Ontology-based Information Extraction and Question Answering – Coming Together

Günter Neumann

LT lab, DFKI, Saarbrücken

What do I mean ?

☆Ontology-based information extraction

- Ontology defines target knowledge structures
  - i.e., type of entities, relations, templates
- IE for identifying and extracting instances
- Merging of partial instances by means of reasoning

What do I mean ?

☆ Question answering from text and Web

- Answering questions about who, what, whom, when, where or why
- Question analysis:
  - "Human carries ontology"
  - Identifies the partially instantiated relation expressed in a Wh-question
  - Identification of the "expected answer type"
- Answer extraction
  - The „information extraction" part of QA
  - Also here: RTE for validating answer candidates (cf. Clef 2007/2008)

Who is Prime Minister of Canada?
-> PM_of(person:X,country:Canada)
-> EAT=person

Stephen Harper was sworn in as Canada's 22nd Prime Minister on February 6, 2006. (Source: http://pm.gc.ca/eng/pm.asp)

Two Possible Approaches of OBIES+QA

☆ Entailment-based QA

- Domain ontology as interface between NL and DB
- Bijective mapping between NL patterns and DB patterns
- Textual entailment for mastering the mapping/reasoning
- EU project QALL ME

☆ Web-based ontology learning using QA

- Unsupervised methods for extracting answers for factoid, list and definition based question
- Basis for large-scale, web-based bottom-up knowledge extraction and ontology population
- BMBF project Hylap

