Question Answering

Günter Neumann

Language Technology Lab at DFKI

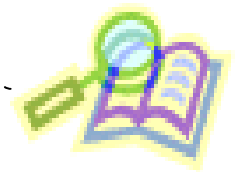
Saarbrücken, Germany

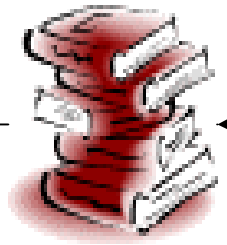
Towards Answer Engines

User Query:
KeyWrds, Wh-Clause, Q-Text

Experienc
ed-based
QA cycles

Search Engines



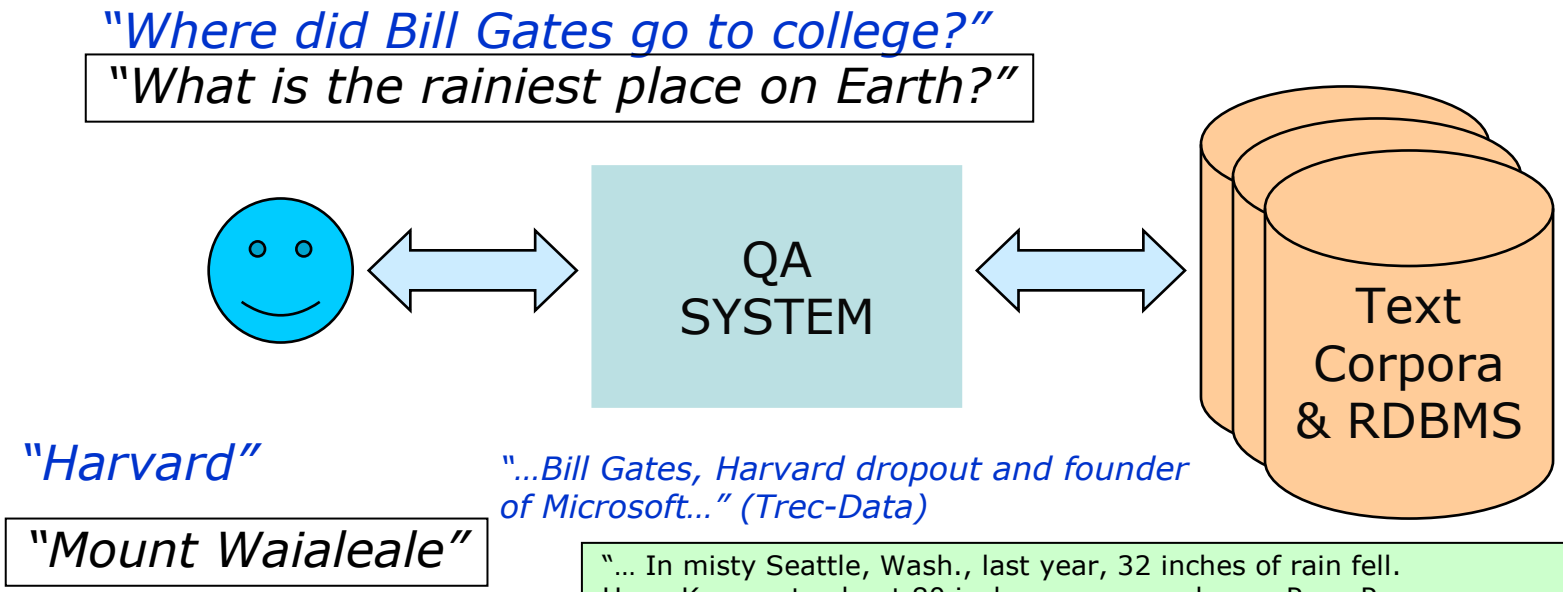


Answer Engines

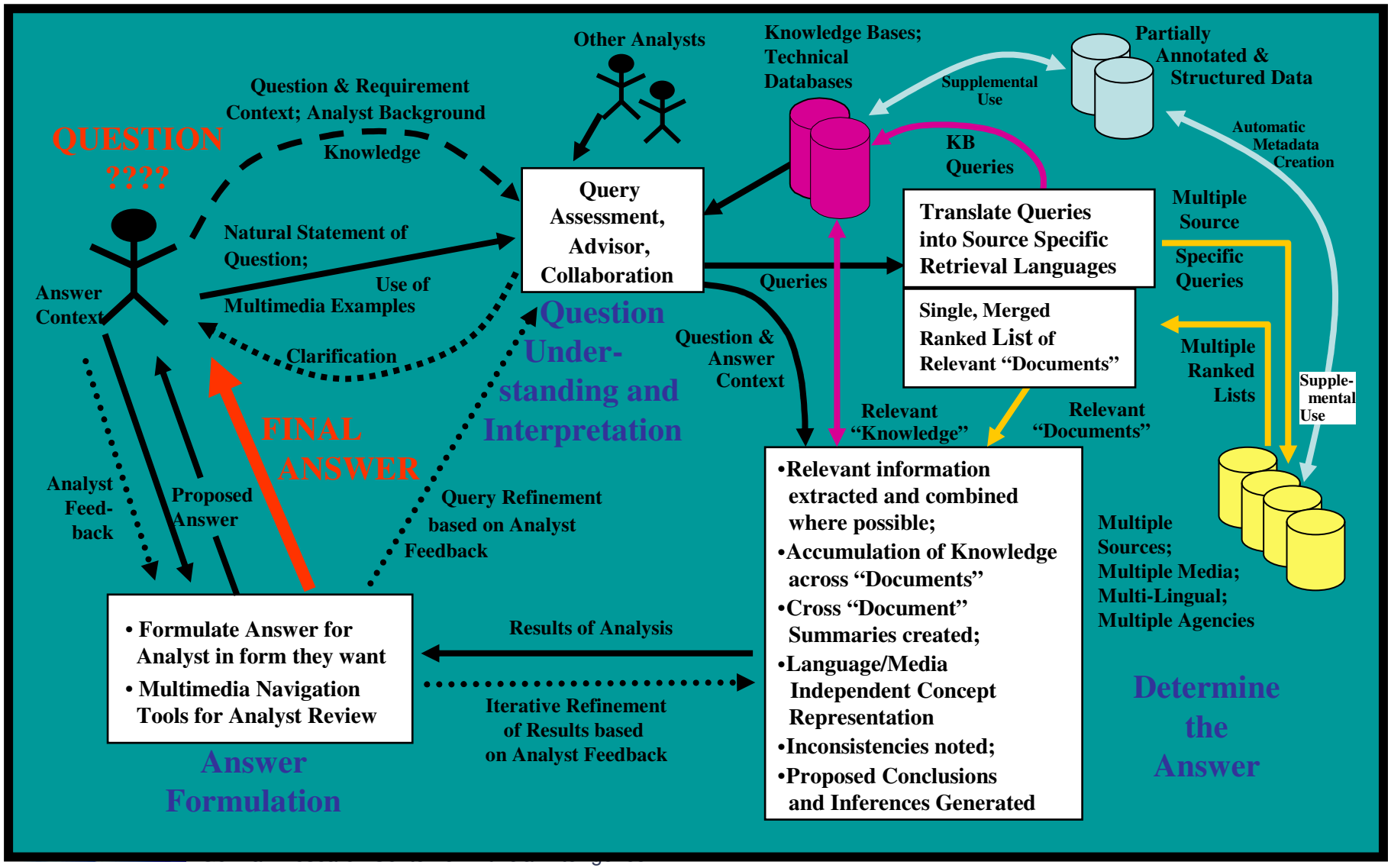
Shift more „interpretation effort" to machines

