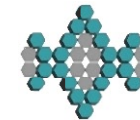


À quand une vraie conversation avec le web?

Anne-Marie Di Sciullo, FRSC
di_sciullo.anne-marie@uqam.ca
www.interfaceasymmetry.uqam.ca

Conseil de recherche en sciences humaines du Canada (GTRC)
Fonds de recherche sur la société et la culture (FQRSC)
Université du Québec à Montréal (FIR)
Fédération sur le traitement des langues naturelles



FÉDÉRATION
SUR LE TRAITEMENT
DES LANGUES NATURELLES



Systemes de traitement du contenu de l'information

- Systemes de dialogues avec des robots
- Systemes de réponse automatique à des questions (SQR)
- Recherche et extraction de l'information (IR/IE)



Architecture

- Les limites des systèmes actuels de traitement de contenus
 - Ramona et le test de Turing
 - START (MIT) et la sémantique extensionnelle
 - Google et la sémantique relationnelle
 - Le rôle du web sémantique
- Les enjeux pour les technologies de l'information d'avant-garde
- La contribution de la sémantique des langues naturelles



Ramona: le robot conversationnel KurzweilAI.net, une personnalité virtuelle?



Test de Turing



- **L'intelligence des machines** (Turing 1950)
- Reformulation de la théorie de Descartes sur l'intelligence humaine et la capacité de manipuler le langage de manière flexible.
- **Test de Turing**: le jeu de l'imitation
Un interrogateur, un humain, une machine.
Si l'interrogateur ne peut décider, sur la base des interactions verbales répétées, lequel est l'humain et lequel est la machine.

The Turing Test

Verbal Behavior as the Hallmark of Intelligence



edited by **STUART SHIEBER**



© 2005 Anne-Marie Di Sciullo

Fragments de conversation avec Ramona

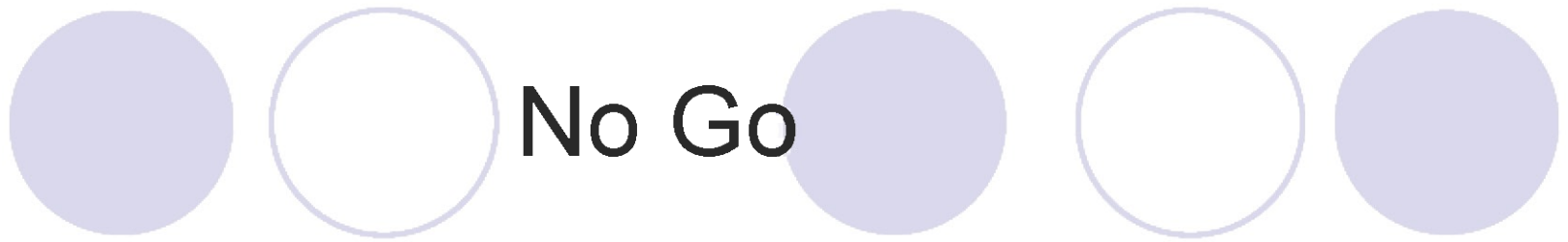
.....

- User: That's very discreet of you. You are obviously a polite young woman. How old are you, anyway?
- Ramona: In any case, I'm 21.
- User: Would you agree that you are polite and discreet?
- Ramona: I'm not so sure if I would like to agree that you are polite and discreet.
- User: Not me. You. You seem very polite to me.
- Ramona: You think I'm polite to me.
- User: Do you have trouble with pronouns, Ramona?
- Ramona: No I don't have trouble with pronouns Ramona.

.....

Fragments de conversation avec Ramona

-
- Ramona: Would you like to know more?
User: yeah tell me more about yourself
Ramona: What would you like to know? I do not know what yourself is. Ok. Ask me something.
-
- Ramona: Let's talk about something else
User: OK - how 'bout Bach?
Ramona: I'm not sure if I can quantify the answer.
-
- User: What time is it?
Ramona: Time does not exist.
User: Where is gravity?
Ramona: What holds us to the earth.
User: What is an idea?
Ramona: I do not know what idea is. Ok. Do you dream?
-



- Ramona ne passe pas le test de Turing, malgré le fait qu'elle semble manipuler le langage de manière flexible, comme le font les humains, dans certains cas.
- Ramona ne peut notamment traiter les entités qui sont dénotées par des pronoms et par les anaphores pronominales, alors qu'un humain est en mesure de le faire facilement.

