Université de Montréal

Question Answering

Guy Lapalme
Université de Montréal

If no one asks me, I know what it is. If I wish to explain it to him who asks me, I do not know.

Confessions of St. Augustine, XI,XIV
cited by Patrick Süskind, *Sur l'amour et la mort*

rali

Question Answering

- The general understanding of a question and its answer is an *AI-complete* problem
- A conceptual theory of answering questions was developed and implemented by Wendy Lenhert (1978) in the QUALM system
  - 13 conceptual categories (*à la* Schank)
    - enablement, causal antecedent, concept completion. ...
  - inference by conceptual dependencies between documents and the question

rali

Simplified approaches

- Data base systems with structured questions
- IR returning documents from keywords combined with logical connectors
- reading understanding systems on specific texts
- matching predefined answers (AskJeeves)
- user-friendly interface with human searchers
  http://answers.google.com

rali

Question Answering
underlying assumptions

- users save time with an answer instead of a reference to a whole document
- better gain when one does not need to look at the context
- deals mainly with *closed* questions
- answer is noun phrase instead of a procedure
- reference date are newspaper type articles (TREC documents)

rali

TREC-QA
examples of questions

- How many calories are there in a Big Mac ?
- What two US biochemists won the Nobel Prize in medecine in 1992 ?
- Where is the Taj Mahal ?
- Where is Rider College situated?
- Name a film in which Jude Law acted.
- What language is commonly used in Bombay ?

TREC-QA
true questions examples

- `What will the US response be if Iran closes the Strait of Hormuz ?`
- `What effect on the price of oil on the international market are likely to result from the terrorist attacks on the Saudi facilities ?`

High Performance Knowledge Bases (HPKB) dealing with the integration of many information sources

rali

QA systems classification criteria

- linguistic resources and general knowledge
  - Word Net, ontologies, gazeeter, encyclopedia
- NLP techniques
  - word matching, syntactic or semantic analysis
- Reasoning methods
  - logic, induction, abduction

rali

QA system types - I

- Answers to factual questions
  - answer is found *verbatim* in documents
  - matching of keywords and named entities
- Simple reasoning
  - current world knowledge need
  - linguistic knowledge (nominalisation, synonyms)
- Information combination from multiple types of documents

rali

QA system types - 2

- Need for advanced reasoning (by analogy, temporal, spatial ...)
  - question is split in disjoint element to search in many sources
    - Is the Fed going to raise interests at the next meeting ?
    - Is the US out of recession ?
    - Is the airline industry out of trouble ?
- Interactive systems
  - questions depending on the context of previous questions and answers
  - allow the disambiguation of questions

rali

QA System Architecture

- Question processing
  - determining the type of expected answer
- Document processing
  - searching for relevant *passages*
- Answer Extraction
  - answer quality evaluation
  - reasoning depending on the question and the whole set of passages