User still carries the major efforts in understanding

- ☆ Input: a question in NL; a set of text and database resources
- ☆ Output: a set of possible answers drawn from the resources



Intelligent information analysts



Challenges for QA

- QA systems should be able to:
  - Timeliness: answer question in real-time, instantly incorporate new data sources.
  - Accuracy: detect no answers if none available.
  - Usability: mine answers regardless of the data source format, deliver answers in any format.
  - Completeness: provide complete coherent answers, allow data fusion, incorporate capabilities of reasoning.
  - Relevance: provide relevant answers in context, interactive to support user dialogs.
  - Credibility: provide criteria about the quality of an answer

Challenges for QA

- ☆ Open-domain questions & answers
- ☆ Information overload
  - How to find a needle in a haystack?
- ☆ Different styles of writing (newspaper, web, Wikipedia, PDF sources,…)
- ☆ Multilinguality
- ☆ Scalability & Adaptibility

Information Overload

"The greatest problem of today is how to teach people to ignore the irrelevant, how to refuse to know things, before they are suffocated. For too many facts are as bad as non at all". (W.H. Auden)

Problems in Information Access?

- ☆ Why is there an issue with regards to information access?
- ☆ Why do we need support in find answers to questions?

- ☆ IA increasingly difficult when we have consider issues such as:
  - the size of collection
  - the presence of duplicate information
  - the presence of misinformation (false information/ inconsistencies)

What is Question Answering ?

- ☆ Natural language questions, not queries
- ☆ Answers, not documents (containing possibly the answer)
- ☆ A resource to address ‘information overload’?
- ☆ Most research so far has focused on fact-based questions:
  - “How tall is Mount Everest?”,
  - “When did Columbus discover America?”,
  - ”Who was Grover Cleveland married to?”.
- ☆ Current focus is towards complex questions
  - List, definition, temporally restricted, event-oriented, why-related, …
  - Contextual questions like “How far is it from here to the Cinestar?”
- ☆ Also support information-seeking dialogs:
  - “Do you mean President Cleveland?”
  - “Yes”.
  - “Francis Folsom married Grover Cleveland in 1886.”
  - “What was the public reaction to the wedding?”

Ancestors of Modern QA

- ☆ Information Retrieval
  - Retrieve relevant documents from a set of keywords; search engines
- ☆ Information Extraction
  - Template filling from text (e.g. event detection); e.g. TIPSTER, MUC
- ☆ Relational QA
  - Translate question to relational DB query; e.g. LUNAR, FRED

☆ Traditional QA Systems (TREC)

- Question treated like keyword query
- Single answers, no understanding

Q: *Who is prime minister of India?*

<find a person name close to *prime*, *minister*, *India* (within 50 bytes)>

A: *John Smith is not prime minister of India*

- **Future QA Systems**
  - System understands questions
  - System understands answers and interprets which are most useful
  - System produces sophisticated answers (list, summarize, evaluate)

What other airports are near Niletown?

Where can helicopters land close to the embassy?

Major Research Challenges

- ☆ Acquiring high-quality, high-coverage lexical resources
- ☆ Improving document retrieval
- ☆ Improving document understanding
- ☆ Expanding to multi-lingual corpora
- ☆ Flexible control structure
  - – “beyond the pipeline”
- ☆ Answer Justification
  - – Why should the user trust the answer?
  - – Is there a better answer out there?

Why NLP is Required

☆Question: "When was Wendy's founded?"

☆Passage candidate:

– "The renowned Murano glassmaking industry, on an island in the Venetian lagoon, has gone through several reincarnations since it was founded in 1291. Three exhibitions of 20th-century Murano glass are coming up in New York. By Wendy Moonan."

☆Answer: 20th Century

Predicate-argument structure

- ☆ **Q336:** *When was Microsoft established?*
- ☆ **Difficult** because Microsoft tends to establish lots of things…

  Microsoft plans to establish manufacturing partnerships in Brazil and Mexico in May.

- ☆ Need to be able to detect sentences in which `Microsoft' is **object** of `establish' or close synonym.
- ☆ Matching sentence:

  Microsoft Corp was founded in the US in 1975, incorporated in 1981, and established in the UK in 1982.

Why Planning is Required

- ☆ Question: *What is the occupation of Bill Clinton's wife?*
  - No documents contain these keywords plus the answer
- ☆ Strategy: decompose into two questions:
  - *Who is Bill Clinton's wife?* = X
  - *What is the occupation of* X*?*

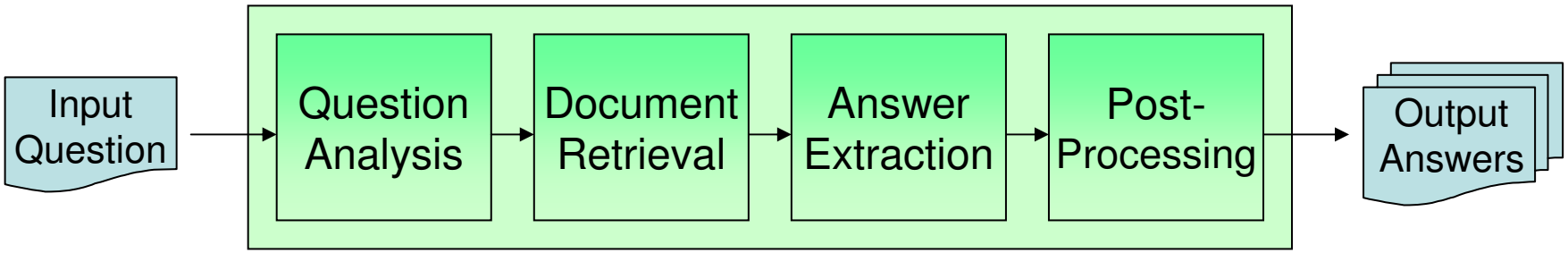
Brief history of QA Systems

- ☆ The focus in the beginning of QA research was on closed-domain QA for different applications:
  - Database: NL front ends to databases
    - BASEBALL (1961), LUNAR (1973)
  - AI: dialog interactive advisory systems
    - SHRLDU (1972), JUPITER (2000)
  - NLP: story comprehension
    - BORIS (1972)
  - NLP: retrieved answers from an encyclopedia
    - MURAX (1993)
- ☆ At late 90th the focus shifted towards open-domain QA
  - TREC's QA track (began in 1999)
  - Clef crosslingual QA track (since 2003)

