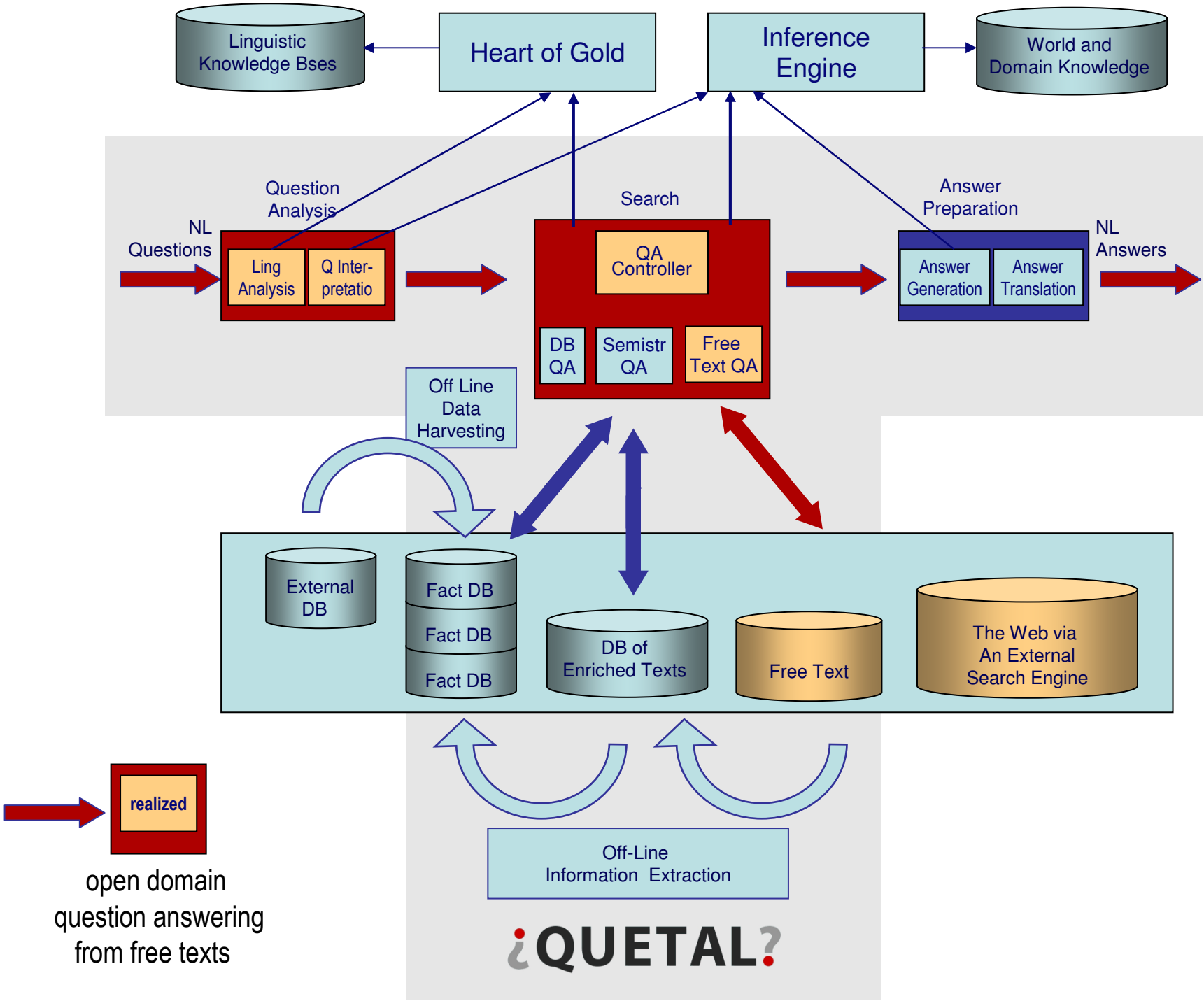
¿QUETAL?