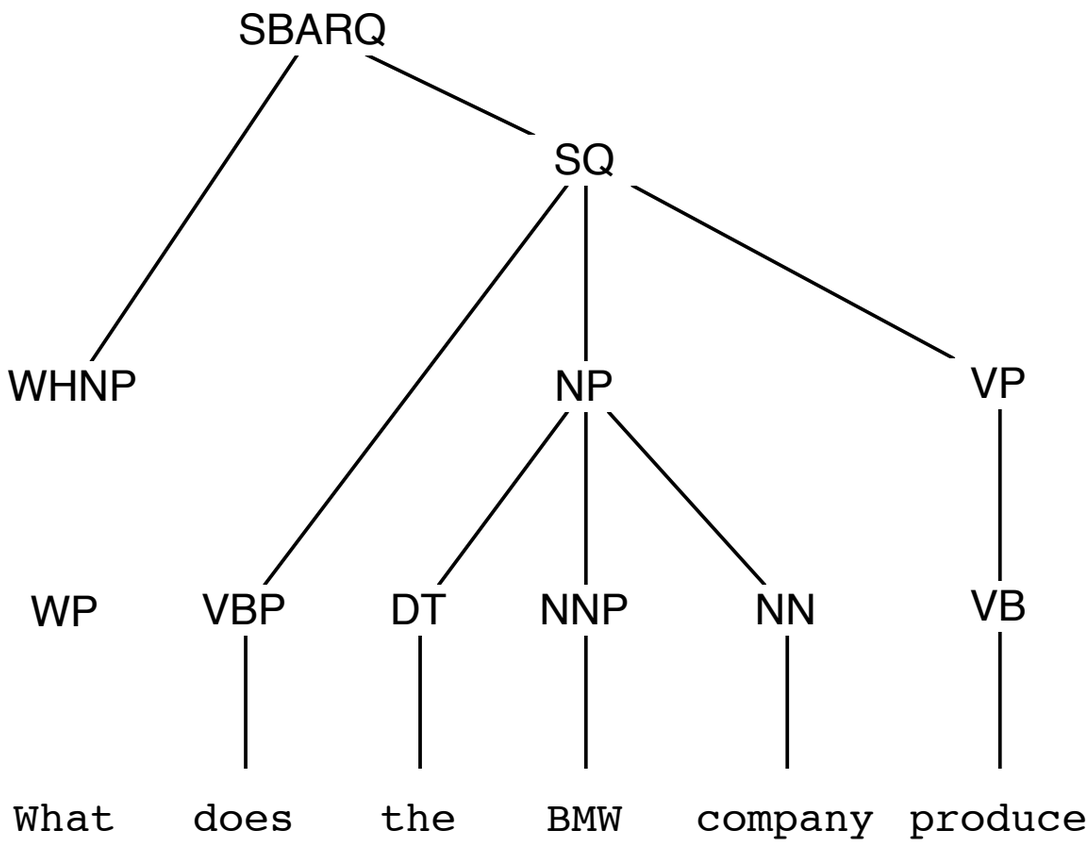
rali

Question processing

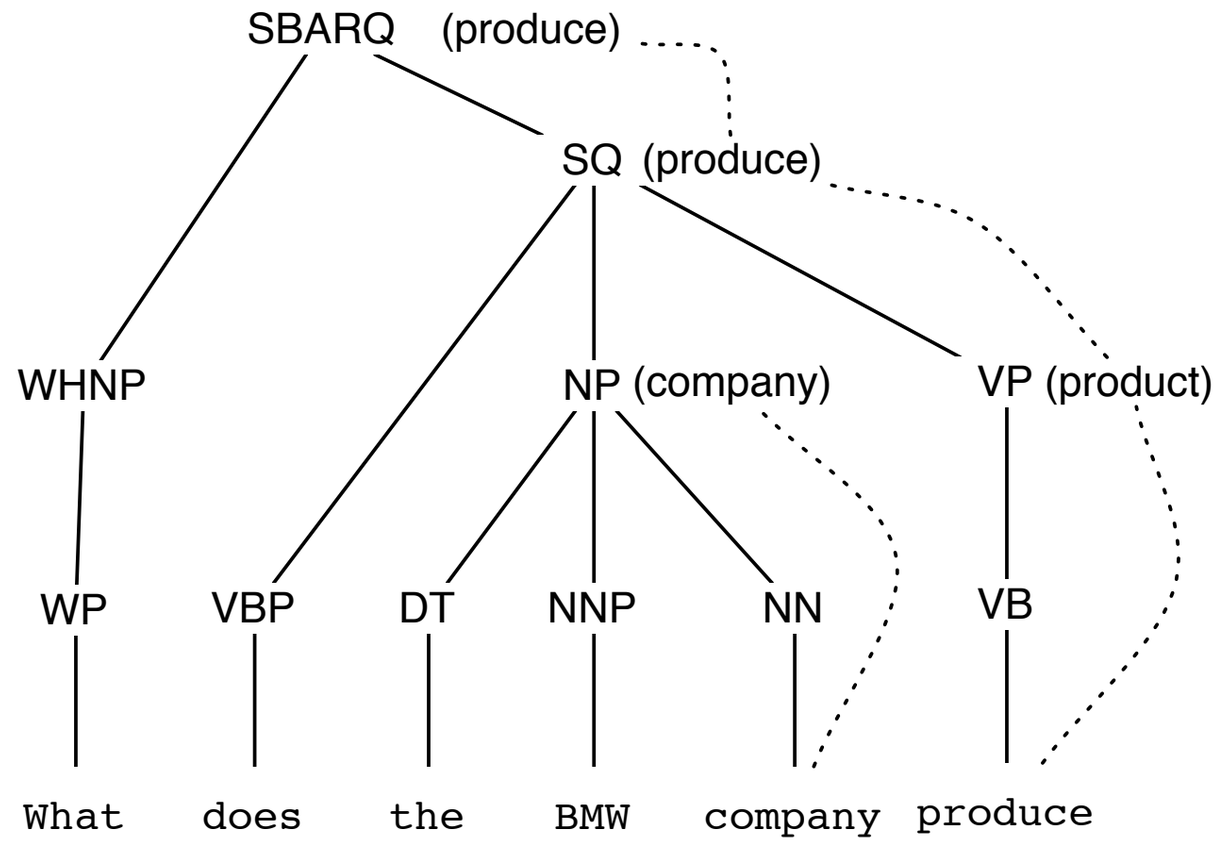
- Determining the type of expected answer
  - categories associated with question words: *who, what, where, how ...?*
  - dependencies extracted from parsing
- Where look for information ? in passage containing
  - the greatest number of representative concepts from the question
  - a concept from the right category

rali

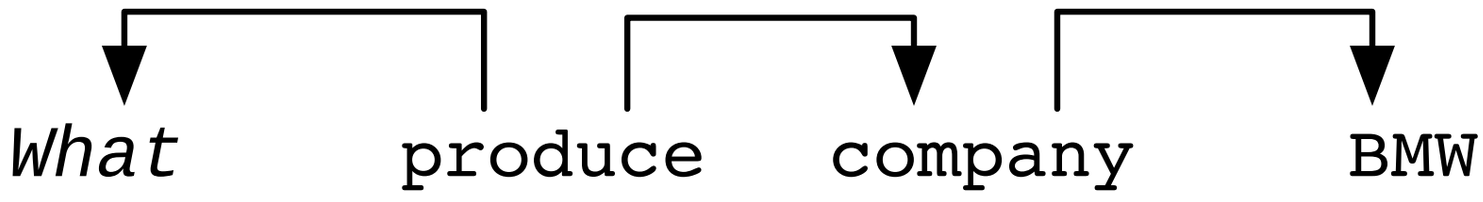
Dependencies extraction example (Harabagiu 03)



