

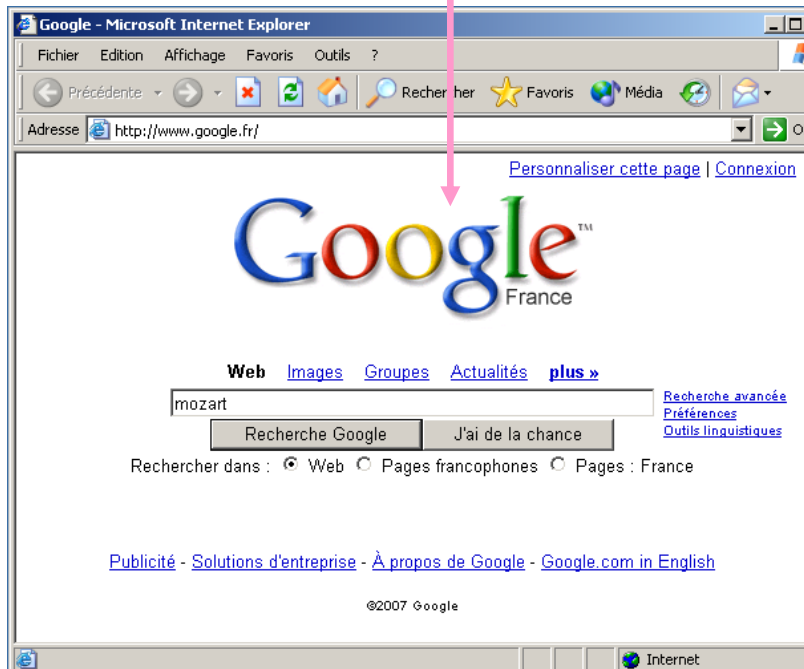
Les systèmes de question-réponse

Contexte général

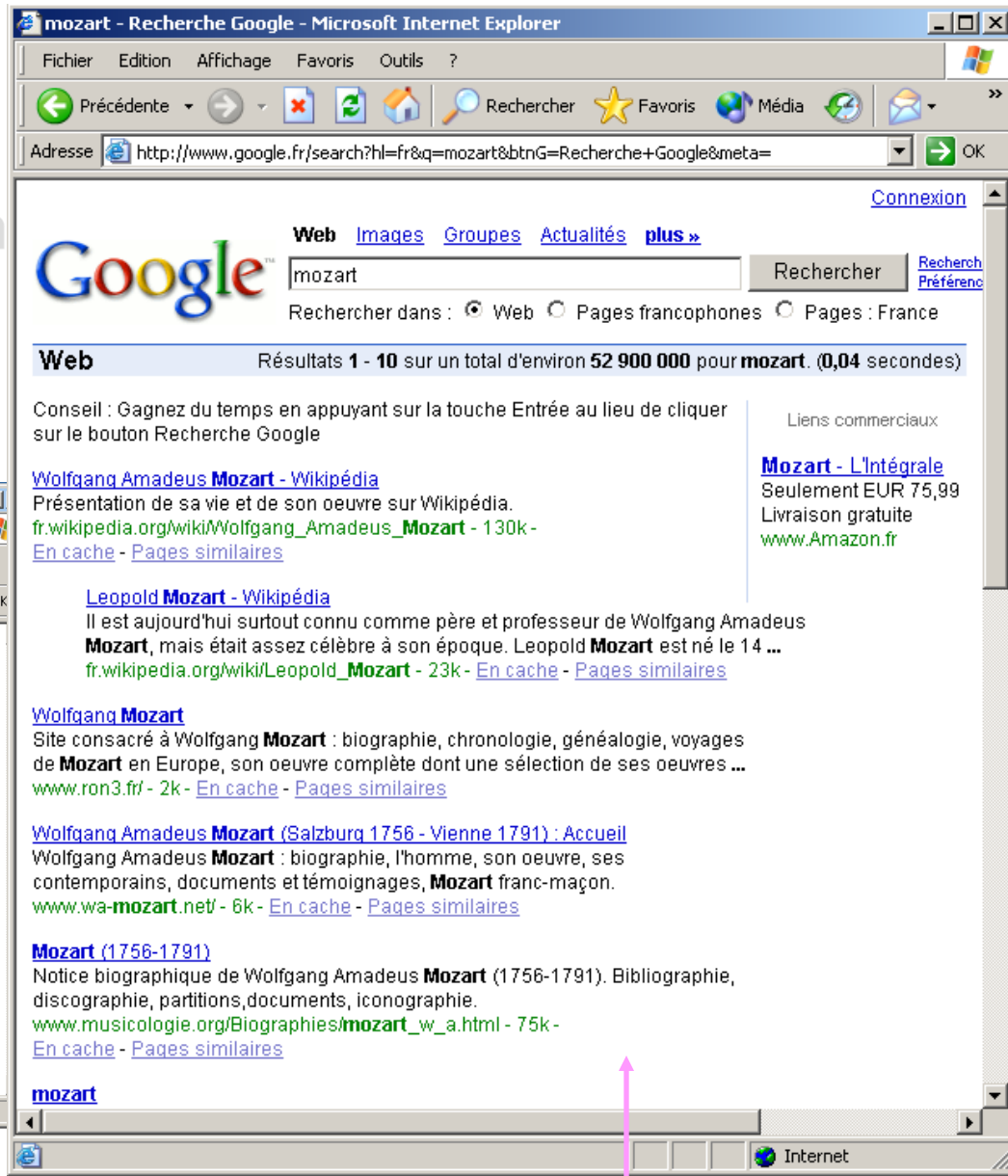
Rechercher une information



Moteurs de recherche



Recherche rapide des documents :
mots-clés de la requête



Liste de liens vers des pages
Web traitant de la requête

Contexte général

Quelques problèmes :

- Pages ne répondant pas à la question
- Incohérence des réponses
- Informations réparties dans des pages différentes
- Etc.



Contexte général



Natural Language Question Answering System

what is the capital of the second largest country in Asia|

Ask Question

Clear

Un système QR
en ligne :

START (Katz, 97)

<http://start.csail.mit.edu>

START's reply

==> what is the capital of the second largest country in Asia

I know that the second largest country in Asia is India (source: [The World Factbook 2006](#)).

Using this information, I determined what the capital of India is:

[New Delhi](#) is the capital of [India](#).

Source: START KB

[India](#)



Capital:  

New Delhi

Source: [The World Factbook 2006](#)