A Multilingual Hybrid Question-Answering System

Cross-Lingual Open-Domain Question Answering

Günter Neumann, Bogdan Sacaleanu


Linguistic Knowledge Bses
Heart of Gold
Inference Engine
World and Domain Knowledge
Question Analysis
NL Questions
Ling Analysis
Q Inter-pretatio
Search
QA Controller
DB QA
Semistr QA
Free Text QA
Answer Preparation
Answer Generation
Answer Translation
NL Answers
Off Line Data Harvesting
External DB
Fact DB
Fact DB
Fact DB
DB of Enriched Texts
Free Text
The Web via An External Search Engine
Off-Line Information Extraction
¿QUETAL?
realized
open domain question answering from free texts

¿QUETAL? Cross-lingual Open-Domain Question-Answering

"Mit wem ist David Beckham verheiratet?"

{person:David Beckham, married, person:?}

German Question

Question Analysis

English Question Object

IR-Query Construction

IR-Google

IR-Lucene/XML

Annotated Corpus

Query Translation

- Online MT-systems
- WSD
- Expansion

Question Object:

- Focus, Scope
- AnswerType

Documents

Passage selection

Passages

"David Beckham, the soccer star engaged to marry Posh Spice, is being blamed for England 's World Cup defeat."

Answer

Answer Selection

Candidates

Answer Extraction

Posh Spice

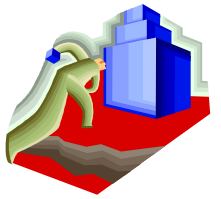
{person:David Beckham, person:Posh Spice}

Challenges for Textual QA

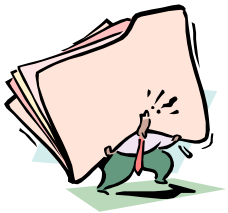
- ☆ Open domain
  - No restriction on the domain and type of question
  - No restriction on document source and style (news text corpus, Web, …)
- ☆ High demands on robustness & efficiency of LT core components
  - From keywords to full NL questions
  - Very large scale sources of free text
  - Trade-off between off-line and on-line annotation
- ☆ Cross-linguality
  - How to exploit MT technology for textual QA ?
- ☆ Reusability & Scalability
  - Same QA framework for heterogenous document sources
  - Incremental bottom-up software development

Our Design Perspective

- ☆ Foster bottom-up system development
  - Data-driven, robustness, scalability
  - From shallow & deep NLP



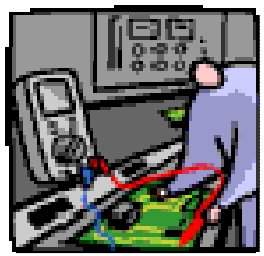
- ☆ Large-scale answer processing
  - Coarse-grained uniform representation of query/documents
  - Text zooming
  - Ranking scheme for answer selection



- ☆ Need-triggered use of knowledge sources
  - Rather exploit data-driven strategies & linguistic structure

- ☆ Common basis for
  - Online Web pages
  - Large textual sources

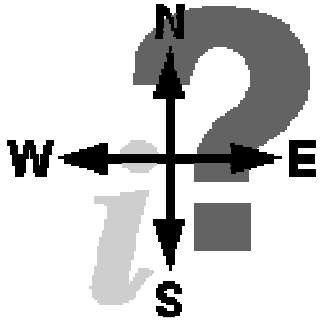
¿QUETAL? Textual QA in Quetal: R&D Results



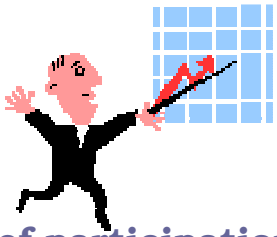
QA-framework Quantico
- Web & XML-annotated documents
- ~ 5-8 sec/QA-cycle

