

La RI sémantique / La RI en domaine de spécialité

Jean-Pierre Chevallet

LIG-MRIM

EARIA Octobre 2014

Plan

- 1 Introduction
 - Documents, information et signal
 - RI Sémantique
 - Questionnements par l'exemple
- 2 Descripteurs : atomes des index
 - Niveau morphologique
 - Niveau terminologique
 - Niveau syntaxique
 - Niveau sémantique et conceptuel
- 3 Ressources
 - Ressources lexicales extraites
 - Ressources conceptuelle non formelles
 - Ressources conceptuelles formelles
- 4 Correspondance sémantique
- 5 Conclusion

Plan

- 1 Introduction
 - Documents, information et signal
 - RI Sémantique
 - Questionnements par l'exemple
- 2 Descripteurs : atomes des index
 - Niveau morphologique
 - Niveau terminologique
 - Niveau syntaxique
 - Niveau sémantique et conceptuel
- 3 Ressources
 - Ressources lexicales extraites
 - Ressources conceptuelle non formelles
 - Ressources conceptuelles formelles
- 4 Correspondance sémantique
- 5 Conclusion

Accéder à de l'information : un processus sémantique

- Information massivement disponible mais rapidement accessible ?



Accéder à de l'information : un processus sémantique

- Information massivement disponible mais rapidement accessible ?
- SRI : un assistant automatisé pour faciliter l'accès à de l'information.



Accéder à de l'information : un processus sémantique

- Information massivement disponible mais rapidement accessible ?
- SRI : un assistant automatisé pour faciliter l'accès à de l'information.
- Information : ce que l'utilisateur peut extraire de documents avec ses propres connaissances.



Accéder à de l'information : un processus sémantique

- Information massivement disponible mais rapidement accessible ?
- SRI : un assistant automatisé pour faciliter l'accès à de l'information.
- Information : ce que l'utilisateur peut extraire de documents avec ses propres connaissances.
- Document : une forme, un contenu, un rôle social.



Accéder à de l'information : un processus sémantique

- Information massivement disponible mais rapidement accessible ?
- SRI : un assistant automatisé pour faciliter l'accès à de l'information.
- Information : ce que l'utilisateur peut extraire de documents avec ses propres connaissances.
- Document : une forme, un contenu, un rôle social.

RI : basée sur le **sens (sémantique)**, mais est-ce utile à un SRI ?



Signal et correspondance

北京 提供北京地区时政新闻和媒体信息。

Signal et correspondance

北京 提供北京地区时政新闻和媒体信息。

Peux-t-on simplement comparer le "signal" du document au "signal" de la requête ?

Signal et correspondance

北京 提供北京地区时政新闻和媒体信息。

Peux-t-on simplement comparer le "signal" du document au "signal" de la requête ?

De quoi la SRI a besoin au minimum ?

SRI : remplacer l'humain par le calcul statistique

Peux-t-on simplement comparer le "signal" du document au "signal" de la requête ?

Oui ... mais ..

- Mettre dans le SRI un minimum de connaissances, comme le type de média, le format, la langue, le découpage des mots, ...



SRI : remplacer l'humain par le calcul statistique

Peux-t-on simplement comparer le "signal" du document au "signal" de la requête ?

Oui ... mais ..

- Mettre dans le SRI un minimum de connaissances, comme le type de média, le format, la langue, le découpage des mots, ...
- Faire des hypothèses, comme l'indépendance entre les mots



SRI : remplacer l'humain par le calcul statistique

Peux-t-on simplement comparer le "signal" du document au "signal" de la requête ?

Oui ... mais ..

- Mettre dans le SRI un minimum de connaissances, comme le type de média, le format, la langue, le découpage des mots, ...
- Faire des hypothèses, comme l'indépendance entre les mots
- Exploiter la redondance statistique : plus la base est grande, plus on a de chance de trouver la réponse à une question sous la forme littérale (ex; "hauteur de la tour Eiffel").



SRI : remplacer l'humain par le calcul statistique

Peux-t-on simplement comparer le "signal" du document au "signal" de la requête ?

Oui ... mais ..

- Mettre dans le SRI un minimum de connaissances, comme le type de média, le format, la langue, le découpage des mots, ...
- Faire des hypothèses, comme l'indépendance entre les mots
- Exploiter la redondance statistique : plus la base est grande, plus on a de chance de trouver la réponse à une question sous la forme littérale (ex; "hauteur de la tour Eiffel").

La RI "classique" se base sur ces quelques contraintes simples.



RI Sémantique : c'est quoi ?

Prenons des exemples :

(Bioasq training dataset 2014 <http://bioasq.lip6.fr>)

- "Is Rheumatoid Arthritis more common in men or women?"
- "Are there any DNMT3 proteins present in plants?"
- "Is alternative splicing of apoptotic genes playing a role in the response to DNA or mitochondrial damage?"

Constats :

- Présence de **termes** très techniques

RI Sémantique : c'est quoi ?

Prenons des exemples :

(Bioasq training dataset 2014 <http://bioasq.lip6.fr>)

- "Is Rheumatoid Arthritis more common in men or women?"
- "Are there any DNMT3 proteins present in plants?"
- "Is alternative splicing of apoptotic genes playing a role in the response to DNA or mitochondrial damage?"

Constats :

- Présence de **termes** très techniques
- Requêtes **factuelles**, plus que thématiques.

RI Sémantique : c'est quoi ?

Prenons des exemples :

(Bioasq training dataset 2014 <http://bioasq.lip6.fr>)

- "Is Rheumatoid Arthritis more common in men or women?"
- "Are there any DNMT3 proteins present in plants?"
- "Is alternative splicing of apoptotic genes playing a role in the response to DNA or mitochondrial damage?"

Constats :

- Présence de **termes** très techniques
- Requêtes **factuelles**, plus que thématiques.
- Forte **dépendance** entre les termes : prédicats, arguments, attributs ...