Question analysis



Extracted dependency model



Semantic form of the question

rali

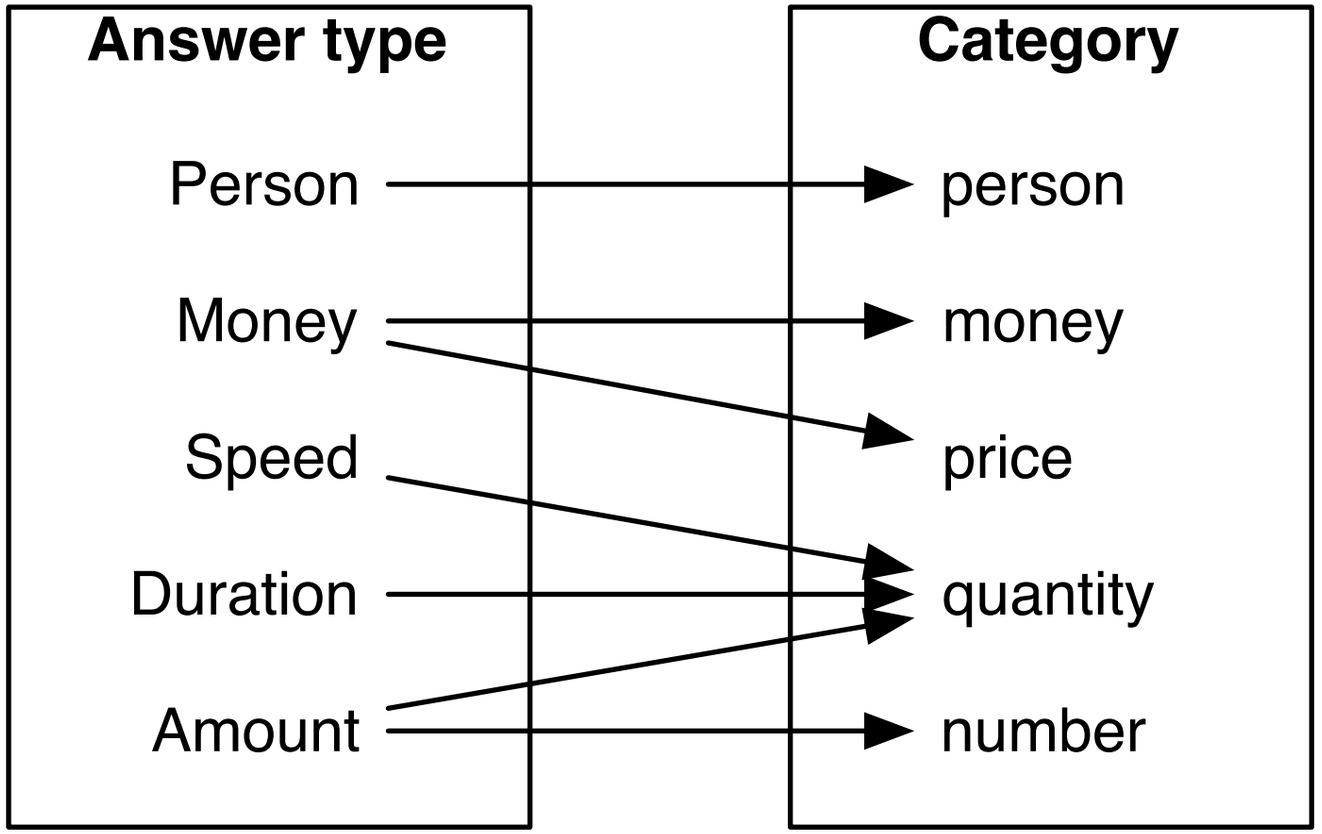
Answer types taxonomies - I

- (Hand) creation of a type hierarchy of interesting answers
  - Reason, Time, Person, Organization, Landmark...
  - Numerical Value
    - Degree, Dimension, Rate, Percentage, Count, Speed...
  - Location
    - University, City, Province, Country...

