EARIA'06

Quelle langue pour la RI ?
Première partie

E. Gaussier, Univ. J. Fourier

Plan

- Illustration (pourquoi ?)
- Traitement automatique des langues (généralités)
- Morphologie et RI
- Syntaxe et RI
- Sémantique et RI

Illustration (1)

- *C* collection de documents (d^1, ..., d^N)
- V vocabulaire d'indexation, formé des mots de la collection

Loi de Mandelbrot : la fréquence *f(n)* est liée à son rang *n* dans l'ordre des fréquences (généralise et corrige la loi de Zipf)

$$f(n)*(a+bn)^c \approx K \quad (f(n)*n \approx K)$$

En pratique, peu de mots très fréquents (50% des documents) et beaucoup de mots peu fréquents (hapax -> 20% des documents)

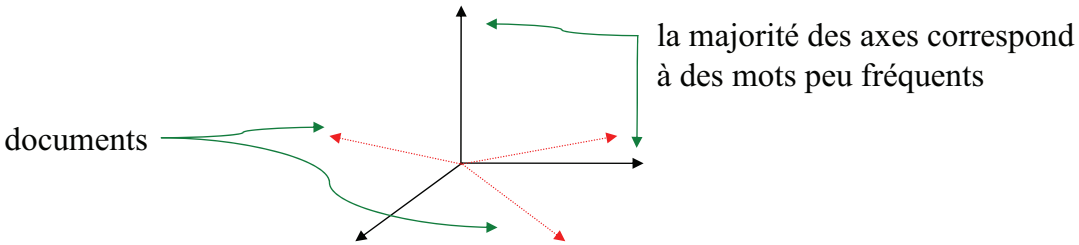
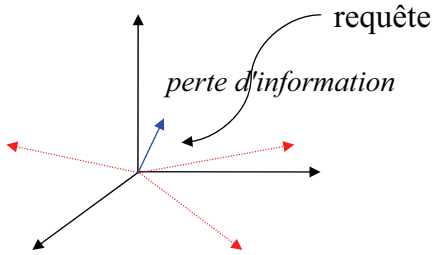


Illustration (2)

- Tous les documents (et les requètes) se retrouvent artificiellement éloignés (*curse of dimensionality*)
- Artificiellement :
  - Variantes morphologiques d'un même concept

    conducteur(s), conductible(s), conductibilité(s), conduire(...)
  - Synonymie: *courant, flot*
- Représentation simple, mais perte d'information peut-être cruciale



TAL – Généralités (1)

- Le TAL est traditionnellement divisé en 4 niveaux interdépendants
  - Morphologie *Etude des formes sous lesquelles se présentent les mots dans une langue. En particulier, étude des relations entre les mots d'une même famille (provenant d'une même racine)*
  - Syntaxe *Etude des relations entre unités lexicales*

    Je_PPS [m'intéresse_V [à_Prep [la recherche_N] [d'information_N]]]
  - Sémantique *Etude du sens des unités lexicales*

    courant (C001:usuel ; C002:flot) - flot (C002:flot)
  - Pragmatique *Etude des unités linguistiques dont la signification ne peut être comprise qu'en contexte*

    Vous avez l'heure ?

TAL – Généralités (2)

- Ces 4 niveaux interviennent dans le processus d'indexation. Cependant, la pragmatique reste confinée à des aspects particuliers (difficulté de description linguistique des différents contextes de recherche)
- Toutefois, la RI impose un certain nombre de contraintes sur les traitements sur lesquels elle s'appuie :
  - Taille des données traitées (complexité en temps raisonnable)
  - Hétérogénéïté des domaines (couverture des traitements)
  - Valeur statistique des traitements (précision et couverture)
  - Indépendance par rapport à la langue (facilité de portage d'une langue à l'autre)

Morphologie et RI (1)

- La morphologie en RI consiste principalement en une étape de *normalisation* des index. Cette normalisation vise à rassembler sous une même forme des formes graphiques différentes mais correspondant au même concept

  conduisons, conduirons, conduis -> conduire, condu-

Morphologie flexionnelle vs. morphologie dérivationnelle

- Deux grands types d'approche :
  - Une approche lexicale : reconnaissance du mot en contexte, rattachement, par le biais d'un dictionnaire électronique, à sa forme normalisée (lemme ou racine)
  - Une approche légère : analyse (sommaire) du mot et racinisation

Morphologie et RI (2)

L'approche lexicale

- Représentation, dans un dictionnaire, de chaque mot du lexique

  (G) courant, (L) courant[C001/2], courir[C003], (R) cour-
  (MS) AdjMS[C001], SubsMS[C002], PP[C003]
  (S) C001, C002, C003

- Etiquetage morpho-syntaxique

  Le courant_SubsMS dans la branche A du réseau ...