Hybrid approach for cross-lingual textual QA

Answer credibility checking



Flexible robust free question analysis

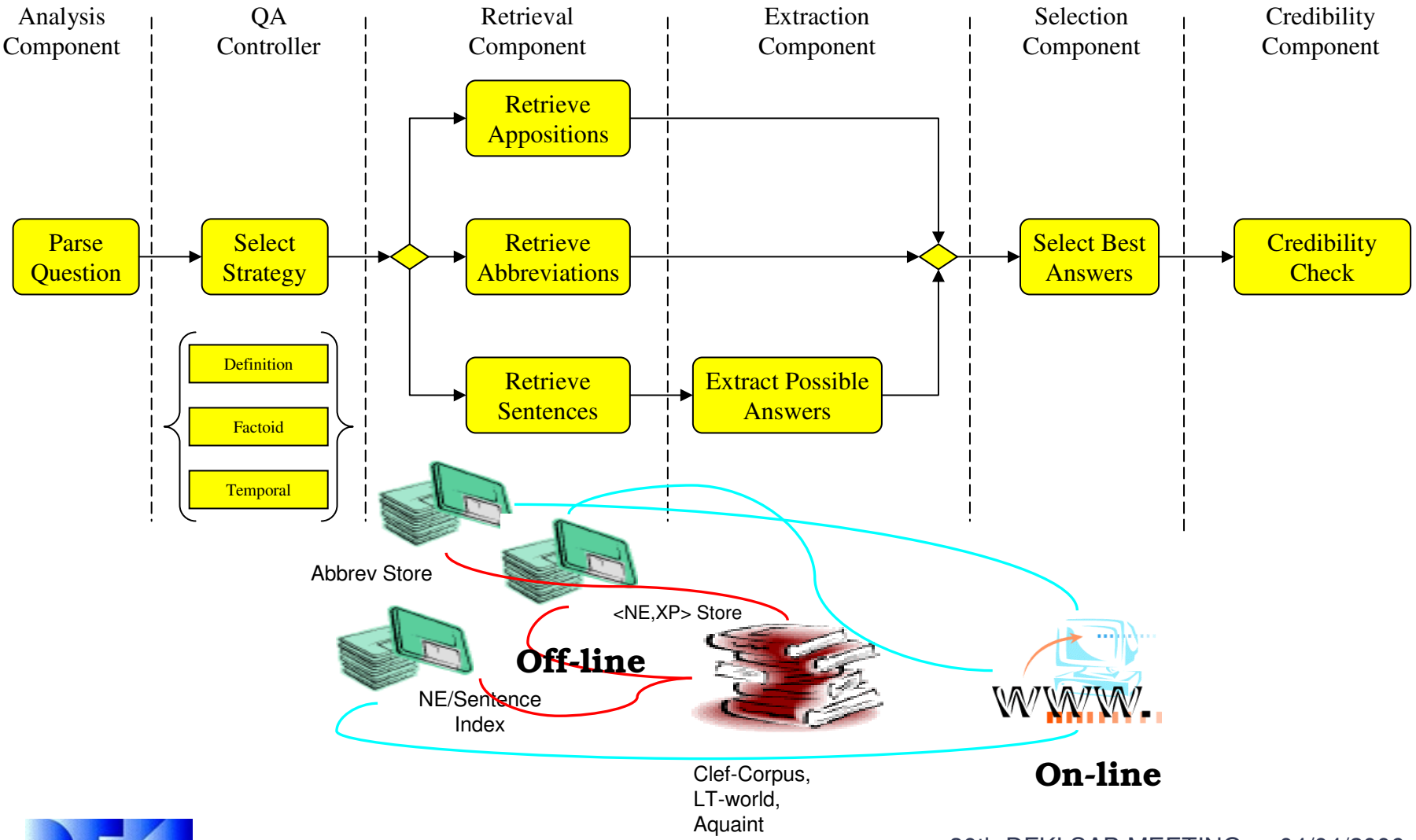


Clef participation:
best results for German & English as target languages (25%DE2EN, 47.5%DE2DE)

Question-type specific selection of answer extraction strategies

Dissemination (projects):
- SmartWeb (BMBF)
- HyLaP (BMBF)
- QALL-ME (EC)
- RASCALLI (EC)
- ...