rali

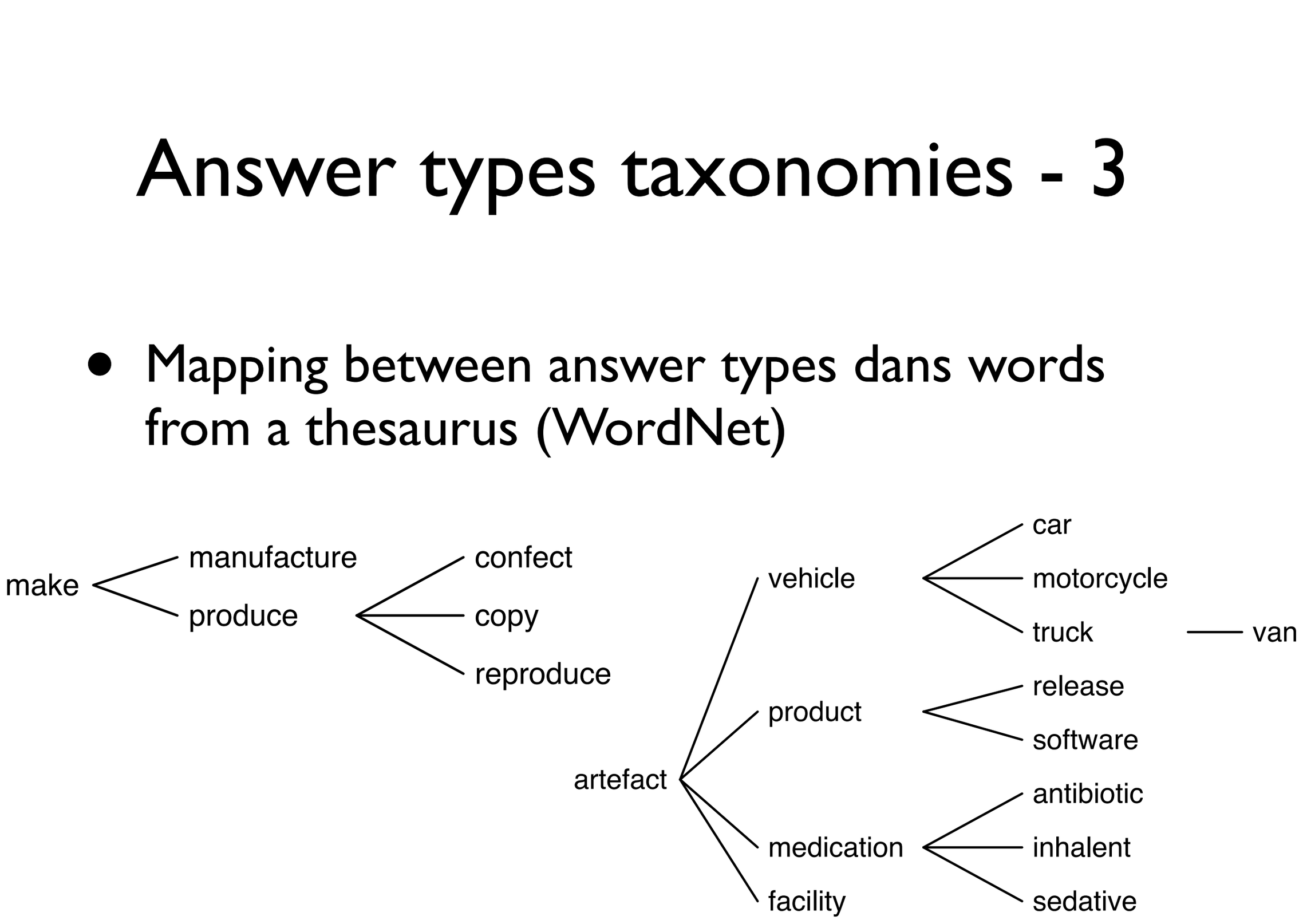
Answer types taxonomies -2

- Assignment of a many to many mapping between answer types and named entities types



Answer types taxonomies - 3

- Mapping between answer types dans words from a thesaurus (WordNet)



Two word hierarchies in WordNet

Keyword extraction from the question

- Content words
  - quoted expressions
  - named entities
  - complex noun phrases
- Nouns and their adjectival modifiers
- Verbs
- Modifications of those
  - morphology
  - synonyms
  - hyperonyms

rali

Document processing

- Passage determination
  - fixed or variable length
  - overlapping or not
- Relevant passage search with IR methods
  - number of similar keywords (in the same order ?)
  - number of different keywords
  - empirical thresholds

