DFKI at QA@Clef 2007

Günter Neumann, Bogdan Sacaleanu,
Christian Spurk, Rui Wang

Language Technology Lab at DFKI

Saarbrücken, Germany

Overview

- DFKI is participating since 2003
  - Focus on German monolingual QA and German/English cross-lingual QA
  - Promising results so far (acc.): DEDE=43,50%, ENDE=32,98%, DEEN=25.50%
- Goal for Clef 2007: increase spectrum of activities
  - Consideration of additional language pairs (ESEN, PTDE)
  - Participation in QAST pilot task
  - Participation in Answer Validation Exercise (AVE)

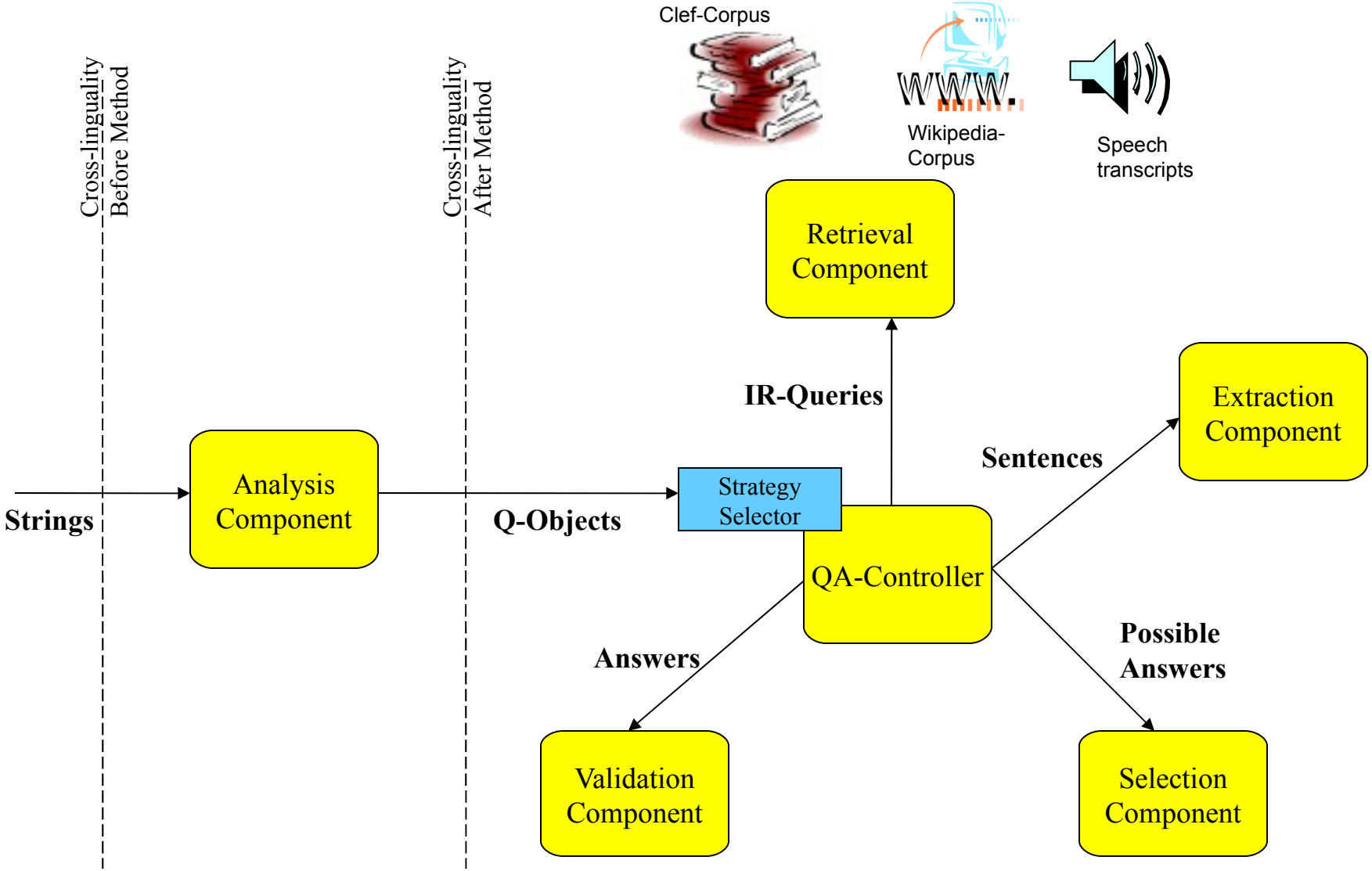
QA architecture - some design issues

- NL question
  - Declarative description of search strategy and control information
  - Analysis should be as complete and accurate as possible
  - Use of full parsing and semantic constraints
- Consider document sources as implicit search space
  - Off-line: Provide question type oriented preprocessing for context selection
  - On-line: Provide question specific preprocessing for answer processing