INFORMATION RETRIEVAL

Où se situe la RI sémantique ?

- RI classique documentaire
- Systèmes à question / réponses
- Base de données
- Base de connaissances (ex: Web sémantique)

Probablement une fusion de tout cela ?

RI Sémantique : la question

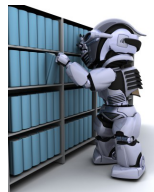
La recherche d'information dans un domaine de spécialité :

- La "signification" prend plus d'importance.
 - *Is Rheumatoid Arthritis more common in men or women ?*
 - *Are there any DNMT3 proteins present in plants ?*
 - *Is alternative splicing of apoptotic genes playing a role in the response to DNA or mitochondrial damage ?*

Question

En quoi un système de RI sémantique est différent d'un système "classique" ?

Système de RI Sémantique



Un système de RI :

Document : sa représentation dans le SRI dans le **modèle du document**(ou requête)

Indexation : produire une **instance** dans le modèle du document (off line)

Correspondance : calculer (matching) une **similarité** entre deux instances du modèle (in line)

NB: pour simplifier, on suppose que documents et requêtes ont le même modèle.

Système de RI Sémantique : exemple

Document

"RA is an autoimmune disease that results in a chronic, systemic inflammatory disorder that may affect many tissues and organs, but principally attacks flexible (synovial) joints."

Requête

"Is Rheumatoid Arthritis more common in men or women?"

Modèle

Vecteur pondéré de termes

Indexation : production des instances du modèle

Traitement classique :

- Découpage des mots
- Filtrage des mots vides (stop words)
- Normalisation des mots
- Comptage

Indexation : production des instances du modèle

Traitement classique :

- Découpage des mots
- Filtrage des mots vides (stop words)
- Normalisation des mots
- Comptage

instance du modèle vectoriel (index du document)

(affect,1) (attack,1) (autoimmune,1) (chronic,1) (disease,1) ...
(ra,1) (result,1)

Requête

(arthritis,1) (common,1) (men,1) (rheumatoid,1) (women,1)

Constats

instance : document

(affect,1) (attack,1) (autoimmune,1) (chronic,1) (disease,1) ...
(ra,1) (result,1)

"Is Rheumatoid Arthritis more common in men or women?"

Requête

(arthritis,1) (common,1) (men,1) (rheumatoid,1) (women,1)

- Perte des relations donc du sens de la requête ... on ne voit pas comment le système peut répondre correctement ..

Constats

instance : document

(affect,1) (attack,1) (autoimmune,1) (chronic,1) (disease,1) ...
(ra,1) (result,1)

"Is Rheumatoid Arthritis more common in men or women?"

Requête

(arthritis,1) (common,1) (men,1) (rheumatoid,1) (women,1)

- Perte des relations donc du sens de la requête ... on ne voit pas comment le système peut répondre correctement ..
- Aucune intersection de termes, les instances sont des vecteurs complètement différents, alors que le document a bien un rapport avec la requête

Constats

instance : document

(affect,1) (attack,1) (autoimmune,1) (chronic,1) (disease,1) ...
(ra,1) (result,1)

"Is Rheumatoid Arthritis more common in men or women?"

Requête

(arthritis,1) (common,1) (men,1) (rheumatoid,1) (women,1)

- Perte des relations donc du sens de la requête ... on ne voit pas comment le système peut répondre correctement ..
- Aucune intersection de termes, les instances sont des vecteurs complètement différents, alors que le document a bien un rapport avec la requête

Où sont les problèmes ?

Problèmes soulevés

- Modèle des documents
 - Doit être adapté aux types de requêtes pour avoir une chance d'y répondre.
 - Dans l'exemple : peu de chance car pas de modélisation de "more common", mais cela peut marcher un peu avec la redondance d'information dans les textes.

Problèmes soulevés

- Modèle des documents
 - Doit être adapté aux types de requêtes pour avoir une chance d'y répondre.
 - Dans l'exemple : peu de chance car pas de modélisation de "more common", mais cela peut marcher un peu avec la redondance d'information dans les textes.
- **Connaissances** pour :
 - 1 réaliser l'indexation
 - 2 réaliser la correspondance.

Exemple: savoir que "RA" est potentiellement un abrégé de "Rheumatoid Arthritis".

Réponses (temporaire)

En quoi un système de RI sémantique est différent d'un système "classique" ?

- Requêtes plus précises : un modèle de requête (document) plus complexe que le traditionnel "sac de mots".
- Termes d'un domaine : besoin de connaissances explicites,
 - pour un indexation correcte,
 - pour rendre possible la correspondance
- Calcul de la correspondance adapté au modèle.
 - Pouvoir comparer des index complexes : arbres, graphes, formules logiques, etc
 - Ordonner les réponses : correspondance partielle.

Réponses (temporaire)

Documents et requêtes en langue naturelle : il semble nécessaire de mettre en place des **traitements automatiques de la langue** :

- Eviter les pièges de la langue : polysémie, synonymie, variation terminologique, etc
- Mieux exploiter la richesse de la langue : constructions,
- Classification des requêtes : factuelles (qui, ou, quand, ...),
- Produire une représentation complexe des requêtes (et documents).

