DES MOTS AUX TEXTES

Terminologie et recherche d'information

BIBLIOGRAPHIE

Fuchs, Catherine, 1993, *Linguistique et traitements automatiques des langues*, Hachette Université.

Jacquemin, Christian, (ed), 2000, *Traitement automatique des langues pour la recherche d'information*, Revue TAL 41-2.

Laporte, Eric, 1997, " Les mots. Un demi-siècle de traitements", in Revue TAL, numéro 38, *Etat de l' Art*pp. 47-68

Pustejoski, James, 1995, *The Generative Lexicon*, Cambridge, MIT Press

INTRODUCTION

Trier l'information pertinente

Les textes sont d'abord et avant tout des mots, des suites de mots.

Quel travail faire sur le mot pour améliorer la recherche d'information ?

Deux axes :

1- la constitution d'outils et de ressources pour le TAL appliqué à la RI

2- l'enrichissement d'outils de RI par des techniques de TAL.

AXE 1 : Constitution d'outils et de ressources pour le TAL appliqué à la RI

a) Mots et racines / familles de sang

« racinisation » ou en anglais « stemming » : associer à un mot une pseudo-racine

- Manuelle (essentiellement suffixation)
- Automatique :
  - Par suppression d'affixes
  - Distance entre chaînes
  - Extraction à partir de dictionnaires électroniques

Evaluation :

- Sur-racinisation : la pseudo-racine est trop large. C'est le cas de *nat-* qui regrouperait *nature* et *nation*.
- Sous-racinisation : la pseudo-racine n'est pas assez large. C'est le cas de *adaptat-* qui distingue abusivement *adaptation* et *adapter*.

Méthodes sont utilisées :

1) la comparaison de bigrammes (pb : *savoir, sais, su*)

2) la méthode proposée par C. Jacquemin en 1997 (liste de termes et corpus)

3) la méthode par élimination de suffixes

Méthode 2 :

active immuniz-ation / active-ly immuniz-ed

sex-ual (sexual partner / sex partner)

different-ial / different diagnostic

acoustic-al / acoustical signal

Méthode 3 :

On dispose d'une liste de terminaisons, d'un ensemble de règles de désuffixage, et d'un ensemble de règles de recodage. Avec ces trois données, on construit par itération une procédure de suppression du suffixe le plus long.

Exemple :

Le patron est [C] (VC) m [V] m pour mesure

Nationalism-----> ational---------> at : ne s'applique pas, racine n-.

Nationalism------> alism ---------> al : s'applique, racine national-

Nationalism--------> al--------->rien : s'applique, racine nation-.

Lemmatiseur ≠ racinisateur. Cf FLEMM, ou Cordial

b) Mots et co-occurences / familles de sens

Le problème de l'homonymie

Etude des contextes et des cooccurences :

- terminologie : *train d'atterrisage, détecteur de mensonges*
- rapports entre Nom et Verbe (cf Bouillon et al.)

Deux points font l'originalité de ce travail :

1- son ouverture linguistique. Pustejoski : des ressources lexicales **enrichies par des informations sémantiques explicatives**

2- une méthode d'apprentissage automatique du lien verbo-nominal.

L'acquisition de ressources lexicales enrichies de connaissances sémantiques

• Le problème : le phénomène d'équivalence sémantique

• Les approches existantes et leurs limites

a) la synonymie et l'hyperonymie, qui plus est restreintes aux noms

b) Le lien entre le nom et le verbe est donc très fort puisqu'il peut rester implicite.

interpréteur de commande – interpréter ; magasins de disques – vendre ; parc à munitions - entreposer...

• L'idée du travail de Bouillon et al.

a) Trois hypothèses de départ

- les ressources lexicales permettent d'améliorer la performance des systèmes de recherche d'information.
- ces ressources doivent être adaptées aux tâches de reformulation et de précision des requêtes.
- et elles doivent être acquises sur corpus.

b) Arrière–plan théorique

Associer à chaque mot une structure interne, qui explicite :

a) les différents prédicats indispensables à sa compréhension

b) la manière de projeter cette structure au niveau syntaxique.

Exemples :

Je commence la symphonie / La symphonie commence

Je prends mon repas avec moi / Pendant le repas, j'ai dormi

Je commence le livre / Je commence à lire le livre

Je fais cuire un gateau / Je fais cuire des pommes de terre

*Le livre se lit bien / *Le livre s'écrit bien.*

*Ce vin se boit facilement / *Ce vin se produit facilement.*

c) Mise en place informatique . Méthode d'acquisition du lien Verbo-nominal

c) Mots construits donc inconnus

un système de génération et d'analyse automatiques d'unités lexicales construites non attestées par les dictionnaires, assorties d'informations sur leur mode de constructions, et leur sens.

un virus, pour être détectable doit ...

Le ministère entend organiser cette traçabilité totale des OGM, di champs jusqu'au produit fini.

Dans les trois exemples cités, on voit que sont indissociables les outils de Tal et les ressources qu'ils génèrent.

AXE 2 : Enrichissement d'outils de RI par des techniques de TAL.

1) Enrichir l'indexation des documents

Exemple 1 : le système DSIR de M. Rajman et al., dimension sémantique qui repose sur des fréquences de coocurrences entre mots

Exemple 2 : le travail de Gaussier et al., qui a pour but d'enrichir l'indexation des documents au moyen d'analyse morphologique et syntaxique à large couverture.

2) Augmenter la qualité du moteur de recherche

(Cf. Patrice Bellot et Marc El Bèze)

L'amélioration des sorties d'un moteur de recherches

Classement des documents en fonction de leur pertinence.

Pb de l'homonymie + Pb de la synonymie

Un continuum entre doc. pertinents et doc. non pertinents.

Autre idée :

1- regrouper les documents en plusieurs classes thématiques.

2- mesurer la pertinence des documents en calculant une distance entre classe.

Requête courte = requête ambiguë.

Nbre de classes = I -0,521 . (taille de la requ^te) + 10,65 I

CONCLUSION

• De la linguistique au TAL. Questions controversées

Les gains linguistiques sont-ils des gains en information ?

Les traitements simplifiés en TAL sont-ils meilleurs que des traitements superficiels ?

Le TAL peut-il être utilisé sur des requêtes pauvres telles que celles exprimées sur un moteur de recherche ?

Le TAL peut-il être utilisé sur des documents sales tels que ceux trouvés sur le web ?

• Du TAL à la linguistique.

La morphologie à choisir pour améliorer la recherche d'information a conduit à :

- redéfinir la distinction entre mots vides et mots pleins.
- redéfinir la notion de mot en l'ouvrant à celle de terme.

"cours des monnaies" (vs "cours de dessin")

"pièce de rechange" (vs "pièces dans un logement")

- repenser la normalisation ou les langues contrôlées (étude de styles, linguistique textuelle)