Le projet *Victoria*

MorphAcq : acquisition automatique des règles morphologiques d'une langue concaténative

Équipe RL

Sommaire

- Qu'est ce que le TALN ?
- Origine et objectifs du projet *Victoria*.
- Acquisition automatique de la morphologie d'une langue concaténative.

Sommaire

- **Qu'est ce que le TALN ?**
- Origine et objectifs du projet *Victoria*.
- Acquisition automatique de la morphologie d'une langue concaténative.

Traitement Automatisé des Langues Naturelles => TALN

Domaine de recherche visant à doter les ordinateurs de compétences linguistiques afin d'analyser ou générer des textes en langues humaines.

Aussi appelé :

➔ **L**inguistique **C**alculatoire (**LC**)

En anglais :

➔ **N**atural **L**anguage **P**rocessing (**NLP**)

• **C**omputational **L**inguistics (**CL**)

Applications pratiques du TALN

- désambiguïsation sémantique,
- résumé automatique,
- extraction d'information,
- systèmes question-réponse,
- traduction,
- etc...

Principales limitations (1)

Les langues naturelles sont ambiguës, c.a.d, plusieurs interprétations possibles des données.

=> Description de la langue difficile
=> Acquisition automatisée difficile.

Exemple :
« Si le prof de droit allemand arrive en retard, il fiche tout en l'air. »

Questions :
- « Si » conjonction de subordination ou réponse affirmative ?
- Qui est allemand ?
- « fiche » nom commun ou verbe ?
- « En l'air », « en retard », est ce une composition ou un tout ?

Principales limitations (2)

Les êtres humains ont une composante « exponentielle »
dans leur façon de s'exprimer (loi de Zipf)

Améliorer la description d'un aspect linguistique
est toujours plus couteux...

Conséquences :

1. peu de ressources linguistiques pour peu de langages,
2. leur couverture et qualité sont améliorables.

Sommaire

- **Qu'est ce que le TALN ?**
- **Origine et objectifs du projet *Victoria*.**
- Acquisition automatique de la morphologie d'une langue concaténative.

Origine du projet *Victoria*

Projet issu de la collaboration entre trois équipes:

- **Projet Alpage** INRIA/Univ Paris 7,
- **Groupe Lys/Cole** Univ de la Corogne,
- **Équipe RL** I3S, Univ de Nice Sophia Antipolis

Objectifs du projet

Simplifier l'acquisition, l'extension et la correction de règles morphologiques, de lexiques et de grammaires.

Objectifs organisés autours de trois axes principaux:

- plate-forme web d'outils collaboratifs,
- stratégies de transfert inter-langues,
- outils automatique ou semi-automatique d'acquisition, d'extension et de correction de ressources linguistiques.

Sommaire

- **Qu'est ce que le TALN ?**
- **Origine et objectifs du projet *Victoria*.**
- **Acquisition automatique de la morphologie d'une langue concaténative.**

MorphAcq : acquisition automatique des règles morphologiques d'une langue concaténative

- **Lionel NICOLAS** (Équipe RL – I3S)
- **Jacques FARRÉ** (Équipe RL - I3S)
- **M.A. MOLINERO** (Groupe Lys - Univ de la Corogne)

Sous-Sommaire

- Introduction
- Le processus, étape par étape
- Quelques exemples de résultats

Sous-Sommaire

- **Introduction**
- Le processus, étape par étape
- Quelques exemples de résultats

Morphologie d'une langue

La morphologie regroupe plusieurs mots (formes) sous une même classe (lemme) .

Morphologie flexionnelle (même catégorie syntaxique)
Exemple: chant -e, -es, -ons, -ez, -ent, -ais, -ait, -ions etc.

Morphologie dérivationnelle (catégorie syntaxique différente)
Exemple: porport -ion, -ionel, -ionnellement etc.

Règles Morphologiques

Règles morpho = liste d'opérations sur chaînes de caractères

Utiles pour :
- les lexiques à deux niveaux,
- l'obtention d'informations simples pour des mots inconnus.

Opérations simple pour une morphologie concaténative:
Concaténation de suffixes et de préfixes.
(pas d'infixes)

Objectif

Obtenir les règles d'une morphologie concaténative flexionnelle et dérivationnelle à partir de texte brut.

Approches existantes

=> Minimum Description Length (MDL).
=> Probabilité des transitions entre caractères.

Problème :
usage intensif de variables et de formules ad-hoc

Sous-Sommaire

- **Introduction**
- **Le processus, étape par étape**
- Quelques exemples de résultats

Construction d'une liste simpliste

Les mots sont rangés dans un arbre dont les transitions sont étiquetées avec les lettres des mots introduits.

Une sous chaîne est un possible affixe si:
- elle apparaît sur un sous arbre présents deux fois dans l'arbre,
- elle apparaît globalement plus que les « possibles stems »,
- elle aparaît un noeud où un mot fréquent est présent,
- elle co-apparait au moins deux fois avec un autre candidat,

Détection de paires

Plus un lemme est utilisé plus ses différentes formes apparaissent dans un corpus.

=> Si un possible affixe appartient à une famille, il aura une co-apparition croissante en fréquence avec d'autres candidats.

=> Pour une langue concaténative, ces co-apparitions seront sur les mêmes nœuds.

Construction de familles

Détection de paires inter-reliées apparaissant sur un même nœud

Filtrage des familles incomplètes ou incorrectes:
- sur les sous familles,
- sur les mots fréquents.

Sous-Sommaire

- ✔ **Introduction**
- ✔ **Le processus, étape par étape**
- ➢ **Quelques exemples de résultats**

Exemples de règles acquises pour l'anglais (1)

Préfixes

#con , #de , #pre , #re , #trans <> ferred# ,fer#
#as , #de , #re <> signing# ,sign# ,signed#
, #mid , #north , #south <> western# ,west#
, #over , #war <> heads# ,saw#
, #con , #pre <> text#
, #un , #under <> developed#
, #dis , #re <> placed#
, #ex , #un <> changed#
, #dis , #un <> like# ,qualified#
...

Exemples de règles acquises pour l'anglais (2)

Suffixes

, ance# , ant# , ants# , ed# , ing# <> #attend, #assist
al# , als# , e# , ed# , es# , ing# <> #approv, #arriv
e# , ed# , ement# , ements# , es# , ing# <> #replac, #achiev,
, ed# , er# , ers# , es# , ing# <> #publish
e# , ed# , ement# , ers# , es# , ing# <> #advanc
al# , ally# , e# , ed# , es# , ing# <> #practic
ate# , ated# , ating# , ation# , ative# , ator# <> #specul
al# , ally# , ated# , ating# , ation# , ations# <> #nomin

...

Merci pour votre attention.

Des questions?