Question-réponse : les objectifs

Poser des questions en langue naturelle

- Analyser la question : quel type ?, ...
- Domaine : ouvert / fermé

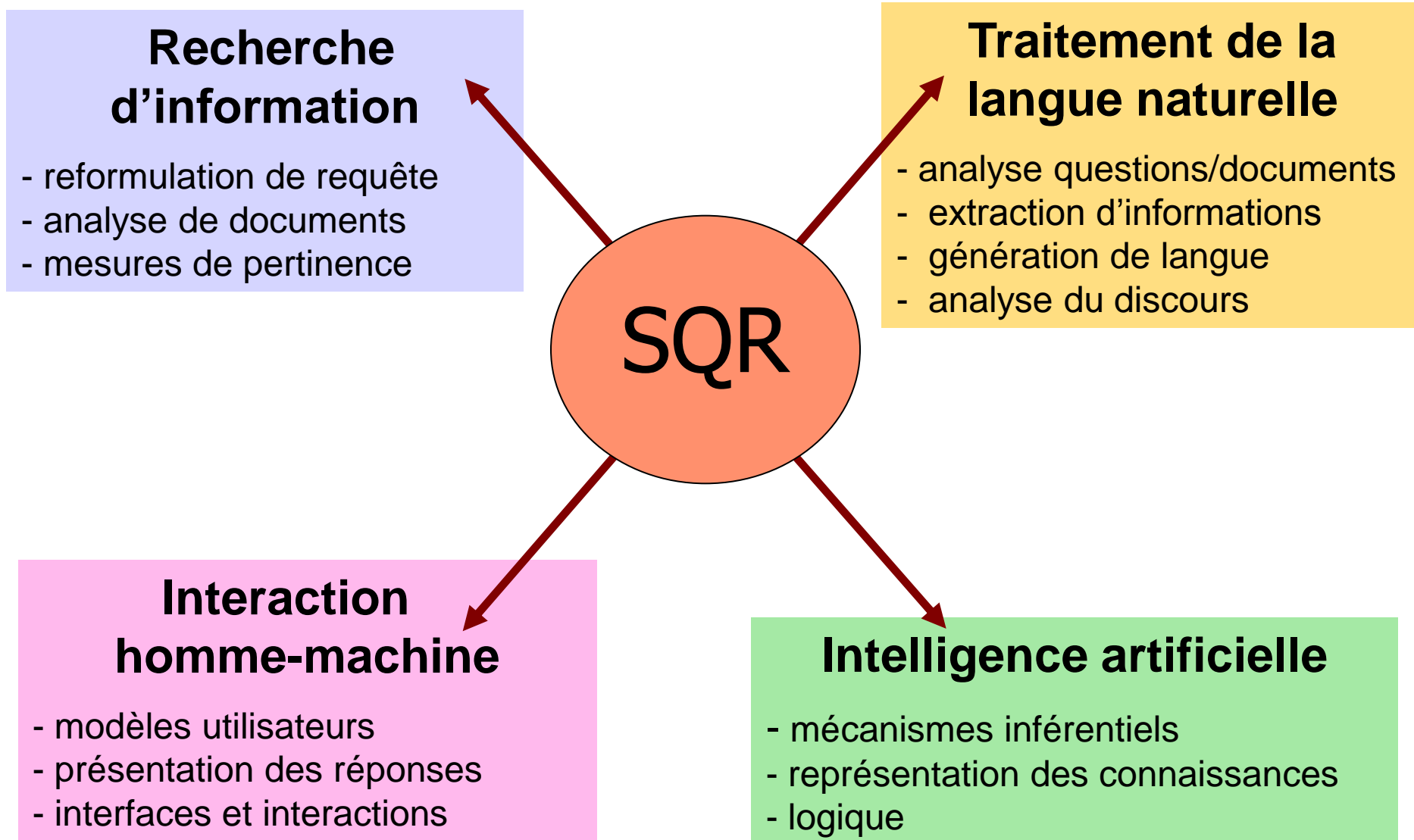
Trouver les informations

- Données structurées → *Database QA*,
- Textes hétérogènes (pages web, documents multimédia, etc)
→ *Text based QA, Web based QA*

Répondre aux questions

- Réponses exactes,
- Passages,
- Multimedia,
- etc.

Question-réponse : les stratégies

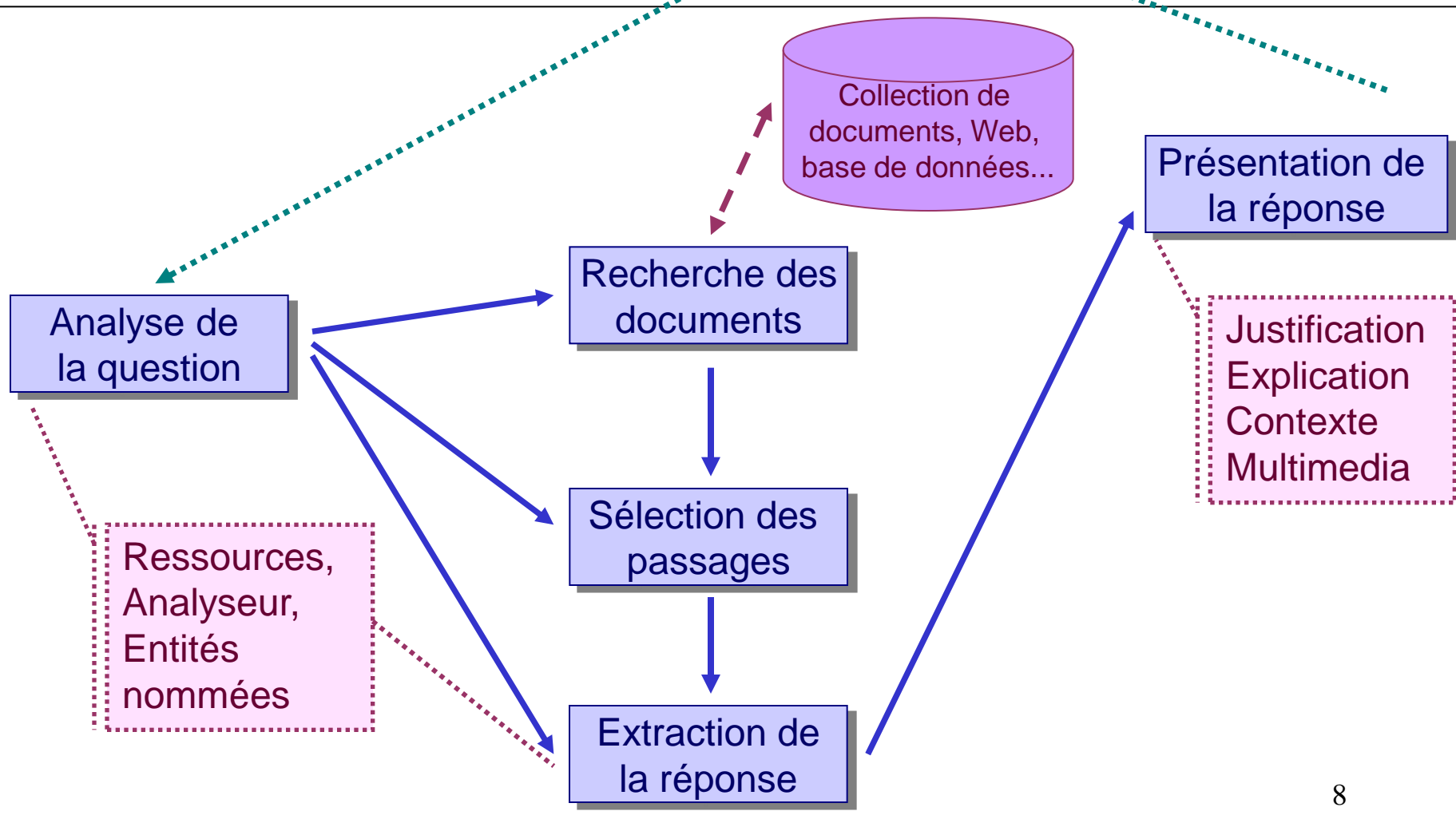


Plan du cours

- Architecture d'un système de question-réponse
 - Analyse de la question
 - Recherche des documents
 - Sélection des passages
 - Extraction de la réponse
- Présentation de quelques systèmes
 - LCC
 - USC-ISI
 - Insight
 - Microsoft
 - IBM
- Evaluation
- L'avenir...

Architecture d'un système QR

Question en
langue naturelle



Un peu de vocabulaire...

Type de la question :

Catégorie de la question qui permet de déclencher des stratégies différentes pour la sélection des passages, l'extraction, etc.

- **Factuelle (factoide)** : Qui a assassiné John Lennon ?
- **Booléenne (oui/non)** : Saddam Hussein est-il vivant ?
- **Définition** : Que signifie TVA ?
- **Cause/conséquence** : Pourquoi la Terre se réchauffe-t-elle ?
- **Procédure** : Comment avoir un visa pour le Canada ?
- **Liste** : Citer 5 villes françaises
- **Evaluative/comparative** : Quelle est la plus grande ville du monde ?
- **Opinion** : Que pensent les Américains de la guerre en Irak ?
- ...

Un peu de vocabulaire...

Type de réponse attendu :

*Type d'objet (généralement le type d'entité nommée reconnue)
ou type de phrase attendu par la question.*

- **PERSONNE** : “Qui ...”, “Quel ministre...”
- **ORGANISATION** : “Qui...”, “Quelle compagnie...”
- **LIEU** : “Où ...”, “Dans quel pays...”
- **DATE** : “Quand ...”, “En quelle année...”
- **NOMBRE** : “Combien ...”, “Quel est le taux...”
- ...

mais aussi

- **EXPLICATION** : “Pourquoi ...”, “Pour quelle raison...”
- **PROCEDURE** : “Comment ...”, “Quelles sont les étapes pour...”
- ...

