion

- Système expérimental mais fonctionnel.
- Apport d'une solution originale pour l'interrogation de moteurs de recherche en langage naturel.

Conclusion

- Système expérimental mais fonctionnel.
- Apport d'une solution originale pour l'interrogation de moteurs de recherche en langage naturel.
- Projet enrichissant qui nous a fait découvrir des perspectives de recherche intéressantes.