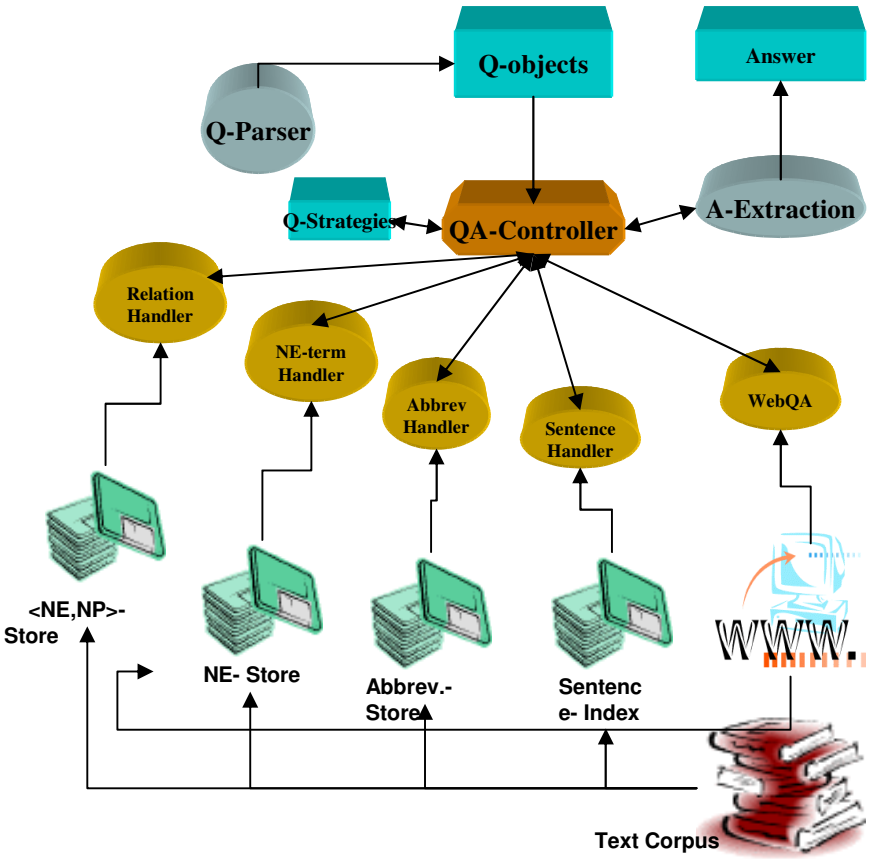
Quantico: Activity Flow



¿QUETAL? Free Question Analysis for Textual QA

- ☆ Query analysis as control information
  - Q-type/A-type/Q-constraints/…
  - Local Wh-grammars + dependency structure for initial (underspecified) Q-info
  - Tree-traversal for determining more specific Q-info
    - Non-local syntactic constraints
    - Coarse-grained lexical semantic consistency checks
    - Semantic types for main noun/verb lemmas

- ☆ Q-type specific Strategy selection



*The implementation was done by Rob Basten as part of his Master Thesis *Answering Open Domain Temporally Restricted Questions in a Multi-Lingual Context*, DFKI & Uni. Twente, NL

¿QUETAL? Temporal Question Strategies*

Examples (1 & 3 from Clef):
What nearly caused the cancellation or postponement of the 1996 European Football Championship?
Name a German tennis player who won Wimbledon between 1980 and 1990?
Whom was Michael Jackson married to before he married Debbie Row?

Core idea:
Process questions of this kind on basis of our existing technology following a divide-and-conquer approach:

☆ **question decomposition**

- A temporally restricted questions Q is decomposed into two sub-questions
- one referring to the "timeless" proposition of Q, and
- the other to the temporally restricting part.