Open-Domain Question Answering

- ☆ Open domain
  - No restrictions on the domain and type of question
  - No restrictions on style and size of document source
- ☆ Combines
  - Information retrieval, Information extraction
  - Text mining, Computational Linguistics
  - Semantic Web, Artificial Intelligence
- ☆ Cross-lingual ODQA
  - Express query in language X
  - Answer from documents in language Y
  - Eventually translate answer in Y to X

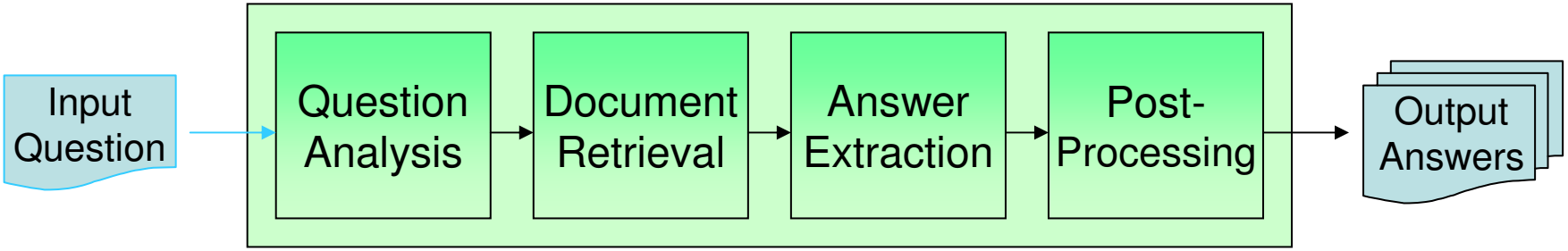
Classic "Pipelined" OD-QA Architecture



☆ A sequence of discrete modules cascaded such that the output of the previous module is the input to the next module.

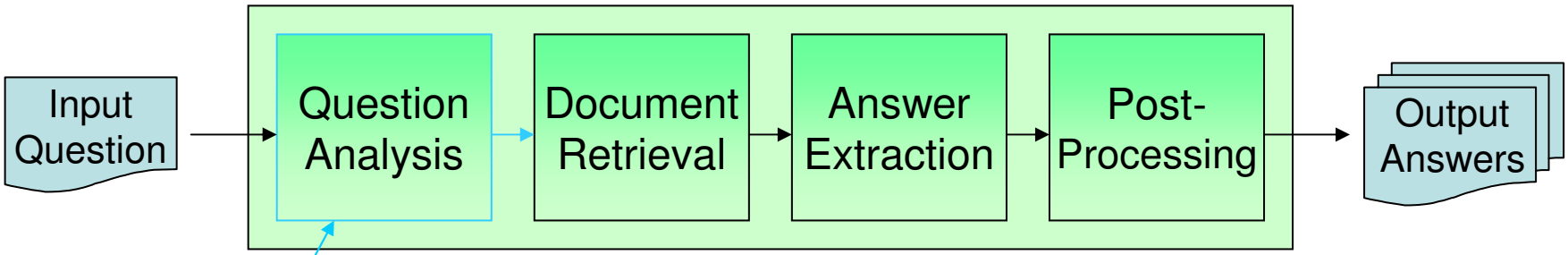
Classic "Pipelined" OD-QA Architecture

"Where was Andy Warhol born?



Classic "Pipelined" OD-QA Architecture

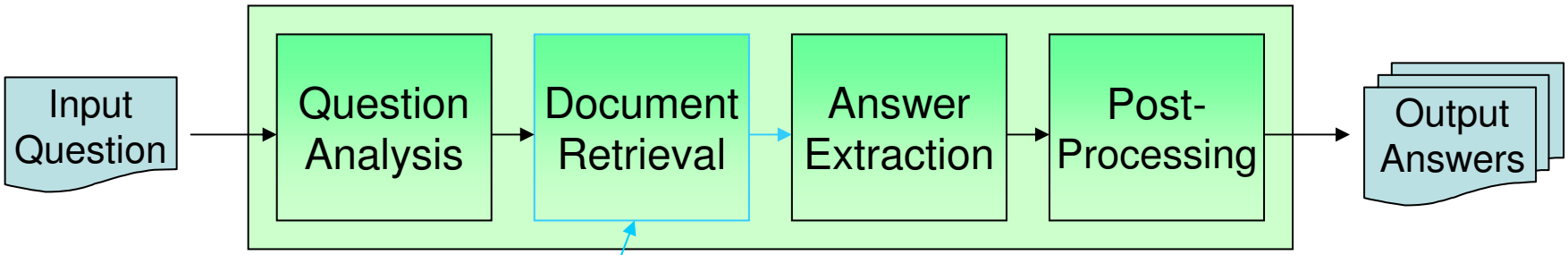
"Where was Andy Warhol born?



Discover keywords in the question, generate alternations, and determine answer type.

