Traitement automatique des langues

de l'extraction d'information au dialogue

Anne-Laure Ligozat
Sophie Rosset
Pierre Zweigenbaum

LIMSI, CNRS

1/7/2019

Traitement automatique des langues (TAL)

Natural Language Processing (NLP)

Définition

Méthodes informatiques
qui visent à traiter (analyser, générer, transformer)
du matériau langagier
d'une manière qui fait sens pour les humains.

Synonyme [informatique]

Traitement automatique du langage naturel (TALN)

« La terminologie d'une science reflète l'histoire de cette science »
John Humbley, atelier TIA, 1/7/2019

Fonctions du traitement automatique des langues

La langue comme moyen de communication

Dialogue personne-machine

La langue comme objet

Correction automatique,
Traduction automatique, etc.

La langue comme support d'information, de connaissance, etc.

Extraction d'information
Découverte de connaissances, etc

Communication

La langue comme moyen de communication

Recherche d'information [textuelle] (*Information Retrieval*)

Étant donné une requête (mots-clés),
fournir les documents qui contiennent ces mots-clés.

Recherche de réponses à des questions (*Question Answering*)

Étant donné une question formulée en langue naturelle,
produire une réponse (en langue naturelle) à cette question.

Dialogue personne-machine (*Human-Machine Dialogue*)

Tenir une conversation suivie,
dans un but précis ou pas.

Support

La langue comme support d'information, de connaissance, etc.

Extraction d'information (*Information Extraction*)

Analyse ciblée d'un texte
pour y détecter des informations
de type prédéfini.

Fouille de textes (*Text Mining*)

Exploration d'un grand nombre de textes
pour faire émerger de nouvelles connaissances.

Pourquoi le TAL est difficile

Données complexes, naturelles, hétérogènes et multidimensionnelles, de grande dimension, d'une grande variété

- non formel
- ambiguïté, variabilité
- implicite, redondance
- grand nombre d'événements rares

Domaine pluridisciplinaire

- Informatique
- Linguistique
- Statistique
- Sciences cognitives, Intelligence artificielle

Les paliers de la langue

0. La phonétique et la phonologie

Comment les mots et les phrases sont liés aux sons qui les réalisent à l'oral

1. La morphologie

Comment les mots sont construits et quels sont leurs rôles dans la phrase

2. La syntaxe

Comment les mots se combinent pour former des syntagmes, puis des propositions et enfin des phrases correctes

3. La sémantique

Comment les mots font du sens lorsqu'ils sont insérés dans une phrase (indépendamment du contexte)

4. La pragmatique

Comment les phrases peuvent être interprétées selon leur contexte d'énonciation (interlocuteurs, phrases précédentes, connaissance commune du monde...)

Le TAL a besoin de connaissances (*ressources*)

Cela détermine deux grandes classes de méthodes

Méthodes à connaissances humaines

Lexiques, grammaires, patrons, règles,
ontologies, bases de connaissances, etc.,
créés par des humains

Méthodes fondées sur les données

Apprentissage automatique des connaissances nécessaires
à partir de données :

Corpus de textes etc., annotés ou pas

Si données annotées, connaissances humaines !

On trouve davantage de ressources pour l'anglais que pour le français