☆ **answer fusion**

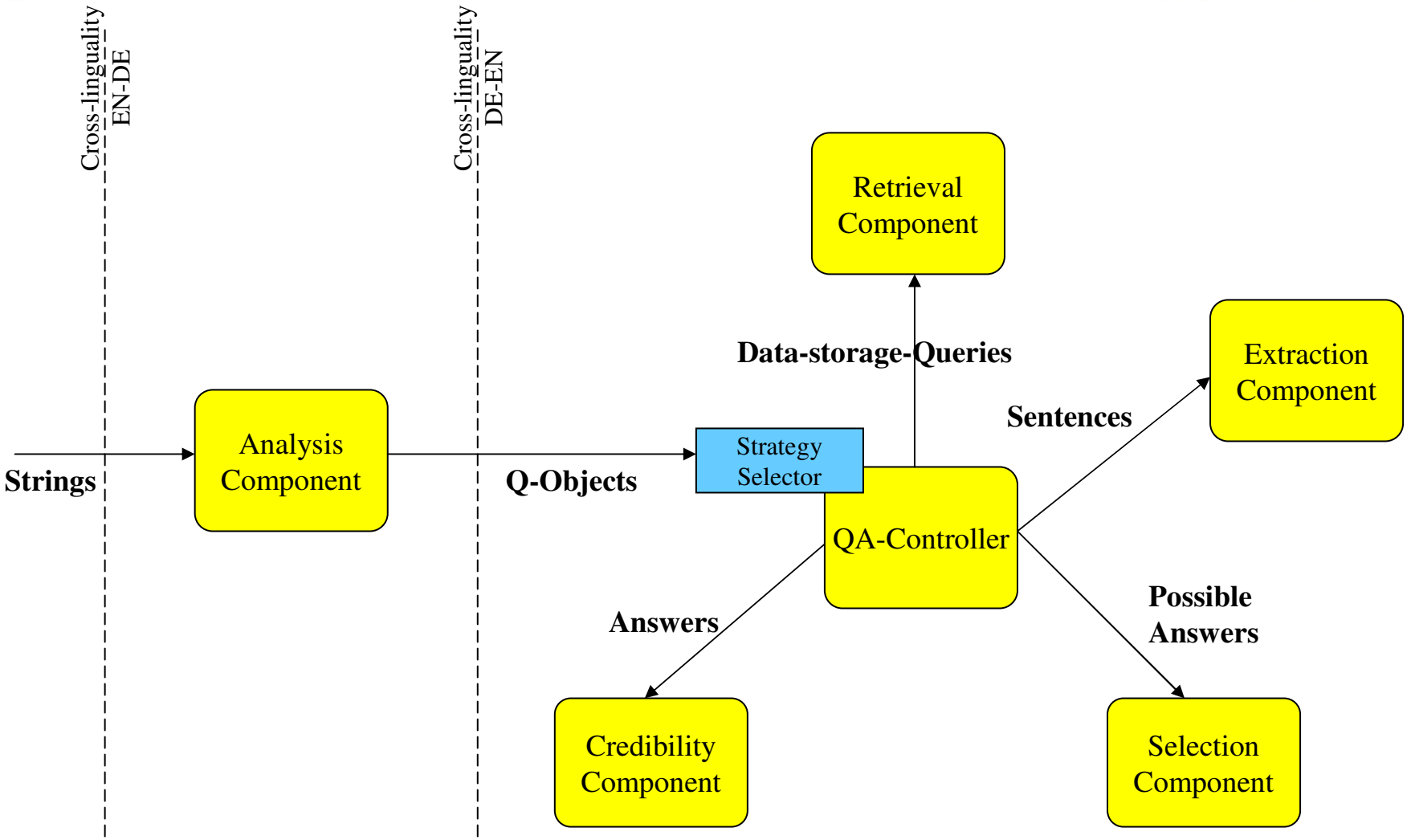
- The answers of both are searched for independently
- but checked for consistency in a follow-up answer fusion step
- the found explicit temporal restriction is used to constrain the "timeless" proposition.

Who was the German Chancellor when the Berlin Wall was opened? ⇒
Who was the German Chancellor ? & When was the Berlin Wall opened?

☆ **Initial/fallback strategy**

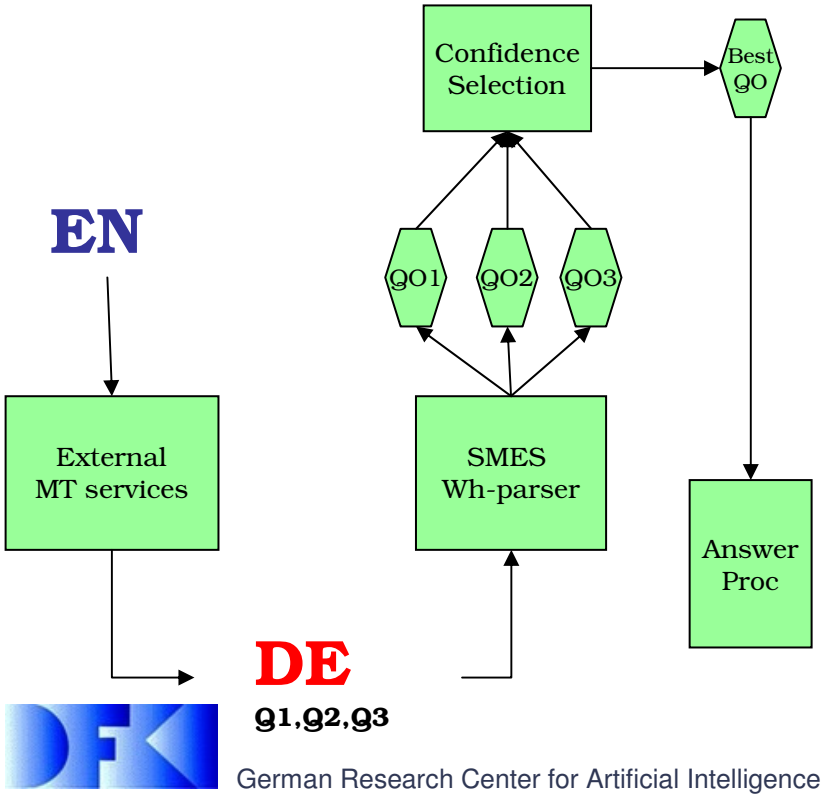
- The existing methods for handling factoid questions are used without change to get initial answer candidates.
- In a follow-up step, the temporal restriction from the question is used to check the answer's temporal consistency.