Keywords: Andy (Andrew), Warhol, born
Answer type: **Location (City)**

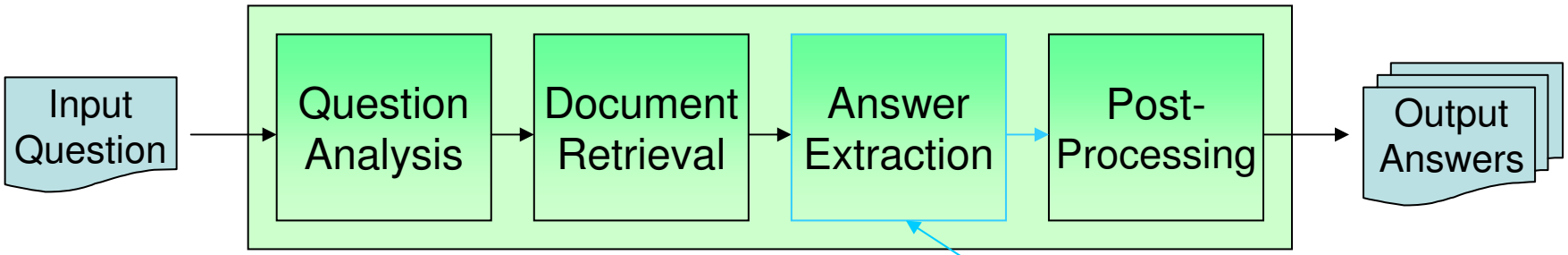
Classic "Pipelined" OD-QA Architecture



Formulate IR queries using the keywords, and retrieve answer-bearing documents

(Andy OR Andrew) AND Warhol AND born

Classic "Pipelined" OD-QA Architecture

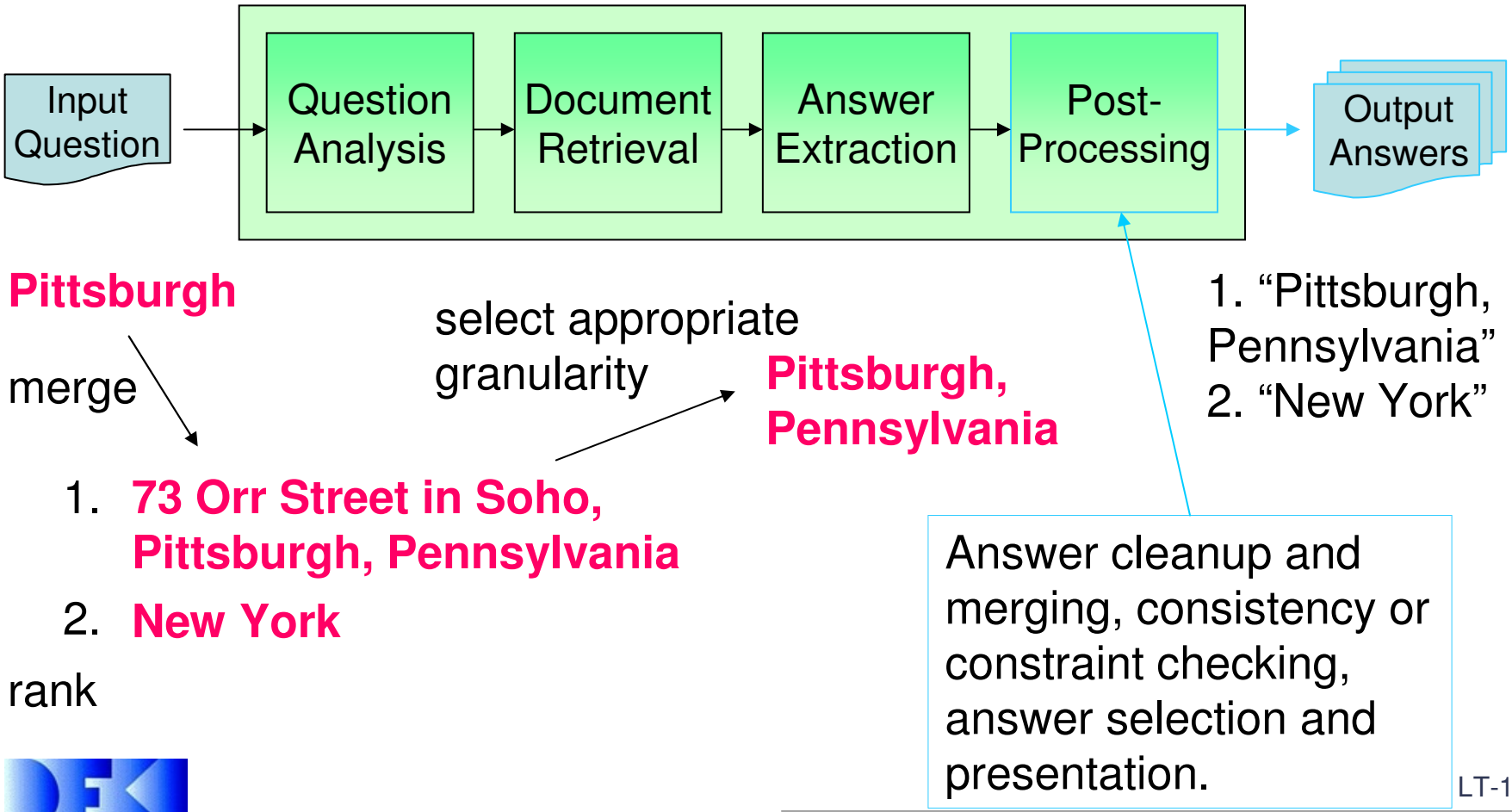


"**Andy Warhol** was born on **August 6, 1928** in **Pittsburgh** and died **February 22, 1927** in **New York**."

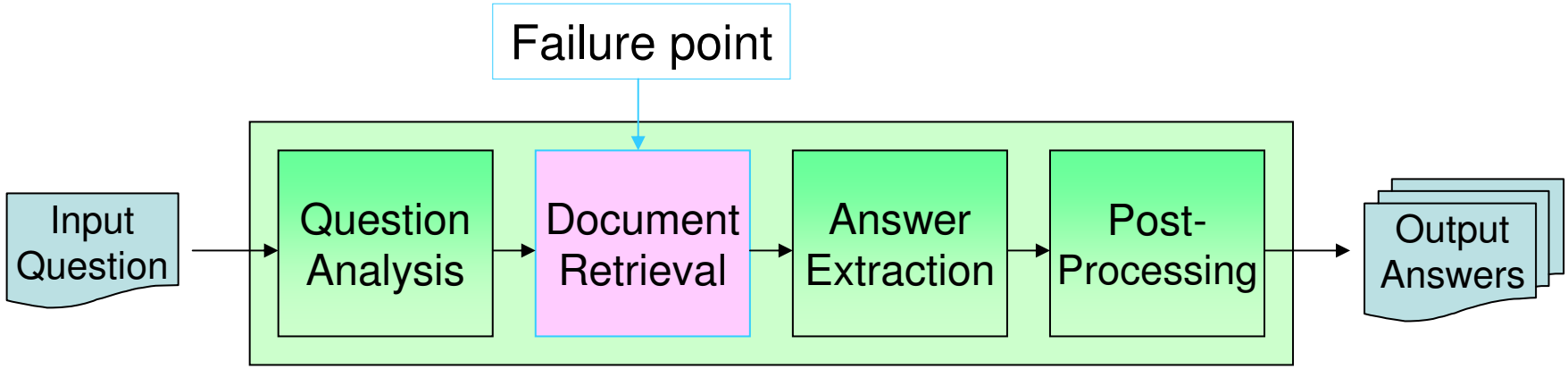
Extract answers of the expected type from retrieved documents.

"**Andy Warhol** was born to Slovak immigrants as **Andrew Warhola** on **August 6, 1928**, on **73 Orr Street in Soho, Pittsburgh, Pennsylvania**."

Classic "Pipelined" OD-QA Architecture



What is the cause of wrong answers?



- ☆ A pipelined QA system is only as good as its weakest module
- ☆ Poor retrieval and/or query formulation can result in low ranks for answer-bearing documents, or no answer-bearing documents retrieved

TREC QA track

☆ What is TREC?

- Text REtrieval Conference is a series of workshops aim at developing research on technologies for IR.
- started: 1992, Sponsored by: NIST, DARPA
- TREC-10 (2001), no. of tracks: 6, no. participants: 87

☆ What is TREC QA track?