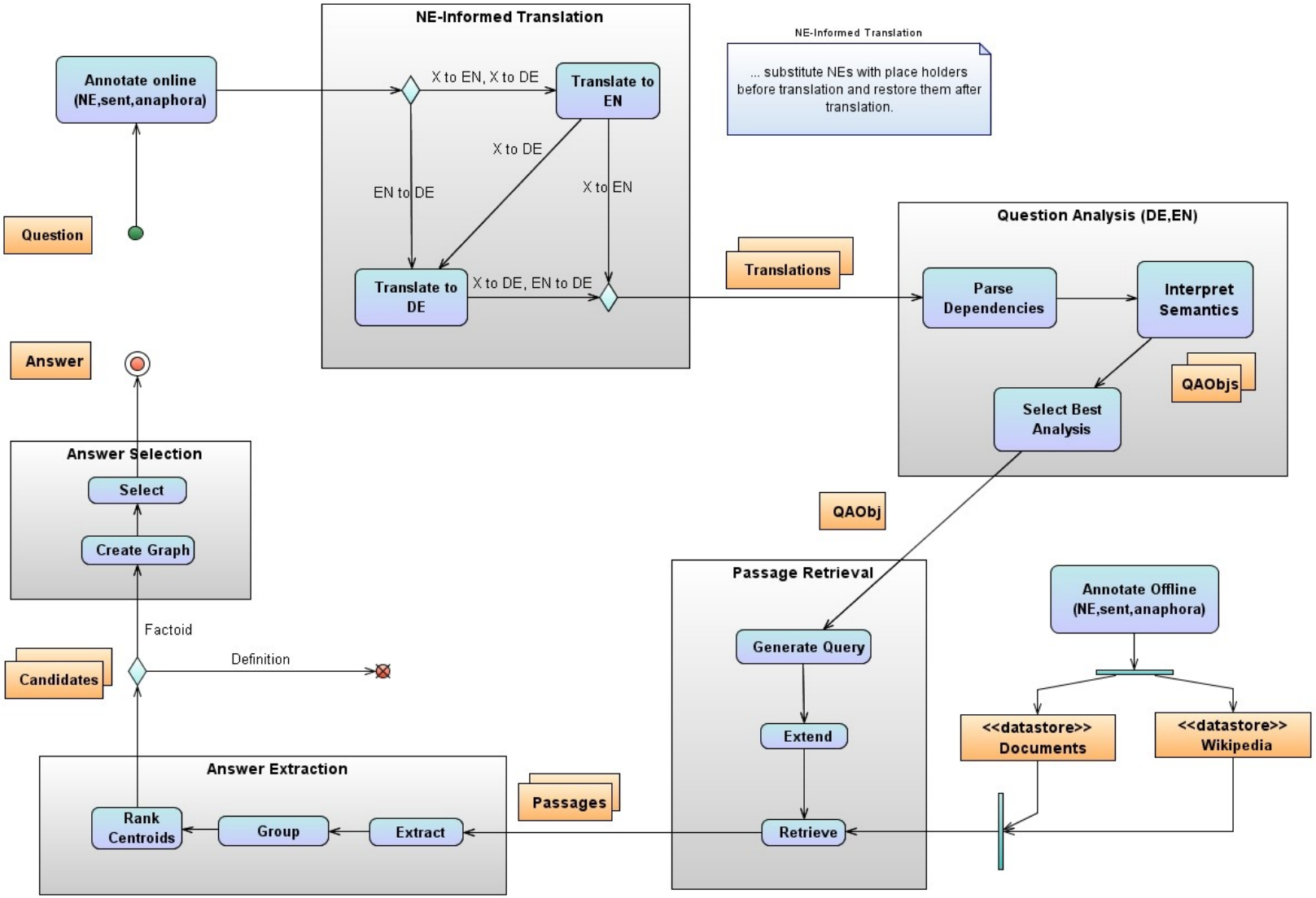
Common architecture for different answer pools

- Answer sources (covered by our technology)
  - Structured sources (DBMS)
  - Linguistically well-formed textual sources (news articles)
  - Well-structured web sources (Wikipedia)
  - Web snippets
  - Speech transcripts, cf. QAST
- Assumption:
  - QA for different answer sources share pool of same components
- Service oriented architecture (SOA) for QA
  - Strong component-oriented approach
  - Basis for open-source QA architecture (cf. EU project QALL-ME)

