Extraction d'Information et Système de dialogue

EI et Compréhension hors contexte

Sophie Rosset

Groupe Traitement du Langage Parlé
Département Communication Homme-Machine
LIMSI - CNRS

Compréhension hors contexte

Kézako ?

Un système de dialogue regroupe une séquence de traitements :

Un utilisateur (*via* ASR, clavier, ...) produit une suite de mots

$\rightarrow$ à un *analyseur*

qui produit une analyse mais SANS connaître ce qui s'est passé dans le dialogue et le transmet à

$\rightarrow$ l'analyseur contextuel (partie du DM) pour qu'il affine l'analyse en tenant compte de ce qui s'est passé etc.

Compréhension hors contexte

Objectif : extraction de l'information pertinente

- repose sur des *ontologies* ou des *taxonomies*
- est guidée par la tâche (recherche d'information par exemple)
- est guidée par le domaine (informations géographiques, résultats foot...)
- Différentes approches/méthodes : statistiques, linguistiques, mixtes

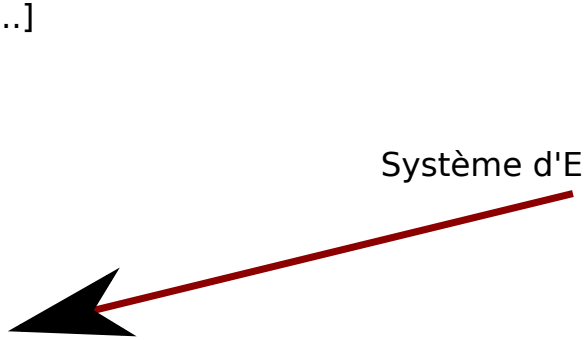
$\rightarrow$ Entités nommées, entités spécifiques, actes de dialogue ...

$\rightarrow$ revient à de l'EI au moins dans la forme

$\rightarrow$ lien fort avec l'EI pour la constitution de la base de données

Compréhension hors contexte

La France est un pays d'Europe de l'Ouest [...] Elle a pour capitale Paris, pour langue officielle le français et pour monnaie l'euro.[...] Peuplée de 66,6 millions d'habitants, [...] Elle est frontalière de la Belgique et du Luxembourg au nord-est[...]



id: France
region: Europe de l'Ouest
capitale: Paris
langue_off: français
monnaie: euro
habitants: 66,6 millions
pays_frontalier:
- pays: Belgique
  localisation: nord-est
- pays: Luxembourg
  localisation: nord-est

Système d'EI pour constituer la base de connaissances

+ détection de focus et type de question

Quelle est la capitale de la France
Quel pays partage une frontière au nord-est

Compréhension hors contexte

- Entités Nommées : dates, lieux, événements, personne (prénom, nom), etc.
- Mais encore ?
  - marqueurs de questions
  - focus
  - actes de dialogue

→ quelle équipe a joué contre le PSG ?

- **PSG** = *une équipe* qu'il faut repérer
- **quelle équipe** = un *marqueur* et un *focus* qui indiquent qu'on cherche une équipe

→ Non pas le PSG

- **pas le PSG** = *une équipe* qu'il faut repérer avec sa *négation*
- **non** = un *acte de dialogue* de type *rejection*

Compréhension hors contexte

Plusieurs approches possibles (inconvénients et avantages...).

- Grammaires sémantiques qui peuvent prendre différentes formes : (probabilistic) context-free grammar (P/CFG), etc. et peuvent être analysées avec des algos comme CKY... (souvent dirigées par la tâche et le domaine)
- Des analyseurs stochastiques (HMM, SVM, CRF, MaxEnt)
- des approches/méthodes/systèmes mixtes : ce qui finalement donne les meilleurs résultats au moindre coût

Les grammaires sont coûteuses à écrire, gèrent assez mal l'ambiguïté. Les méthodes stats sont coûteuses en terme de taille de corpus nécessaire pour l'apprentissage.

Compréhension hors contexte

Comment allez-vous faire ?

- Un moteur d'analyse : **Dark** (/home/tp-home001/tlaverg/eisd)
- Des grammaires : à vous de les écrire
- Un langage de programmation : Lua