- focuses on the evaluation of systems, in a competition-based manner, that answer questions in unrestricted domains.
- started: TREC-8 (1999), no. participants: 20
- Homepage: http://trec.nist.gov/data/qamain.html

History of QA at TREC

- ☆ QA Track first introduced at TREC 8 (Voorhees, 1999)
  - 200 fact-based short-answer questions
  - Questions mainly back formulated from documents
  - Answers could be 50-byte or 250-bytes snippets
  - 5 answers could be returned for each question
  - Best systems could answer over 2/3 of the questions (Moldovan et al., 1999; Srihari and Li, 1999).
- ☆ TREC 10 (Voorhees, 2001) introduced:
  - List questions such as *"Name 20 countries that produce coffee"*
    - *Best 3 systems: 0.76%, 0.45%, 0.34% average accuracy* *(computed as the number of distinct instances divided by the target number instances)*
    - *Average for all 9 systems: 0.33 %*
  - Questions which don't have an answer in the collection (NIL answers)

History of QA at TREC

- ☆ In TREC 11 (Voorhees, 2002):
  - Answers had to be exact
  - Only one answer could be returned per question
  - Best 3 systems: 83%, 58%, 54.2%, accuracy on 500 questions
  - Next systems: 38.4%, 36.8%, 35,8%, 28.4%, …
- ☆ TREC 12 (Voorhees, 2003) Introduced definition questions:
  - Define a target such as “aspirin” or “Aaron Copland”
  - A definition should contain a number of important facts (vital nuggets)
  - Can also include other associated information (non-vital nuggets)
  - Evaluated using a length based precision metric which penalizes long answers containing few nuggets.
    - Performance for the best systems: 0.555, 0.473, 0.461, 0.442, 0.338, 0.318
  - Final scores (fact, list, def questions) for best systems:
    - 0.559, 0479, 0.363, 0.313, 0.266, 0.256

History of QA at TREC

- ☆ TREC 13 (Voorhees, 2004) combines the three question types into scenarios around targets. For instance
  - Target: Hale Bopp Comet
  - Factoid: When was the comet discovered?
  - Factoid: How often does it approach the earth?
  - List: In what countries was the comet visible on it's last return?
  - Other: Tell me anything else not covered by the above questions
- ☆ Performance of best systems:
  - 0.601, 0.545, 0.386, 0.278

- ☆ Questions were based around 75 targets
  - 19 people
  - 19 organizations
  - 19 things
  - 18 events
- ☆ The series of targets contained a total of:
  - 362 factoid questions
  - 93 list questions
  - 75 (one per target) other questions
- ☆ All answers had to be with reference to a document in the AQUAINT collection of newswire texts.

Example Scenarios

- ☆ AMWAY
  - F: When was AMWAY founded?
  - F: Where is it headquartered?
  - F: Who is president of the company
  - L: Name the officials of the company
  - F: What is the name “AMWAY” short for?
  - O:
- ☆ return of Hong Kong to Chinese sovereignty
  - F: What is Hong Kong’s population?
  - F: When was Hong Kong returned to Chinese sovereignty?
  - F: Who was the Chinese President at the time of the return?
  - F: Who was the British Foreign Secretary at the time?
  - L: What other countries formally congratulated China on the return?
  - O:

Example Scenarios

- Shiite
  - F: Who was the first Imam of the Shiite sect of Islam?
  - F: Where is his tomb?
  - F: What was this person's relationship to the Prophet Mohammad?
  - F: Who was the third Imam of Shiite Muslims?
  - F: When did he die?
  - F: What portion of Muslims are Shiite?
  - L: What Shiite leaders were killed in Pakistan?
  - O:

Evaluation Metrics

- For factoid questions the metric is accuracy
  - Only exact supported answers and correct NIL responses are counted
- For list questions the metric is F-measure (β = 1)
  - Only exact supported answers are counted
  - Set of correct answers (for recall purposes) is the union of all correct answers across all submitted runs plus any instances found during question development.
- For other questions the metric F-measure (β = 3)
  - Recall is the proportion of vital nuggets returned
  - Precision is a length based penalty, where each valid nugget allows 100 non-whitespace characters to be returned.
- These are combined to give a weighted score per target
  - Weighted Score = 0.5xFactoid + 0.25xListAvgF + 0.25xOtherAvgF
- Performance of the best systems:
  - 0.543, 0.464, 0.246, 0.241, 0.222, 0.205, 0.201, 0.187

TREC QA tracks

- ☆ Trec 2006 and 2007 QA main tracks
  - Quite similar to Trec 2005
  - More difficult questions whose answering required more reasoning
  - Additional text corpora, e.g., blogs in case of 2007
- ☆ Interactive QA – ciQA
  - Home page: http://www.umiacs.umd.edu/~jimmylin/ciqa/
  - Idea: given a information need in form of a template and a short NL description, provide a web-based QA system that can be used to do QA cycles
  - **Template:** What evidence is there for transport of [drugs] from [Bonaire] to [the United States]?
    Narrative: The analyst would like to know of efforts made to discourage narco traffickers from using Bonaire as a transit point for drugs to the United States. Specifically, the analyst would like to know of any efforts by local authorities as well as the international community.
  - Systems are evaluated by organizers by using a system for 5 minutes to process such a information need