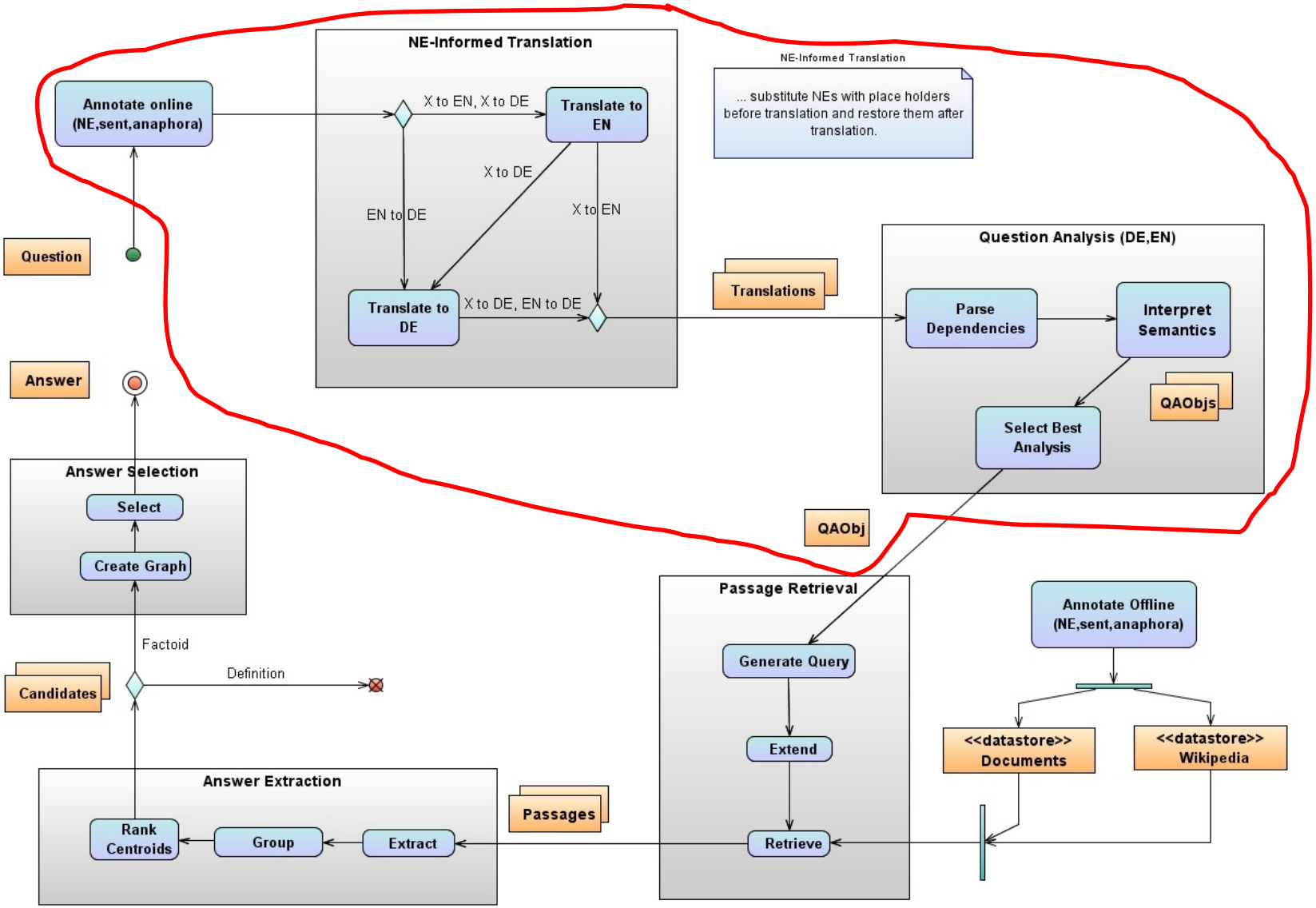
Overview QA architecture



LT-Lab System Architecture for Clef 2007



Query processing components



Before Method

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

Assumption: the better the query analysis of a translated question is done the better was the translation being made

Completeness wrt.
-Parse tree
-major semantic Wh-types

Source Question
(DE/EN/ES/PT)

Possibly Via English

External MT services

German/English Questions

Q1,Q2,Q3

German/English Wh-parser

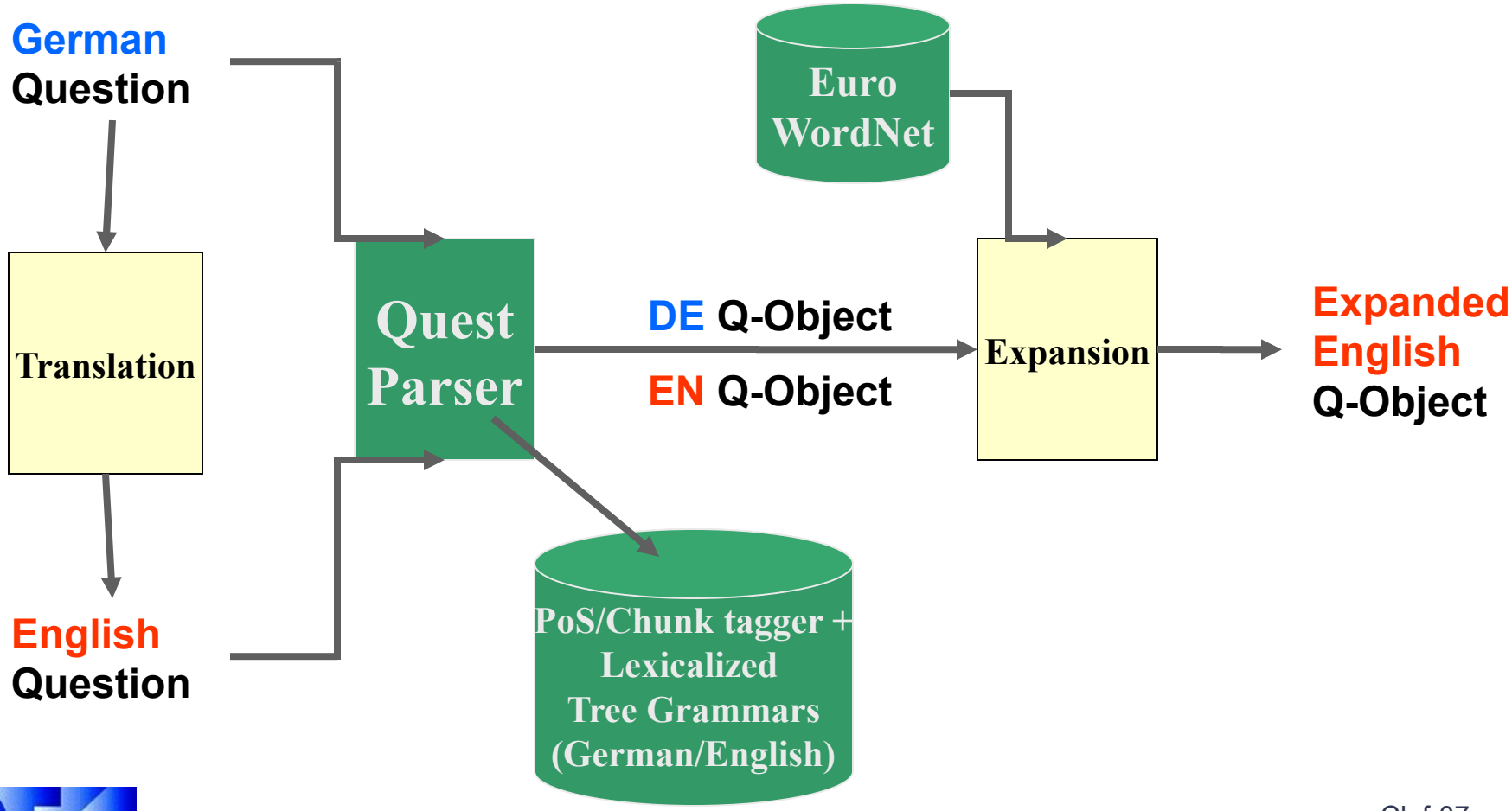
Confidence Selection

QO1 QO2 QO3

Best QO

Answer Proc

Cross-language Query Analysis: After Method



Example

1. Translation services for Word Sense Disambiguation (WSD)

Wo wurde das Militärflugzeug Strike Eagles 1990 *eingesetzt*?