- L'étiquetage morpho-syntaxique détermine la catégorie grammaticale (partie du discours, *part-of-speech*) en contexte (*part-of-speech tagging*)

Morphologie et RI (3)

L'approche lexicale

- Deux approches pour l'étiquetage morpho-syntaxique
  - Grammaire

    PPS le V -> PPS le_PC V (je le vois)
  - *Nécessite un codage manuel par un linguiste*
  - *Les règles sont différentes d'une langue à l'autre et peuvent varier d'un domaine à un autre*
  - *Peu de règles suffisent en général (env. 100 pour le français courant)*
  - *Possibilité de développer des règles pour des domaines particuliers (biologie)*

[INTEX - www.atala.org ; Amrani, Kodratoff, Matte-Taillez *archivesic.ccsd.cnrs.fr/sic_00001260/en/]*

Morphologie et RI (4)

L'approche lexicale

- Apprentissage, essentiellement supervisé, à partir de corpus étiquetés manuellement – approche *n-gram* de mots (chaîne de Markov cachée)

  P(le_PC|PPS)*P(V|le_PC) -> P(le_D|PPS)*P(V|le_D) (je le vois)
  - Nécessite des corpus annotés (the more the better)
  - Beaucoup d'initiatives pour développer et mettre à disposition des corpus annotés
  - Possibilité de se reposer sur des corpus non annotés (moins bonne performance)

[WIKIPEDIA en.wikipedia.org/wiki/Part-of-speech_tagging; Brill *Some advances in rule-based part-of-speech tagging (95)]*

Morphologie et RI (5)

L'approche lexicale

- Une fois l'étiquetage morpo-syntaxique réalisé, on substitue à chaque forme graphique son lemme

  N -> MS

  Adj -> MS

  V -> Infinitif

  D -> MS

Cette opération se fait par application des données contenues dans le dictionnaire (processus de génération associé au dictionnaire).

- Ces étapes constituent une première normalisation (morphologie flexionnelle) précise et robuste

La petite ferme le voile

Morphologie et RI (6)

L'approche lexicale

- Passage du lemme à la racine

  – A partir de dictionnaires existants

    Dictionnaires difficiles à construire car les relations entre mots d'une même famille sont en partie irrégulières (*plomb/plombier* ; allomorphie *conduc-/condui-*)

  – En utilisant des procédures de troncature à partir des formes lemmatisées *[*Gaussier *Unsupervised Learning of Derivational Morphology from Inflectional Lexicons (99)]*

Morphologie et RI (7)

L'approche légère

- La racinisation par ré-écriture (l'algorithme de Porter)
  - Deux types de règles, avec liste d'exceptions
    - Suppression de suffixes : *s ->* *cats -> cat*
    - Ré-écriture de certaines suffixes

      *(*v*) y -> i* *happy -> happi*
  - Prise en compte du contexte immédiat (e.g. le mot contient une voyelle) dans l'application des règles
  - Les règles (env. 50) sont ordonnées, et appliquées en séquence

Versions en plusieurs langues [www.tartarus.org/~martin/PorterStemmer ; Savoy *A stemming procedure and stopword list for general French corpora (99)*]

Morphologie et RI (8)

Comparaison des deux approches

- L'approche lexicale fournit des résultats plus précis (en termes linguistiques)
- L'approche légère a une plus large couverture (autres langues, domaines spécialisés)
- Les performances en RI sont grosso modo identiques
- La morphologie flexionnelle semble plus efficace (gain en précision sans perte de rappel), alors que la morphologie dérivationnelle peut introduire une perte de rappel

Apport de la morphologie

- Gain démontré sur les langues occidentales - une réduction de l'espace intéressante (8 variantes flex. en moy. par mot en fr.)
- Les résultats varient d'une langue à l'autre (voir la deuxième partie)

Syntaxe et RI (1)

Les paires adjacentes

- [Salton & McGill *Introduction to Modern Information Retrieval - 1983*] introduisent le concept de paires adjacentes pour mieux rendre compte du contenu d'un document

  [George Bush] is the [US president]

- En anglais, ces paires permettent de rendre compte des termes élémentaires (Adj N, N N), de certaines entités nommées et de certaines relations sujet-verbe et verbe objet

- En français

  [George Bush] est le [président des EU]

- Les termes techniques suivent un autre mode de formation (de type roman) que l'anglais (de type germanique)