Architectures of QA Systems

DB-QA

NL Question

Text-QA

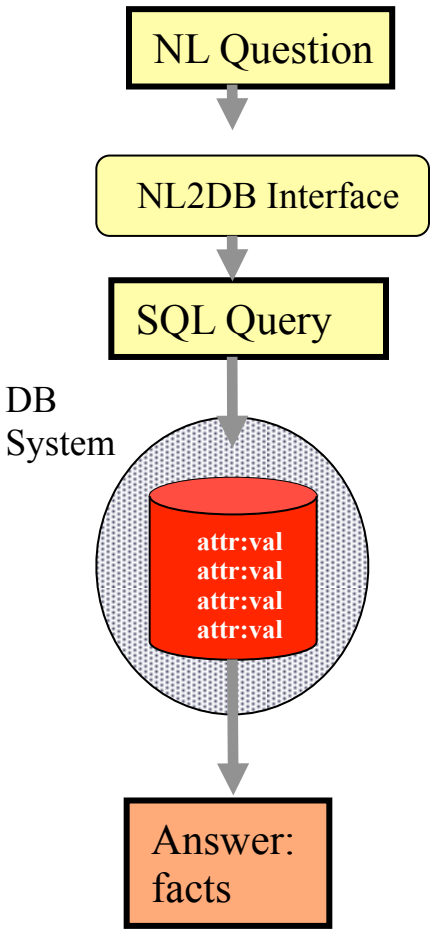
NL Question

Hybrid-QA

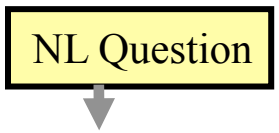
NL Question

Architectures of QA Systems

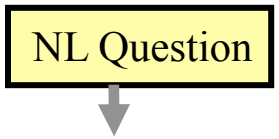
DB-QA



Text-QA

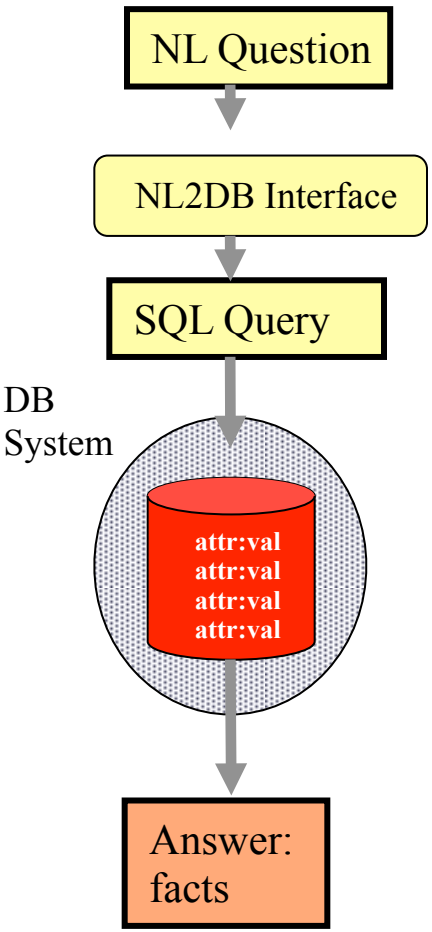


Hybrid-QA

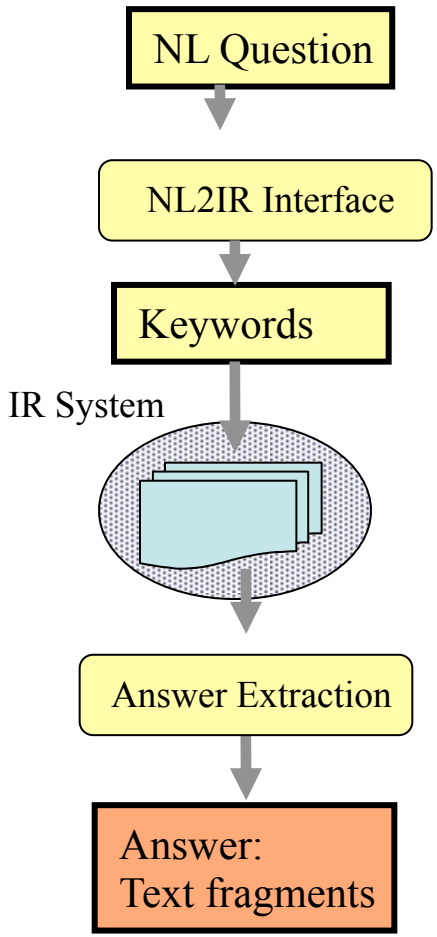


Architectures of QA Systems

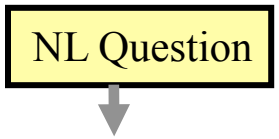
DB-QA



Text-QA

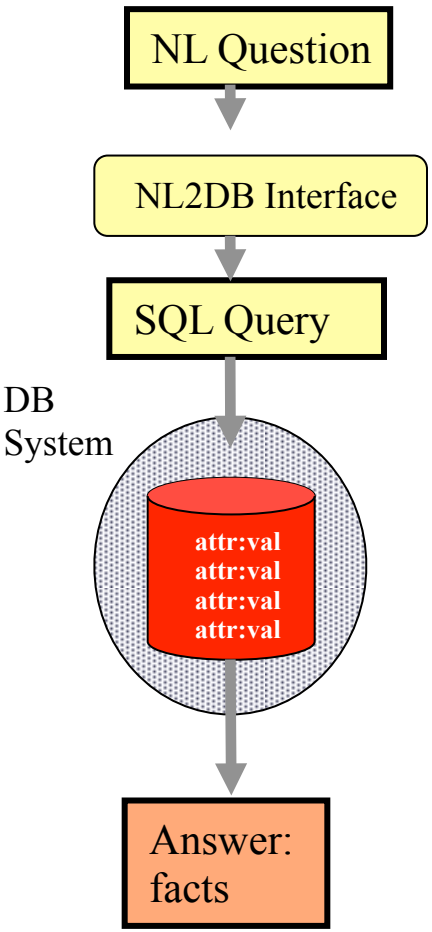


Hybrid-QA

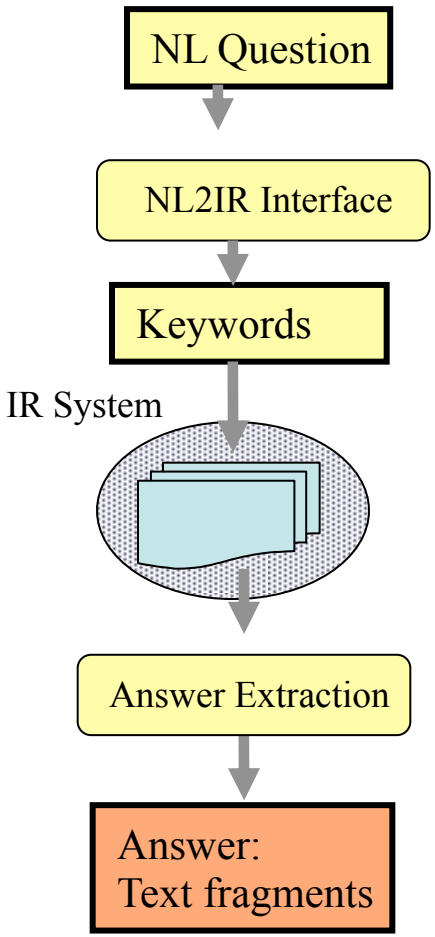


Architectures of QA Systems

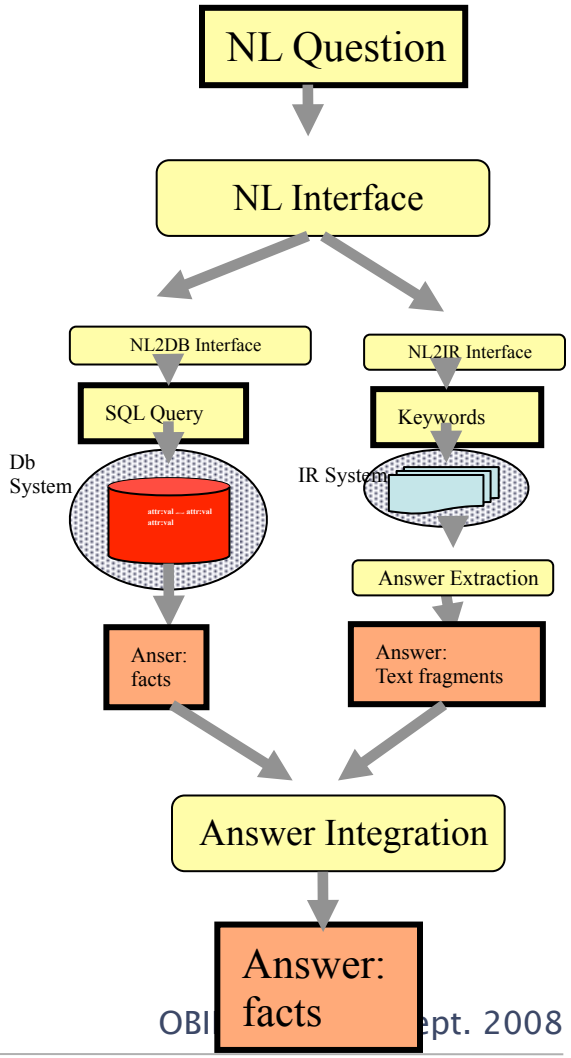
DB-QA



Text-QA

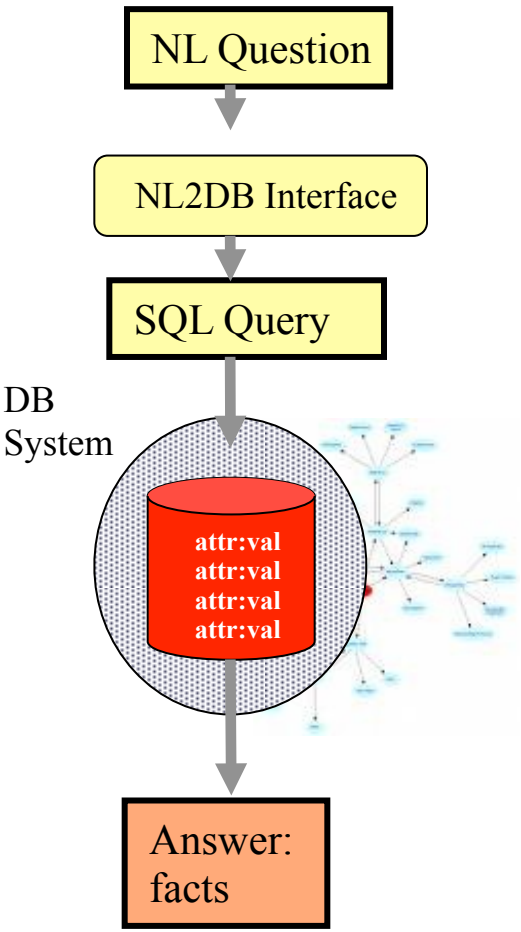


Hybrid-QA

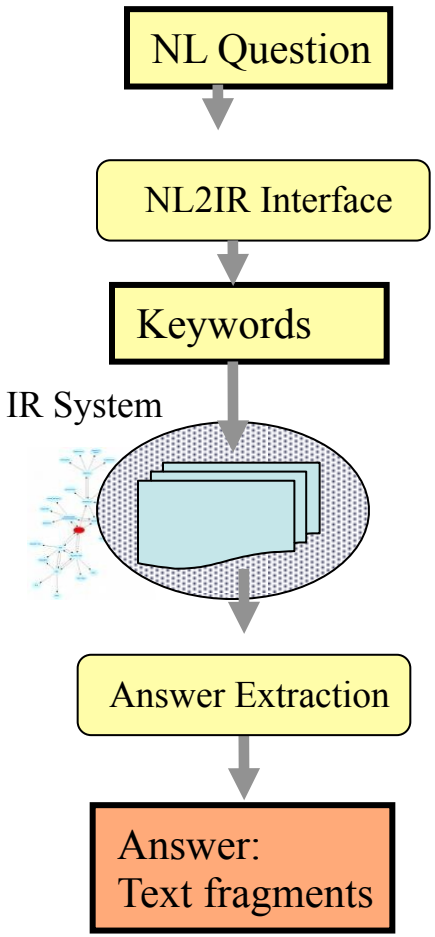


Architectures of QA Systems

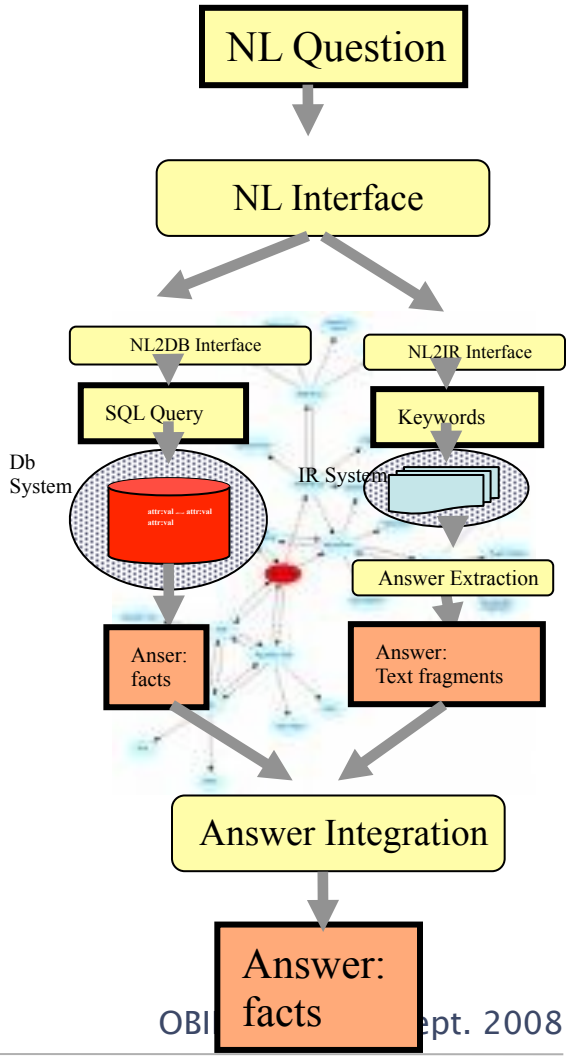
DB-QA



Text-QA



Hybrid-QA



The QA bottleneck

☆ Hybrid QA:

- Increase of semantic structure (Semantic Web, Web 2.0) ⇒ Fusion of ontology-based DBMS and information extraction from text
- Dynamics and interactivity of Web requests for additional **new** complexity of the NL interface.