FreeTranslation: *Where did the military airplane become would strike used Eagles 1990?*

Systran: *Where was the military aircraft Strike Eagle used 1990?*

Logos: *Where was the soldier airplane Strike Eagles installed in 1990?*

BoO_{EN} := {soldier, airplane, strike, eagle, install, 1990, military, become, strike, *use*, aircraft}

2. Query expansion using EuroWordNet

$\forall x \in BoO_{EN}$: lookup(EuroWN);
If x is unambiguous: extend BoO_{EN}
Else $\forall$ readings(x):
get its aligned German readings &
Look them up in BoO_{GN}
If successfully then add English terms to BoO_{EN}

Reading-697925
EN: {handle, *use*, wield}
DE: {handhaben, hantieren}

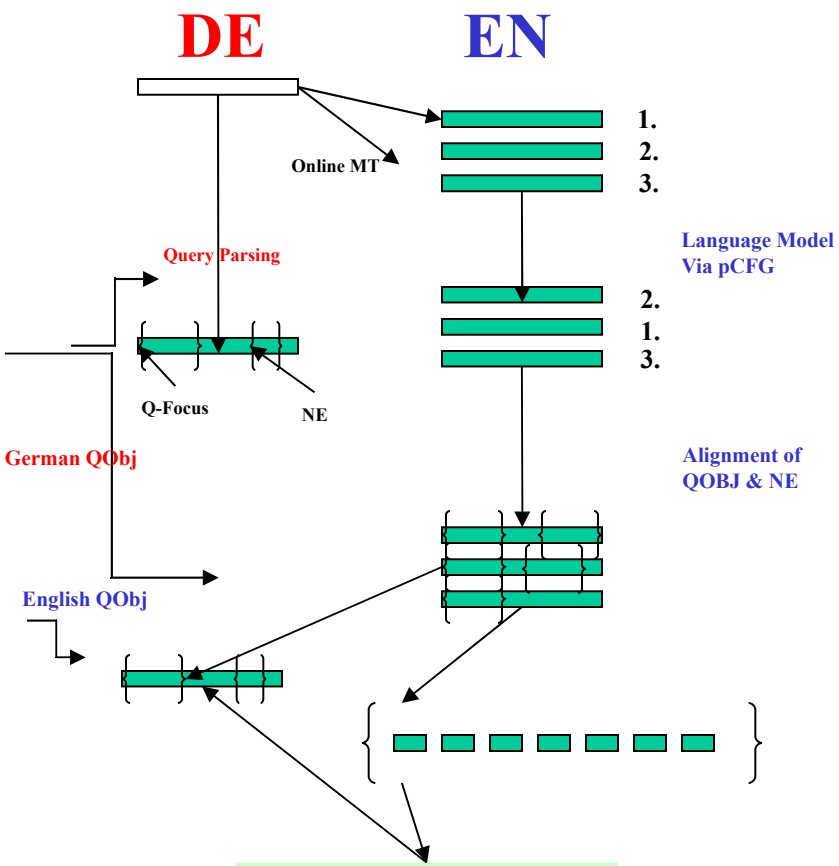
Reading-1453934:
EN: {behave toward, use}
DE: not aligned

Reading-658243:
EN: {apply, employ, make use of, put to *use*, use, utilise, utilize}
DE: {anbringen, anwenden, bedienen, benutzen, *einsetzen*, ...}

Hybrid NL-Query Translation for Cross-lingual ODQA

Improvements

- Language Model
  - translations from the on-line MT systems are ranked according to a language model
  - pCFG extracted from document corpus
    ⇒ *corpus-sensible ranking of translations*
- Allignment of Query-Information
  - based on several filters (dictionary, PoS & string similarity)
    ⇒ "transformation" of DE-QueryObject (Q-Focus) onto to EN-translation
    ⇒ *no need of parsing on English side*
- NE-specific alignment
  - Not person names
  - but organizations, locations



Question analysis

(translated) NL questions

SMES for DE&EN
- Morphology
- Dependency trees
- Shallow&Deep Proc.

SMES for
- Wh-attachment
- Q-type, A-type, Q-focus

IA-schema
- Generated Wordforms
- NE-types/Concepts
- Weights

Topic processing → **Syntactic analysis** → **Semantic analysis** → **IA proto query construction**