Autres exemple de requêtes

CLEF eHealth 2014

(<http://clefehealth2014.dcu.ie>)

- *Right upper lobe pneumonia with cavitary lesion*
- *Traumatic brain injury, status post fall*

TREC Medical reports task

(<http://trec.nist.gov/data/medical.html>)

- *Patients who had positron emission tomography (PET), magnetic resonance imaging (MRI), or computed tomography (CT) for staging or monitoring of cancer*
- *Patients admitted for hip or knee surgery who were treated with anti-coagulant medications post-op*

Autres exemple de requêtes

Patients who had positron emission tomography (PET), magnetic resonance imaging (MRI), or computed tomography (CT) for staging or monitoring of cancer

- On voit apparaître une structure dans requête :
 - Type d'objet recherché : *Patients*
 - Référence à un documents multimédia : des images médicales
 - Modalités des images : *PET, MRI, CT*
 - Pathologie : *cancer*
 - Diagnostique : *staging or monitoring*

Plan

- 1 Introduction
 - Documents, information et signal
 - RI Sémantique
 - Questionnements par l'exemple
- 2 Descripteurs : atomes des index
 - Niveau morphologique
 - Niveau terminologique
 - Niveau syntaxique
 - Niveau sémantique et conceptuel
- 3 Ressources
 - Ressources lexicales extraites
 - Ressources conceptuelle non formelles
 - Ressources conceptuelles formelles
- 4 Correspondance sémantique
- 5 Conclusion

Le descripteur : atome de l'index

- Indexer : choisir des "descripteurs" pour représenter le document par un index.
- Définir / comprendre les composants des textes, les mots, la composition des mots, les structures comme les termes, la notion de concepts.
- Choisir des composants des textes comme descripteurs.
- Besoin de ressources extérieures au corpus de textes pour permettre l'indexation.

Descripteur et Modèle de document

- Un index : un ensemble de descripteurs pour le document
- Descripteur : atomique ou composés (structurés)
- Indexation contrôlée : sélection des descripteurs limitée (ex: liste d'autorité)
- Indexation libre : pas de liste à priori, mais des contraintes tout de même (ex: mots pleins)

Paramètres pour l'indexation

- Choix des descripteurs
- Choix de la structure des index : Modèle, langage d'indexation
- Choix de la structure des requêtes : Modèle, langage d'interrogation
- Choix de la fonction de correspondance : En fonction des modèles
- Pour l'indexation automatique : choix du processus de sélection des descripteurs et de la construction des index

Indexation : choix des descripteurs

Descripteurs : libre ou contrôlé ?

- Libres : pas de contrainte dans le choix, mais rarement utile de tout indexer
- Contrôlés : construction d'un référentiel (ex: liste d'autorité), manuel ou automatique

Nature du contrôle :

- Choix des index les plus représentatifs du contenu thématique
- Une fois la collection indexée, la requête est contrainte au vocabulaire d'indexation ?

Correspondance par identité des descripteurs : s'échapper ?

- La coordination
- Des liens entre descripteurs : ressource terminologique.

Variation des descripteurs

Contrôle : décider des descripteurs référenciels, puis les détecter dans les textes

Exemple:

"marée noire" : pollution marine par le pétrole / mer polluée par des hydrocarbures

Variation / normalisation :

- Descripteurs différents mais sens similaires
- Variation terminologique, morphosyntaxe
- Ex: nominalisation *polluer*, *pollution*, *polluant*, *pollueur*, ...
- Variation sémantique : hyperonymes, hyponymes (pétrole / hydrocarbure / combustible fossile), synonymes contextuels ou partiels (pétrole / or noir, pétrole / hydrocarbure)

Coordination des descripteurs

Coordination

Composer les descripteurs simples, expressions plus complexes

Pré-coordination : coordination à l'indexation

- Dans le langage d'indexation
- Combinaison des descripteurs à l'indexation, mots composés, expressions, augmente la précision (exactitude)
- Ex: "Garage" / "réparation de voiture"

Coordination des descripteurs

Coordination

Composer les descripteurs simples, expressions plus complexes

Pré-coordination : coordination à l'indexation

- Dans le langage d'indexation
- Combinaison des descripteurs à l'indexation, mots composés, expressions, augmente la précision (exactitude)
- Ex: "Garage" / "réparation de voiture"

Post-coordination : coordination à l'interrogation

- Coordination des termes simples dans le langage requête
- Ex: langage booléen, *reparation* $\vee$ *voiture*

Liens entre les descripteurs

- Thésaurus : relier les descripteurs
Relations : hyperonymie (inclusion, inv. =hyponymie),
synonymie, holonymie (partie de, inv. =meronymie),
antonymie (contraire), eponymie (surnom), lien ("voir aussi")
- Dictionnaire : décrire, relier les mots
- Base de donnée lexicale : relier les lemmes (Wordnet)
- Terminologie : relier les termes
- Ontologie : relier des concepts
Formalisation à l'aide d'une logique (ex: CYC)
- "Ontologie lexicale" = Terminologie

TL pour produire les descripteurs

Le but d'un SRI n'est pas de comprendre les textes mais de satisfaire un utilisateur.

Critères pour l'usage du TALN en RI :

- Utilité/efficacité : effet sur la qualité des réponses
- Exactitude : ex. les racinisations approximatives
- Nature du traitement : syntagmes, analyse de surface, POS, ex: verbe vs nom
- Intégration au SRI : au niveau du modèle de correspondance ex: dérivation de termes

Beaucoup de travaux : [Salton, Smeaton, Allan, Fagan, Jacquemin, Krovetz ...] peu de réussites.

TL pour produire les descripteurs

Etapas des traitements :

- La construction des mots : morphologie.
- La construction des termes : terminologie.
- La construction des phrases : syntaxe.
- La construction du sens : sémantique.

Découpage plus détaillé

- Morphologie : morphème, mot
- Morpho-syntaxe : les parties du discours (PoS)
- Terminologie : terme dans un domaine, composé de mots, stabilité.
- Syntagme : rôle dans la phrase.
- Syntaxe : phrase complète.
- Sémantique : sens, signification.
- Concept : abstraction du sens.
- Pragmatique : contexte, savoir faire.

Niveau morphologique

Le plus bas niveau d'analyse ...