Syntaxe et RI (2)

Termes techniques en anglais et en français

- Patrons élémentaires en anglais

  Adj N (probabilistic retrieval)
  N N (information retrieval)

- Patrons élémentaires en français

  N Adj (recherche probabiliste)
  N (Prep) (D) N (recherche d'information)

- La prise en compte d'une préposition soulève le problème du découpage des termes (pb qui n'existe pas en anglais)

- En pratique, expressions régulières qui encodent les séquences admissibles de parties du discours. Les entités nommées sont traitées de la même façon (souvent avec des lexiques spécifiques)

Syntaxe et RI (3)

Prise en compte dans un système de RI

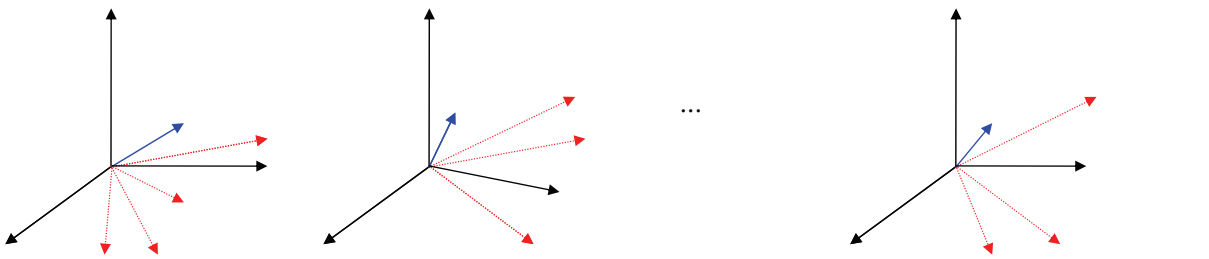
- La prise en compte de nouveaux index pose le problème de l'accroissement de dimension de l'espace vectoriel sous-jacent (malédiction des grandes dimensions)
- De plus, une indexation trop précise risque de faire disparaître des similarités entre document (US president vs president)
- Solution
  - Les index composés (termes, entités nommées, relations syntaxiques) constituent un ou des jeu(x) d'index additionnel(s)
  - Plusieurs espaces vectoriels sont considérés puis les résultats sont fusionnés

Syntaxe et RI (4)

Prise en compte dans un système de RI

$$cos(q,d) = a_1\, cos_{EV1}(q,d) + ... + a_p\, cos_{EVp}(q,d)$$

où *Evi* désigne l'espace vectoriel associé au type *i* d'index (termes, noms de personne, ...) et où les a_i sont des poids indiquant la confiance que l'on a dans la similarité obtenue dans chacun des sous-espaces



index std

termes techniques

relations S-V

Syntaxe et RI (5)

Evaluation

	ref	lem	stem (0.66)	np	synt (0.33)
R-P moy					
à 0.00	0.49	0.56	0.54	0.55	0.54
à 1.00	0.067	0.068	0.073	0.071	0.076
Prec. moy.	0.20	0.24	0.24	0.25	0.25

Expériences réalisées sur le français (collection AMARYLLIS – 1ère campagne) – modèle vectoriel

Seule la différence avec l'indexation de référence (sans traitement) est significative

Syntaxe et RI (6)

Conclusion

- Une histoire de résultats mitigés *[*Fagan *(87) Experiments in automatic phrase indexing for document retrieval;* Gaussier *et al. Recherche d'information en français et traitement automatique des langues, TAL]*
- Relativement peu de nouveaux apports dans ce domaine en RI ad hoc
- Applications dérivées où l'analyse syntaxique joue un rôle important :
  - extraction d'information
  - question/réponse (AskJeeves)
  - RI dans des domaines de spécialité ou sur des indexations complexes (graphes conceptuels)

*[*I. Ounis, Y. Chiaramella, L. Kefi, C. Berrut *(MRIM) ; Assistance intelligente à la recherche d'information, Hermès 2003]*

Sémantique et RI (1)

Quel type d'impact ?

- Polysémie et synonymie : deux phénomènes opposés
  - Synonymie : deux mots différents renvoient au même concept

  Des dimensions doivent être fusionnées -> réduction de dimension
  - Polysémie : un mot renvoie à des concepts/sens différents

  Autant de dimensions que de concepts -> accroissement de dimension

  Pas nécessairement : les concepts différents peuvent exister par ailleurs (désambiguïsation naturelle)

- Relations sémantique de type thésaurique (générique/spécifique – hyperonymie/hyponymie - méronymie/holonymie)

  Permettent de regrouper des termes sous un même concept -> réduction de dimension

Sémantique et RI (2)

Quelles approches ?