Un peu de vocabulaire...

Focus :

Propriété ou entité recherchée par la question.

- “Dans quel pays se trouve le Grand Canyon?”
- “Quelle est la hauteur du Mont-Blanc ?”

Thème de la question (*topic*) :

Objet sur lequel porte la question.

- “Quelle est la hauteur du Mont-Blanc ?”
 - Focus : hauteur
 - Topic : Mont-Blanc

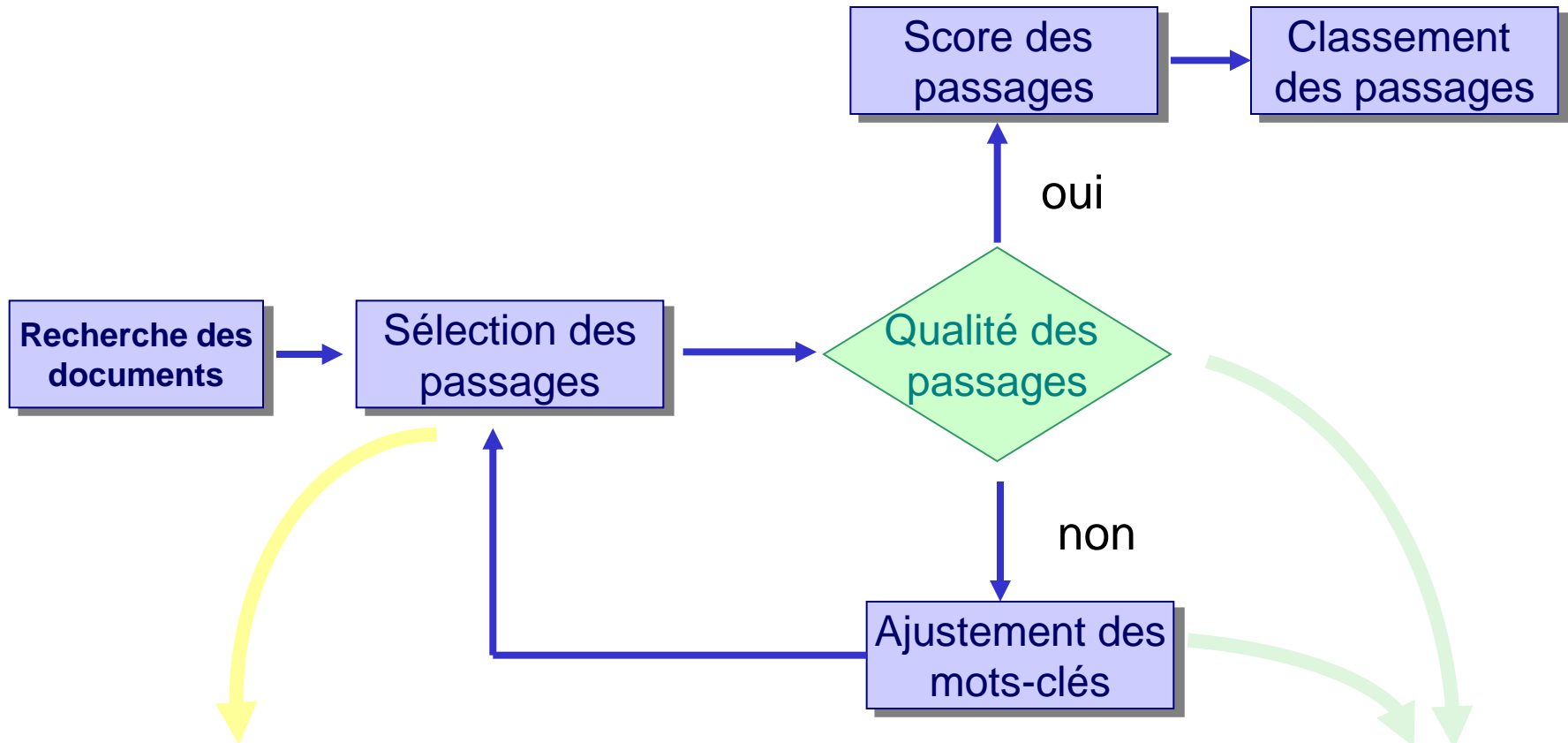
Construire la requête

Extraction des mots-clés : *sac de mots* en RI

Questions	Mots-clés
Qui est Hugo Chavez ?	Hugo Chavez
Que fabrique l'entreprise Peugeot ?	Peugeot, entreprise, fabriquer
Quel pays l'Irak a-t-il envahi en 1990 ?	Irak, envahir, 1990
Combien d'habitants y a-t-il en France en 2007 ?	France, habitants, 2007

Analyse des questions
Recherche des documents
Sélection des passages
Extraction des réponses

Les passages...



- type de réponse attendu,
- mots-clés de la question,
- entités nommées de la question

- $\text{Nb Passage} < \text{seuil}$ → expansion de requête
- $\text{Nb Passage} > \text{seuil}$ → suppression d'un/ plusieurs mots clés

Un peu de vocabulaire...

Passages candidats :

Passages/textes (d'une phrase à un document) sélectionnés par le moteur de recherche

- Selon la requête et l'index utilisé, un passage candidat peut contenir ou PAS des réponses candidates.
- Généralement classés selon un score par le moteur de recherche

Réponses candidates :

Mot ou groupe de mots généralement du même type que le type de réponse attendu

Un peu de vocabulaire...

Un exemple :

Qui est le père de la reine Elizabeth II ? → Type personne

Passage candidat n°1 :

François Mitterrand et la reine Elizabeth II inaugurent le tunnel sous la Manche.

Passage candidat n°2 :

La reine Elizabeth II a évoqué l'intervention de son père, le roi George VI, 48 ans plus tôt devant le Parlement.

Les techniques :

Patrons d'extraction :

- Recherche d'une **séquence calque de la question** où l'élément cherché correspond à une variable.
 - *Quelle est la capitale des USA ? → X est la capitale des USA*
- **Approche héritée des techniques d'extraction d'informations**
- Gestion de variantes pour dépasser la notion de calque (**variations linguistiques**)
 - *Quelle est la capitale de X ?*
 - *X a pour capitale Y*
 - *Y est la plus grande ville de X. Elle est la capitale de X*

Les techniques :

Patrons d'extraction :

Quelques variations linguistiques (*réécriture*) :

- Morphologique : Où se trouve la capitale de l'Europe / capitale européenne ?
- Lexicale : Comment s'appelle la reine de Hollande / des Pays-Bas ?
- Syntaxique :
Moscou compte 9 millions d'habitants / Les 9 millions d'habitants de Moscou
- Sémantique :
Qui a acheté le Phocée ? → Bernard Tapie a vendu le Phocée à Mouna Ayoub
Comment Adolf Hitler est-il mort ? → Adolf Hitler s'est suicidé

Les techniques :

Patrons d'extraction :

Limite : tâche longue et coûteuse quand faite à la main

→ *Alternative possible : automatiser l'acquisition via des méthodes d'apprentissage.*

Techniques d'apprentissage à partir de corpus :

- Apprentissage supervisé (corpus annoté)
- Apprentissage semi-supervisé (corpus en partie annoté)

Les techniques :

Patrons d'extraction :

Acquisition de patrons par apprentissage :

- Partir d'un patron déterminé,
- Remplacer un élément du patron (généralement prédicat ou un des arguments) par une variable,
- Extraire d'un corpus (web, corpus spécialisé) un ensemble d'éléments pouvant instancier le patron.

Les techniques :

Patrons d'extraction :

- Patron de base :

<company>IBM</company> acheter <company>Lotus </company>

IBM a acheté Lotus

Lotus a été acquis par IBM

L'achat de Lotus par IBM

acheter(IBM, Lotus)
acquérir(IBM, Lotus)
achat(IBM, Lotus)

- Patrons acquis :

⇒ IBM <acheter, acquérir> Lotus

⇒ <achat de> Lotus <par> IBM

Les techniques :

Les patrons d'extraction : oui mais...

- Les anaphores posent problème :

Andorre est un petit pays enclavé entre la France et l'Espagne dans le sud ouest de l'Europe,

.....

