

**Recherche d'information
et extraction
automatique
d'informations**

Dominic Forest, Ph.D.
Ecole de bibliothéconomie et des sciences de l'information
Université de Montréal

SCI6136 – Fouille de documents
Cours 7, 27 octobre 2009

--	--

Rappel

- Terme complexe
 - Terme constitué de plusieurs formes graphiques
- Plusieurs stratégies pouvant assister l'identification des termes complexes
- Segments répétés
 - "Toutes les suites d'occurrences non séparées par un délimiteur de séquence sont des occurrences de segments [...]. Les segments dont la fréquence est supérieure ou égale à 2 dans le corpus sont des segments répétés dans le corpus" (Lebart et Salem, 1994, p. 60)

--	--

Rappel

- Extraction de termes et assistance à l'indexation
 - Démarche proposée par Moens (2000, chap. 4)
 - Extraction du lexique
 - Suppression des mots fonctionnels
 - Amputation des terminaisons (optionnel)
 - Identification des groupes de mots (*phrases*) (optionnel)
 - Remplacement des mots par des catégories (optionnel)
 - Pondération de la représentativité des mots

--	--

Rappel

- Extraction de termes et assistance à l'indexation
 - Principaux avantages des techniques d'extraction de termes pour assister l'indexation
 - Simples
 - Rapides
 - Efficaces (dans de nombreux cas)
 - Principales limites des techniques d'extraction de termes pour assister l'indexation
 - Trop générales, trop spécifiques
 - Paramétrage (seuil arbitraire)
 - Insensibles aux parties du discours et aux parties des documents
 - Souvent limitées aux unités lexicales du corpus

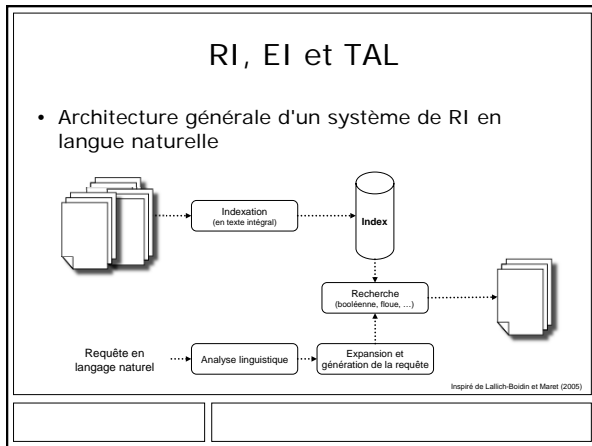
RI, EI et TAL

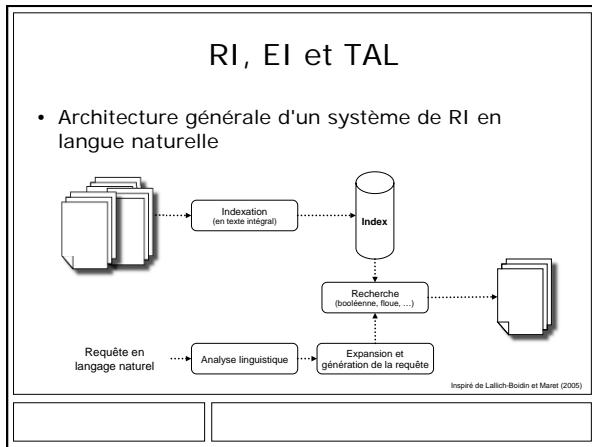
- Objectifs :
 - Comprendre l'importance de la fouille de textes et du traitement automatique des langues (TAL) dans le processus de recherche d'information (RI)
 - Identifier et comprendre les principaux modules des systèmes de RI dans lesquels peuvent intervenir des opérations d'extraction automatique d'information et de TAL
 - Évaluer les avantages et les inconvénients des processus de fouille et de TAL à l'intérieur d'une architecture de système de RI

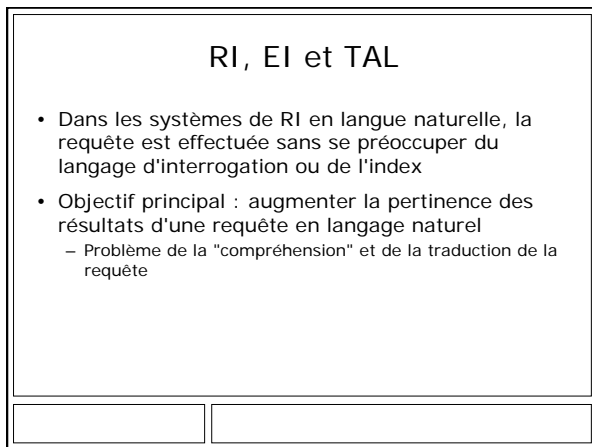
RI, EI et TAL

- Architecture générale d'un système de RI en langage naturel

Inspiré de Lallich-Boidin et Maret (2005)