Unités atomiques des langues

Phonologie & Ecriture

- Phonèmes : unité phonétique d'une langue, unifier des sons différents (accents)
- Graphèmes : élément graphique atomique d'une écriture, unifier des visuels différents (variantes graphiques A,a)
- Pictogrammes (concret), idéogrammes (abstrait), phonogrammes (sonore, prononciation, lettre)
- Plusieurs graphèmes pour un phonème : digramme (en), trigramme (eau), n-gramme
- Morphogrammes : plus de rôle sonore mais grammatical (joliE, chanteNT,grandES)

Phonèmes & Morphème

- Phonèmes : pas de signification car reliés uniquement au sons d'une langue.
- Morphème : unité de signification atomique d'une langue

pollutio	salissure en latin	pollu-
-tion	le résultat de l'action	pollu-tion
-eur	l'agent (humain) de l'action	pollu-eur
-ant	participe présent, action simultanée	en polluant
-ant	nominalisation du participe	un polluant
	agent (non humain)	
dé-	invers	dépollution

Les mots

Mot

Combinaison de morphème, la plus petite forme linguistique ayant une autonomie

Propriétés des morphèmes

- La substitution : fum-er, fum-oir, fum-eur, fum-age, fum-iste
- La signification : chap-eau, bat-eau (mais -eau n'a pas de sens)

Composition des morphème complexe et cohérence :

- -oir : lieu, couloir, fumoir, abreuvoir, observatoire , (mais peignoir ...)
- mourir, la mort, un mouroir, un mourant, un meurtre

Morphèmes homonymes

- poire/poir-ier (arbre), plomberie/plomb-ier (personne), sucre / sucr-ier (contenant)

Morphèmes allomorphes :

- poire/poir-ier orange/orang-er

types de morphèmes

Morphèmes lexicaux ou lexèmes

classe ouverte, sémantique "forte" : chant, pollu-, lav-

Morphèmes grammaticaux

classe fermée, sémantique "faible" modifie le lexeme : -eur, -oir, -iste

Morphèmes libres

il peut seul constituer un mot : chant

Morphème lié

N'apparaît pas seul, et n'est pas un lexeme

Décomposition des mots

Morphologie

Analyse de la décomposition des mots en morphèmes

Un mot est soit :

- Un morphème libre
- Une composition de morphème en base et affixe
- Affixe : base + suffixe, préfixe + base, infixes (dans la base)
- La base n'est pas forcément un morphème libre
- en-clencher, dé-clencher
- irréalisable = ir- + réalisable = ir- + (réalis- + -able)

La base non réductible est appelé le radical, ou racine si moins de 3 lettres

réal- allomorphe de réel

Autres constructions des mots

L'emprunt d'une autre langue

- Suit alors une autre morphologie ou un mélange de deux morphologies superposées
- redingote (anglais riding-coat), adagio (italien), assassin (haschashin : les gens qui fument le haschisch)
- " 10.000 mots français sont passés dans la langue anglaise au moment de sa formation (Middle English, 1150-1500)", ex: combatant, enemy, memoir, homage, department, mortgage [J.D Paolini]

La composition de morphèmes libres : choux-fleur, pomme de terre, contredire, ...

Morphème et langue

Grande variété de construction des mots.

Français

Morphèmes sans "espace" : mots facile à identifier. Langue flexionnelle : des morphèmes grammaticaux bien séparés des morphèmes lexicaux

Vietnamien

Morphèmes avec des espaces (1 min). Langue "isolante" : mots invariables, pas de morphèmes grammaticaux. Difficile à repérer correctement : séparer correctement des mots

Allemand

Langue agglutinnante. Agglutination des morphèmes en mots, et des mots en termes ! Problème inverse.

Choisir le bon descripteur pour le modèle des documents

Phonèmes ? non car pas lié au sens

Morphèmes ? Possible, mais problème du découpage correct
(analyseur+ ressource lexicale)

Solution simplifiée : des n-grames découpés sans contraintes

Choix du morphème :

- Morphème lexicaux (lexemes) : porteur de sens, regroupe des variations de mots.
- Morphèmes grammaticaux : sens plus difficile à cerner (-oir dans couloir)

Extraction du descripteur

Segmenter un texte en unités linguistiques

- paragraphe, phrase, mot
- Imbrication syntaxe et sens : la ponctuation est un délimiteur de phases mais apparaît aussi dans les abréviations.

Qu'est-ce qu'un mot ?

- Combinaison de morphème, la plus petite forme linguistique ayant une autonomie, groupe de lettre ayant une unité de sens, couper ce mot fait perdre du sens, ex: "Hot dog"
- Découpage non évident dans les langues agglutinatives :

Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz
(loi sur le transfert des obligations de surveillance de l'étiquetage de la viande bovine)

Traiter la variation : les mots

La racinisation

Réduire le mot à une racine

Approximation par un algorithme (ex: Porter)

- Algorithme approximatifs sans analyse morphologique
- 2 étapes : 1) traitement des affixes 2) normalisation du lemme
- Réduction quelque soit la catégorie
- un polluant, en polluant, pollueur, pollution, polluer → pollu
- Eliminer les flexions : les suffixes

Augmente le rappel, diminue l'ensemble des descripteurs

Traiter la variation

Variation morphologique

- Si on connaît la catégorie (Partie du discours) possible de dériver le mot vers une forme unique sans changer de catégorie.
- Lemmatisation : le nombre, le genre (nom, adjectif), le mode le temps, la personne (verbe)
- " Les pollueurs doivent enlever les hydrocarbures"
- " le pollueur devoir enlever le hydrocarbure"

Erreurs liées à la segmentation ou la catégorisation

- " La bande dessinée dans la presse"
- " la bande dessiner dans le presser"

Segmentation des mots

La qualité de la segmentation conditionne le reste de l'indexation, et donc le SRI lui même.

Méthode de segmentation

- ressource explicite : un lexique
- Analyseur morphologique + statistique : la forme permet de prévoir la catégorie des mots, la statistique permet de détecter les stabilités susceptible d'être liée à un mot
- Parfois besoin de la syntaxe ou de la sémantique : "il la note", "la note est un la"

Morphologie : Utilité en RI

- Connaître la morphologie de la langue pour définir clairement le descripteur
- Analyser la morphologie avec exactitude : Analyseur morphosyntaxique
- Analyser la morphologie approximativement : Enracineur algorithmique
- Des améliorations sur certaines requêtes, mais pas de différence en moyenne
- Dépend fortement des langues [GAU00,SAV99]

Les entités nommées

Unité lexicale particulière

- Fait référence à une entité unique du monde concret
- domaines spécifiques notamment humains, sociaux, politiques, économiques ou géographiques

Noms propres ou Valeurs typées

- Organisations, lieux, personnes, ...
- Poids, longueurs, monnaies, ...