LingPipe for
- NER
- Coreference Resolution

alias-i

Q-Object

Sequence of NE resolved Wh-questions

IA proto query

Output example of query analysis

Which Jewish painter lived from

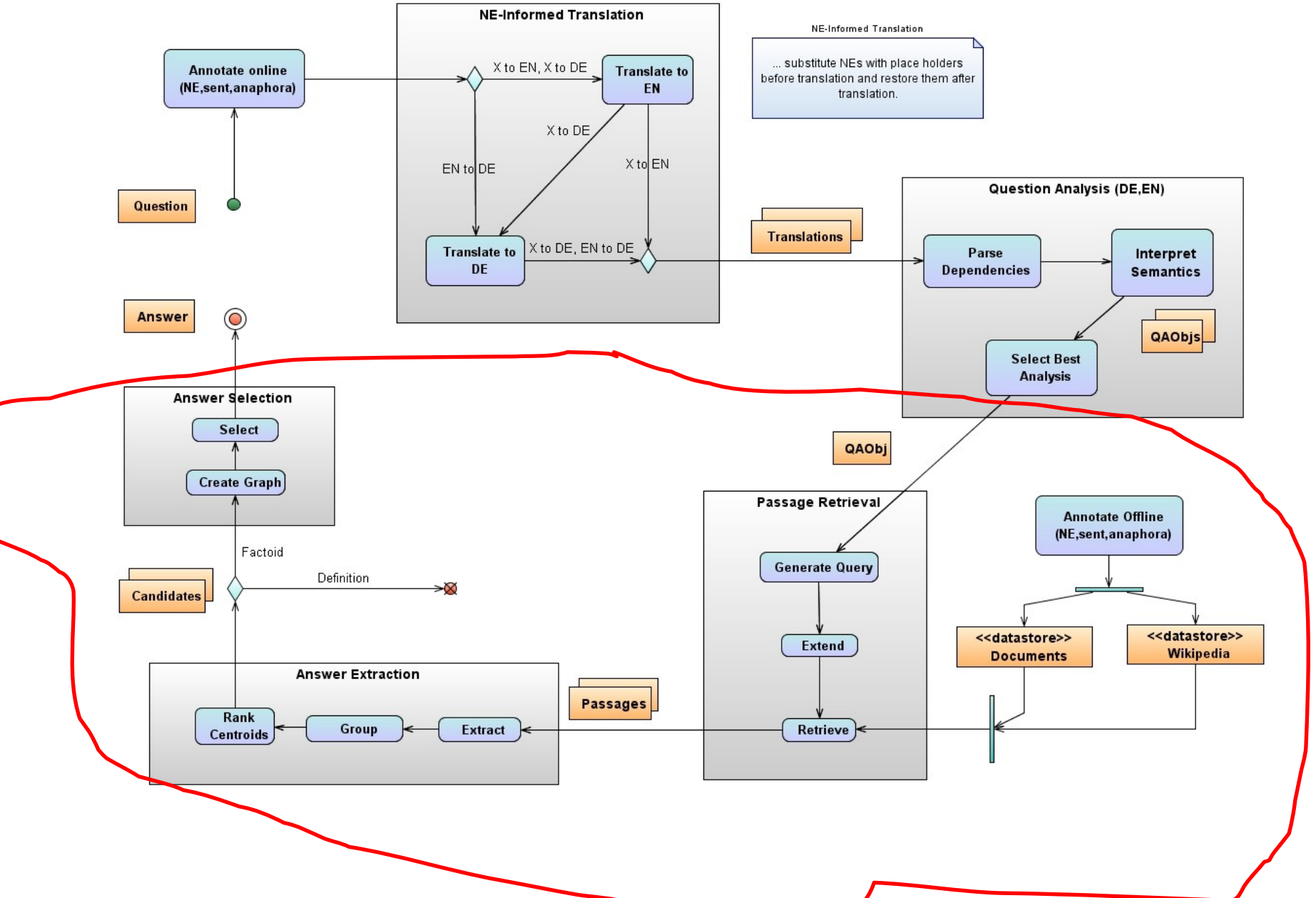
Exploiting Natural Language Generation

```
<QOBJ msg="quest" id="qId0" lang="DE" score="1">
 <NL-STRING id="qId0">
  <SOURCE id="qId0" lang="DE">Welche juedischen Maler l
von 1904-1944?</SOURCE>
  <TARGETS/>
 </NL-STRING>
 <QA-control>
  <Q-FOCUS>Maler</Q-FOCUS>
  <Q-SCOPE>leb</Q-SCOPE>
  <Q-TYPE restriction="TEMP">C-COMPLETION</Q-TYPE>
  <A-TYPE type="list:SOME">NUMBER</A-TYPE>
 </QA-control>
 <KEYWORDS>
  <KEYWORD id="kw0" type="UNIQUE">
   <TK pos="V" stem="leb">lebten</TK>
  </KEYWORD>
  <KEYWORD id="kw1" type="UNIQUE">
   <TK pos="A" stem="juedisch">juedischen</TK>
                …
</KEYWORD>
 </KEYWORDS>
 <EXPANDED-KEYWORDS/>
 <NE-LIST>
  <NE id="ne0" type="DATE">1944</NE>
  <NE id="ne1" type="DATE">1904</NE>
 </NE-LIST>
</QOBJ>
```

IA query created for Lucene

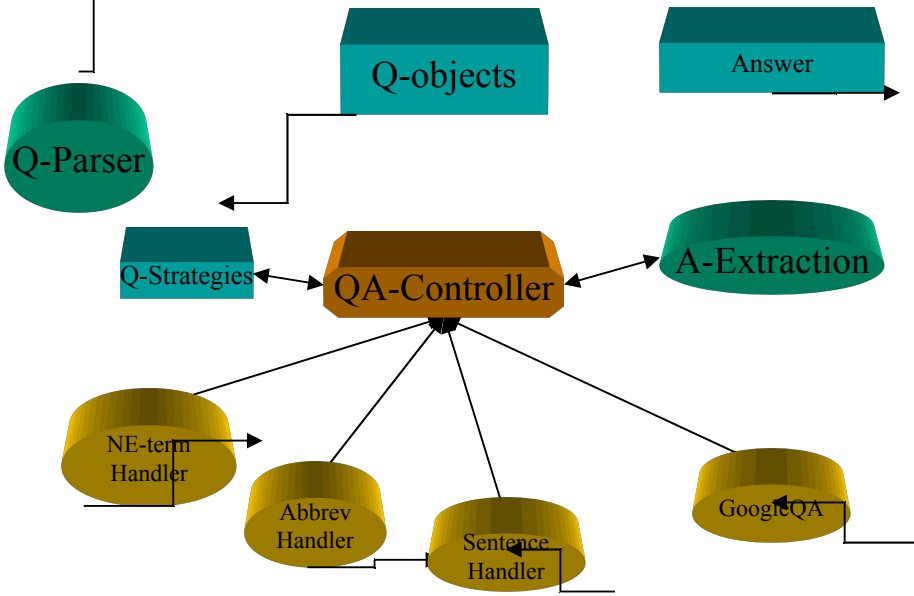
+neTypes:NUMBER
AND
("lebten" OR "lebte" OR "gelebt" OR "leben" OR "lebt")
AND +maler^4
AND jüdisch^1
AND 1944^1
AND 1904^1