Cross-linguality in QA



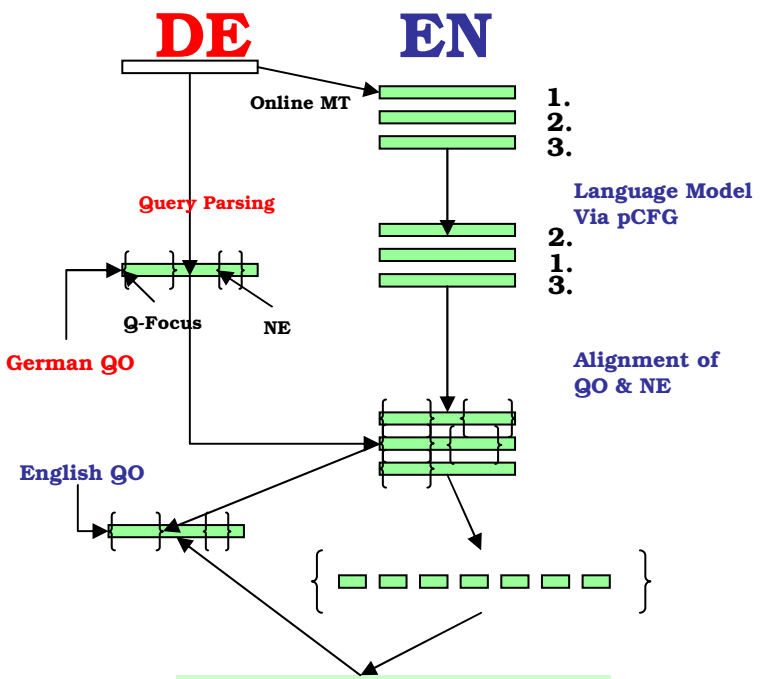
Before Method EN-DE

- Question translation
- Translations processing -> QObjects
- QObject selection



After Method DE-EN

- Question processing -> QObject
- Question translation + alignment
- QObject alignment



SAB Recommendation

The SAB recommended to take into account the dimension of credibility of the answer

- ☆ There exists very few work in the area of textual QA, e.g., Lita et al. (CMU), AAAI-2005
- ☆ Credibility in QA:
  - Provide criteria about the assumed quality of an answer
  - Determine the credibility of the answer source
  - Incorporate a measure of credibility in computing the answer confidence
- ☆ Examples of meta information
  - Table of trusted links per question topic
  - Information from URL (last update, semantic relationship of link name with answers)
  - Textual information (style, fingerprints, discourse markers)