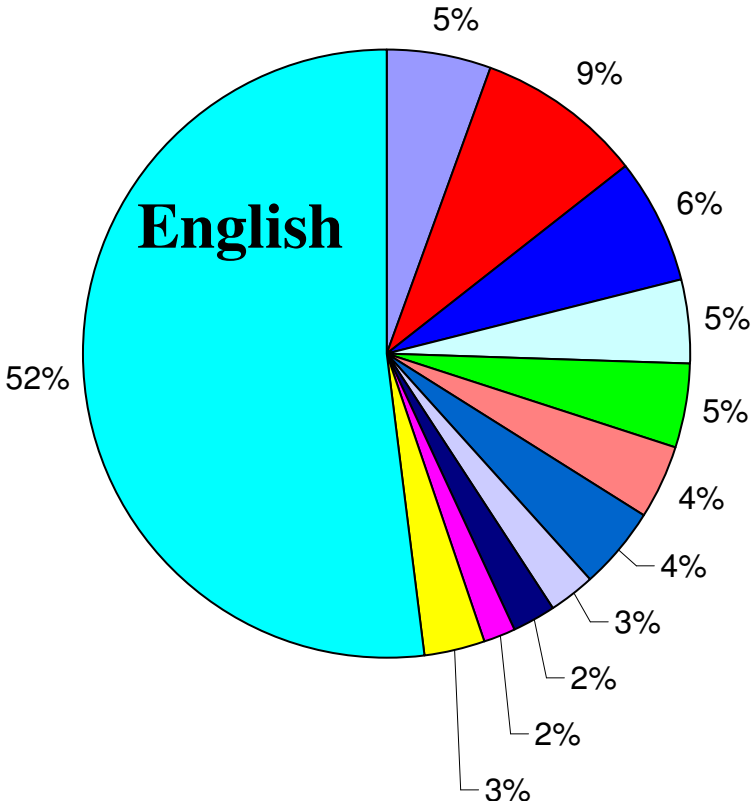
Find documents written in any language

– Using queries expressed in a single language

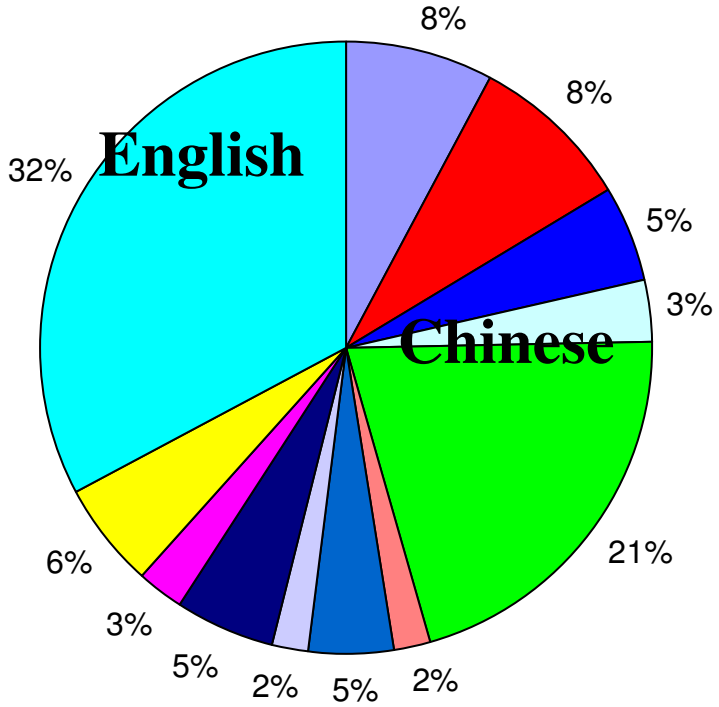
Source: D. Oard Cross-Language IR presentation

Why is CLIR important?

2000



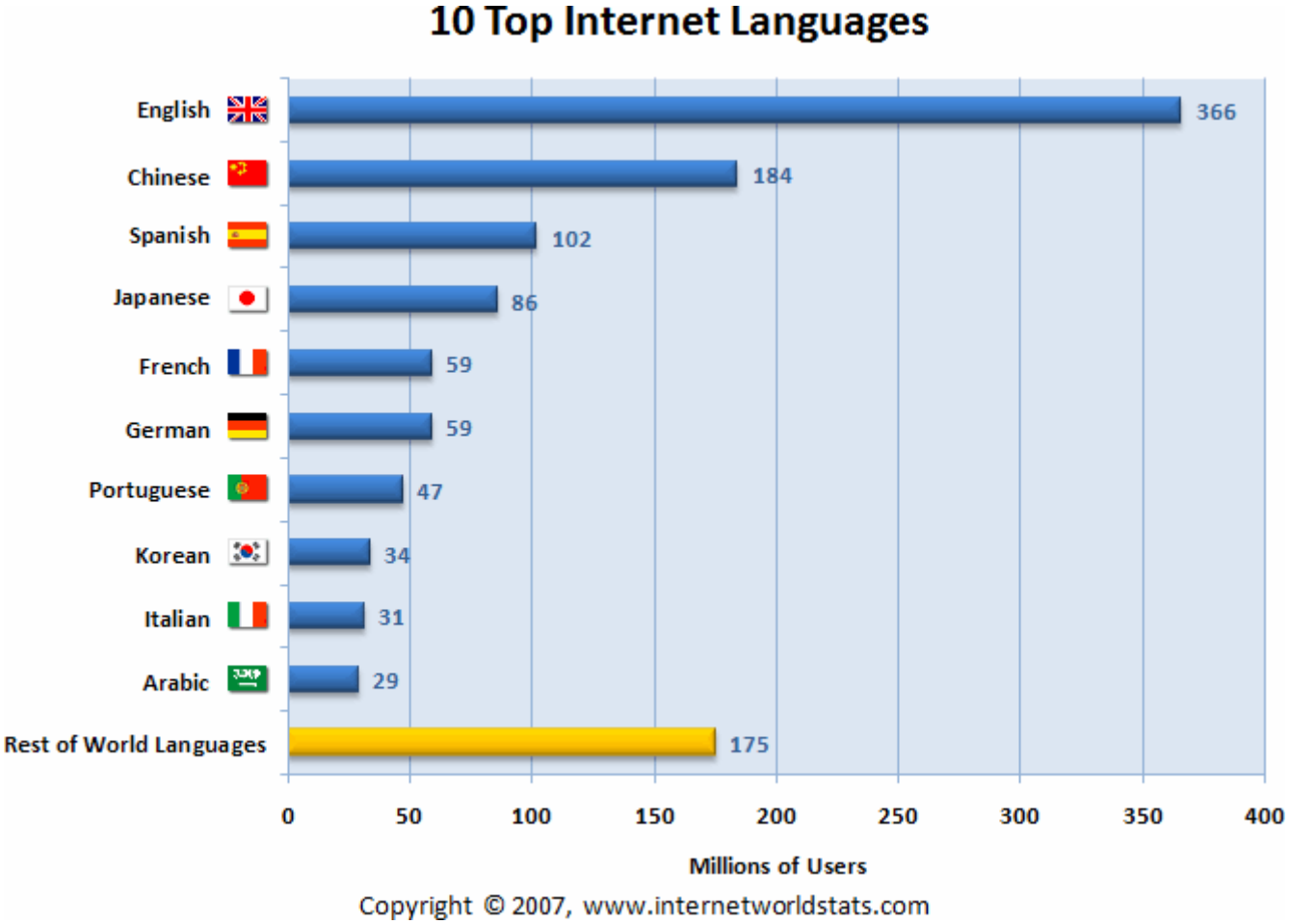
2005



Spanish | Japanese | German | French
Chinese | Scandanavian | Italian | Dutch
Korean | Portuguese | Other | English

Source: Global Reach

http://www.internetworldstats.com/stats7.htm



More details

CLIR Conferences

- ☆ Cross Language Evaluation Forum (CLEF)
  - CLIR using European languages.
    - Bulgarian, Danish, Dutch, English, Finnish, French, German, Italian, Portuguese, Spanish, Swedish, Russian
    - http://clef.iei.pi.cnr.it/

- ☆ NTCIR (NII-NACSIS Test Collection for IR Systems) Project
  - CLIR in Asian Languages
    - Chinese, Japanese, and Korean
    - http://research.nii.ac.jp/ntcir/index-en.html