„Who wrote the script of Saw III?"

Complex linguistic & knowledge-based reasoning

=

```
SELECT DISTINCT ?writerName  WHERE
{ ?movie name "Saw III"^^string . ?movie
hasWriter ?writer . ?writer name ?writerName . }
```

„Who is the author of the script of the movie Saw III?"

Possible approaches

☆ Full computation (inference)

- ⇒ AI complete; especially, if incomplete/wrong queries are allowed

☆ Controlled sublanguage

- A user may only express questions using a constrained grammar and with unambiguous meaning
- ⇒ cognitive burden is not acceptable

☆ Controlled mapping

- One-to-one mapping between NL patterns and DB-query patterns
- Flexible use of NL possible through methods of textual inference

Textual Inference

- ☆ Motivation: textual variability of semantic expressions
- ☆ Idea: for two text expressions T & H:
  - Does text T justify an inference of hypothesis H?
  - Is H semantically entailed in T?

Prof. Clever, full professor at Bostford University, published a new paper.

?

Prof. Clever works at Bostford University.

- ☆ PASCAL Recognizing Textual Entailment (RTE) Challenge
  - since 2005, cf. Dagan et al.
  - 2008: 4th RTE (at TAC), 26 groups (two subtasks)
- ☆ RTE is considered as a core technology for a number of text based applications:
  - QA, IE, semantic search, text summarization, …

Textual Inference for QA

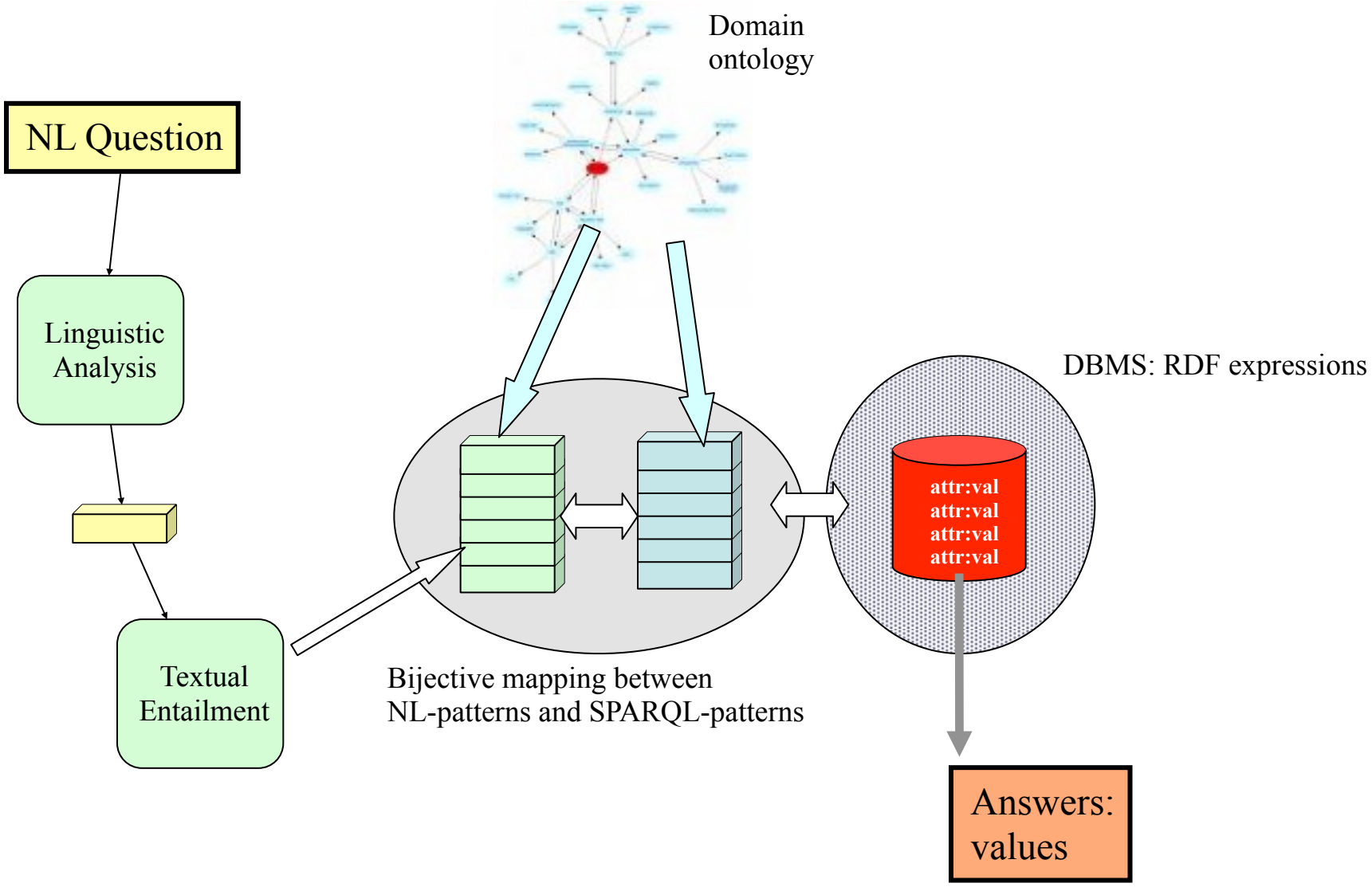
☆ RTE successfully applied to answer validation

- Example
  - Q: „In which country was Edouard Balladur born?", A: "France"
  - T: „*Paris, Wednesday CONSERVATIVE Prime Minister Edouard Balladur, defeated in France's presidential election, resigned today clearing the way for President-elect Jacques Chirac to form his own new government…*"
- Entailed(Q+A, T) ⇒ YES/NO ?
- Clef 2008, AVE task ⇒ DFKI best results for English and German

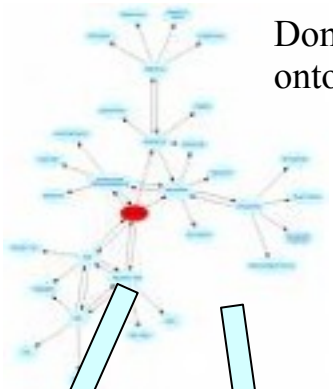
☆ New: RTE for semantic search

- Does question X entail an (already answered) question Y ?