Le tourisme est le secteur majeur de **ce** petit pays avec une économie saine, qui compte pour environ 80% de son PNB....

Quel est le secteur économique majeur de l'Andorre ?

→ ??

Les techniques :

Les patrons d'extraction : oui mais...

- Les expressions temporelles compliquent l'analyse :

...**L'Iran a envahi l'Irak** le 22 septembre 1980. **Après 8 longues années de combat**, la résolution 598 des Nations Unies a mis fin à la guerre...

Combien de temps a duré la guerre Iran-Irak ?

→ ??

Les techniques :

Les patrons d'extraction : oui mais...

- La reconnaissance d'entités nommées ne suffit pas :

Lee Harvey Oswald, l'homme qui a assassiné le président **John F. Kennedy**, a été capturé puis tué par **Jack Ruby**.

Qui a tué Lee Harvey Oswald ?

→ John F. Kennedy !!!!

Les techniques :

Les patrons d'extraction : oui mais...

- Modalité (discours déclaré, opinion, négation)
- Informations sur le web, blagues, ...

.....George Bush pense que **Steven Spielberg**
est le premier ministre d'Israël.....

Qui est le premier ministre d'Israël ?

→ Steven Spielberg

Les techniques :

Utiliser des critères : *calculer un score pour chaque réponse candidate*

Bon contexte global : pertinence du passage d'où la réponse candidate a été extraite.

- Nombre de mots-clés présents dans le passage
- Nombre de mots communs à la question et au passage
- Passage bien classé par le moteur de recherche
- ...

Bon contexte local : adéquation du passage par rapport à la question.

- Distance moyenne entre la réponse candidate et les mots-clés présents dans le passage
- Nombre de mots de la réponse candidate qui ne sont pas des mots-clés de la question
- ...

Type sémantique correct : le *type sémantique* de la réponse candidate est soit le même soit un sous-type du type de réponse attendu.

Redondance : présence de la réponse dans les passages sélectionnés.

Les techniques :

Par les relations syntaxiques :

Analyse syntaxique de la question et des passages candidats

Extraction des relations prédicat-argument :

- Question : Qui a écrit la Déclaration d'Indépendance ?

Relations : [X, écrire], [écrire, Déclaration d'Indépendance]

- Réponse candidate : Jefferson a écrit la Déclaration d'Indépendance.

Relations : [Jefferson, écrire], [écrire, Déclaration d'Indépendance]

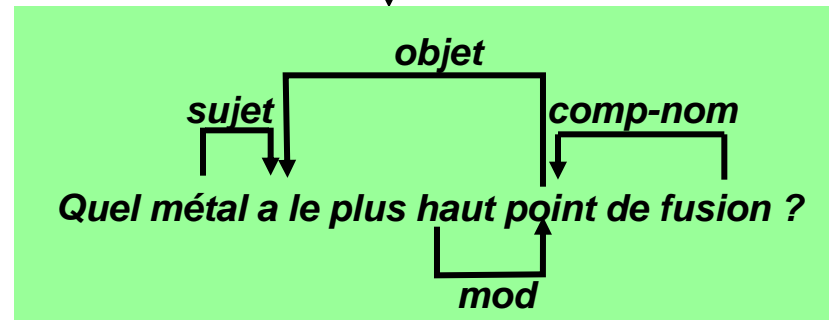
⇒ X = Jefferson

Les techniques :

Par les relations syntaxiques: arbres de dépendance et calcul de distance

Quel métal a le plus haut point de fusion ?

1. Analyse de la question



2. Extraction de mots clés

3. Type de réponse attendu

METAL

4. Extraction des relations

<METAL, a, sujet>

<point, a, objet>

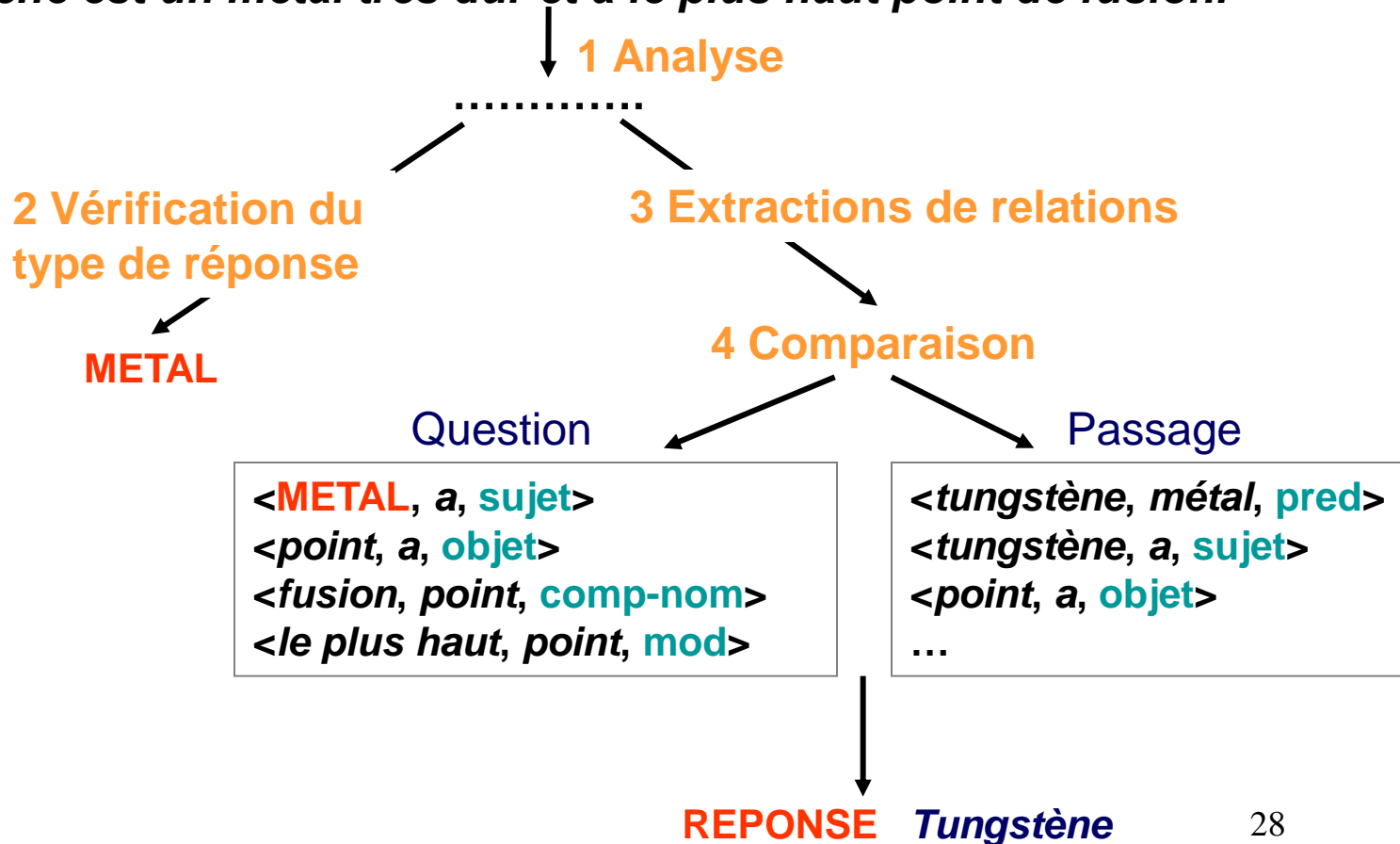
<fusion, point, comp-nom>

<le plus haut, point, mod>

Les techniques :

Par les relations syntaxiques :

“Tungstène est un métal très dur et a le plus haut point de fusion.”



Les techniques :

Avec la logique :

Preuve / démonstration logique :

- Convertir la question en but,
- Convertir le passage en formule logique,
- Ajouter des prédicats représentant des règles de subsomption, des connaissances, etc. ,
- Prouver le but / la question à partir du passage.