Cross Language QA

- ☆ Similar task as TREC QA but with Questions and documents in different languages.
- ☆ CLEF
  - – Multiple Languages QA
    - 2003 preliminary task
    - 2004, 2005, 2006, 2007
- ☆ NTCIR
  - – Question Answering Challenge:
    - NTCIR 3 (QAC1 Oct 2001-Oct 2002)
    - NTCIR 4 (QAC2 Apr 2003 – June 2004)
    - NTCIR 5 (QAC3 Nov 2004 – June 2005)

Multilingual QA Track at Clef

	2003	2004	2005	2006	2007
Target languages	3	7	8	9	10
Collections	News 1994		+News 1995		+Wikipedia Nov. 2006
Type of questions	200 Factoid		+ temporal restrictions + Definitions	-Type of questions + Lists	+ Linked questions + Closed lists
Supporting information	Doc.	Doc.	Doc.	Snippet	Snippet
Pilots and exercises		-Temporal restrictions - Lists		-AVE - RealTime - WiQA	- AVE - QAST

Clef 2006: 200 Questions

- FACTOID (150): loc, mea, org, oth, per, tim
- DEFINITION (40): per, org, object, oth
  - Person: Who is Josef Paul Kleihues?
  - Object: What is a router?
  - Other: What is a tsunami?
- LIST (10): “Name works by Tolstoy.”
- Temporally restricted (40): by date, by period, by event
- NIL questions (without known answer in the collection
- Input format: question type (F, D, L) not indicated

Clef 2007: Clef 2006 plus

- ☆ Closed lists:
  - Who were the components of the Beatles?
  - Who were the **last three** presidents of Italy?
- ☆ Linked questions
  - Topic: Otto von Bismarck
    - Who was called the “Iron-Chancellor”?
    - When was he born?
    - Who was his first wife?
  - Topics
    - Person or Event
    - Not provided to participants
    - Only a portion of the questions (from 15% depending on the languages)

Run format

- Clef 2006:
  - Multiple answers: from one to ten *exact* answers per question
  - *exact* = neither more nor less than the information required
  - each answer has to be supported by
    - docid
    - one to ten text snippets justifying the answer (substrings of the specified document giving the actual context)
- Clef 2007:
  - News articles
  - Wikipedia dump from November 2006 (→ caused critical decrase of performance)

Clef 2007: Activated Tasks

(at least one registered participant)

- 10 Source languages (11 in 2006, 10 in 2005)
- 9 Target languages (8 in 2006, 9 in 2005)

S / T	BG	DE	EN	ES	FR	IN	IT	NL	PT	RO
BG										
DE										
EN										
ES										
FR										
IT										
NL										
PT										
RO										

Activated Tasks

- questions were not translated in all the languages
- **Gold Standard**: questions in multiple languages only for tasks were there was at least one registered participant

	MONOLINGUAL	CROSS-LINGUAL	TOTAL
CLEF 2003	3	5	8
CLEF 2004	6	13	19
CLEF 2005	8	15	23
CLEF 2006	7	17	24
CLEF 2007	8	29	37

Participants

	America	Europe	Asia	TOTAL	Registered participants	New comers	Veterans
CLEF 2003	3	5	-	8			
CLEF 2004	1	17	-	18 (+125%)	22	13	5
CLEF 2005	1	22	1	24 (+33%)	27	9	15
CLEF 2006	4	24	2	30 (+25%)	36	10	20
CLEF 2007				22 (-26%)	29	8	14

LT-Lab

List of participants (2006, 2007)

Acronym	NAME	Contry
SYNAPSE	SYNAPSE Developpement	France
Ling-Comp	U.Rome-La Sapienza	Italy
Alicante	U.Alicante- Informatica	Spain
Hagen	U.Hagen-Informatics	Germany
Daedalus	Daedalus Consortium	Spain
Jaen	U.Jaen-Intell.Systems	Spain
ISLA	U.Amsterdam	Netherlands
INAOE	Inst.Astrophysics,Optics&Electronics	Mexico
DEPOK	U.Indonesia-Comp.Sci.	Indonesia
DFKI	DFKI-Lang.Tech.	Germany
FURUI Lab.	Tokyo Inst Technology	Japan
Linguateca	Linguateca-Sintef	Norway
LIC2M-CEA	Centre CEA Saclay	France
LINA	U.Nantes-LINA	France
Priberam	Priberam Informatica	Portugal
U.Porto	U.Porto- AI	Portugal
U.Groningen	U.Groningen-Letters	Netherlands
	Univ. of Evora	Portugal
	Univ. Poli. De Catalunay	Spain

