Introduction au TAL

Pascal Amsili

Novembre 2020

Plan

Sorbonne Nouvelle

Domaine du TAL

La linguistique informatique et le traitement automatique des langues (TAL) désignent l'**application de programmes et techniques informatiques à tous les aspects du langage humain** depuis la reconnaissance de la parole jusqu'à l'analyse sémantique du contenu d'un texte.

- Interaction avec les humains (parlant/écrivant/signant) (dans les deux directions)
- Manipulation de la « matière linguistique »

Sorbonne Nouvelle

Positionnement du domaine

- Domaine technologique
  organisé autour d'un certain nombre d'applications, dont les plus connues sont la traduction automatique (historiquement la première application, dès les années 1950), la correction orthographique, la recherche d'information, le résumé de texte, la génération de texte, la synthèse de la parole, la reconnaissance vocale.
- Domaine de recherche
  en relation avec la linguistique, l'informatique, l'intelligence artificielle, les sciences cognitives, l'ingénierie.

Disciplines connexes

Linguistique computationnelle *Etudes des propriétés calculatoires de la langue*

Linguistique outillée *Recherche linguistique utilisant les ressources et programmes de TAL pour ses propres objectifs*

Linguistique formelle *Description mathématisée des langues et de la langue*

Sciences cognitives *Etude fonctionnelle et biologique de la fonction langage*
(cf. neuro-sciences computationnelles)

Informatique fondamentale *Théorie des langages formels, compilation, complexité*

Sorbonne Nouvelle

Plan

Sorbonne Nouvelle

Applications

Les applications qui utilisent du TAL (mais pas seulement)

- Assistants personnels
- Commande vocale
- Moteurs de recherche
- Outils de gestion documentaire
- ...

Les applications de TAL pur

- Traduction automatique
- Résumé automatique
- Q&A
- Synthèse vocale (TTS)
- Reco vocale/de la parole (AST – *Automatic Speech Recognition*)

Sorbonne Nouvelle

Tâches

Moteur de la recherche en ingénierie : tâches normalisées

- définition précise (entrées/sorties) souvent simplifiée
- données de référence et mesures d'évaluation des performances
- communauté de chercheurs et supports de publication

Quelques exemples :

- Coref
- IR (= moteurs de recherche sur requete)
- IE (remplissage de formulaires)
- NER
- Text similarity
- Analyse de sentiment (opinion mining)
- WSD
- SRL
- RTE/NLI

Sorbonne Nouvelle

Modules

La plupart des applications reposent sur une série de manipulations qui identifient les objets aux différents niveaux linguistiques.

- segmentation
- tokenisation
- repérage des expressions multi-mots (idiomes, mots-composés,...)
- lemmatisation
- POS-tagging
- WSD/SRL
- chunking
- parsing (dépendance vs constituant)
- ...

Sorbonne Nouvelle

Plan

Sorbonne Nouvelle

Jalons historiques

- années 40 : préhistoire ; oppositions entre approches symboliques (discrètes) et approches statistiques
- années 50-70 : premiers espoirs déçus. Invention de la linguistique formelle (Chomsky) et des langages de haut niveau (LISP) ; invention de Eliza ; rapport ALPAC
- années 70-90 : arrivée des statistiques qui viennent surpasser les systèmes existants à base de règles
- années 2000 : éclatement de la bulle du Nasdaq ; apparition des méthodes d'apprentissage « profond » et puissance de calcul ; systèmes de dialogue à maturité technologique, de même que les systèmes de traduction