Important en RI car très discriminant !

Reconnaissance des entités nommées

- Liste, ou patrons syntaxique
- Variation importante : abréviations, acronymes (CGT, orthographe (transcription autre écriture), métaphore ("big blue", "Matignon")
- Peut se muter en substantif : "S.M.I.C", "le smic", "un smicard"

Terme & Syntagme

Indexation par syntagme (phrase indexing)

- Unité syntaxique de taille inférieur à la phrase
- Syntagme nominal pour descripteur

Termes : syntagme nominal avec une signification précise (unique) dans un domaine (ex: "système expert")

- Plus précis pour l'indexation
- Sujet à des variations importantes

Listes des termes possibles : terminologie, avec une structuration similaire à un thésaurus.

Terminologie calculée : Sélection des syntagmes dans un corpus cohérent d'un domaine

La variation

Comme les morphèmes, les syntagmes sont sujets à variation:

- Morphologique (cf. précédemment).
- Lexicale : gasoil / diesel, pollution de l'air / pollution atmosphérique
- Syntaxique : pollution de l'air par les moteurs diesels / les diesels qui polluent l'air

Important : prendre en compte cette variation sinon le gain en précision est compensé par une perte en rappel

Niveau syntaxique

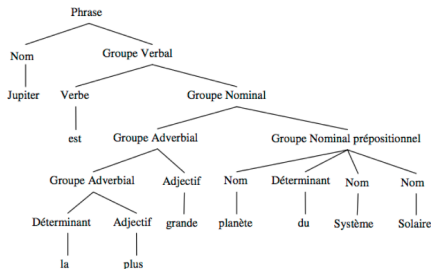
Structure globale (distribution des rôles) des expressions : les phrases, les requêtes, ...

Analyse de surface

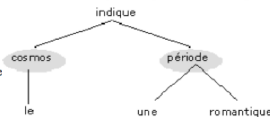
- Analyse partielle : pas l'ensemble de la phrase, ex: syntagmes
- Analyse robuste : toujours donner un résultat
- Améliorations dépendent des requêtes et de la collection
- Résolution de phénomènes langagiers
 - Ex: anaphore grammaticale = reprise sémantique par un syntagme d'un énoncé précédent : "La sonde thermique ... cette sonde ... ce capteur"
 - Anaphore pronominale : "La sonde thermique elle"

Types de grammaires

- Grammaires Symboliques (vs: Statistiques)
- Grammaires de constituants (ou syntagmatiques)
règle décrivant la construction/analyse d'une phrase en constituants
- Grammaires de dépendances
Relation de dépendances en les mots



"Le cosmos indique une période romantique."



But de l'analyse syntaxique pour la RI

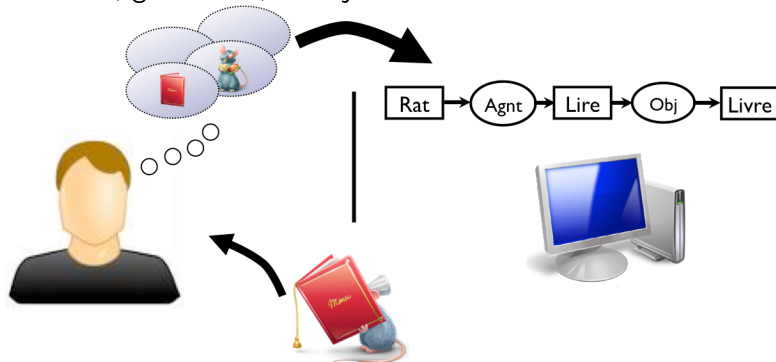
- Si la syntaxe influence la recherche
 - Le cas pour les requêtes. Ex : syntaxe langage d'interrogation.
 - Pas le cas en général ... [Maisonnasse06]

But de l'analyse syntaxique pour la RI

- Si la syntaxe influence la recherche
 - Le cas pour les requêtes. Ex : syntaxe langage d'interrogation.
 - Pas le cas en général ... [Maisonnasse06]
- Si la syntaxe permet de détecter les descripteurs :
 - Les termes *fracture du fémure*, généralement avec des ressources
 - Les syntagmes : si le modèle de RI s'y prête (voir plus loin).
 - Si les descripteurs sont des structures sémantiques (Logique de descriptions, graphes conceptuels)

Concepts

Abstraire, généraliser, ... objectiver dans un formalisme.



Concepts vs Acception

Acception

Signification d'un mot à partir de son usage dans la langue (Ex: synsets de Wordnet)

Concepts

Entité abstraite qui unifie et résume un ensemble d'objets concrets ou mentaux par abstraction de traits communs pertinents (Ex: concepts de CYC)

Plan

- 1 Introduction
 - Documents, information et signal
 - RI Sémantique
 - Questionnements par l'exemple
- 2 Descripteurs : atomes des index
 - Niveau morphologique
 - Niveau terminologique
 - Niveau syntaxique
 - Niveau sémantique et conceptuel
- 3 **Ressources**
 - Ressources lexicales extraites
 - Ressources conceptuelle non formelles
 - Ressources conceptuelles formelles
- 4 Correspondance sémantique
- 5 Conclusion

Ressources

Besoins de ressources :

- Lexicales : avec une structure (ex: WordNet)
- Conceptuelle : avec un lien vers le lexique (les termes)

Origine des ressources

- Explicites : construites manuellement (WordNet, UMLS, CYC)
- Semi-Explicite : à partir de ressources non formalisée (DBpedia, à partir des info-box)
- Extraites automatiquement : la couverture d'une ressource est toujours incomplète.