- Approche lexicale : fondée sur des ressources sémantiques élaborées par des équipes de linguistes (précision vs couverture)

- Approche implicite : vise à découvrir, sur la base d'une analyse statistique des collections, les concepts sous-jacents, et à relier chaque terme à un ou plusieurs concepts (couverture vs précision)

Sémantique et RI (3)

Approche lexicale

- Ressources à disposition :
  - Dictionnaires généraux utilisés pour la désambiguïsation sémantique (Oxford-Hachette)
  - Ressources sémantiques lexicales :
    - Wordnet (wordnet.princeton.edu), EDR (www.ijnet.or.jp/edr)
    - Thésaurus spécialisés : UMLS (méta-thésaurus et réseau sémantique), MeSH (thésaurus), Medline
  - Ontologies : OpenCyc (www.cyc.com/cycdoc/upperont-diagram.html)

Sémantique et RI (4)

Approche lexicale

- Désambiguïsation automatique :
  - Affecter un sens à un mot dans son contexte (liste de sens prédéfinis) – Senseval (www.senseval.org)
  - Processus implicite en RI

    napoléon -> cognac ? empereur ?

    empereur napoléon

    cognac napoléon
  - Apport en RI discuté, fortement dépendant du type de requêtes

    *[*Shutze & Pedersen *Information retrieval based on word senses (95)]*

Sémantique et RI (5)

Approche lexicale

- Utilisation de ressources sémantiques pour normalisation synonymique et/ou hyperonymique
  - Question : où s'arrêter dans la normalisation ?

```
dog, domestic dog, Canis familiaris
   => canine, canid
      => carnivore
        => placental, placental mammal, eutherian, eutherian mammal
          => mammal
            => vertebrate, craniate
              => chordate
                => animal, animate being, beast, brute, creature, fauna
                  => ...
```

  - Réponse : ???

Sémantique et RI (6)

Approche lexicale

- Réponse :
  - On ne remonte que d'un nombre fixé de niveau *[*Kiryakov & Simov *Ontologically supported semantic matching (99);* Bruandet & Chevallet *Utilisation et construction de bases de connaissances pour la recherche d'information (03)]*
  - On définit une distance sur les DAGs qui permet de prendre en compte tous les concepts mais avec des poids qui dépendent de la distance au terme d'origine *[*Siolas & Dalche-Buc *Support vector machines based on a semantic kernel for text cateorization (00)]*
  - On cherche des critères qui fournissent une coupe « optimale » dans le réseau sémantique *[*Seydoux *Exploitation de connaissances externes dans les représentations vectorielles en recherche documentaire (05)]*

Sémantique et RI (7)

Approche implicite

- On cherche des concepts latents consitués à partir des termes retenus :
  - Combinaison linéaire (analyse sémantique latente)

$$D \approx U \sum V^t$$

*[*Deerwester *et al. Indexing by latent semantic analysis (90)]*

  - Distribution de probabilité (combinaison linéaire positive)

$$P(w,d) = P(d) \sum_d P(c|d)\, P(w|c)$$

*[*Hoffman *Probabilistic latent semantic analysis (99)]*

- Ces concepts définissent un nouvel espace (réduit) sur lequel la collection est projetée

Sémantique et RI (8)

Comparaison

- Approche lexicale précise mais nécessite des données difficiles à élaborer
  - Effort en cours dans beaucoup de langues (toutes ne sont pas au même niveau)
  - Applications à des domaines spécifiques (tous ne sont pas au même niveau)
- Approche implicite plus facile à déployer mais les notions de sens sous-jacentes sont parfois difficiles à interpréter

Impact

- Positif dans beaucoup d'expériences (e.g. R-precision)

	CISI	CRAN	MED	CACM
Tfldf	0.18	0.23	0.42	0.21
Noyau impl.	0.22	0.31	0.48	0.21

Conclusion

Dichotomie

- Développement et exploitation de ressources existantes
  - Dictionnaires morphologiques : largement répandus, mais statut différent suivant les langues
  - Analyseurs syntaxiques (de surface) : tendent à se répandre (langues occidentales)
  - Ressources sémantiques : plusieurs initiatives dans ce sens, mais l'effort à entreprendre est considérable
- Acquisition automatique de ressources
  - Permet de pallier le manque de ressources
  - Des résultats en fait satisfaisants (aucune ressource ne sera adaptée à toutes les collections)
- Adaptation (et enrichissement) de ressources existantes aux collections considérées