Current Control Flow



Current Control Flow

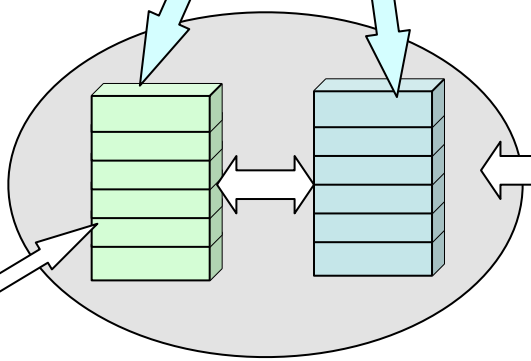


Domain ontology

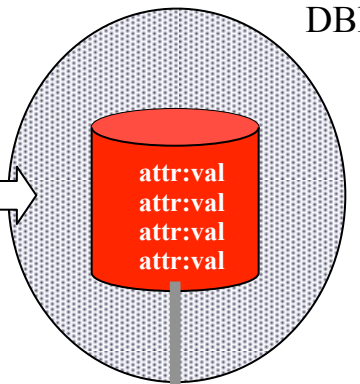
NL Question

Linguistic Analysis

Textual Entailment



Bijective mapping between NL-patterns and SPARQL-patterns



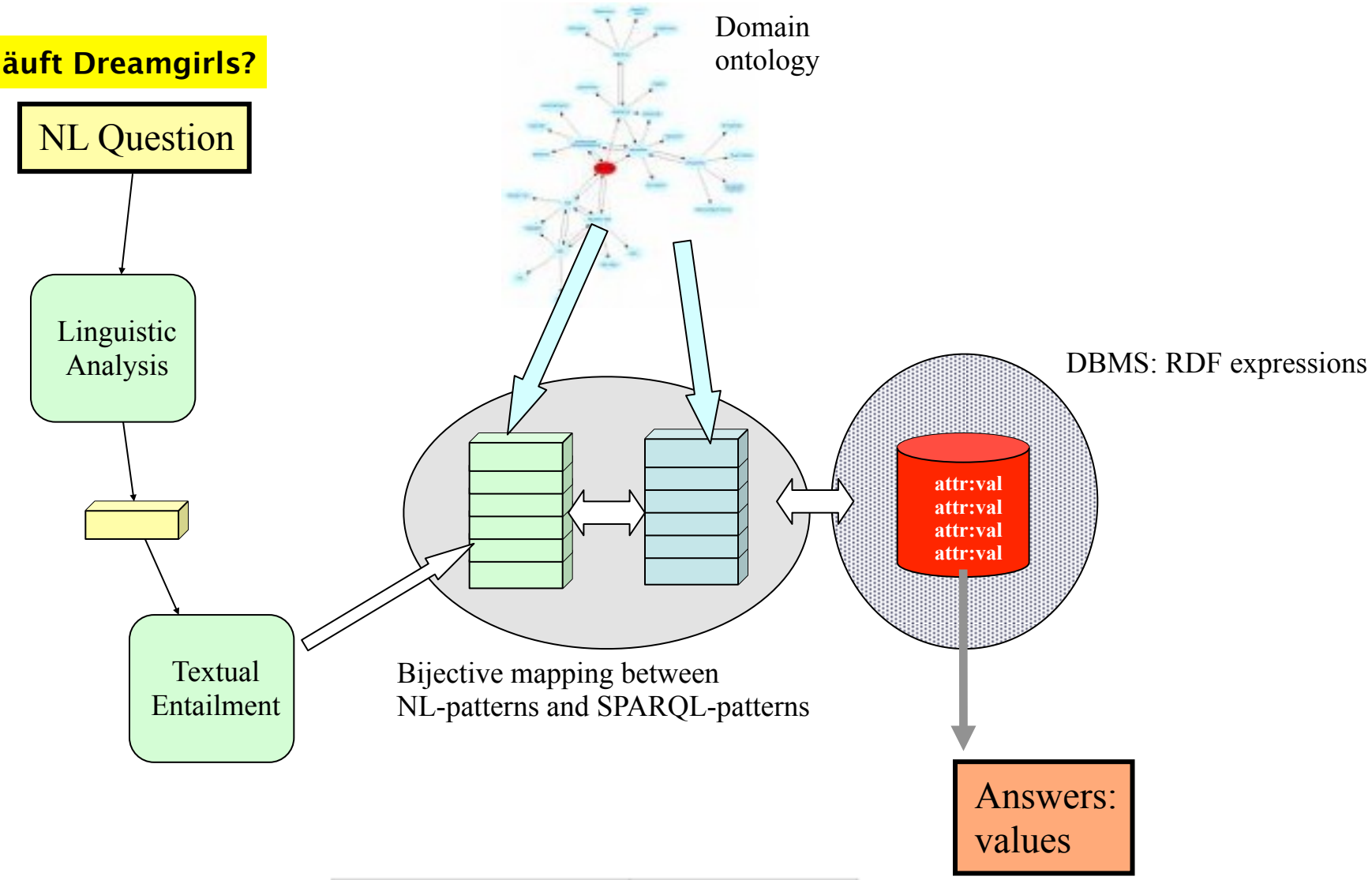
DBMS: RDF expressions

Answers: values

Frage Muster	DB-Anfrage Muster (Ausschnitte)
In welchem Kino kann man [MOVIE] sehen?	SELECT ?cinema ...
Wo ist das Kino [CINEMA]?	SELECT ?location ...
Wer führte bei dem Film [MOVIE] die Regie?	SELECT ?director ...

Current Control Flow

Wo läuft Dreamgirls?



Frage Muster	DB-Anfrage Muster (Ausschnitte)
In welchem Kino kann man [MOVIE] sehen?	SELECT ?cinema ...
Wo ist das Kino [CINEMA]?	SELECT ?location ...
Wer führte bei dem Film [MOVIE] die Regie?	SELECT ?director ...

Current Control Flow

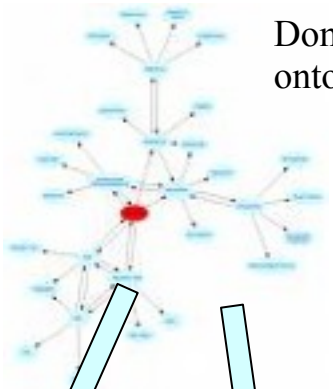
Wo läuft Dreamgirls?

NL Question

Linguistic Analysis

Wo läuft [movie]?

Textual Entailment



Domain ontology

DBMS: RDF expressions

attr:val
attr:val
attr:val
attr:val

Bijective mapping between NL-patterns and SPARQL-patterns

Answers: values

Frage Muster	DB-Anfrage Muster (Ausschnitte)
In welchem Kino kann man [MOVIE] sehen?	SELECT ?cinema ...
Wo ist das Kino [CINEMA]?	SELECT ?location ...
Wer führte bei dem Film [MOVIE] die Regie?	SELECT ?director ...

Current Control Flow

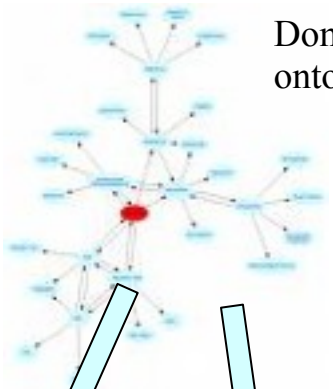
Wo läuft Dreamgirls?

NL Question

Linguistic Analysis

Wo läuft [movie]?

Textual Entailment



Domain ontology

DBMS: RDF expressions

attr:val
attr:val
attr:val
attr:val

Bijective mapping between NL-patterns and SPARQL-patterns

Answers: values

Frage Muster	DB-Anfrage Muster (Ausschnitte)
In welchem Kino kann man [MOVIE] sehen?	SELECT ?cinema ...
Wo ist das Kino [CINEMA]?	SELECT ?location ...
Wer führte bei dem Film [MOVIE] die Regie?	SELECT ?director ...