Les techniques :

Avec la logique :

- Question : *Qui a assassiné John Lennon ?*
- Question sous forme logique :

`personne(X1) & assassiner(X1, X2) & John_Lennon(X2)`

- Réponse candidate : *Mark Chapman a abattu John Lennon en 1980,...*

`Mark_Chapman(X1) & abattre(X1, X2) & John_Lennon(X2)`

- Exemple d'inférence lexicale :
`abattre → assassiner`

Synthèse

Techniques vs. QA modules	Question analysis	Document retriever and passage retriever	Answer extraction	Answer presentation
Information extraction techniques			<i>surface extraction patterns</i>	
Information retrieval techniques		<ul style="list-style-type: none"> • <i>Search engines,</i> • <i>Indexing (n-grams, paragraph indexing),</i> • <i>Passage retrieval</i> 		
NLP/ linguistic techniques	<ul style="list-style-type: none"> • <i>Part of speech (POS) tagging,</i> • <i>Shallow parsing</i> • <i>Named entity recognition (NER),</i> • <i>PP attachments,</i> • <i>Word sense disambiguation (WSD)</i> 	<i>Conceptual indexing</i>	<ul style="list-style-type: none"> • <i>POS tagging</i> • <i>Shallow parsing,</i> • <i>NER,</i> • <i>Anaphora resolution,</i> • <i>Collocation,</i> • <i>WSD,</i> • <i>PP attachments</i> 	<i>Natural language generation</i>
Machine learning techniques	<ul style="list-style-type: none"> • <i>Learn answer type taxonomy,</i> • <i>Learn query terms from LN question</i> 	<i>Learning to score passages</i>	<ul style="list-style-type: none"> • <i>Learning surface pattern rules,</i> • <i>Answer scoring</i> 	
Knowledge mining and data mining techniques			<ul style="list-style-type: none"> • <i>Leverage data redundancy,</i> • <i>bridge the gap between question and answer,</i> • <i>Improve answer quality</i> 	
Artificial intelligence techniques	<ul style="list-style-type: none"> • <i>Query semantic representation,</i> • <i>transformation rules of NL question into a logic form</i> 		<i>Logic prover to go beyond word matching, answer fusion</i>	<i>Answer justification</i>

Plan du cours

- Architecture d'un système de question-réponse
 - Analyse de la question
 - Recherche des documents
 - Sélection des passages
 - Extraction de la réponse
- Présentation de quelques systèmes
 - LCC
 - USC-ISI
 - Insight
 - Microsoft
 - IBM
- Evaluation
- L'avenir...

Quelques outils de QR en ligne

- Ask.com <http://www.ask.com>
 - Recherche question proche dans une base de FAQ
 - Si recherche infructueuse, recherche d'information standard
 - Réponses = documents
- AnswerBus <http://answerbus.coli.uni-saarland.de>
 - Réponses = phrases
- LCC <http://www.languagecomputer.com/>
 - Réponses = courts passages (snippets)
- Powerset <http://www.powerset.com/>

LCC

(Moldovan & Rus, 2001)

Présentation :

- Utilise un **démonstrateur logique** pour justifier les réponses :
 - Questions sous forme logique,
 - Réponses candidates sous forme logique,
 - Synonymes WordNet,
 - Axiomes,
 - Inférences lexicales,
- Le **moteur d'inférence** essaie de valider la réponse en passant la question en forme négative et en prouvant la contradiction,
- Si la preuve échoue, les prédicats de la question sont **graduellement relaxés** jusqu'à ce que la preuve soit démontrée.

LCC

(Moldovan & Rus, 2001)

Inférences lexicales :

Question : What year did Marco Polo travel to Asia?

Réponse : Marco Polo divulged the truth after returning in 1292 from his travels, which included several months on Sumatra

Inférences lexicales :

- (1) travel_to:v → GLOSS → travel:v → RGLOSS → travel:n
- (2) travel_to → GLOSS → travel:v → HYPONYM → return:v
- (3) Sumatra:n → ISPART → Indonesia:n → ISPART → Southeast_Asia:n → ISPART → Asia:n

Question : What is the legal age to vote in Argentina?

Réponse : Voting is mandatory for all Argentines aged over 18.

Inférences lexicales :

- (1) legal:a → GLOSS → rule:n → RGLOSS → mandatory:a
- (2) age:n → RGLOSS → aged:a
- (3) Argentine:a → GLOSS → Argentina:n

LCC

(Moldovan & Rus, 2001)

Le "logic prover" :

Question : *Which company created the Internet Browser Mosaic?*

- (`_organization_AT(x2)`) & `company_NN(x2)` & `create_VB(e1,x2,x6)` & `Internet_NN(x3)` & `browser_NN(x4)` & `Mosaic_NN(x5)` & `nn_NNC(x6,x3,x4,x5)`

Réponse candidate : *... Mosaic, developed by the National Center for Supercomputing Applications at the University of Illinois...*

- ... `Mosaic_NN(x2)` & `develop_VB(e2,x2,x8)` & `by_IN(e2,x8)` & `National_NN(x3)` & `Center_NN(x4)` & `for_NN(x5)` & `Supercomputing_NN(x6)` & `application_NN(x7)` & `nn_NNC(x8,x3,x4,x5,x6,x7)` & `at_IN(e2,x15)` & `University_NN(x10)` & `of_NN(x11)` & `Illinois_NN(x12)` & `nn_NNC(x15,x10,x11,x12)` ...

Inférences lexicales : `develop` ↔ `make` et `make` ↔ `create`

- $\exists x_2 x_3 x_4, \forall e_2 x_1 x_7$ (`develop_vb(e2,x7,x1)` ↔ `make_vb(e2,x7,x1)` & `something_nn(x1)` & `new_jj(x1)` & `such_jj(x1)` & `product_nn(x2)` & `or_cc(x4,x1,x3)` & `mental_jj(x3)` & `artistic_jj(x3)` & `creation_nn(x3)`).
- $\forall e_1 x_1 x_2$, (`make_vb(e1,x1,x2)` ↔ `create_vb(e1,x1,x2)` & `manufacture_vb(e1,x1,x2)` & `man-made_jj(x2)` & `product_nn(x2)`).

Axiomes : $\forall x_0$ (`mosaic_nn(x0)` → `internet_nn(x0)` & `browser_nn(x0)`)₃₆

USC - ISI (Ravichandran and Hovy, 2002) (Hermjakob et al. 2003)

Patrons de surface :

When was X born →

- Mozart was born in 1756
- Gandhi (1869-1948)

Peuvent être repérés avec les patrons :

- <NAME> was born in <BIRTHDATE>
- <NAME> (<BIRTHDATE> -

USC - ISI

Apprendre les patrons de surface : bootstrapping

- Pour un type de question donné (*When was X born?*), commencer avec des réponses connues pour différentes valeurs de X :
 - *Mozart 1756*
 - *Gandhi 1869*
 - *Newton 1642*
- Requêtes sur le Web : *“Mozart +1756”*
- Collecter les 1000 premiers documents,
- Filtrer, ...
- Trouver les meilleures sous-chaînes : *Mozart (1756-1791)*
- Filtrer : *Mozart (1756-*
- Remplacer les termes de la requête par des balises : *<NAME>
<ANSWER>*
- Déterminer la précision de chaque patron

USC - ISI

Trouver les réponses :

- Déterminer le type de la question,
- Lancer la requête,
- Segmenter en phrases,
- Remplacer les termes de la question par les balises
 - i.e. remplacer Mozart par <NAME>
- Rechercher les instances de patrons associés au type de la question,
- Sélectionner les mots qui correspondent à <ANSWER> ,
- Donner un score en fonction de la précision du patron₃₉

Insight

(Soubbotin, 2002)

Encore des patrons :

Exemple :

a == {countries} b == {official posts}
 w == {proper names (first and last)}
 e == {titles or honorifics}

Patrons pour *Who is the President (Prime Minister) of given country?*

- a b e w w
- e w w d b, a
- b, a e w w

Questions de définition : (A : mot-clé, X : réponse)

What is **Moulin Rouge** ?

- <A; comma; [a/an/the]; X; [comma/period]>
- **Moulin Rouge**, a cabaret

What is **methane** ?