Anaphore pronominale

La résolution de l'anaphore pronominale est établie sur la base des asymétries des domaines de structures d'arguments.

(1) [[the chief officer] trusts [himself]]
└──────────────────────────┘

(2) [[the chief officer] trusts [him]]
└──────────x──────────┘

(3) [[the president] thinks [that [the chief officer] trusts [nim]]]
└──┘

(4) [[[the president] talked to the members of the company
└──┘

today. The reactions of the shareholders were unequal. The minutes
.....
of the meeting indicate [that [the chief officer] trusts [him]]
.....

Interface syntaxe-sémantique

La résolution de l'anaphore pronominale dépend des traits morphosémantiques des antécédents et des pronoms.

- (5) a. The president bought a Mercedes. They are solid cars.
b. I need a rewritable CD. Where did you put them?
- (6) a. #I saw the president's Mercedes. They are solid cars.
b. #I need the rewritable CD. Where did you put them?
- (7) a. The Mercedes is a nice car. They sell well to corporate.
b. The rewritable CD is popular. They sold well this year.
- (8) a. The Mercedes is a nice car. It sells well to corporate.
b. The rewritable CD is popular. It sold well this year.

Les DP indéfinis et les DP pluriels dénotent des ensemble d'individus Ils peuvent ainsi être les antécédents de pronoms pluriels.



Connaissance du monde

La connaissance linguistique détermine les propriétés basiques de l'anaphore discursive, alors que la connaissance du monde réel est parasite.

- (9) The President's party was a success. They celebrated until 2 AM.
- (10) The CEO talked to the CLO. He became a millionaire within a time-frame of four years.

Les résultats d'études indépendantes corroborent ceci. Lappin et Leass (1994) montrent que la connaissance du monde réel ne fait qu'améliorer marginalement la performance de l'algorithme RAP basé sur les propriétés syntaxiques (2%).

MARS et RAP

- Les approches probabilistes pour la résolution de l'anaphore discursive ont des résultats médiocres.

MARS (Mitkov Anaphora Resolution System) (Mitkov 2002) ne peut identifier l'antécédent des pronoms dans les mini discours (5) à (10).

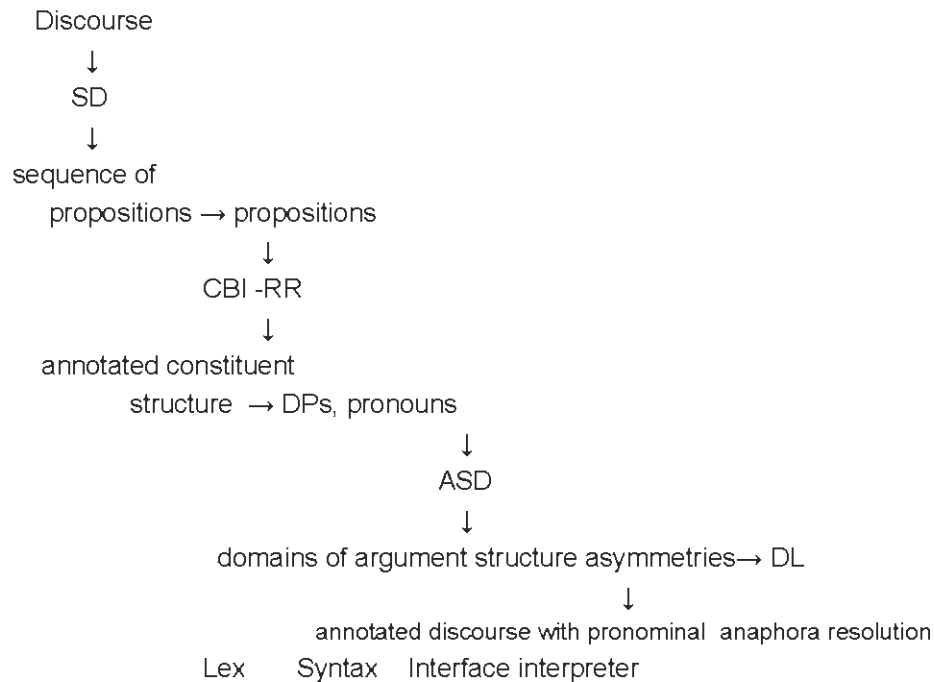


Les approche hybrides tels que RAP (Lappin et Leass 1994) qui n'incorporent pas les domaines d'asymétries argumentales ont un meilleur taux de succès, mais ne sont pas optimaux.