Extraction de ressources

Basé sur du texte plein : pas seulement les données des "info-boxes".

Exemple : thésaurus d'association.

- Basé sur la notion de contexte : deux termes ont un lien s'ils sont souvent utilisés dans le même contexte
- Contexte Documentaire : partager des zones de textes ou le terme apparait Taille du contexte : fenêtre, phrase, paragraphe, document
- Contexte lexical : Partager la proximité, l'usage entre mots
Ex: les sujet de certains verbes

Contexte documentaire

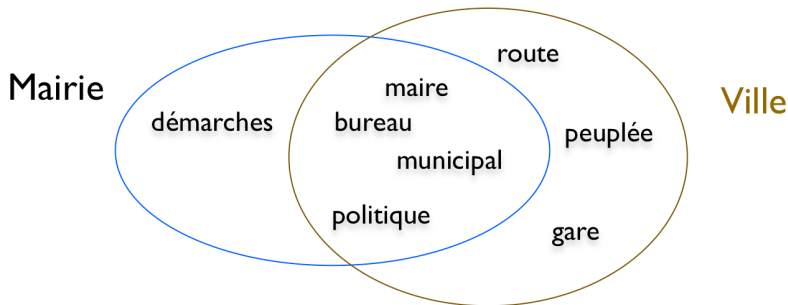
L'association entre deux termes est proportionnelle au contexte partagé.



Paramètres

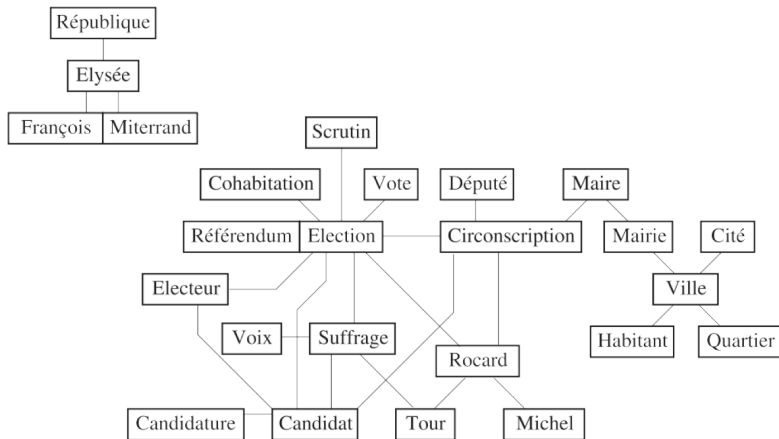
- Portée : taille du document
- Sélection : filtres appliqués aux contextes (ex: linguistiques)

Contexte lexical

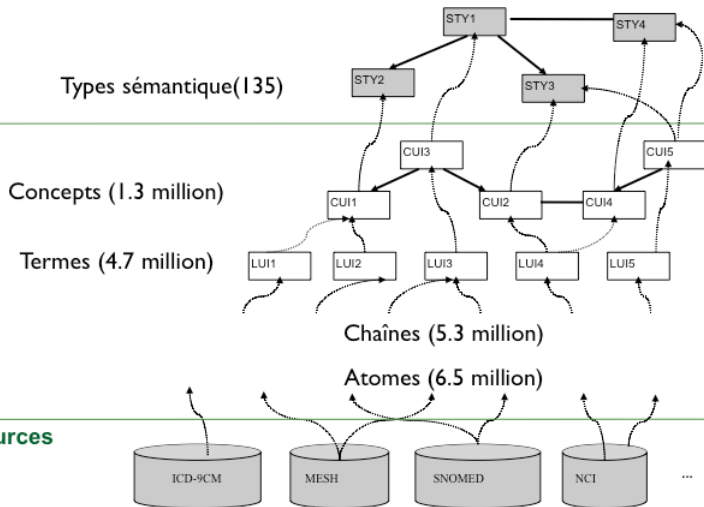


Mesure de l'association à base de "confiance/support" (data mining)

Exemple de résultat



Méta-thésaurus : UMLS



Constructions de ressources conceptuelle non formelles

- Cout très élevé de leur construction
- Ex: ULMS, mise en commun de multiples ressources construites séparément.
- Besoin d'une mise à jour constante, gestion et correction des erreurs (UMLS, 2 versions par an)
- Effort limité à certains domaine ou les finances sont disponibles (ex: domaine médical).

Utilisations des ressources conceptuelle non formelles

Malgré tous les efforts ... structures trop incomplète pour pouvoir être utilisé pleinement.

- Relations non formelles dans le thésaurus : quel sens ?
- Beaucoup d'incohérence, ex: boucles dans le "isa"
- Relations sémantiques globalement appliquées aux classes de termes (UMLS)
- Termes (concepts) construits mais schéma de construction non explicite.
 - Ex: *fracture du fémur* reliée à *fracture*, puis à *pathologie*, mais rien d'explicite concernant un lien vers l'anatomie.

Utilisations des ressources conceptuelle non formelles

- Pour palier à ces problèmes
 - Solutions hybrides : index conceptuel, index terminologique, index morphologique
 - Travailler sur le formalise et la fonction de correspondance.

Ou aller vers une expression formelle des descripteurs et des ressources (ontologies formelles).

Logique de descriptions (Web sémantique)

Modélisation des connaissances dans un formalisme logique.

- Logique de description
- Compromis entre puissance de description, et complexité des calculs.
- Fragment de la logique du premier ordre.

Base de connaissance : une ontologie peuplée, car avec un ensemble de faits.

- Sémantique d'un domaine : concepts, rôles (relations), axiomes, attributs et individus.

