Introduction au traitement automatique du langage naturel

Benoit Crabbé

2011-2012

Administration

- 12 $\times$ 2h
- bcrabbe@linguist.jussieu.fr
- Support : langage python et librairie NLTK.

Compréhension du langage naturel

- Le vieux but du TAL est de produire des systèmes de compréhension du langage naturel.
- Qu'est-ce que ça veut dire ?
- Répondre à des questions posées à des moteurs de recherche :
  - Quels sites touristiques puis-je visiter entre Athènes et Delphes avec un budget limité ?
  - Quels médicaments pour la gorge **ne** donnent **pas** mal à l'estomac ?

Challenges

Demande de mener une série de tâches de TAL : Extraction d'information (dans la base de données, page web, etc), inférence, résumé de texte, etc. **Actuellement hors de portée de la technologie sur un domaine général**

Désambiguisation lexicale

- Première tâche : désambiguisation des mots (WSD)

  Il a **servi** *le* **plat** *à ses invités*

- Un simple lookup dans un dictionnaire ne nous donne pas la solution, car on y trouvera par exemple (à supposer qu'on réduise les formes observées à leurs **lemmes** :
  - **servir :**
    1. être au service de qqun
    2. mettre la balle en jeu au tennis
    3. apporter à manger...
  - **plat:**
    1. surface plate, (comme adjectif)
    2. récipient destiné à contenir des aliments (comme nom)
- Souvent la désambiguisation des mots demande d'explorer le contexte dans lequel les mots apparaissent pour choisir le bon sens.
- Parfois plus difficile : identifier de quoi on parle !

Désambiguisation lexicale

- Considérons la préposition *dans* :
  1. Les enfants sont perdus **dans** les montagnes
  2. Les invités viendront **dans** l'après-midi
  3. Pierre est **dans** les choux
- On a trois usages : (1) locatif (2) temporel (3) figé.
- $\Rightarrow$ La tâche de désambiguisation lexicale cherche à déterminer pour un mot en contexte le sens approprié.

Qui a fait quoi à qui ?

- Plus difficile : **identifier les relations** entre les entités dans la phrase ou dans le texte :
  1. Les voleurs ont volé les tableaux. **Ils** ont été vendus plus tard. (il = les tableaux ? les voleurs vendus par leurs complices ?)
  2. Les voleurs ont volé les tableaux. **Ils** ont été pris plus tard. (il = les voleurs ? les tableaux ?)

Résolution d'anaphores

La résolution d'anaphore consiste à trouver les antécédents des pronoms.

Identification de relations thématiques

Dans la première phrase, ce sont les voleurs qui agissent (AGENTS) et les tableaux qui subissent (PATIENTS)
Dans la seconde phrase le sujet subit l'événement (PATIENT) et l'agent reste non exprimé.

Génération et dialogue

- Résoudre les deux problèmes précédents permet d'envisager de faire des applications de dialogue ou de génération :
  - Humain : Qui a pris les tableaux ?
  - Robot : les voleurs
- Ou encore:
  - Humain : Qu'est-ce qui a été volé ?
  - Robot : Les tableaux
- Notons au passage que pour faire de la traduction automatique, il faut pouvoir résoudre ce genre de problèmes :

  The thieves stole the paintings. They were subsequently found.

  Ici on a deux traductions possibles selon l'antécédent choisi:
  - Les voleurs ont volé les peintures. Elles ont été retrouvées plus tard
  - Les voleurs ont volé les peintures. Ils ont été retrouvés plus tard

Exemple de système de MT

Installer python et nltk sur votre système

```
> python
> from nltk.import *
> nltk.download() #choose the nltk book package
> from nltk.book import *
> babelize_shell()
Babel>The thieves stole the paintings. They were subsequently found.
Babel> french
Babel> run
```

Traduit et retraduit le texte de anglais à français sur ses propres sorties, on peut observer que le système fait des erreurs qui s'amplifient avec les itérations

```
0> The thieves stole the paintings. They were subsequently found.
1> Les voleurs ont vole les peintures. Ils ont ete plus tard trouves
2> The robbers stole paintings. They were found later.
3> Les peintures d'etole de voleurs. Elles ont ete trouvees plus tard.
```

Le dialogue et le test de Turing

- Pendant longtemps on a considéré qu'une machine pourrait simuler une activité "intelligente" si elle est capable de passer le **test de Turing** : c'est-à-dire qu'un interlocuteur aveugle ne puisse pas déterminer si il a affaire à un interlocuteur artificiel ou humain.
- Aucune machine n'a passé le test à ce jour.
- Cela marche à peu près pour des systèmes à **domaines** limités (applications pour systèmes embarqués; horaires de trains,réservation cinéma...)
- Cela demande un mécanisme de compréhension du langage, de génération du langage et un système de raisonnement
- Une question posée par les systèmes de dialogue est celle de la **pragmatique** :

> — *Pourriez-vous me dire à quelle heure le train va partir ?*
> — *Oui.*

Tester des hypothèses à partir de textes

- Il peut être tentant de vouloir acquérir/tester des connaissances à partir de textes
- Exemple : déterminer le nombre de livres écrits par un auteur:
- Le textual entailment essaye de déterminer si un text confirme ou pas une hypothèse.
  - Hypothèse : BHL a écrit 53 livres
  - Texte : Bernard Henri Levy est l'éditeur et l'auteur de 53 livres, également de 150 réponses, articles et critiques de livres.
- Pour répondre à la question, un système doit pouvoir tirer un nombre important d'inférences: (1) BHL = Bernard Henri Levy (2) tirer l'inférence qu'un auteur de livre écrit des livres. (3) qu'un éditeur n'a pas nécessairement écrit un livre et (4) qu'en tant qu'auteur **et** éditeur, il n'a pas nécessairement écrit 53 livres.

Première conclusion

- Le problème de la compréhension du langage est un problème qui tombe dans la classe des problèmes de l'intelligence artificielle.
- L'évolution historique du domaine nous dit que l'on préfère travailler avec des systèmes simplifiés plutôt que d'encoder des bases de données de faits représentants l'ensemble des connaissances du monde.
- Nous allons voir dans le cours un certain nombre de techniques approximatives qui apportent des solutions robustes mais parfois incorrectes aux problèmes les plus classiques du TAL.