rali

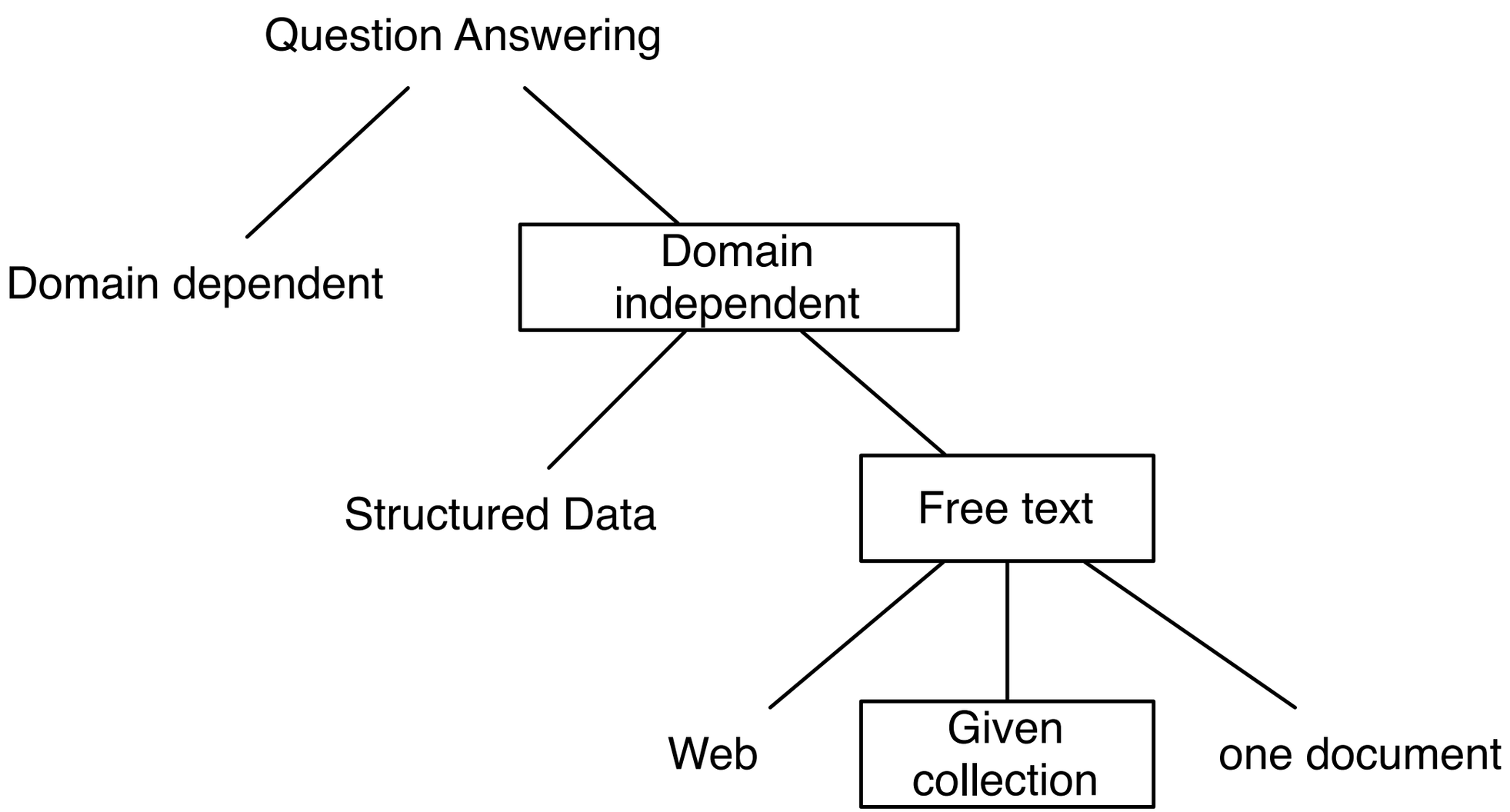
Answer extraction

- shallow parsing and hand-crafted pattern applications
- distance computation between passage subsequences and keywords
  - weight score optimization
  - entropy maximization
  - matching function by learning

rali

Types of QA

TREC



rali

TREC-QA Competitions
Evolution - I

- TREC-8 (1999)
  - documents: 1,9 GB (528 000 docs, ~3 600B/doc)
  - questions : 200 short questions with a guaranteed answer in the corpus
  - for each question: 5 answers of 50 or 250 bytes with a supporting document number
  - answer can be an extract or be generated
  - score : Mean Reciprocal Rank (MRR)

rali

TREC-QA Competitions Evolution - 2

- TREC-9 (2000) (same task as TREC-8)
  - documents: 3 GB (979 000 docs, ~3 000B/doc)
  - questions
    - 500 questions extracted from logs (Encarta et Excite)
      - harder : `Who is Colin Powell ?`
      - answer do not necessarily use the words of the question
    - 193 variants with same meaning as 54 questions for checking robustness
      - `What is the tallest mountain ?`
      - `What it the highest mountain in the world ?`
      - `Name the highest mountain.`

TREC-QA Competitions Evolution - 3

- TREC-2001 (same documents as TREC-9)
  - questions
    - 500 questions out of logs (MSNSearch et AskJeeves)
      - automaticaly and manually filtered
      - some questions do not have answers in the document (NIL)
      - definitions questions
    - list questions (cardinality given)
      - `Name 4 US cities that have “Schubert” theater.`
      - `Who are 6 actors who have played Tevye in “Fiddler on the Roof” ?`
      - `What are 9 novels written by John Updike ?`

TREC-QA Competitions Evolution - 4

- TREC-2002
  - documents : AQUAINT corpus of English News
  - questions
    - 500 questions out of logs (MSNSearch et AskJeeves)
      - automatically and manually filtered
      - some questions do not have an answer in the documents (NIL)
      - no definition question
    - a single *exact* answer by question
    - each answer must be assigned a confidence score that the system has in this answer

TREC-QA Competitions
Evolution - 5

- TREC-2003
  - documents : AQUAINT corpus of English News
  - questions
    - combine closed, list and definition questions
    - list of undefined cardinality
    - each question is typed and its scoring depends on it
    - final score is weighted mean of question scores

rali

Compétitions TREC-QA
Évolution-6

- TREC-2004 (same documents as 2003)
  - questions around a same subject simili-dialog
    - `Club Med`
      - `FACTOID : How many Club Med vacation spots are there world wide?`
      - `LIST : List the spots in the United States.`
      - `FACTOID : Where is an adults-only Club Med?`
      - `OTHER`
  - 23 PERSON, 25 ORGANIZATION, 17 THING
  - 230 FACTOID, 56 LIST, 65 OTHER

rali

TREC-QA Competitions Evolution - 7

- TREC-2005 (same task as 2004)
  - questions about a single topic, simili-dialog
  - new category: EVENT
  - systems must return a sorted list of their relevant documents which will be evaluated for their relevance

rali

TREC-QA Competitions Conclusion

- Evolution from factoid isolated questions towards a more realistic *analyst type* use context
- With a closed collection but the web can act as an *inspiration source*
- Different evaluation measures according to the question types
- Established the domain, created corpora and evaluation tools
- Still far from having deals with the *real* problem of understanding

rali

Other QA Competitions - I

- CLEF-QA (2003-)
  - questions and documents in 8 languages other than English
  - questions in one language and documents in another (73 combinations in 2005)
  - 200 questions (factual, description, mean, object and NIL) with exact answers (2004-)
- NTCIR-QAC (NII Test Collection for IR Systems)
  - question and answer in Japanese
  - multilingual (English, Chinese and Japanese) in 2005

rali

Other QA Competitions - 2

- EQUER (2004)
  - 500 questions in French
    - definition, factual, list, boolean, NIL
  - answers
    - passages of at most 250 characters
    - *short* answers
  - tasks
    - general : 1,5GB newspapers and official (French Senate)
    - medical : 140 MB scientific articles and practical recommandations

rali

Question Answering
Recent approaches

- Question analysis
  - category
  - expected answer type
  - semantic constraints
  - important words
- Answer searching
  - passage searching with important words
  - named entity detection according to the expected type
  - find the entity satisfying these constraints
- Confirmation with external resources ?