Current Control Flow

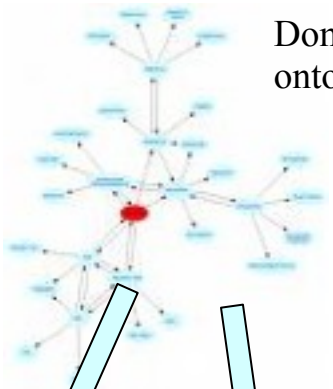
Wo läuft Dreamgirls?

NL Question

Linguistic Analysis

Wo läuft [movie]?

Textual Entailment



Domain ontology

DBMS: RDF expressions

attr:val
attr:val
attr:val
attr:val

Bijective mapping between NL-patterns and SPARQL-patterns

**"SELECT ?cinema …
WHERE ?movie name Dreamgirls …"**

Answers: values

Frage Muster	DB-Anfrage Muster (Ausschnitte)
In welchem Kino kann man [MOVIE] sehen?	SELECT ?cinema …
Wo ist das Kino [CINEMA]?	SELECT ?location …
Wer führte bei dem Film [MOVIE] die Regie?	SELECT ?director …

Current Control Flow

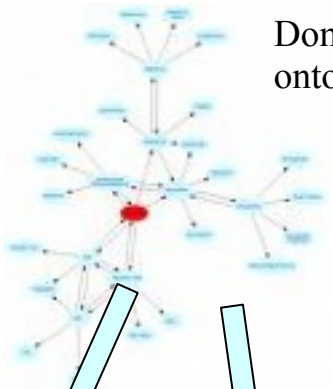
Wo läuft Dreamgirls?

NL Question

Linguistic Analysis

Wo läuft [movie]?

Textual Entailment



Domain ontology

DBMS: RDF expressions

attr:val
attr:val
attr:val
attr:val

Bijective mapping between NL-patterns and SPARQL-patterns

**"SELECT ?cinema …
WHERE ?movie name Dreamgirls …"**

Answers: values

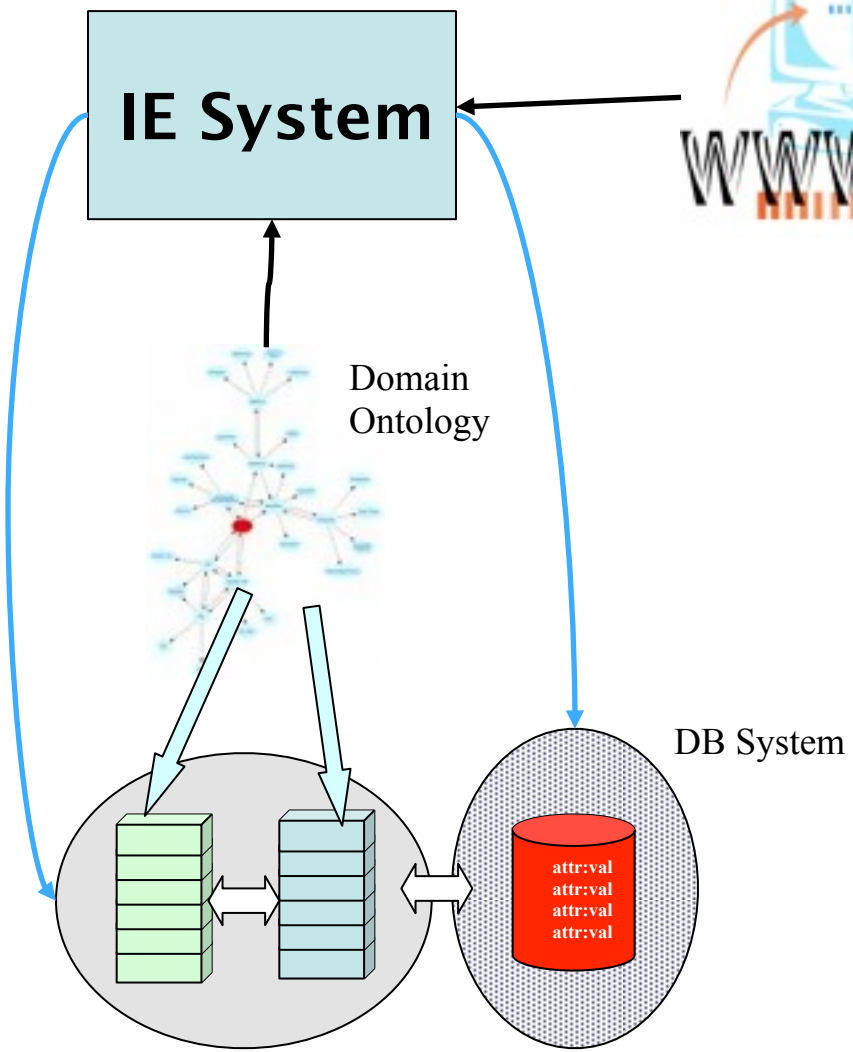
Xanadu

Frage Muster	DB-Anfrage Muster (Ausschnitte)
In welchem Kino kann man [MOVIE] sehen?	SELECT ?cinema …
Wo ist das Kino [CINEMA]?	SELECT ?location …
Wer führte bei dem Film [MOVIE] die Regie?	SELECT ?director …

Advantages

- ☆ Inference remains on the linguistic level
- ☆ RTE method are by definition robust ⇒
  supports processing of underspecified/illspecified requests
- ☆ Good interplay with ontology-based DB
- ☆ Opens up possibility to automatically learn mappings via ontology-based information extraction

- ☆ Extraction of relevant information from textual sources (Web pages)
- ☆ Integration of the extracted data into current DB
- ☆ Domain ontology as starting point:
  - Relevance
  - Normalization
  - Mapping



Possible approaches

☆ Bootstrapping an ontology

- Basic components for handling IE-specific subtasks expressed as Wh-questions
- Unsupervised, language-indepdent approaches
- Populating/extending domain ontology

☆ Interactive dynamic information extraction

- Topic-based web crawling
- IE system mines for all possible relevant entities and relations
- See talk on Eichler et al., Friday, 13:30

Unsupervised Web-based Question Answering for ontology bootstrapping

☆ Our goal:

- Development of ML-based strategies for complete **end-to-end** answer extraction for different types of questions and the open domain.

☆ Our perspective:

- Extract exact answers for different types of questions **only** from web snippets
- Use strong data-driven strategies
- Evaluate them with Trec/Clef Q-A pairs

☆ Our current results:

- ML-based strategies for open domain **factoid**, **definition** and **list** questions
- Question type specific query expansion for controlling web search
- **Unsupervised** learning for answer extraction
- Promising performance (~ 0.5 MRR on Trec/Clef data)

Unsupervised Web-based Question Answering for ontology bootstrapping

☆ Our goal:

- Development of ML-based strategies for complete **end-to-end** answer extraction for different types of questions and the open domain.

☆ Our perspective:

- Extract exact answers for different types of questions **only** from web snippets
- Use strong data-driven strategies
- Evaluate them with Trec/Clef Q-A pairs

F: When was Madonna born?
D: What is Ubuntu?
L: What movies did James Dean appear in?