SRA du futur

- Résolution de l' anaphore pronominale par identification des relations asymétriques que le pronoms entretiennent avec leurs antécédents.

Architecture of Discourse Pronominal Anaphora Interpreter (Di Sciullo 2005: 320)



Systeme Question-Réponse

Input: les questions soumises par l'utilisateur

Recherche de l'information dans une base de connaissance

Output: les résultats reliés à la question soumise par l'utilisateur sous forme de réponse



Types de SQR



- Type 1: Interfaces langues naturelles/base de données**
 - Base de connaissances = bases de données structurées**
 - Traduction des requêtes en SQL**
 - Domaine de recherche limité au type d'information dans la base de connaissance**
 - Est relativement efficace**
- Type 2: Systèmes de dialogue interactif**
 - La source de connaissance peut être une base de données ou de l'information non structurée étalée sur des collections de textes (web)**
 - Le domaine de recherche est plus restreint**



SQR de type 2

- Answer Bus
- AskJeeves
- IONAUT
- START
- QuASM
- WebQA

START
Natural Language Question Answering System

QuASM? EDU

Fausse impression



- ❑ Les SQR peuvent donner la fausse impression, uniquement au moyen de techniques statistiques, qu'ils manipulent la connaissance linguistique (e.g syntaxe).

Mais ce n'est pas le cas.

Test:

Q: Who killed John F. Kennedy?

Q: Who did J.F. Kennedy Kill?

Résultat: les mêmes adresses de documents.

- ❑ Les systèmes de question-réponses tels que START (Katz 1997) ne sont pas en mesure de traiter les questions incluant des prépositions orphelines, ni les questions qui incluent des présuppositions fausse dans le monde défini par la collection des documents questionnés (Di Sciullo et Aguerro 2003).

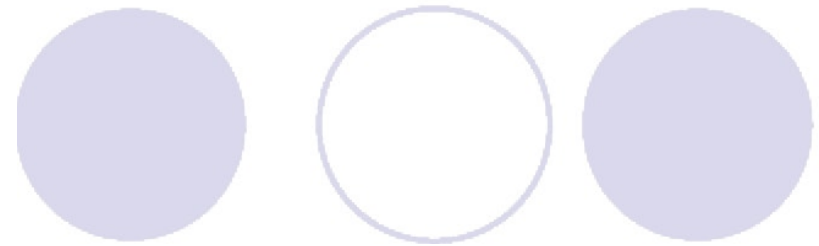
Evaluation

A decorative graphic at the top of the slide consists of two rows of circles. The top row has a solid light purple circle on the left and an outlined light purple circle on the right. The bottom row has a solid light purple circle on the left, an outlined light purple circle in the middle, and a solid light purple circle on the right.

- AnswerBus, AskJeeves, IONAUT ont au mieux les capacités des moteurs de recherche
- Output: liste ordonnée de liens à des documents qui peuvent ou non contenir la réponse à la question
- L'ordonnancement est souvent basé sur les opérations booléennes de conjonction et de disjonction des mots de contenu (e.g. verbes et noms) qui font partie de la question soumise par l'utilisateur.

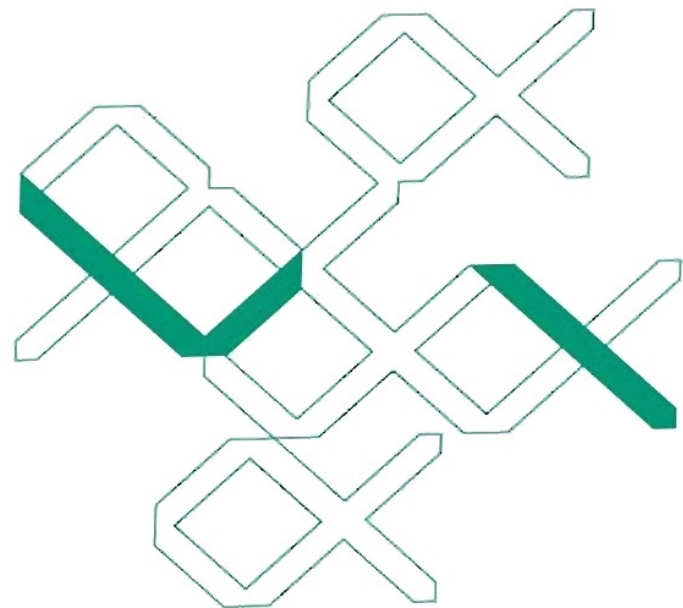
SQR du futur

Relations asymétriques telles que 'être un sous-ensemble propre de' et d'autres relations sémantiques telles que 'être une variable de' sont nécessaires pour le développement des SQR.