Our starting point

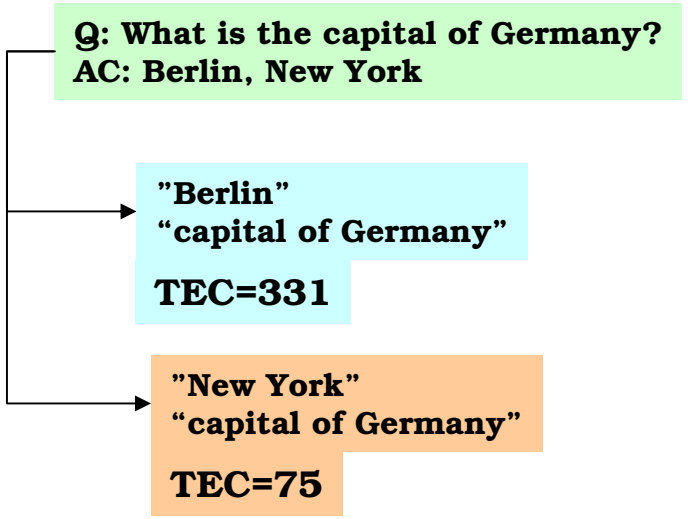
- ☆ It is known that redundancy plays an important role for Web-based/textual QA
  - Answers get higher rank, if they are mentioned more often in different documents.
- ☆ So seen, redundancy is already a measure of credibility
- ☆ But, how to collect further information that supports an answer?
  - Use a list of trusted links to filter document sources
  - Select the document that mostly supports the answer

¿QUETAL? Two methods have been investigated

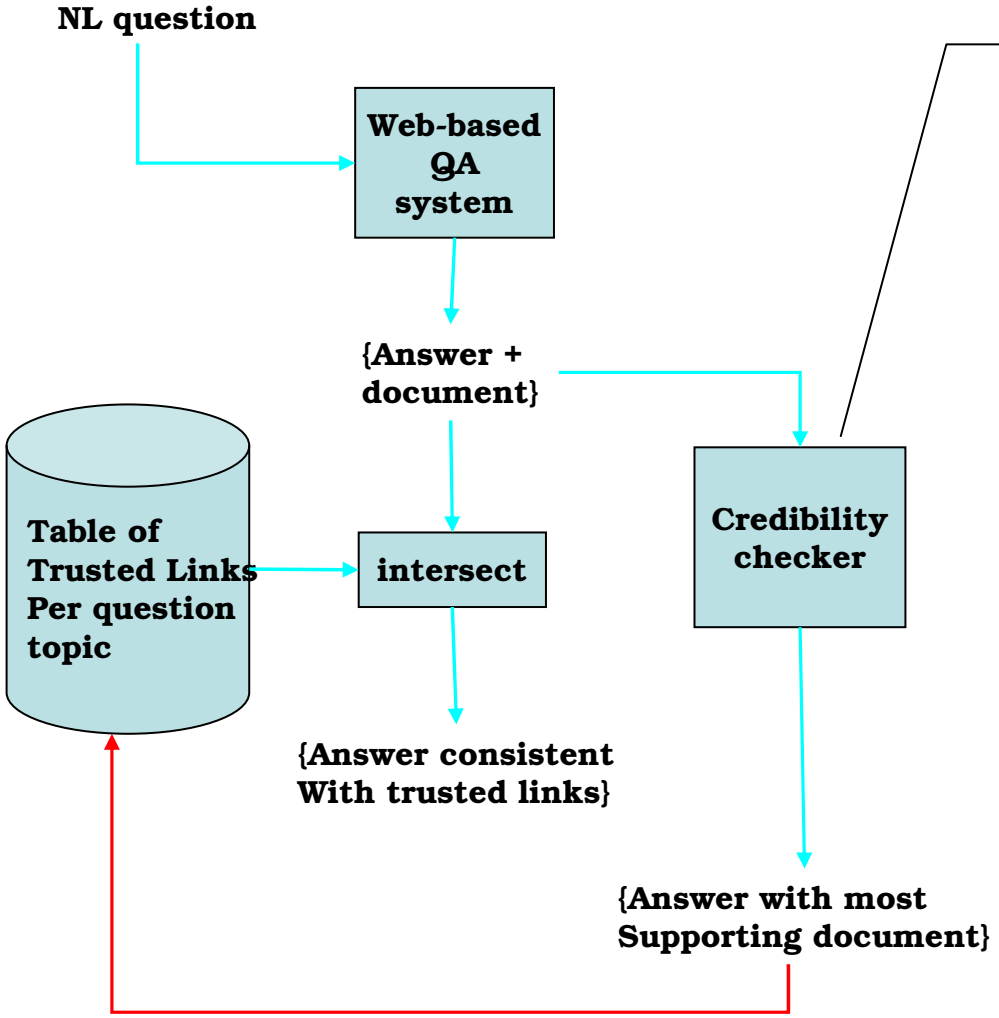
- ☆ Google's total frequency counts
  - For answers extracted from a (small) text corpus, exploit their external Web redundancy
- ☆ More general model that integrates
  - Table of trusted links
  - Automatic determination of credibility for Web document sources