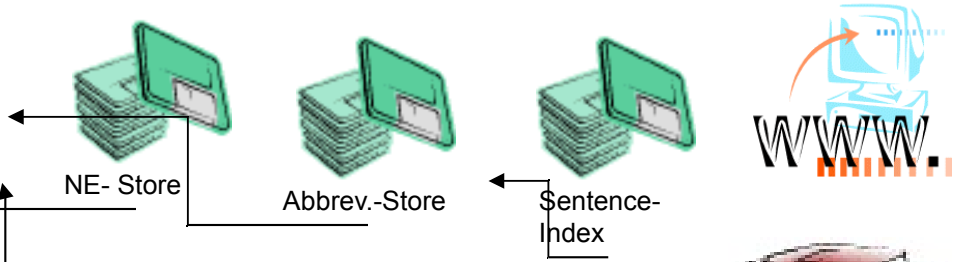
Answer processing components



Open-domain Question Answering: Multi-dimensional annotation

- Idea: off-line annotation of the data collection, which support
  - Query-specific indexing (Q-strategies), and
  - Answer extraction
- Sentence-level pre-processing proved valuable
  - Sentences-boundary
  - Named Entity + Co-reference
  - Abbreviations
  - NE-lists (+ context)





Why multi-dimensional annotation of un-structured text?

- The assumption is that a structural analysis of un-structured texts towards the type of information that can be the focus of questions, will support the retrieval of relevant small textual information units through informative IR-queries.
  - From candidate document retrieval to candidate answer retrieval.

- However, since we cannot foresee all the different user's interests/questions, a challenging research question is:
  - How detailed can the structural analysis be made without putting over a "straitjacket" of a particular view on the un-structured source?

- The assumptions here are:
  - Questions and answers are somewhat related ("questions influence the information geometry and hence, the information view and access", see also Rijsbergen, 2004)
  - There is a bias between off-line and on-line answer extraction.

Some initial experiments

We have performed some experiments focusing on the relationship between the size of information units and answer containment (using the QA-test set from Clef-2003).

#N Unit-Type	1	5	10	20	30	40	50	100
*Sentences**	37.9	58.2	65.8	69.6	70.8	72.1	74	75.9
Sentences	28.4	53.1	60.1	67	70.2	72.7	72.7	74.6
*Passages**	39.8	63.2	68.3	73.4	74	75.3	76.5	77.8
Passages	31.6	60.7	67.7	71.5	74.6	77.2	77.2	80.3
*Documents**	47.4	69.6	76.5	80.3	81.6	82.9	82.9	83.5
Documents	46.2	68.3	77.8	82.2	82.2	83.5	84.1	85.4

As a result we hyphothesized that it is reasonable to use NE-annotated sentences as major retrieval units for the IR-engine

⇒

Simplified answer extraction process & no need of special passage extraction methods

Precision of retrieval for different unit types and top N units retrieved, namely documents, passages, sentences –
and their NE-annotated correspondents (marked by *).

Experiments & Results

Run ID	Right		W	X	U
	#	%	#	#	#
dfki061dede$_M$	60	30	121	14	5
dfki061ende$_C$	37	18.5	144	18	1
dfki061deen$_C$	14	7	178	6	2
dfki062esen$_C$	10	5	180	10	0
dfki062ptde$_C$	5	2.5	189	4	2

Performance still ok although some lost

Coverage problems of English Wh-parser

BUG in NE-Informed Translation (used DE-based recognizer)

Problems with MT online services (PT-EN-DE)

Remarks

- Online MT services are still insufficient
  - Develop own MT solutions, cf. EU project EuroMatrix
- Bad coverage of our English Wh-parser
  - First prototype for Clef 2007
- Answer extraction currently robust enough for different answer sources
  - Similar performance for newspaper and Wikipedia
- Need more semantic analysis on answer side without lost of coverage and domain-independency
  - We are exploring cognitive semantics (cf. Talmy, 1987)
- Number of QA components also used in QAST pilot task and AVE

DFKI at QAST and AVE

- QAST pilot task
  - For given written factoid question
  - Extract answer from manual or automatic speech transcripts
- Answer Validation Exercise
  - Given a triple of form (question, answer, supporting text)
  - Decide whether the answer to the question is correct and
  - Is supported or not according to the given supporting text

Result (encouraging)