☆ Our current results:

- ML-based strategies for open domain **factoid**, **definition** and **list** questions
- Question type specific query expansion for controlling web search
- **Unsupervised** learning for answer extraction
- Promising performance (~ 0.5 MRR on Trec/Clef data)

Current ML-based Web-QA System

List-WQA

List Extraction

List context

Def-WQA

NL-Question

Surface E-patterns

Definition context

Definition Extraction

Clusters of Potential senses

…

Snippets

GA-QA

NL-string(s)

Genetic Algorithms

Lexico-syntactic patterns

Exact answ 1
Exact answ 2
…

Snippets

QA-History

(feedback Loops)

Snippets

Answer Prediction

Answer Context

Extraction via Trivial patterns

Factoid-WQA

German Research Center for Artificial Intelligence

Factoid-WQA - Technology

http://amasci.com/tesla/tradio.txt TESLA **INVENTED RADIO? ... He invented modern radio, but made such serious business mistakes that the recognition (to say ...**

- ☆ Consult only snippets
  - Submit NL question string (no query refinement, expansion, reformulation, …)
- ☆ Goal
  - Identify smallest possible phrases from snippets that contain exact answers (AP phrases)
  - Do not make use of any smoothing technology or pre-specified window sizes or length of phrases

The prime minister Tony Blair said that

- ☆ Answer extraction
  - Use only very trivial patterns for extracting exact answers from AP phrases
  - Only Wh-keywords, distinguish type of tokens, punctuatio symbols for sentence splitting

The prime minister Tony Blair said that

Who → Person; When → Time

Experiments

Results for each question type over all languages.

CA	Total	MRR	NAG(%)	WAG(%)	NAF(%)	1(%)	2(%)	3(%)
WHEN	218	0.60	25.11	10.96	21.46	35.16	5.02	1.8
WHERE	232	0.57	10.77	24.14	20.68	30.60	9.91	3.87
WHO	439	0.38	11.39	27.56	32.57	18.90	6.83	2.73

Distribution of answer candidates (all languages).

CA	NAF(%)	1(%)	2(%)	3(%)
WHEN	33.82	55.42	7.91	2.84
WHERE	31.86	47.00	15.23	5.95
WHO	53.37	30.97	11.19	4.47

Experiments

The results for the individual languages.

CA(EN)	Total	MRR	NAG(%)	WAG(%)	NAF(%)	1(%)	2(%)	3(%)
when	69	0.69	15.69	15.69	17.65	45.10	3.92	1.96
where	64	0.74	7.81	12.5	15.62	53.12	10.93	0
who	148	0.50	7.43	12.83	32.43	33.78	10.14	3.38
CA(DE)	**Total**	**MRR**	**NAG(%)**	**WAG(%)**	**NAF(%)**	**1(%)**	**2(%)**	**3(%)**
Wann	58	0.45	36.20	12.07	27.59	22.03	1.17	0
Wo	58	0.46	9.37	18.75	23.43	20.31	12.5	6.25
CA(ES)	**Total**	**MRR**	**NAG(%)**	**WAG(%)**	**NAF(%)**	**1(%)**	**2(%)**	**3(%)**
Cuándo	59	0.55	16.64	11.86	23.73	32.20	10.17	11.86
Dónde	63	0.59	10.93	31.25	15.62	26.56	10.93	3.21
Quién	86	0.27	9.65	40.68	28.96	11.72	6.21	2.75
CA(PT)	**Total**	**MRR**	**NAG(%)**	**WAG(%)**	**NAF(%)**	**1(%)**	**2(%)**	**3(%)**
Quando	56	0.04	30.76	12.30	42.45	3.08	1.54	0
Onde	47	0.18	10.93	25	20.31	10.93	1.56	4.68
Quem	146	0.14	17.12	29.45	36.30	10.95	4.11	2.05

ML for Definition Questions – Def-WQA

☆ Questions such as:

- *What is a prism?*
- *Who is Ben Hur?*
- *What is the BMZ?*

☆ Answering consists in collecting as much descriptive information as possible (*nuggets*):

- *The distinction of relevant information*
- *Multiple sources*
- *Redundancy*

☆ Exploit ***only*** web snippets:

- Avoid processing and downloading a wealth of documents.
- Avoid specialized wrappers (for dictionaries and encyclopedias)
- Snippets are automatically "anchored" around questions terms → Q-A proximity
- Considering N-best snippets → redundancy via implicit multi-document approach
- Extend the coverage by boosting the number of sources through simple surface patterns (also here: KB poor approach).

Determining descriptive phrases from snippets

Determining descriptive phrases from snippets

- Surface patterns, e.g., “What is the DFKI?”
  - “DFKI is a” OR “DFKI is an” OR “DFKI is the” OR “DFKI are a”…
  - “DFKI, or ”.
  - “(DFKI)”
  - “DFKI becomes” OR “DFKI become” OR “DFKI became”

Determining descriptive phrases from snippets

- ☆ Surface patterns, e.g., “What is the DFKI?”
  - “DFKI is a” OR “DFKI is an” OR “DFKI is the” OR “DFKI are a”…
  - “DFKI, or ”.
  - “(DFKI)”
  - “DFKI becomes” OR “DFKI become” OR “DFKI became”
- ☆ Some fetched sentences:
  - “**DFKI is the** German Research Center for Artificial Intelligence”.
  - “The **DFKI is a** young and dynamic research consortium”
  - “Our partner **DFKI is an** example of excellence in this field.”
  - “the **DFKI, or** Deutsches Forschungszentrum für Künstliche **...**”
  - “German Research Center for Artificial

Determining descriptive phrases from snippets

- ☆ Surface patterns, e.g., “What is the DFKI?”
  - “DFKI is a” OR “DFKI is an” OR “DFKI is the” OR “DFKI are a”…
  - “DFKI, or ”.
  - “(DFKI)”
  - “DFKI becomes” OR “DFKI become” OR “DFKI became”
- ☆ Some fetched sentences:
  - “**DFKI is the** German Research Center for Artificial Intelligence”.
  - “The **DFKI is a** young and dynamic research consortium”
  - “Our partner **DFKI is an** example of excellence in this field.”
  - “the **DFKI, or** Deutsches Forschungszentrum für Künstliche **...**”
  - “German Research Center for Artificial
- ☆ LSA-based clustering into potential senses
  - Determine semantically similar words/ substrings
  - Define different clusters/potential senses on basis of non-membership in sentences
- ☆ Ex: What is Question Answering ?
  - **SEARCHING**: Question Answering is a computer-based activity that involves searching large quantities of text and understanding both questions and textual passages to the degree necessary to. ...
  - **INFORMATION**: Question-answering is the well-known application that goes one step further than document retrieval and provides the specific information asked for in a natural language question. ...
  - …

Example Output: What is epilepsy ?

☆ Our system's answer in terms of clustered senses:

---------- **Cluster STRANGE** ----------
0<->In epilepsy, the normal pattern of neuronal activity becomes disturbed, causing strange...
---------- **Cluster SEIZURES** ----------
0<->Epilepsy, which is found in the Alaskan malamute, is the occurrence of repeated seizures.
1<->Epilepsy is a disorder characterized by recurring seizures, which are caused by electrical disturbances in the nerve cells in a section of the brain.
2<->Temporal lobe epilepsy is a form of epilepsy, a chronic neurological condition characterized by recurrent seizures.
---------- **Cluster ORGANIZATION** ----------
0<->The Epilepsy Foundation is a national, charitable organization, founded in 1968 as the Epilepsy Foundation of America.
---------- **Cluster NERVOUS** ----------
0<->Epilepsy is an ongoing disorder of the nervous system that produces sudden, intense bursts of electrical activity in the brain.
...