Web-based Answer Validation

- ☆ Assume, answers have been extracted from some text corpus
- ☆ Web-based answer plausibility check
  - direct_answer_string := question + answer;
  - Google's Total Estimated Counts (TEC) for ranking answer candidates
- ☆ Presupposes an independency between answer candidates ⇒ method seems to be useful (cf. Clef 2005)
- ☆ In case of "hidden semantic relationship" (e.g., is-a), method is not suited/sufficient.



General Model



Answer not via trusted links ->
Automatically determine trusted documents -> "credibility assessment"

Currently used checkers:
1. **LSA + URL-content**
2. **Update info of URL**
3. **Discourse markers**
4. **W3C HTML quality**
5. **Spelling**

Current major problem:
How to evaluate credibility checks?

Plausible:
Via user feedback.

Fogg et al. 2002 "How do people evaluate a Web Site's credibility?"

¿QUETAL? What information to consider ?

Topic	Percent (2440 com.)	Comment Topics
1	46.1	Design Look
2	28.5	Information Design/Structure
3	25.1	Information Focus
4	15.5	Company Motive
5	14.8	Information Usefulness
6	14.3	Information Accuracy
7	14.1	Name Recognition & Reputation
8	13.8	Advertising
9	11.6	Information Bias

Topic	Percent (2440 com.)	Comment Topics
10	9.0	Writing Tone
11	8.8	Identify of Site Operator
12	8.6	Site Functionality
13	6.4	Customer Service
14	4.6	Past Experience with Site
15	3.7	Information Clarity
16	3.6	Performance on Test by User
17	3.6	Readability
18	3.4	Affiliations

Semantic checker

Discourse checker

W3C HTML quality

List of trusted links

Site server (update info)

Spelling/Grammar checker

QA@Clef 2005

- ☆ Motivation of participation
  - External evaluation
  - Foster development of software infrastructure
  - International research community
  - Makes fun
- ☆ Additional increase in participants and languages
  - 24 groups
  - 9 source/10 target languages (8 monlingual/73 crosslingual tasks)
- ☆ Task
  - Corpus: newspaper articles from 1994/1995, in case of DE/EN ~ 500MB
  - 200 questions:
    120 factoid (F), 50 definitions (D), 30 temporally restricted (T), 20 NIL
  - Return single best exact answer for each question

¿QUETAL? DFKI Results for Clef-2005

DFKI@QA@Clef-2004:
DE2DE: 25.38%
DE2EN: 23.5%
EN2DE: NOT

	Run/200 Questions	Right #	Right %	Wrong	IneXact	Right %	Right % D	Right % T
monolingual	dfki051dede	87	43.50	100	13	35.83	66.00	36.67
monolingual	dfki052dede*	54	27.00	127		15.00	52.00	33.33
cross-lingual	dfki051ende	46	23.00	141	12	17.67	50.00	3.33
cross-lingual	dfki052ende*	31	15.50	159	8	8.33	42.00	0
cross-lingual	dfki051deen	51	25.50	141	8	18.18	50.00	13.79

* dfki052xxde = dfki051xxde + WebValidation

We achieved best results for target languages:

- **German (one other group DE2DE: 36%, one other EN2DE: 5%)**
- **English (12 runs; 2nd system: 23.5%, 3rd system: 19%)**

… concerning the performance decrease when using Web validation

- ☆ Error sources:
  - Lack of redundancy in case of number of German Web pages
  - The correct Clef-answer might be “spoiled down”
  - Timeline of Clef corpus (1994/1995) problematic for validating “non-historically” related Q
  - Errors through the translation of complex and long questions had a negative effect on the recall of the web search (EN2DE)
- ☆ However, after detailed analysis of German runs:
  - 51 different assignments for runs without & with validation
  - 13 questions (of which 8 are definition questions) are now answered correctly
  - 28 questions are now answered wrongly, but
  - 14 of them because of different timeline
- ☆ Needed:
  - Integration of contextual and situational information into QA cycle taking into account user feedback
  - -> HyLaP, QALL-ME