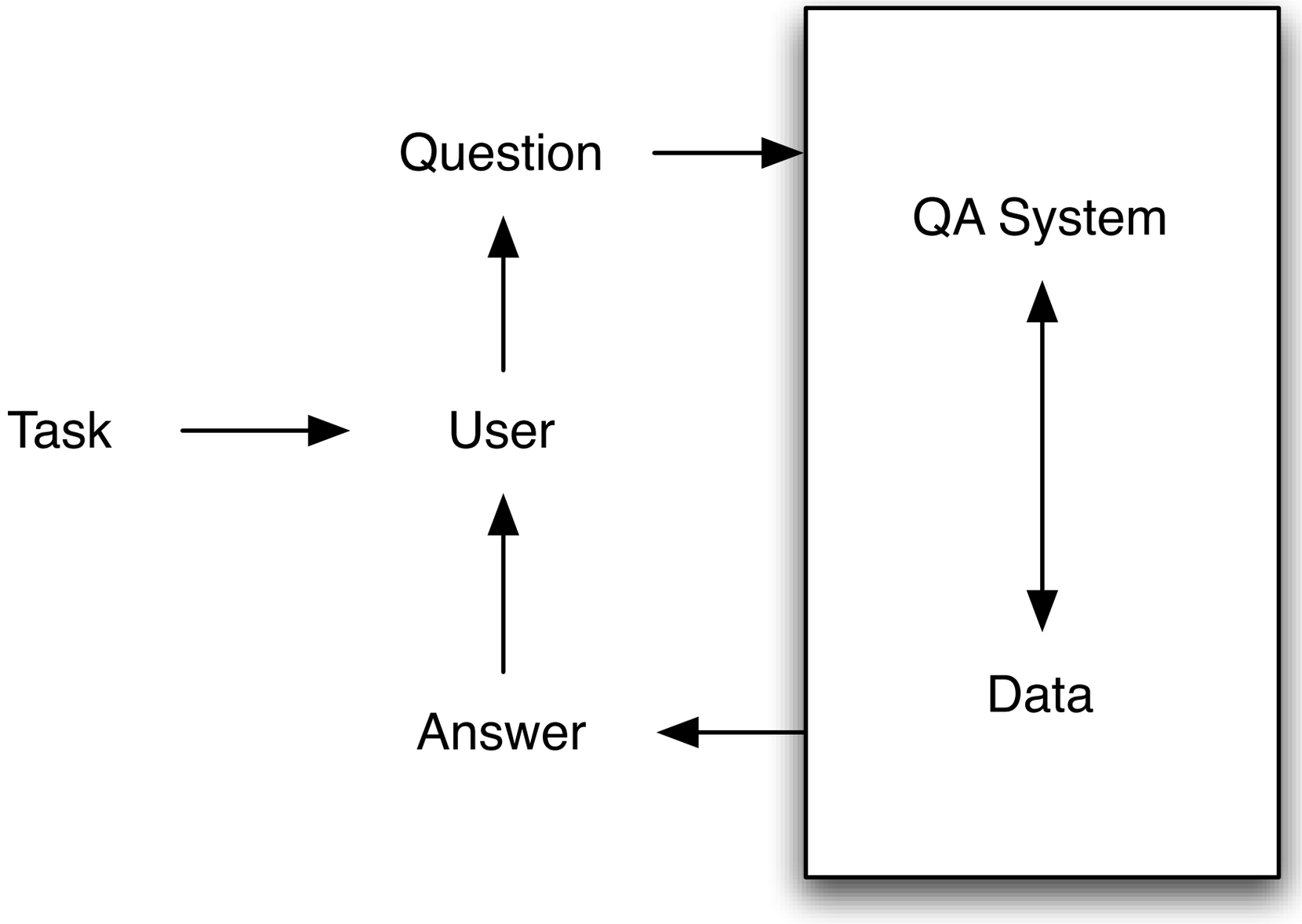
rali

QA - Evaluation

- Answer (including NIL)
  - correct (all the answer and nothing else)
  - inexact (right but imprecise or with supplementary information)
  - wrong
  - not supported by the document (correct *lenient* evaluation)
- Delicate task even for factoid questions
- Painstaking manual process

rali

QA-Evaluation
Issues to consider



rali

Evaluation
issues to consider I

- user
  - age, language level, social and cultural environment
- question
  - for a given task or to test the system ...
  - granularity
  - what is an easy or hard question ?
- system
  - application area
  - application type: commerce, customer support, welcome
  - data bases
  - internal working

rali

Evaluation
issues to consider 2

- data
  - size and heterogeneity
- answer
  - depends on the database content!!!
  - also depends on the question
    - `how many pounds in a ton ?`
    - `what is the average body temperature ?`
- task
  - fill lack of knowledge, curiosity or interest
  - fill lack of information in order to achieve an action

rali

Evaluation examples questions EQUER - 1

- Qui a épousé Bill Gates à Hawaii ?
  Who married Bill Gates in Hawaii ?
  - `Melinda` (correct if justified)
  - `Melinda French` (correct if justified)
  - `sa collègue Melinda French` (correct if justified)
  - `a épousé sa collègue Melinda French` (wrong)
- Quel âge a l'abbé Pierre ? *How old is Abbé Pierre?*
  - `81 ans` (correct if justified)
  - `quatre-vingt-cinq ans` (correct if justified)
- Quel est le record du monde du 100 mètres ?*What is the 100 meter world record ?*
  - `48s42` (correct if justified, on foot or swimming)