Logique de descriptions : T-Box

Définition d'un ensemble de concepts C . Un concept est interprété comme un sous ensemble de l'ensemble d'interprétation. C'est un prédicat unaire

Ex: $Personne(a)$, $Peau(p)$

Un rôle est un prédicat binaire : $\exists cancerDe.Peau$

T-Box (Terminological Box)

L'ensemble des concepts, des rôles et leurs axiomes.

Ex: $Maladie, Artrose \sqsubseteq Maladie$, $Homme \sqsubseteq Personne$,
 $Femme \sqsubseteq Personne$,

Construction d'une T-Box

- Inclusion de concepts : *Artrose* $\sqsubseteq$ *Maladie*
- Equivalence de concepts *Melanome* $\equiv \exists \text{cancerDe}.\text{Peau}$
- Equivalence et intersection : *Jambe* $\equiv \text{Membre} \sqcap \text{Inferieur}$
- Inclusion de rôles : *cancerDe* $\sqsubseteq$ *LesionDe*

Sémantique (dénotationnelle):

- Un concept est un ensemble d'individu, un sous ensemble de Δ
- Un rôle : un ensemble de relations, sous ensemble de $\Delta \times \Delta$

Détail de la sémantique

Fonction d'interprétation I :

$$I(C) \subseteq \Delta$$

$$I(\top) = \Delta$$

$$I(\perp) = \emptyset$$

$$I(\neg C) = \Delta \setminus I(C)$$

$$I(C \sqcap D) = I(C) \cap I(D)$$

$$I(C \sqcup D) = I(C) \cup I(D)$$

$$I(\exists R.C) = \{a \in \Delta : \exists b \in \Delta \text{ tel que } b \in I(C) \text{ et } (a, b) \in I(R)\}$$

$$I(\forall R.C) = \{a \in \Delta : \forall b \in \Delta \text{ si } (a, b) \in I(R) \text{ alors } b \in I(C)\}$$

Assertion-Box (A-Box)

- La T-Box est l'ontologie, les concepts et leur organisation.
- La A-Box déclare des instances.

Exemple : *Personne(marie)*, *Peau(p1)*, *Maladie(m1)*,
cancerDe(m1, p1)

La base de connaissance : union de la T-Box et la A-Box.

Le gros intérêt de la Logique de description : le raisonnement !

- Inférence de propriétés d'individus :
 $Maladie \equiv \exists cancerDe.\top$, si $cancerDe(x, y)$ alors $Maladie(x)$.
- Inférence de relations entre individus.
- Subsumption de concepts :
 $Malade \equiv \exists atteindPar.Maladie$,
 $\exists atteindPar.Lesion \sqsubseteq Personne$, $Lesion \sqsubseteq Maladie$ alors on a
 $Malade \sqsubseteq Personne$.
- Calcul automatique d'une hiérarchie de subsumption.
- Test de cohérence de l'ontologie (T-Box), ne don redondance , etc.

Avantage / inconvénients de la logique de description

- Logique : sémantique explicite, fragment de la logique des prédicats (ordre 1)
- Modélisation explicites des relations : pas possible dans le logique des propositions (ordre 0)
- En fonction des opérateurs : beaucoup de choix de logiques possibles.
- Modulation : complétude expressions vs complexité du raisonnement
- Raisonnement : décidable (en général)

linconvénients de la logique de description

Mais :

- Relations : choisir un sens privilégié de la relation, l'interprétation d'un rôle est toujours un sous ensemble de δ .
- Logique expressives : complexité Non déterministe Polynomiales (NP), ou Exponentielles (EXP), ou même Non déterministe Exponentielles (NEXP) : impossible de passer à l'échelle pour un SRI.

Comment faire le la RI ...

... avec des concepts ...

... des logiques de descriptions ...



Plan

- 1 Introduction
 - Documents, information et signal
 - RI Sémantique
 - Questionnements par l'exemple
- 2 Descripteurs : atomes des index
 - Niveau morphologique
 - Niveau terminologique
 - Niveau syntaxique
 - Niveau sémantique et conceptuel
- 3 Ressources
 - Ressources lexicales extraites
 - Ressources conceptuelle non formelles
 - Ressources conceptuelles formelles
- 4 Correspondance sémantique
- 5 Conclusion

SRI à base de concepts

- Etape de désambigüisation : Sanderson montre d'un taux de désambigüisation inférieur à 90% dégrade le SRI
- Difficile de faire fonctionner seul un SRI à base d'acceptions
Ex: Baziz qui est forcé de fusionner son index avec des descripteurs classiques à base de mots

SRI à base de concepts

- Grande complexité du traitement : temps important d'analyse.
- Risque d'erreurs à toutes les étapes : risque de contre sens, perte de rappel
- Des outils de TLN plus dédiée à la RI : rapidité vs précision, conserver les alternatives,

Exemple d'outil : Metamap (Mélange, Java C et Prolog)

SRI à base de concepts

Combinaison complexes de descripteurs :

- Terme structuré en tête + arguments
- Arbre de dépendance syntaxique
- Terme d'une logique de description
- Graphes conceptuels [Sowa]

Revoir le processus de correspondance sac de concepts peu performant ...

Influence sémantique sur les modèles et la correspondance

Modèle et correspondance, évoluer pour prendre en compte :

- Le sens véhiculé par les requêtes et les documents.
- Des informations complémentaires non présente dans les documents mais indispensable pour en comprendre le sens.

Influence sémantique sur les modèles et la correspondance

Modèle et correspondance, évoluer pour prendre en compte :

- Le sens véhiculé par les requêtes et les documents.
- Des informations complémentaires non présente dans les documents mais indispensable pour en comprendre le sens.

La correspondance basée sur la statistique de l'intersection de "sac d'index" n'est peut être plus suffisante.

- Des représentations des documents et requêtes avec une structure (dépendances)
- Une opération de correspondance avec une notion de déduction.

Dédution du document vers la requête, pas nouveau ...