- <X; [comma]; [also] called; A [comma]>
- naturally occurring gas called methane

Microsoft AskMSR (Brill et al., 2002)

Présentation :

- Approche qui s'appuie sur l'hypothèse que quelqu'un a déjà répondu **sur le Web** à la question.
- **Objectifs** : éviter de traiter les problèmes de :
 - Relations lexicales, syntaxiques, sémantiques,
 - Anaphores,
 - Synonymie,
 - Réponses indirectes.
- **Avantage** : redondance sur le Web

Microsoft AskMSR (Brill et al., 2002)

Principe :

- Formuler **plusieurs requêtes** (chaque réécriture a un score). Par exemple :
What is relative humidity?
 - [“+is relative humidity”, LEFT, 5]
 - [“relative +is humidity”, RIGHT, 5]
 - [“relative humidity +is”, RIGHT, 5]
 - [“relative humidity”, NULL, 2]
 - [“relative” AND “humidity”, NULL, 1]
- Récupérer les 100 premiers documents de Google,
- **Extraire des n-grammes** (n=1,2,3) à partir des **snippets**,
- Pondérer les n-grammes en additionnant les scores des réécritures utilisées,
- Filtrer les n-grammes en utilisant des expressions régulières associées au type de question,
- Rechercher un document qui confirme la réponse dans un corpus (TREC).

Microsoft AskMSR (Brill et al., 2002)

Exemple :

Question : *What is the rainiest place on Earth ?*

Réponse sur le Web : *Mount Waialeale*

Passage du corpus TREC :

... In misty Seattle, Wash., last year, 32 inches of rain fell. Hong Kong gets about 80 inches a year, and even Pago Pago, noted for its prodigious showers, gets only about 196 inches annually. (The **titleholder**, according to the National Geographic Society, is **Mount Waialeale** in Hawaii, where about 460 inches of rain falls each year.) ...

IBM Statistical QA

(Ittycheriah, 2001)

Principe :

$$\begin{aligned} p(c|q,a) &= S_e p(c,e|q,a) \\ &= S_e p(c|e,q,a) p(e|q,a) \end{aligned}$$

q = question
 a = answer
 c = “correctness”
 e = answer type

$p(e|q,a)$: modèle du type de réponse (ATM)

$p(c|e,q,a)$: modèle de sélection de la réponse (ASM)

- ATM **prédit**, à partir de la question et d'une réponse proposée, le **type de réponse** qui satisfait les 2,
- Etant donnés une question, une réponse, et le type de réponse prévu, ASM essaie de modéliser si cette configuration est correcte,
- Distributions : maximum d'entropie,
- Données pour l'entraînement = jugements humains.

IBM Statistical QA (Ittycheriah, 2001)

Principe :

- Analyse de la question par ATM,
- Recherche des documents :
 - Expansion de requête (par analyse du contexte local),
 - 1000 premiers documents,
- Extraction de passage : 100 premiers passages qui ont :
 - Le plus grand nombre de mots en commun avec la question,
 - Type de réponse attendu correct,
 - Faible dispersion des mots de la question,
 - Structure syntaxique similaire à celle de la question.
- Extraction de la réponse :
 - Réponses candidates classées par ASM

IBM PIQUANT

(Prager, 2003)

Annotation prédictive :

- **Objectif** : être sûr que les passages sélectionnés par le moteur contiennent au moins une réponse candidate,
- Reconnaître qu'une réponse candidate a le **bon type** (correspond à un label généré par la reconnaissance d'entité nommée),
- **Annoter** le corpus et **indexer les labels**,
- Identifier les types de réponse dans les questions et inclure les labels correspondants dans les requêtes.

IBM PIQUANT

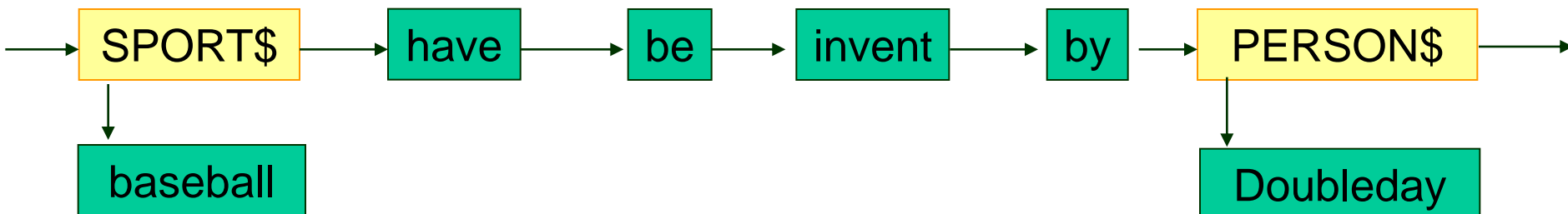
(Prager, 2003)

Annotation prédictive :

Question : *Who invented baseball ?*

- “Who” correspond à **PERSON\$** ou **ORGANIZATION\$**
- On suppose que seulement les personnes inventent...
- D'où *Who invented baseball ?* → {**PERSON\$ invent baseball**}

Le texte : “... *but its conclusion was based largely on the recollections of a man named Abner Graves, an elderly mining engineer, who reported that **baseball had been "invented" by Doubleday** between 1839 and 1841.*”

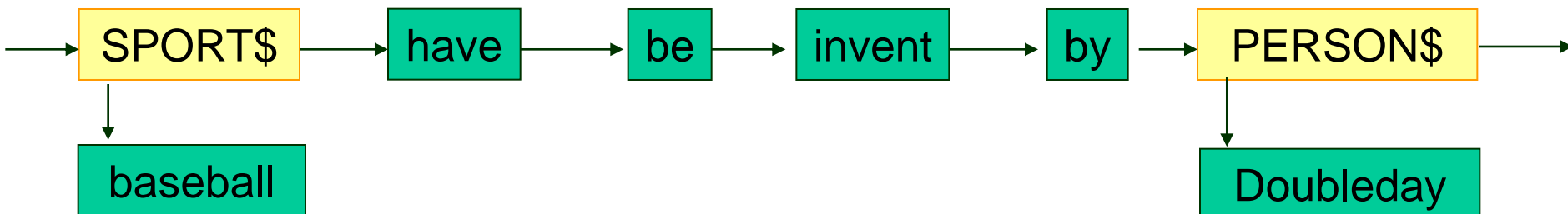


IBM PIQUANT

(Prager, 2003)

Annotation prédictive :

- Exemple précédent :
 - *Who invented baseball ?* → {PERSON\$ invent baseball}
- La **même structure** répond aussi à :
 - *What sport did Doubleday invent ?* → {SPORT\$ invent Doubleday}



IBM PIQUANT

(Prager, 2003)

Subsumption & Disjonction :

Par exemple : New York peut être soit **CITY** soit **PLACE**

- Pour répondre à *Where did John Lennon die?* : type PLACE
- Pour répondre à *In what city is the Empire State Building?* : type CITY

Deux possibilités :

1. **Expansion du type de réponse** et utiliser l'annotation la plus spécifique

1A { (CITY PLACE) John_Lennon die} correspond à CITY

1B {CITY Empire_State_Building} correspond à CITY

Ou

2. Utiliser le **type de réponse le plus générique**

2A {PLACE John_Lennon die} correspond à (CITY PLACE)

2B {CITY Empire_State_Building} correspond à (CITY PLACE)

⇒ Choix 2

IBM PIQUANT

(Prager, 2003)

Annotation prédictive :

- N'importe quelle structure d'un texte peut être annotée :
 - Citations,
 - Explications,
 - Procédures,
 - Opinions,
 - ...
- N'importe quel label sémantique peut être indexé et utilisé pour la recherche.

RESPO
GENE
RATORDocument id: 2619; Sentence no: 9
Question: Why is the sky blue

Exemple d'annotation

PLACE
REGION
LK/MT/RIV
CONTINENT
COUNTRY
STATE
CITY
CAPITAL
NATIONAL
LANGUAGE
MISC
COMPOSITION
MEDICAL
COLLEGE
SPORTS
NAME
NICKNAME

Tyndall, John

John Tyndall born Aug. 2, 1820, Leighlin Bridge, County Carlow, Ire.
John Tyndall died Dec. 4, 1893, Hindhead, Surrey, Eng.