Def-WQA: Results

Corpus	# Questions	# Answered Def-WQA/Baseline	# nuggets Def-WQA/Baseline
TREC 2003	50	50/38	14.14/7.7
CLEF 2006	152	136/102	13.13/5.43
CLEF 2005	185	173/160	13.86/11.08
TREC 2001	133	133/81	18.98/7.35
CLEF 2004	86	78/67	13.91/5.47

Corpus	F-score (β=5)
Trec 2003	0.52

Trec 2003 best systems (on newspaper articles): 0.5 – 0.56

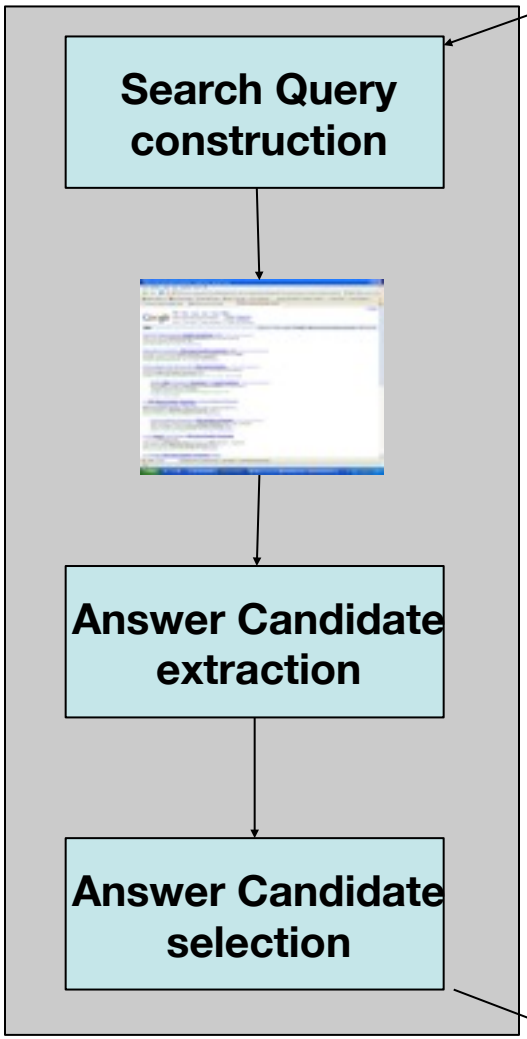
Notes:

- **we prefer sentences instead of nuggets (readability)**
- **we need no predefined window size for nuggets (~ 125 characters)**

- **Def-WQA as a basis for more applications, e.g.,**
  - **list-based questions, web person identification, ontology learning**
- **Still missing: merging/splitting of partitions (evtl. using KBs and authority)**

List-WQA – Overview

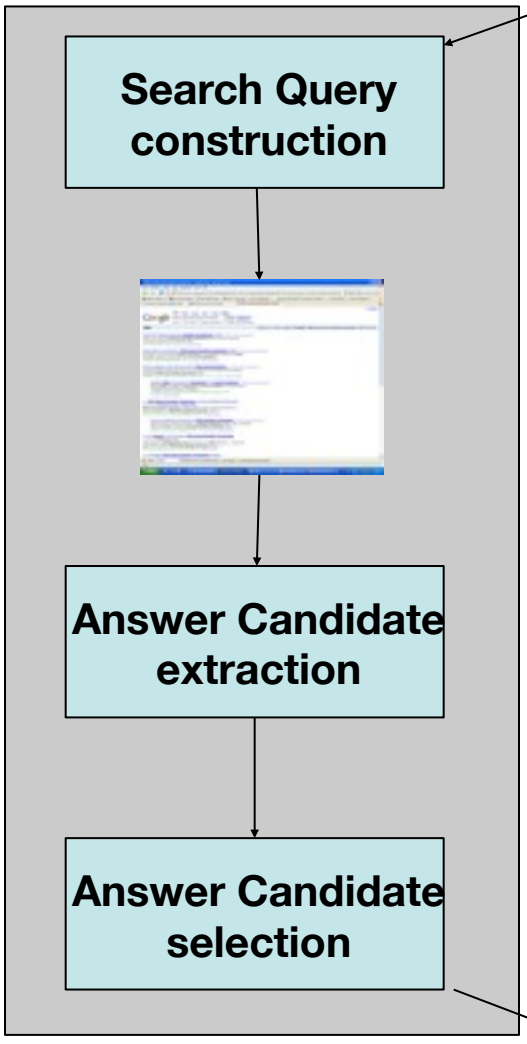
"What are 9 works written by Judith Wright?"



The Moving Image, Woman to Man, The Gateway, The Two Fires, Birds, The Other Half, City Sunrise, The Flame three and Shadow.

List-WQA – Overview

"What are 9 works written by Judith Wright?"

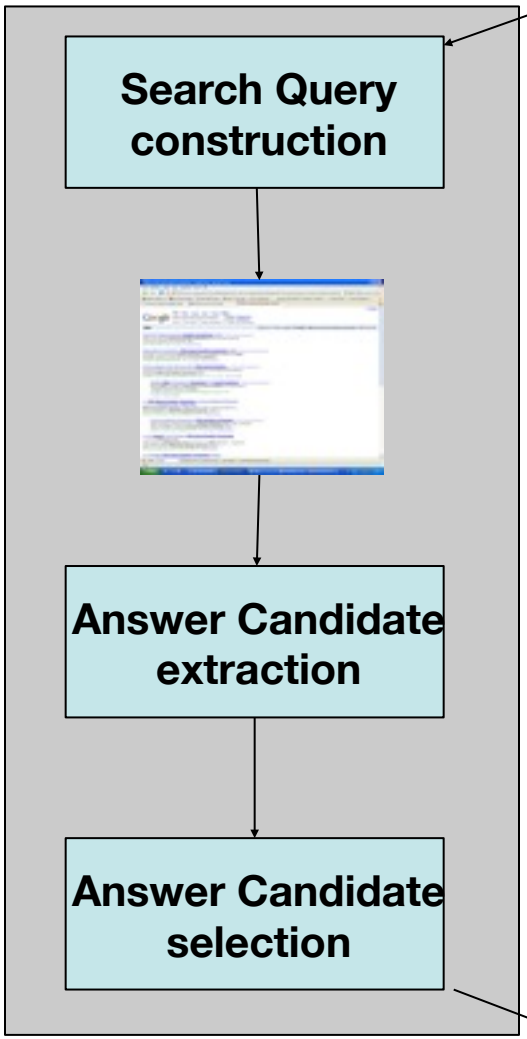


Qfocus → inbody
NPs → intitle
Apply 4 patterns Qi

The Moving Image, Woman to Man, The Gateway, The Two Fires, Birds, The Other Half, City Sunrise, The Flame three and Shadow.

List-WQA – Overview

"What are 9 works written by Judith Wright?"



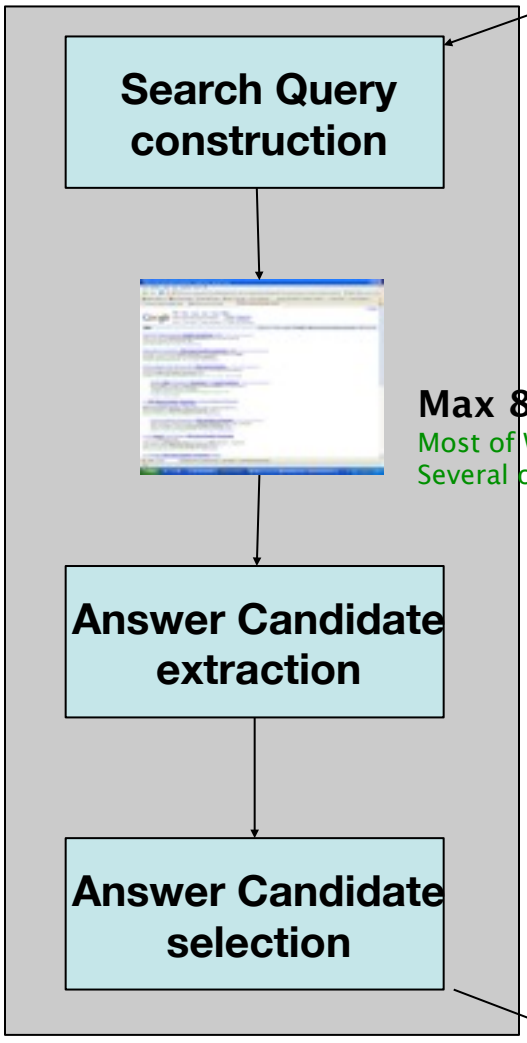
Qfocus → inbody
NPs → intitle
Apply 4 patterns Qi

Q1: (intitle:"Judith Wright") AND (inbody:"works" OR inbody:"written")

The Moving Image, Woman to Man, The Gateway, The Two Fires, Birds, The Other Half, City Sunrise, The Flame three and Shadow.