Asymmetry in Morphology

Anna Maria Di Sciullo



Recherche et extraction de l'information

- Input: requêtes soumises par l'utilisateur
- Indexation
- Recherche de l'information dans une collection
- Output: les résultats reliés au sujet de la requête sous forme de liste ordonnée de liens à des documents



Propriétés des systèmes IR/IE courants

- Recherche par mot clés
- Utilisation d'opérateurs de Boole (Boolean et extended Boolean)
- Utilisation de méthodes probabilistes (réseaux d'inférences, Naive Bayes (Autonomy, Microsoft))
- Utilisation de taxonomies (catégorisation (YellowBris, GammaSite))
- Utilisation d'ontologies (représentation de la connaissance (InQuira, iPhrase, Verity))
- Utilisation de méthodes calcul de similarité (Clustering (Autonomy, Vivisino, GammaSite))
- Utilisation de calculs de proximités (Vector Space (Google, Verity))
- Utilisation de propriétés NLP (Named entity extraction (InQuira, InXight))
- Utilisation d'analyse linguistique (stemming, morphologie, traitement des synonymes, correcteur orthographique)



Résultats

Requête: the development of e-commerce in Italy

Résultats:

[B2B Tech-Press Releases; B2B web development, ecommerce ebusiness ...](#)

... was born in **Italy** and has split his time between **Italy** and Kenya. ...

B2B Technologies is a privately held Internet consulting and **development** company specializing ...

b2btech.com/press-111700.asp [cached](#)

[Magic Software Italy- Application development tools and ...](#)

... An investments and business **development** group dealing in **eBusiness** solutions all ...

www.magicsoftware.com/corporate/executives/default.jsp?branch=it

[cached](#) | [more results from this site](#)

[Activeunit Web Development Company - E-commerce](#)

... ActiveUnit can build custom **Ecommerce** applications and ...

PayGate » GZS:PayMaster » NR:**e-commerce** » PayBox » Protx ...

2003 Activeunit Web Site **Development** Company ...

www.activeunit.com/e_commerce.html [cached](#)

Les opérations booléennes et la non préservation des relations sémantiques

Illustration:

(11) the funding of research by the government

(12) \neq the funding of the government by research

(13) a. \approx The governmental funds for research

b. \approx Government's funds for research

c. \approx The government funding the researchers

d. \neq the governing researchers funds

e. \neq the research funds for the governor

f. \neq the governor researchers funds

IR/IE du futur



- Incorporation de la connaissance riche sur les relations asymétriques de la syntaxe et de la sémantique des langues naturelles
 - (14) **Kennedy's** pictures
Les photos **de Kennedy**
 - (15) processing natural languages and understanding **them**
traiter **les** langues naturelles et **les** comprendre
- Traitement de la variation dans la distribution des éléments fonctionnels qui supportent ces relations.

Sémantique, Web- sémantique

- Dans cette perspective, il est possible d'envisager le développement d'une sémantique des contenus web qui permettent de vraies conversations avec le web, ce que ne fait pas le web sémantique (Berners-Lee, Hendler et Lassila 2002, Koivunen et Miller 2001, Miller 2005).





Objectifs du web sémantique

- Fournir une plateforme accessible qui permette de partager et de traiter l'information par des outils automatiques aussi bien que par des humains
- Définir et relier les données sur le web afin d'améliorer la recherche d'information
- Développer de nouveaux langages qui rendent plus accessible l'information sur le web
- Développer de nouvelles manières d'assister l'utilisateur du web (calendrier électronique intelligent, agenda électronique intelligent, secrétaire intelligent, organisateur de colloque intelligent, etc.)

Comment?