John Tyndall, British physicist who demonstrated why the sky is blue.

In 1853 Tyndall was chosen professor of natural philosophy at the Royal Institution, London, where he became a colleague and friend of Michael Faraday. His early work was concerned with the magnetic properties of crystals, but in 1859 he began investigating the ability of various gases to absorb and radiate heat. He established that humid air absorbs heat with little change in temperature, a phenomenon of meteorological importance.

Tyndall studied the diffusion of light by large molecules and dust, known as the Tyndall effect; he also performed experiments demonstrating that the sky's blue colour results from the scattering of the Sun's rays by molecules in the atmosphere, a phenomenon which was later explained theoretically by Lord Rayleigh.

In 1881 he delivered the final blow to the long-held idea of spontaneous generation by proving that germ-free air does not lead to food decay. Tyndall's publications, more than 16 books and 145 papers, include *Heat Considered as a Mode of Motion* (1862), *Six Lectures on Light* (1873), and *Forms of Water* (1872).

IBM PIQUANT

(Prager, 2003)

Un exemple :

Question : *Where is Belize ?*

- *Where* correspond à :
(CONTINENT\$, WORLDREGION\$, COUNTRY\$, STATE\$, CITY\$, CAPITAL\$, LAKE\$, RIVER\$...).
- On sait que Belize est un pays d'où :
Where is Belize ? → (hyperonymes)
{(CONTINENT\$ WORLDREGION\$) Belize}
- Belize apparaît 1068 fois dans le corpus TREC,
- Belize et PLACE\$ co-occurent dans 537 phrases,
- Belize et CONTINENT\$ ou WORLDREGION\$ co-occurent dans 128 phrases.

IBM Research -- Knowledge Structures Group

piQuAnt using RESPO^{NSE}_{GENE}RATOR 

User query #0202: **Where is Belize located?**

Best guess from LQA = **Central America** from RegularQuery #1, score = 0.7281811

Linguistic Query Agent					
Score	Query	Pos	Corpus	Docid	Fragment
0.7281811	RegularQuery	1	aquaint	XIE20000818.0260	Central America
0.5322438	RegularQuery	6	aquaint	APW19981026.0795	Yucatan peninsula
0.16764504	RegularQuery	2	aquaint	XIE19960510.0179	Central American
0.15251632	RegularQuery	4	aquaint	NYT20000718.0216	Caribbean
0.07603362	RegularQuery	4	aquaint	NYT20000718.0216	Europe

Passages for Linguistic Query Agent, Corpus = acquaint

RegularQuery, QFrame #0

@dynwin(3 @weight(20.00010 @syn(*Belize *Belizean Belizean Belize)) @weight(40.00100 @syn(REGION\$ CONTIN\$)))

Article Title / Answer Passage

1 British Soldiers Injured While Training With Prince William

Two soldiers were seriously hurt and 11 others cut and bruised when a track collapsed under their truck in a training course in Belize , Central America , the report said.

2 Tourism to Be Central America's Revenue Source

The other two signatories of the document were Vice-president Enrique Borgo Bustamante of El Salvador , and Tourism and Environment Minister Henry Young of Belize . The Central American heads of state or government also expressed their congratulations for the advances toward a peace agreement in Guatemala , "specially the recently signed agreement on socio-economic aspects and agrarian " last Monday .

3 Nicaragua's Tourism Grows

In Central America , Belize ranks second in the growth of tourism with 13.2 percent a year , followed by Costa Rica with 11.7 percent and Guatemala 8.4 percent .

4 YOUR TALL SHIPS DREAMS CAN COME TRUE

Windstar Cruises (800-258-7245) offers sailngs in the Caribbean , Europe , Belize , and Costa Rica .

5 YOUR TALL SHIPS DREAMS CAN COME TRUE

Windstar Cruises (800-258-7245) offers sailngs in the Caribbean , Europe , Belize , and Costa Rica .

Plan du cours

- Architecture d'un système de question-réponse
 - Analyse de la question
 - Recherche des documents
 - Sélection des passages
 - Extraction de la réponse
- Présentation de quelques systèmes
 - LCC
 - USC-ISI
 - Insight
 - Microsoft
 - IBM
- Evaluation
- L'avenir...

Les campagnes d'évaluation

- Principales campagnes d'évaluation des SQR :
 - **TREC** (depuis 1999) : anglais,
 - **CLEF** (depuis 2003) : multilingue,
 - **EQUER** (depuis 2004) : français,
 - **NTCIR** (depuis 2002) : japonais,
 - **QUAERO** (depuis 2008) : français, anglais
- Caractéristiques :
 - **Domaine ouvert** (pour EQUER : ouvert + médical),
 - **Collection fermée** (en général, articles de journaux),
 - Principaux types de questions : factoiide et définition.

Les campagnes d'évaluation

- TREC-8 (1999) & TREC-9 (2000)
 - Les systèmes doivent donner les 5 meilleures réponses,
 - Mean Reciprocal Rank (MRR) :
 - 1 point si la 1ère réponse est correcte, sinon
 - 0.5 point si la 2ème réponse est correcte, sinon
 - 0.2 point si la 5ème réponse est correcte, sinon 0
- TREC-10 (2001)
 - Questions sans réponse (NIL)
- TREC-11 (2002)
 - Les systèmes doivent donner LA meilleure réponse,
 - La réponse doit être “exacte”,
 - Disparition des questions de définition (“Qu'est-ce que/qui est X?”),
 - Resultats classés par le système selon un score de confiance,

Les campagnes d'évaluation

- TREC-12 (2003)
 - Retour des questions de définition,
 - Questions de type *liste* (les réponses doivent être exactes),
 - Questions de définition et liste évaluées par F-mesure
- QUAERO (2008...)
 - Questions factuelles, listes, complexes
 - Sur le Web

Les campagnes d'évaluation

Quelques statistiques sur le nombre de participants dans les différentes campagnes d'évaluation

Campaigns	America	Europe	Asia	Australia	Total
TREC-8	13	3	3	1	20
TREC-9	14	7	6	-	27
TREC-10	19	8	8	-	35
TREC-11	16	10	6	-	32
TREC-12	13	8	4	-	25
NTCIR-3 (QAC-1)	1	-	15	-	16
CLEF 2003	3	5	-	-	8
CLEF 2004	1	17	-	-	18
EQUER 2004	0	9 (France and swizland)	-	-	9

Juger les réponses

- **Automatique**, par Mean Reciprocal Rank (MRR)
- **Jugement humain** selon 4 critères :
 - Correcte ou exacte,
 - Non exacte,
 - Incorrecte,
 - Non justifiée.

Juger les réponses

Problème : Qu'est-ce qu'une réponse exacte ?

- Quand il y a une erreur dans la question ?
 - Exemple : *Quand est mort Jacques Chirac ?*
 - Réponse : NIL ?
- Quelle granularité ?
 - Exemple : *Où se trouve Los Angeles ?*
 - En Californie ?
 - Aux Etats-Unis ?
 - En Amérique ?
 - Besoin d'un modèle utilisateur ?

Juger les réponses

Les questions sans réponse :

- **3 possibilités :**

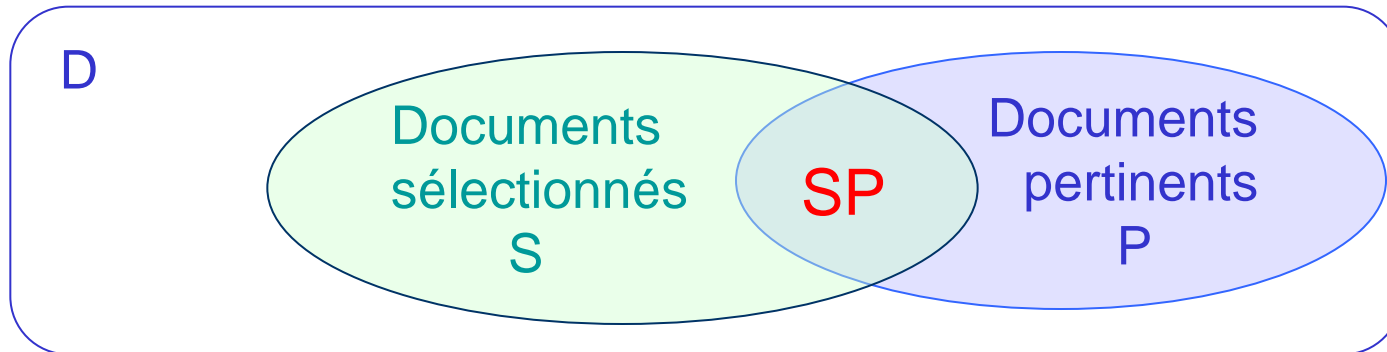
1. La question n'a pas de réponse dans la collection,
2. La question n'a pas de réponse du tout,
3. Le système n'a pas trouvé de réponse.