rali

Evaluation examples questions EQUER - 2

- Que signifie CGT ? *What does CGT mean ?*
  - `Confédération générale du travail`
  - `Confédération générale des travailleurs`
- Quelle est la définition de la chimiothérapie ?
  What is the definition of chemotherapy ?
  - `Thérapeutique par les substances chimiques`
  - `un traitement complémentaire au traitement chirurgical`
- Citez quatre infractions militaires
  Give four military infractions ?
  - `insoumission, capitulation, insubordination, désertion`

rali

Evaluation examples questions EQUER - 3

- Existe-t-il une ligne de TGV Valenciennes-Paris?
  Is there a TGV line between Valenciennes and Paris ?
  - yes if the supporting passage is :
    les 170 passagers d'un TGV Valenciennes-Paris en avaient été quitte pour la peur de leur vie le 21 décembre vers 07H30 du matin.
    the 170 passengers of a Valenciennes-Paris TGV had a frightening experience on December 21th at about 7:30 AM.

Question Classification TREC-2000 (Lavenus02)

rg	%RV	No	Questions	F	H	tal	def	den	EN	log
1	37%	1110	What is sodium chloride?	X			X			
2	26%	1112	How many pounds in a ton?	X					X	
3	23%	1113	What is influenza?		X		X			
4	19%	1100	Who was the 23rd president of the United States?	X					X	
5	17%	1101	What is the average body temperature?	X					X	
6	16%	1095	What city's newspaper is called "The Enquirer"?	X					X	
7	15%	1094	What is the name of the satellite that the Soviet Union sent into space in 1957?	X				X		
8	14%	1102	What does a defibrillator do?		X					X
9	14%	1106	What province is Montreal in?	X					X	
10	12%	1105	How fast is the speed of light?	X					X	
11	7%	1107	What New York City structure is also known as the Twin Towers?			X				
12	5%	1097	What are the animals that don't have backbones called?			X		X		

F : Factual

H: Heterogeneous

tal: raising NLP problems

def : definitions

den : denominations

EN : Named entities

log: logical links

rali

MRR (mean reciprocal rank)

- reciprocal of the rank of the first right answer
- 0 if it does not appear in the first 5
- total score is the mean of all MRRs

	# Systems	MRR 1	MRR 5	# not found	
TREC-8	20	0.66	0.32	27 %	56 %
TREC-9	28	0.58	0.29	34 %	58 %
TREC-2001	36	0.68	0.43	31%	43%

rali

Confidence-Weighted Score

- measures a system's ability to acknowledge its confidence in its answers

$$\frac{1}{Q}\sum_{i=1}^{Q}\frac{\text{number of right answers within the first } i}{i}$$

	# Systems	Score 1	Score 5	# judged correct	
TREC-2002	34	0.86	0.59	83%	36%

List evaluation

- TREC 2003: lists evaluated with F-measure

$$
\begin{aligned}
S &= \text{number known answers} \\
D &= \text{number of distinct right answers} \\
N &= \text{total number of answers} \\
IP &= D/N \\
IRe &= D/S \\
F &= \frac{2 \times IP \times IRe}{IP + IRe}
\end{aligned}
$$

Definitions evaluation

- TREC 2003: definitions evaluated with *nuggets* of information

$$
\begin{aligned}
r &= \text{number of essential nuggets in an answer} \\
a &= \text{number of right acceptable nuggets in an answer} \\
R &= \text{total number of essential nuggets for this question} \\
long &= \text{number of characters in an answer} \\
Recall &= r/R \\
Provision &= 100 \times (r + a) \\
Precision &= \begin{cases} 1 & \text{if } long < Provision \\ 1 - \frac{long - Provision}{long} & \text{otherwise} \end{cases} \\
F(\beta = 5) &= \frac{26 \times Precision \times Recall}{25 \times Precision + Recall}
\end{aligned}
$$

Results of these measures
25 groups at TREC 2003

	Closed classes	List	Defini-tions	Final
1	0.70	0.39	0.44	0.55
5	0.20	0.09	0.55	0.27

rali

Evaluations

- Surprising difficult even for *simple* tasks
- Validations are needed to check for stability between assessors
- Validations for checking question sensibility

rali

Commercial Applications - 1 (Maybury 2004, p 17)

Year	System	Method used	Users
1993	MIT Start	NL Interface for the Web, tuple matching	Web
1993	EasyAsk	QA access to product catalogs	Talbot's, Land's End
1994	Broad Mind	FAQ searching	Kodak, SEC,...
1996	MURAX	NL Interface to an encyclopaedia	Web
1996	Ask Jeeves	Matching of new questions to already answered ones	Web
1996	Kanisa	QA Shell with question and answer models	
1998	Mercado	QA on different source of product catalogs	Macy's, Sears,..
1999	LingoMotors	Logical forms and discourse structure (Pustejovsky)	

rali

Commercial Applications - 2
(Maybury 2004, p 17)

Year	System	Method used	Users
1999	IPLearn	Question transformation in internal commands	
2000	FactCity	QA from tabular data	
2002	Inquira	*Rich* indexing of document	Bank of Am., Honda

Projet Merkure

Université de Montréal 99-03

http://rali.iro.umontreal.ca/ExtractionDInformation/index.fr.html#merkure

- Automatic e-mail processing for investor relationships (BCE)
  - classification
  - case-based reasoning
  - question answering