List-WQA – Overview

"What are 9 works written by Judith Wright?"



Qfocus → inbody
NPs → intitle
Apply 4 patterns Qi

Q1: (intitle:"Judith Wright") AND (inbody:"works" OR inbody:"written")

Max 80 snippets:

Most of Wright's poetry was **written** in the mountains of southern Queensland. ...
Several of her early **works such as** 'Bullocky' and 'Woman to Man' became standard ...

The Moving Image, Woman to Man, The Gateway, The Two Fires, Birds, The Other Half, City Sunrise, The Flame three and Shadow.

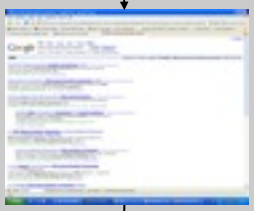
List-WQA – Overview

"What are 9 works written by Judith Wright?"

Search Query construction

Qfocus → inbody
NPs → intitle
Apply 4 patterns Qi

Q1: (intitle:"Judith Wright") AND (inbody:"works" OR inbody:"written")



Max 80 snippets:

Most of Wright's poetry was **written** in the mountains of southern Queensland. ...
Several of her early **works such as** 'Bullocky' and 'Woman to Man' became standard ...

Answer Candidate extraction

Apply 8 patterns πi (hyponym, possessive, copula, quoting, etc.)

Answer Candidate selection

The Moving Image, Woman to Man, The Gateway, The Two Fires, Birds, The Other Half, City Sunrise, The Flame three and Shadow.

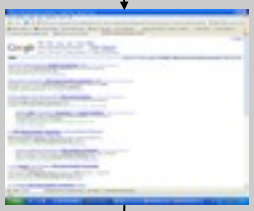
List-WQA – Overview

"What are 9 works written by Judith Wright?"

Search Query construction

Qfocus → inbody
NPs → intitle
Apply 4 patterns Qi

Q1: (intitle:"Judith Wright") AND (inbody:"works" OR inbody:"written")



Max 80 snippets:

Most of Wright's poetry was **written** in the mountains of southern Queensland. ...
Several of her early **works such as** 'Bullocky' and 'Woman to Man' became standard ...

Answer Candidate extraction

Apply 8 patterns πi (hyponym, possessive, copula, quoting, etc.)

π4: **entity** is \w+ **qfocus** \w*

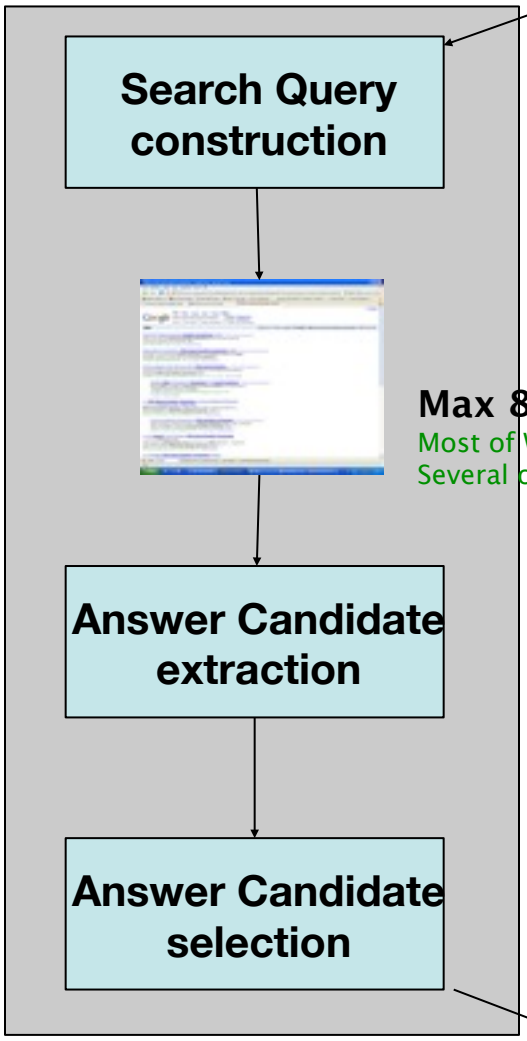
Chubby Hubby is …. Ben and Jerry's **ice cream** brand.

Answer Candidate selection

The Moving Image, Woman to Man, The Gateway, The Two Fires, Birds, The Other Half, City Sunrise, The Flame three and Shadow.

List-WQA – Overview

"What are 9 works written by Judith Wright?"



Qfocus → inbody
NPs → intitle
Apply 4 patterns Qi

Q1: (intitle:"Judith Wright") AND (inbody:"works" OR inbody:"written")

Max 80 snippets:

Most of Wright's poetry was **written** in the mountains of southern Queensland. ...
Several of her early **works such as** 'Bullocky' and 'Woman to Man' became standard ...

Apply 8 patterns πi (hyponym, possessive, copula, quoting, etc.)

π4: **entity** is \w+ **qfocus** \w*

Chubby Hubby is …. Ben and Jerry's **ice cream** brand.

Use Semantic kernel & Google N-grams

The Moving Image, Woman to Man, The Gateway, The Two Fires, Birds, The Other Half, City Sunrise, The Flame three and Shadow.

List-WQA: Results

☆ Answer Selection:

- Two measures **Accuracy** and **F_1** score.
- Two values
  - All questions
  - Only questions where at least one answer was found in the fetched snippets.
- Duplicate answers have also an impact on the performance. For instance:
  - "*Maybelline*" (also found as "*Maybellene*" and "*Maybeline*").
  - John Updike's novel "*The Poorhouse Fair*" was also found as "*Poorhouse Fair*".

Systems\Trec	2001	2002	2003	2004
ListWebQA(F_1)	0.35/0.46	0.34/0.37	0.22/0.28	0.30/0.40
ListWebQA(Acc)	0.5/0.65	0.58/0.63	0.43/0.55	0.47/0.58
Top one(Acc.)	0.76	0.65	-	-
Top two(Acc.)	0.45	0.15	-	-
Top three(Acc.)	0.34	0.11	-	-
Top one(F_1)	-	-	0.396	0.622
Top two(F_1)	-	-	0.319	0.486
Top three(F_1)	-	-	0.134	0.258
Yang & Chua 04 (F_1)	-	-	.464 ~. 469	-

We conclude:
Encouraging results, competes well with 2nd best;
Still creates too much noise;

Web QA and Information Extraction

☆ WebQA:

- Combining generic lexico-syntactic patterns with unsupervised answer extraction from Snippets only
- Language independent and multilingual
- Our approach has a close relationship to the new approach of unsupervised IE, e.g., Etzioni et al. , Weikum et al., Rosenfeld & Feldman

☆ Information extraction

- WebQA as a generic tool for web-based bottom-up knowledge extraction and ontology population
- Ontology-based clustering for unsupervised information extraction
  - Use ontology for generating QA requests -> ontology-driven active QA
  - Use web QA for populating and extending ontology
- Interactive dynamic information extraction