- **Stratégies :**

- Quand 5 meilleures réponses autorisées :
 - Toujours répondre NIL en position 2,3,4,5
 - Fixer un seuil et mettre NIL à la position correspondante
- Quand 1 seule réponse autorisée :
 - Fixer un seuil T et mettre NIL si le score de la réponse $< T$

Evaluer les systèmes

Documents



- **Rappel** : mesure la proportion de document pertinents sélectionnés parmi tous les documents pertinents.

$$\frac{\text{documents pertinents sélectionnés } SP}{\text{documents pertinents } P}$$

- **Précision** : mesure la proportion de document pertinents sélectionnés parmi tous les documents sélectionnés.

$$\frac{\text{documents pertinents sélectionnés } SP}{\text{documents sélectionnés } S}$$

QR à TREC

- Répondre à **500 questions** (qui changent chaque année) de type factoïde et définition,
- **Documents interrogés :**
 - Articles de presses (1 million environ) : LA Times, San Jose Mercury News, Wall Street Journal, NY Times.
 - Lexicalement, syntaxiquement et sémantiquement bien formés.
- **Tâches :**
 - Extraire les 5 meilleures réponses candidates : longue et courte.
 - Extraire la meilleure réponse.
 - Plusieurs autres sous-tâches:
 - questions de type liste,
 - bibliographie, etc.

QR à TREC

- **Meilleur système** : LCC (31 participants)
- Les approches sont principalement hybrides :
 - Statistique + techniques TAL approfondies.
- Les systèmes statistiques ont atteint leurs limites :
 - Besoin de plus de sémantique et de techniques de TAL complexes,
 - Besoin de techniques avancées.
- Utilisation de techniques d'apprentissage dans tout le processus de QR

Plan du cours

- Architecture d'un système de question-réponse
 - Analyse de la question
 - Recherche des documents
 - Sélection des passages
 - Extraction de la réponse
- Présentation de quelques systèmes
 - LCC
 - USC-ISI
 - Insight
 - Microsoft
 - IBM
- Evaluation
- L'avenir...

(Burger et al, 2001)

Productive, Quality QA**Measurable Progress****Composable Toolkit for QA****Personalized QA**

Related Fields:
HPKB, TDT, DB,
Virtual Ref Desk,
User Modeling

What If Questions
Multimodal Answers 2006

Why Questions

Script/Template Answers

How Questions

Fact Answers

Factoid Questions

Answer Justification

Collaborative QA

Multimodal QA

Crosslingual QA

Multilingual QA

Temporal QA

Multisessional QA
(including change detection)

Answer Fusion

Interoperability
Quality Assurance

Resource Selection

Constrained QA
(Resource/Solution)

Stereotypical and
Individualized QA

Question/Answer

Typologies

QA Reuse
QA as Planning

Wizard of Oz
QA Sets

Create QA Sets

Semi-structured Data

Task Model**Reusable Test Collection****User Expectations****Reuse across sessions****2003**

START, FaqFinder, Web Services
lonaut, QANDA (e.g., Google API)

**Systems
(Performance & Eval)**

Methods & Algorithms

**Resources
(Development, Evaluation)**

**Speed Limit
Robust NLP**

**Speed Limit
Inference**

Requirements
Determination

Feasibility
Testing

Answer Toplogy**User Typology**

Task Modeling
Empirical Studies

Collect QA Logs

Public Taxonomies
(e.g., OpenDirectory in RDF)

TIMEBANK

TIMEML

"Perspective USC/ISIS Question
BANK" Typology

TREC QA

trec.nist.gov/data/qa.html

Les QR : aller plus loin...

- Questions qui nécessitent **plusieurs ressources/documents** pour construire la réponse,
- **Décomposition de question :**
 - Comment s'appelle la femme du président des Etats-Unis ?
⇒ Qui est le président des Etats-Unis ?
⇒ Comment s'appelle la femme de George Bush ?
- **QR interlingue :**
 - question en anglais / recherche de la réponse dans des documents en français.
- **Modèle utilisateur**

Les réponses...

- Questions qui ont 0, 1 ou plusieurs réponses,
- Quand dire que 2 réponses sont identiques / différentes ?
 - Variation naturelle : ex : *taille d'un homme*,
 - Estimation : ex : *population*,
 - Variation dans le temps : ex : *populations, fonctions, ...*
- Trouver le bon **format de présentation** :
 - Listes,
 - Graphiques,
 - Dialogue, ...

La langue...

Les problèmes de la langue à traiter :

- Anaphores,
- Synonymes, paraphrases,
- Métonymies,
- Négation et quantificateurs,
- Figures de style,
- ...

La langue...

La négation :

Q: Qui a inventé la guitare électrique ?

R: Même si Mr. Fender **n'a pas** inventé la guitare électrique, il l'a révolutionné.

Citer une entreprise qui fabrique des voitures.

Citer une entreprise qui **ne** fabrique **pas** de voitures.

La langue...

Quelle est la **plus grande** ville de France ?

- **Matching :**
 - Trouver *Paris est la plus grande ville de France* (ou une paraphrase).
- **Recherche *superlative* :**
 - Trouver la liste des villes françaises et leur population et trier !
 - Trouver les plus grandes villes du monde et choisir les villes françaises,
 - Trouver la population du plus grand nombre possible de villes françaises et choisir la plus grande.
- **Heuristique :**
 - Paris est la capitale de la France. (pas de garantie...)
- **Inférence complexe :**
 - Lille est la 2ème plus grande ville de France; Londres est plus grand que Lille; Paris est plus grand que Londres; Paris est en France.
 - Paris > Londres > Lille

Modèle utilisateur...

Exemple :

Où se trouve Chicago ?

- Qu'est-ce que l'utilisateur attend comme réponse ?
 - La ville : mais quelle granularité ?
 - La comédie musicale ?
 - Le film ?
 - ...

Opinions, hypothèses, croyances...

Exemples :

Que pensent les Américains de la guerre en Irak ?

La banquise va-t-elle disparaître ?

- “X **a déclaré** que la banquise va disparaître...”
- “Le professeur Y **pense** que la banquise va disparaître...”

Combien de pays le Pape a-t-il visité en 1990 ?

- “ ... la visite **prévue** du Pape en Argentine ...”

Les questions sur des relations...

Exemples :

Quelles ont été les relations entre Ben Laden et le Soudan ?

Qu'est-ce que Paris et Londres ont en commun ?

Quel est le lien entre Tom Cruise et Nicole Kidman ?

Implication textuelle...

Qu'est-ce que c'est ? :

(ou *Textual entailment*) :

Relation entre deux textes ou expressions,
Texte T et hypothèse H .

T implique H si le sens de H peut être inféré à partir du sens de T .

T = Depuis sa formation en 1948, Israël a été impliqué dans plusieurs guerres...

H = Israël a été fondé en 1948.

Implication textuelle...

Pour quoi faire ?

- En question-réponse :
 - Est-ce qu'une réponse candidate implique la question ?

Question : Que fabrique Renault ?

Passage : Le président Chirac a visité la nouvelle usine
de voiture de Renault

Réponse : Renault fabrique des voitures

- Reconnaissance de paraphrases :
 - Résumé automatique,
 - ...

Implication textuelle...

Les campagnes :

- RTE (Recognising Textual Entailment)
- CLEF-AVE (Answer Validation Exercise)