Task	#Q	#A	MRR	ACC
T1	98	19	0.17	0.15
T2	98	9	0.09	0.09

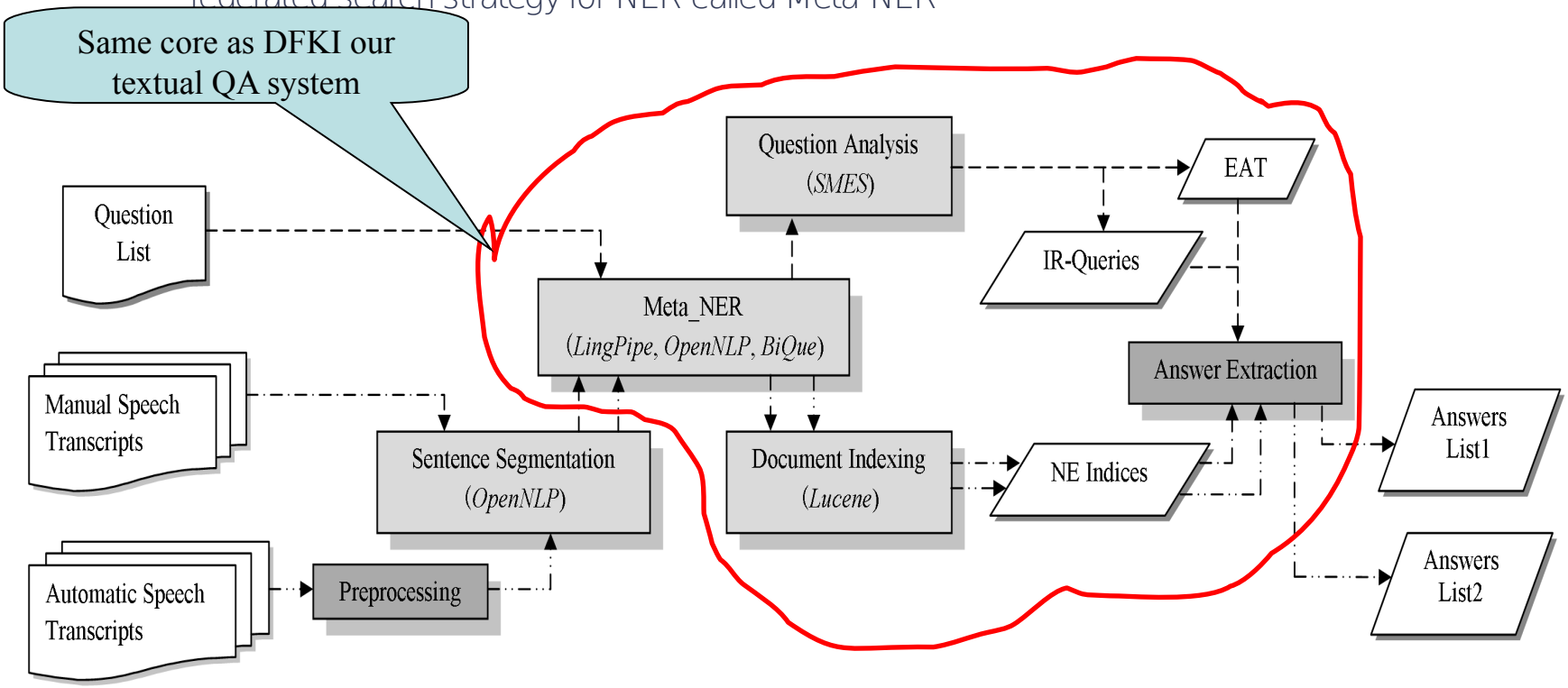
T1 = Chill corpus manual
T2 = Chill corpus automatic

Result (really encouraging)

Runs	Recall	Precision	F-measure	QA Accuracy
dfki07-run1	0.62	0.37	0.46	0.16
dfki07-run2	0.71	0.44	**0.55**	0.21

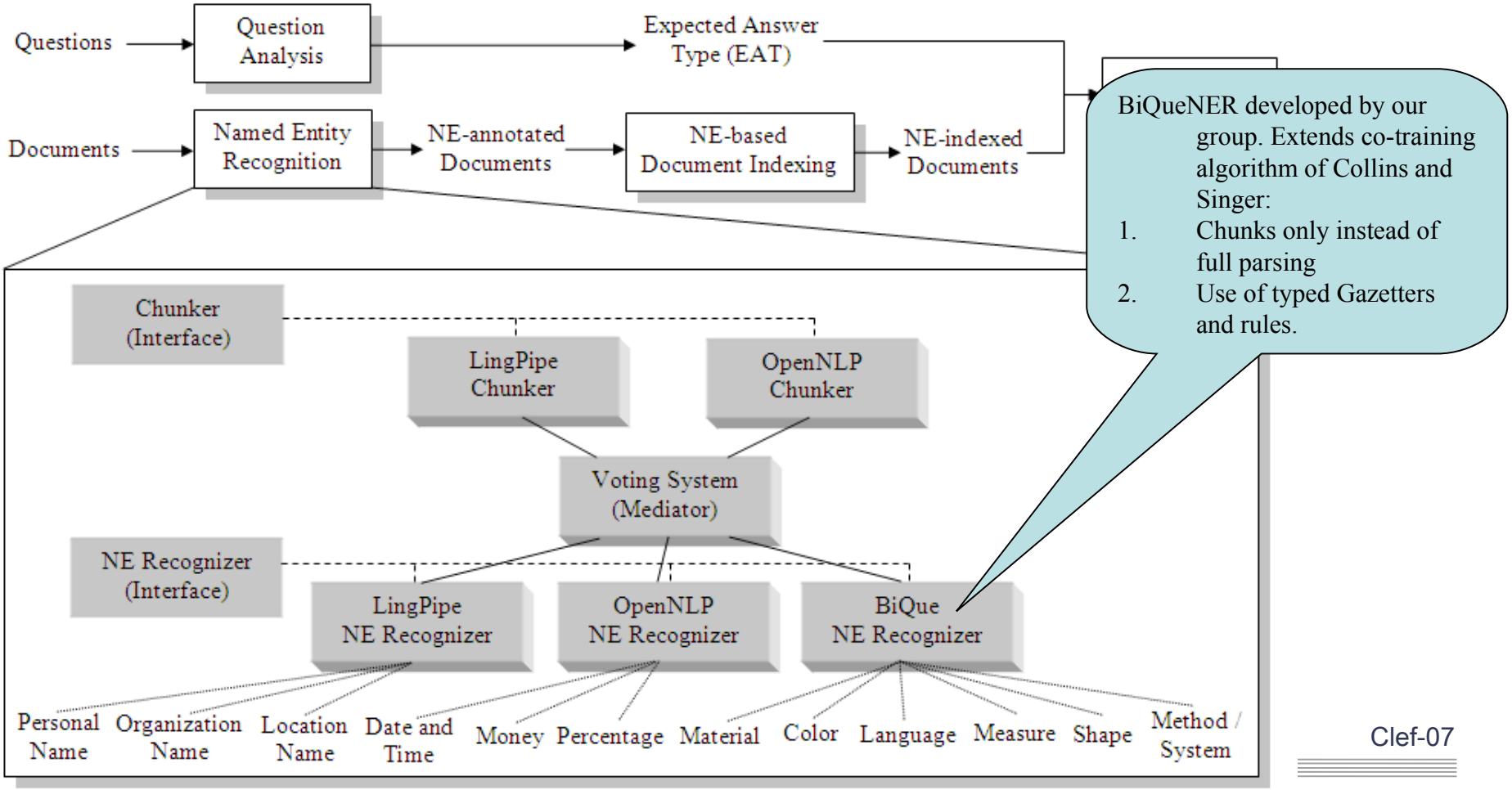
DFKI at QAST pilot task

- Goals
  - Get experience with this sort of answer sources
  - Adapt our text-based open-domain QA system that we used for the Clef main tasks
  - Since QAST required different set of expected answer types we developed a federated search strategy for NER called Meta-NER