QRISTAL

Questions-Réponses Intégrant un Système de Traitement Automatique des Langues

La réponse à toutes vos questions

SYNAPSE DEVELOPPEMENT

PAGES

- Accueil
- Présentation
- Téléchargement
- Support - Conseil
- Enregistrement

LIENS

- Site de Synapse
- Boutique en ligne
- Revue de Presse

AUTRES

- Intégration - SDK
- Nous contacter
- Exemples
- Qui sommes-nous

Le Site de Rolland

Logiciel de Questions-Réponses sur disque dur et Internet

● QRISTAL : Moteur de recherche multilingue intégrant un Système de Traitement Automatique des Langues. Si vous avez déjà recherché des réponses sur le Web sans les trouver, si vous avez déjà passé des heures à rechercher le courrier envoyé à Untel ou le document sur tel sujet, **QRISTAL est fait pour vous !**

Démo - Acheter

Posez une question avec Qristal

● Posez votre question en langage normal : Qristal s'appuie sur la technologie d'analyse du langage Cordial permettant d'effectuer une analyse syntaxique et sémantique des phrases. Vous posez normalement vos questions, le logiciel se charge de leur analyse. ...

● Sur Internet : Qristal intègre un méta-moteur de recherche Internet puis applique sur les pages résultats des moteurs ses techniques d'analyse avancée pour extraire les réponses. ...

● Sur disque dur : Qristal indexe votre disque dur puis vous permet de rechercher directement sur les documents présents sur votre ordinateur. ...

● Réponses : Une réponse simple et précise vous est proposée en plus des listes de documents classiques des moteurs de recherche. ...

Exemples sur Internet :

"Est-ce que Arsenal est un club de Londres?" Réponse Qristal -> "Oui" ...

"Quel âge a Nicolas Sarkozy ?", Qristal -> "49 ans (date de naissance : 28 janvier 1955)" ...

● Recherche multilingue : Qristal permet de poser vos questions et d'obtenir vos réponses en français, anglais, italien, portugais ou polonais. ...

Intégration de QRISTAL

● SDK : Vous pouvez intégrer le moteur de recherche Qristal directement dans vos applications en entreprise. Le moteur est disponible sous la forme de SDK et peut très rapidement s'intégrer dans un applicatif client, un site Internet, un système Intranet, etc. ...

Synapse Développement

● Editeur : Qristal est le nouveau logiciel de l'éditeur Synapse Développement, spécialiste des technologie de traitement avancé de textes. ...

rali

http://www.qristal.fr

Commercial results

- For the moment, no great *success story*
- More experimental tests and system merging
- 75% of sites with QA could not answer to these questions
  - `What is your product return policy?`
  - `How do I contact customer support ?`
  - `Do you sell gift certificate of less than $50 ?`

Lessons learned on QA systems

- Based on shallow parsing methods
- Clients want to control their own content
- Buyers of these technologies would like to use them to better understand their clients
- Not all requests are equal: a few frequent questions make up for the majority (Zipf law)
- Would need to access to a mix of heterogeneous data and answer can be of different types

rali

Conclusion

- Hot topic but the TREC competition is a bit artificial
  - *easy* evaluation
  - comparable results
  - reusable corpus and methodology
- Far from solving the *true understanding* problem

rali

QA and *user-friendly* interfaces

- similar to question answering
- user should be induced to think that the computer is more *intelligent* than it is
- similar problem with voice interfaces

rali

Other challenges of QA

- Temporal events
- Multimedia
- Multiple sources combination
- Confidence measures in the answers
- Cooperative answers

rali

Cooperative answers exemples

- **How students passed the AI course this summer ?**
  - none
  - the AI course was not given this summer
- **Where is the photocopier ?**
  - in room 1340
  - it is broken
- **Where are the gills on a whale ?**
  - a whale does not have gills, it breathes with its lungs

Cooperative answers methods

- Difference detection between what the user think of the DB and what it really contains
- Integrity constraints use
  - all course must have less than 100 students
  - somebody who teaches must be a professor
- Explanation of any difference
  - integrity constraint violation
  - request contains redundant informations

rali

General Conclusion on IE and QA

- Evolution
  - from deep understanding motivated by cognitive problems
  - towards *practical* applications
- Decision aid system based on unstructured informations
- We are only starting to deal with the *real* problem
- Fuzzy boundary between these areas

rali

References

- M. El-Bèze, *Systèmes de question réponses*, chap 3.3 of *Compréhension des langues et interactions*, Hermès, 2006
- S. Harabagiu, D. Moldovan, Question Answering, chap 31, p 560-582 of *The Oxford Handbook of Computational Linguistics*, R. Mitkov, ed, 2003.
- M.T. Maybury, *New Directions in Question Answering*, MIT Press, 2004.
- K. Lavenus, G. Lapalme. *Evaluation des systèmes de question réponse. Aspects méthodologiques.* Traitement automatique des langues, vol. 43, n. 3, p. 181-208, jan 2002.

rali

References
Evaluation competitions

- E.M. Vorhees, D. K. Harman, *TREC: Experiment and Evaluation in Information Retrieval*, MIT Press, 2005, Chap 10, p 233-257.
- CLEF, Cross-Language Evaluation Forum, http://www.clef-campaign.org/
- NTCIR, Cross-Language Question Answering, http://www.slt.atr.jp/CLQA/
- C. Ayache, B. Grau, EVALDA-EQUER, http://www.technolangue.net/article195.html

rali