RI, EI et TAL

- Objectif secondaire: passer d'une chaîne de caractères à une représentation linguistique en supprimant l'ambiguïté sémantique et en "structurant" le plus possible la requête pour en permettre l'expansion (enrichissement), la dégradation et la traduction dans un langage formel d'interrogation
- Opération préliminaire : segmentation par mots de la requête en LN, *tokenisation*

RI, EI et TAL

- Problèmes orthographiques et grammaticaux
 - Variations et erreurs orthographiques
 - *Categorization, categorisation*
 - Google.com (ANG) : 9 490 000, 3 560 000
 - Erreurs orthographiques difficilement prédictibles
 - *Catgorisation, ctegorisation*
 - Solutions possibles : liste de mots synonymes, identification de proximité graphique entre mots

RI, EI et TAL

- Problèmes orthographiques et grammaticaux
 - Variations grammaticales
 - Lien entre la pertinence des résultats d'une requête et l'identification de la catégorie grammaticale des mots de la requête
 - L'ambiguïté sémantique de plusieurs mots peut être levée en identifiant la partie du discours des mots en question
 - Ambiguïté des mots vides
 - Importance de l'identification des parties du discours pour la lemmatisation
 - » Avions → AVOIR ← *liste de mots fonctionnels*
 - Question en suspend : ordre d'application des traitements
 - Extraction des entités nommées (pondération)

RI, EI et TAL

- Problèmes morphologiques
 - Langues flexionnelles
 - Morphologie flexionnelle
 - Avaient, avait, avais, avions
 - Morphologie dérivationnelle (formation de nouveaux mots, affixation)
 - Faire, défaire, refaire
 - Maison, maisonnette
 - Solutions possibles : *stemming*, lemmatisation

RI, EI et TAL

- Problèmes sémantiques
 - Problème de la polysémie des mots de la requête
 - Ex. : Souris, avocat, ...
 - Impact : diminution de la précision
 - Module de désambiguïsation sémantique
 - Augmentation de la pertinence des résultats
 - Nécessaire pour l'expansion de requêtes
 - Problème de la synonymie des mots de la requête et des documents
 - Impact : diminution du rappel
 - Banque, institution financière, ...
 - Dictionnaire de synonymes

RI, EI et TAL

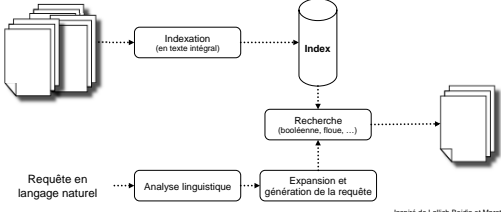
- Problèmes sémantiques
 - Solution possible : ressources lexicales externes
 - WordNet (Démon)
 - Distinctions sémantiques fines, trop fines pour la RI
 - Problème des liens pragmatiques et thématiques
 - » *Tennis problem*
 - » [WordNet::Similarity](#)

RI, EI et TAL

- Problèmes sémantiques
 - Solution possible : techniques statistiques et numériques
 - Désambiguïsation sur la base de données statistiques
 - Ensemble d'apprentissage
 - Été : 85% [nom], 15% [verbe]
 - Technique du "sac de mots" (Inspiré de Lallich-Boidin et Maret (2005))
 - Distinctions sémantiques fondées sur la proximité d'occurrence
 - » Souris, élevage de souris, souris pour ordinateur, élevage de souris par ordinateur
 - Désambiguïsation par vecteurs thématiques
 - » Distinctions pas suffisamment fines
 - Grökker (Démon)

RI, EI et TAL

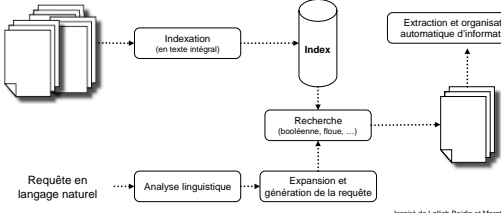
- Architecture générale d'un système de RI en langue naturelle



Inspiré de Lallich-Boidin et Maret (2005)

RI, EI et TAL

- Architecture générale d'un système de RI en langue naturelle



Inspiré de Lallich-Boidin et Maret (2005)

RI, EI et TAL

- Perspectives
 - RI et multilinguisme
 - Recherche d'information translinguistique (CLIR)
 - Intervention de l'utilisateur
 - Processus de découverte d'information ponctué d'interactions ciblées