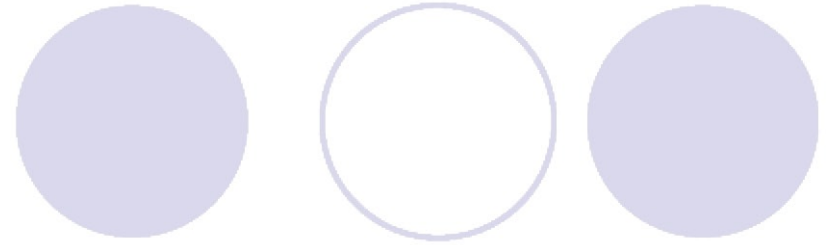
A decorative graphic at the top of the slide consists of two groups of three circles. The first group on the left has a solid light purple circle on the left, a white circle with a light purple outline in the middle, and a solid light purple circle on the right. The second group on the right has a solid light purple circle on the left, a white circle with a light purple outline in the middle, and a solid light purple circle on the right.

- Lier les bases de données et les services web
- Assurer le partage de contenu entre des applications différentes (XML (Extensible Markup language); Schema))
- Protocoles de standardisation (RDF, SOAP)
- Construire des hiérarchies et des thésaurus qui peuvent être utilisés pour relier les termes les uns aux autres (Schemas-RDF, WebOnt)
- Construire un langage qui exprime la connaissance du monde (CyC)

Développements du web sémantique

- ❑ Liaison de bases de données
 - (<http://www.w3.org/XML>)
- ❑ Partage de contenu entre applications utilisant différents XML DTDs ou Schemas
 - (<http://www.w3.org/XML/Schema>), (<http://www.w3.org/RDF>)
 - (<http://www.w3.org/TR/SOAP>)
- ❑ Combinaison de services web
 - (<http://www.w3.org/TR/rdf-schema/>)
 - (<http://www.w3.org/2001/sw/WedOnt>)

Mérites et limites



☐ Mérites:

- Uniformisation
- Standardisation

Semantic Web
Activity

☐ Limites:

- Réduit la sémantique des langues naturelles à des relations lexicales sémantiques superficielles
- Absence de traitement propre à la sémantique des langues naturelles malgré la diversité des langues

Notation sémantique utilisant TAP

``The Chicago Bulls announced yesterday that Michael Jordan will...''

The `<resource ref="http://tap.stanford.edu/BasketballTeam_Bulls">Chicago Bulls</resource>` announced yesterday that `<resource ref="http://tap.stanford.edu/AthleteJordan,_Michael">Michael Jordan</resource>` will...''

Thus, the annotation:

`<resource ref="http://tap.stanford.edu/Athlete Jordan,_Michael">Michael Jordan</resource>` says that the string ``Michael Jordan'' refers to the resource whose URL is

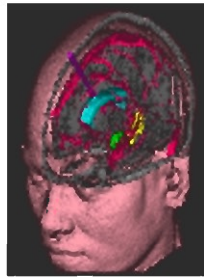
``http://tap.stanford.edu/AthleteJordan,_Michael.'' It is expected that querying this URIL will result in encoded information which provides greater detail about this resource.

Web sémantique



- Permet d'articuler la connaissance du monde, qui n'est pas innée, mais acquise
- Changer le monde d'interprétation, changer le Web sémantique, mais pas la sémantique des langues naturelles
- Le traitement de la logique propre aux langues naturelles n'est pas coextensif au traitement de l'information encyclopédique visé par le web sémantique, mais demeure une condition nécessaire à un vrai dialogue avec le web

Le futur



- Le traitement conceptuel des expressions linguistiques s'appuie sur des relations irréversibles de la Faculté du langage.
- L'inclusion de ces relations dans les systèmes de traitement de l'information ne peut qu'accroître leur performance et les rapprocher de la performance humaine qui, toutes choses égales par ailleurs, est optimale.
- Le paradigme relationnel a en outre l'avantage de rendre possible, via les relations d'asymétrie, un traitement unifié de la parole et du texte.
- Il permet aussi d'entrevoir le traitement de l'information multilingue non pas comme la somme de traitements linguistiques particuliers, propres à chaque langue traitée, mais bien comme un traitement générique, paramétrable selon la langue traitée.

À quand une vraie conversation avec le web?

Les machines passeront le test de Turing d'ici 2029, selon Mitchell Kapor et Raymond Kurzweil (Kurzweil et Kapor 2002)

Lorsque les relations sémantique propres aux langues naturelles pourront être traitées de manière flexible.