META-NER

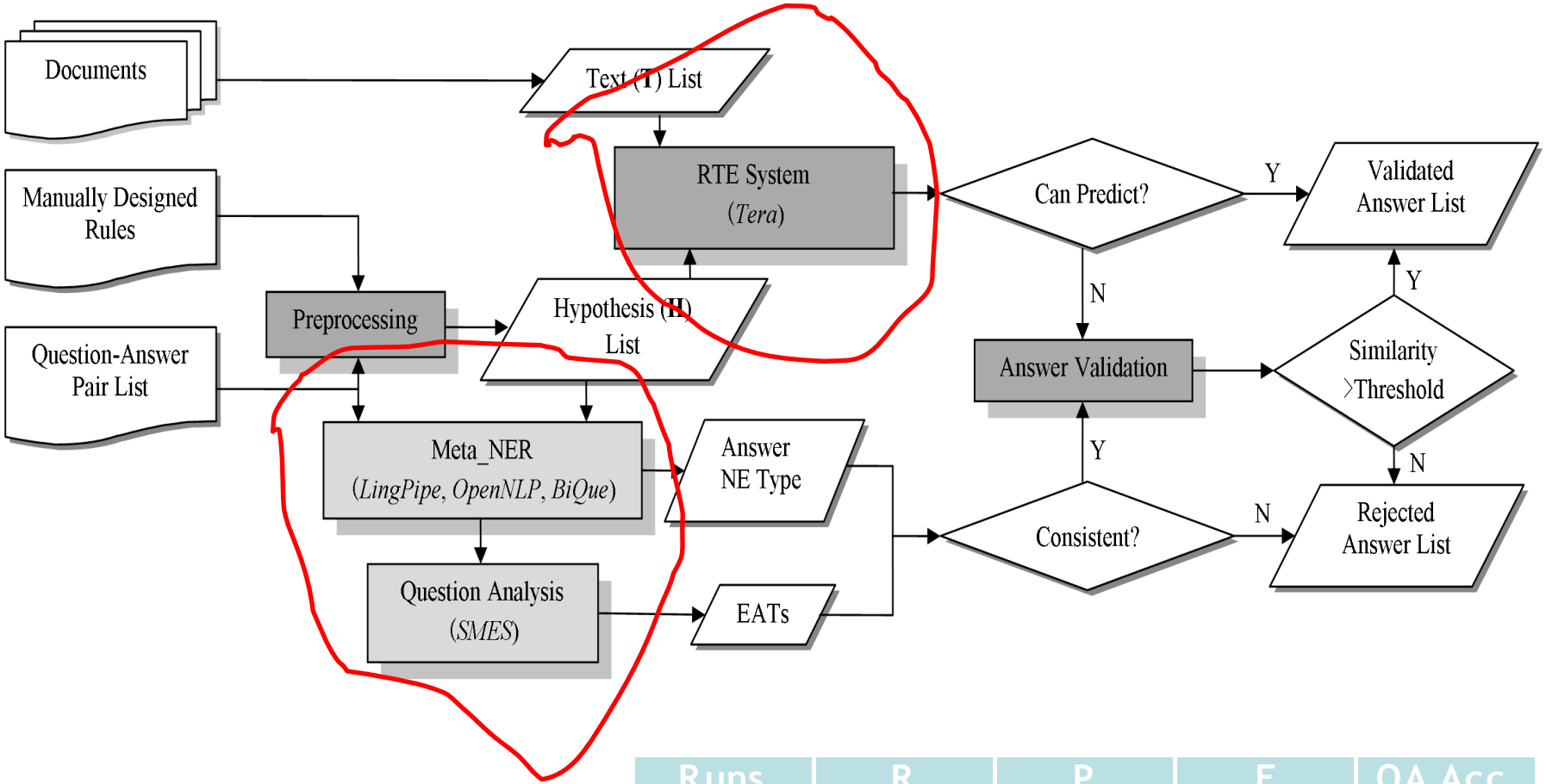
- Call several NER in parallel
- Merge results by a voting strategy



DFKI's AVE System

- AVE System is based on our RTE system (cf. Wang & Neumann, AAAI-2007, RTE-3 challenge)
- RTE method already demonstrated good results for QA task
  - RTE-3 (only QA): 81.5 %, Trec-2003 QA: 65.7 %
- RTE Method: Novel sentence level Kernel method
  - Subtree alignment on syntactic level
    - Check similarity between tree of H and relevant subtree in T
  - Subsequence kernel
    - Consider all possible subsequence of spine (path) of difference pairs
    - SVM for classification

AVE architecture



Runs	R	P	F	QA Acc.
run1	0.62	0.37	0.46	0.16
run2	0.71	0.44	**0.55**	0.21

Error Analysis

- Supporting text from web documents cause parsing problems
- Violation of some of our RTE system's assumptions
  - Required: H should be "verbally" smaller than T
  - Violated by: Q-A made patterns are too long
  - impact on recall
- If supporting text is very long (a complete document) then our RTE system is misleaded
  - Impact on precision

Thanks!