Propositions de Van Rijsbergen :

- Voir la correspondance RI comme une déduction logique :
 $D \rightarrow Q$
- Estimer cette plausibilité : $U(D \rightarrow Q)$
- Suppose que l'on utilise des probabilités : $P(D \rightarrow Q)$

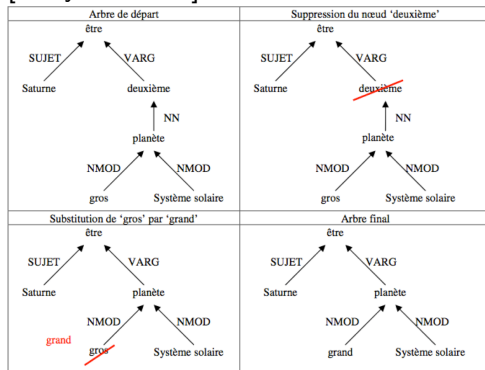
Sa proposition informelle :

Given any two sentences x and y ; a measure of the uncertainty of $y \rightarrow x$ relative to a given data set is determined by the minimal extent to which we have to add information to the data set, to establish the truth of $y \rightarrow x$.

Correspondance par dérivation d'arbre

Saturne est la deuxième plus grosse planète du Système Solaire.
 Jupiter est la plus grande planète du Système Solaire.

[Punyakanok04]



Correspondance par dérivation de termes

"La séparation de la république fédérale tchèque"

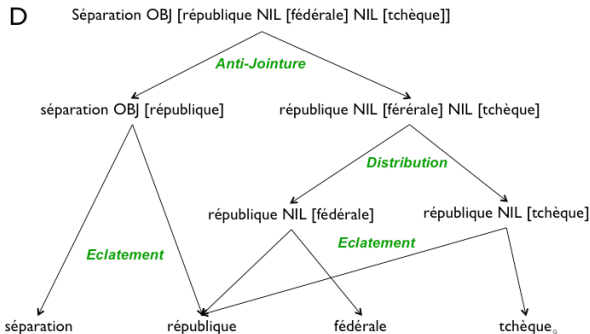
- Calcul de probabilité de correspondance sur l'arbre de dérivation
- inspiré des réseaux Bayésiens
- Probabilité d'un terme composé t , en fonction d'une combinaison Π des composants S de t : $p(t|\Pi(S))$.

Correspondance par dérivation de termes

$D = \dots$ la séparation de la république fédérale tchèque $\dots$

$Q = \text{"république tchèque"} + \text{"séparation"}$

Prouver $D \rightarrow Q$, puis évaluer cette dérivation



Correspondance dans les Logiques de descriptions

Recherche sur des "faits" :

- Interrogation "exacte" avec des requêtes structurée dans une logique formelle.
- Documents sous la forme de triplets.
- Basé sur un moteur d'inférence.

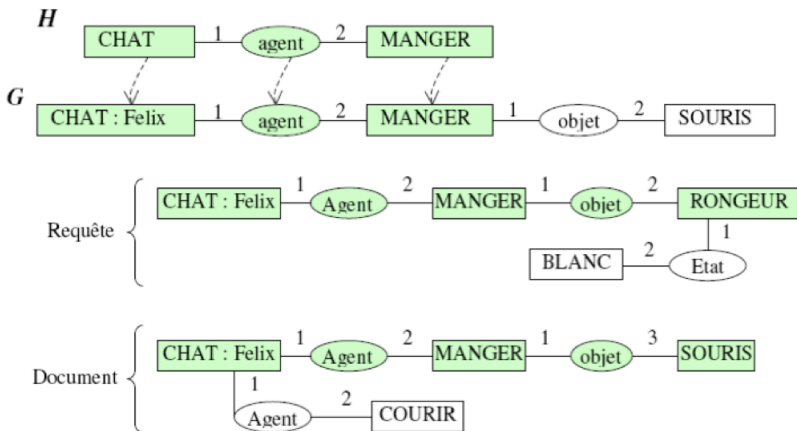
Exemple :

- RDFS : transitivité de propriétés (subClassOf), raisonnement à base de règles.
- Web Ontology Language (OWL) : raisonnement déductifs (concept vers instances, subsomption) ou inductifs (classification, requêtes conjonctives).
- OWL-Lite : raisonnement simple sur les propriétés, inverse, transitivité, symétrie, ..
- OWL-DL : en plus raisonnement sur les classes, intersection,

Graphes conceptuels

- Construction à base d'opérateur graphiques : jointure, copie, spécialisation du type de concepts.
- Equivalence avec le logique des prédicats (ordre 1)
- Opérateur de projection : équivalent à l'implications logique.
- Calcul de la projection : NP en général
- Pas de solution facile pour évaluation de la projection : projection partielle.

Graphes conceptuels



Plan

- 1 Introduction
 - Documents, information et signal
 - RI Sémantique
 - Questionnements par l'exemple
- 2 Descripteurs : atomes des index
 - Niveau morphologique
 - Niveau terminologique
 - Niveau syntaxique
 - Niveau sémantique et conceptuel
- 3 Ressources
 - Ressources lexicales extraites
 - Ressources conceptuelle non formelles
 - Ressources conceptuelles formelles
- 4 Correspondance sémantique
- 5 Conclusion

Conclusion

- Concepts formels : la meilleure solution théorique
- ... mais :
 - Comment obtenir les structures conceptuels avec des texte en LN ?
 - Quel est la modélisation du calcul de la correspondance ?
 $P(D \rightarrow Q)$?
 - Quel type d'évaluation pour classer les résultats ?
 - Comment obtenir une large ontologie formelle ? ex:
transformer UMLS en OWL ?
 - Comment avoir des temps de calcul compatible avec la RI interactive ?

Avancées très lente de ce domaine

Conclusion

- Utilisation de concepts non formels
- Dédution had-hoc : utilisation des relations sémantiques pour étendre documents ou requêtes
- Beaucoup d'utilisation de modèles de RI classiques : modèle vectoriel
- Compenser les erreurs et incomplétudes par une RI classique à base de mots.

Encore du travail de recherche