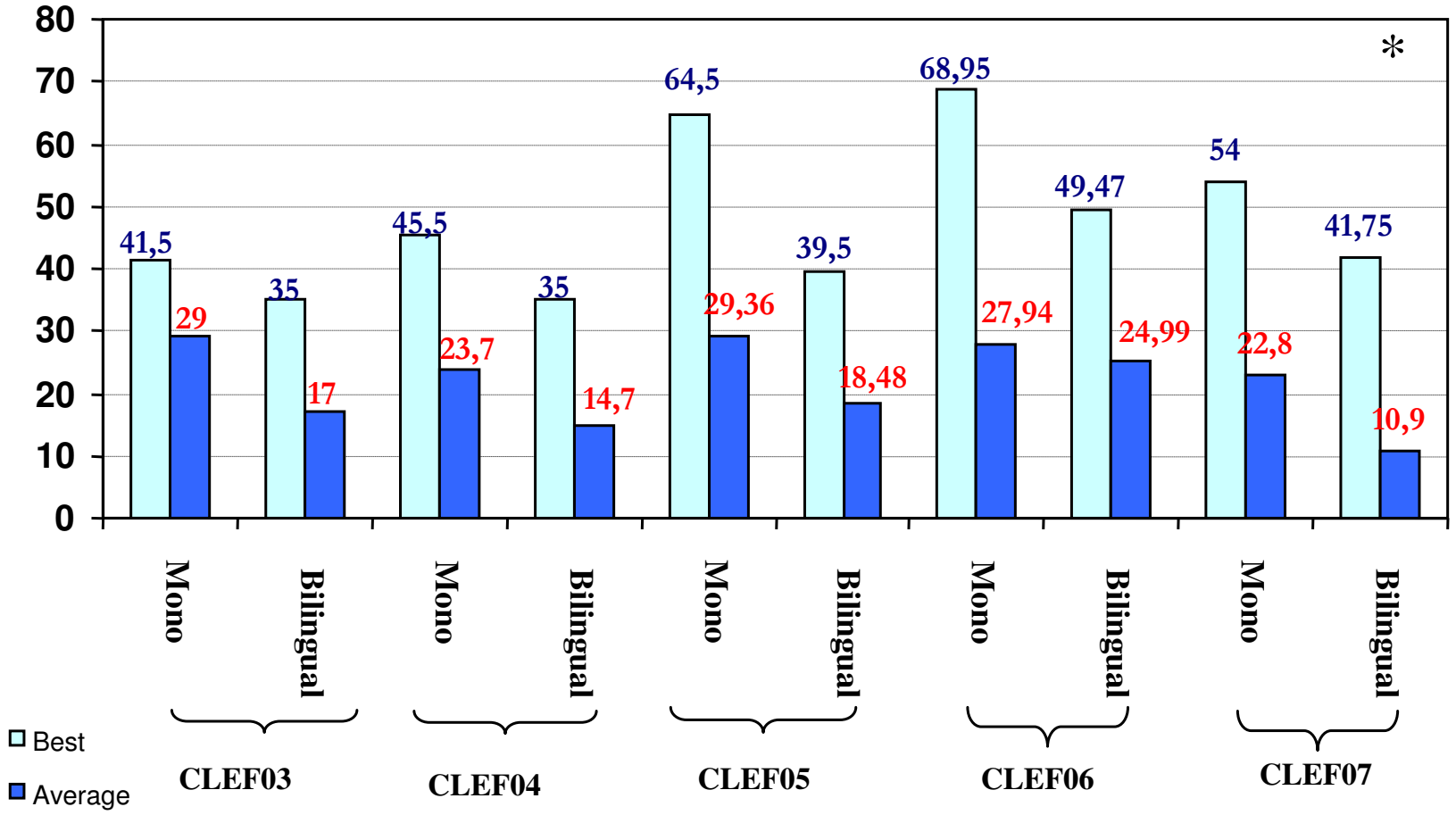
Acronym	NAME	Country
Lab.Inf.D' Avignon	Lab.Inf. D'Avignon	France
U.Sao Paulo	U.Sao Paulo – Math	Brazil
Vanguard	Vanguard Engineering	Mexico
LCC	Language Comp. Corp.	USA
UAIC	U.AI.I Cuza" Iasi	Romania
Wroclaw U.	Wroclaw U.of Tech	Poland
RFIA-UPV	Univ.Politècnica de Valencia	Spain
LIMSI	CNRS Lab-Orsay Cedex	France
U.Stuttgart	U.Stuttgart-NLP	Germany
FBK	FBK-IRST	Italy
JRC-ISPRA	Institute for the Protection and the Security of the Citizen	Italy
BTB	BulTreeBank Project	Bulgaria
dltg	University of Limerick	Ireland
	INESC-ID	Portugal
	Univ. Wolverhampton	UK
	Cindi Group	Canada
	Macquarie University	Australia
	RACAI	Romania

BLUE=Industrial Companies, GREEN=2006 + 2007, RED=not 2007, BLACK=new 2007

Submitted runs

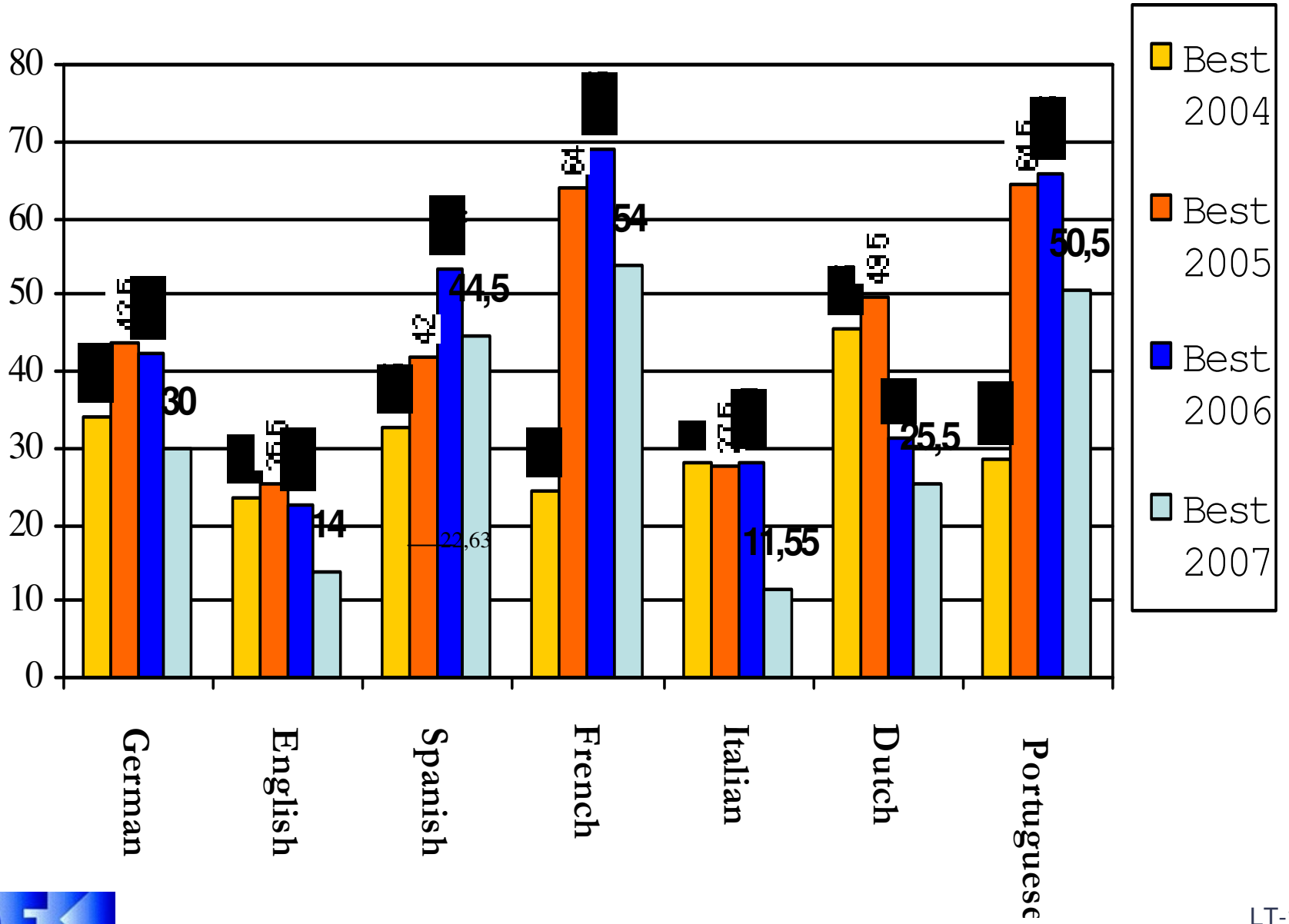
	Submitted runs #	Monolingual #	Cross-lingual #
CLEF 2003	17	6	11
CLEF 2004	48 (+182%)	20	28
CLEF 2005	67 (+39.5%)	43	24
CLEF 2006	77 (+13%)	42	35
CLEF 2007	37 (-52%)	20	17

Results: Best and Average scores

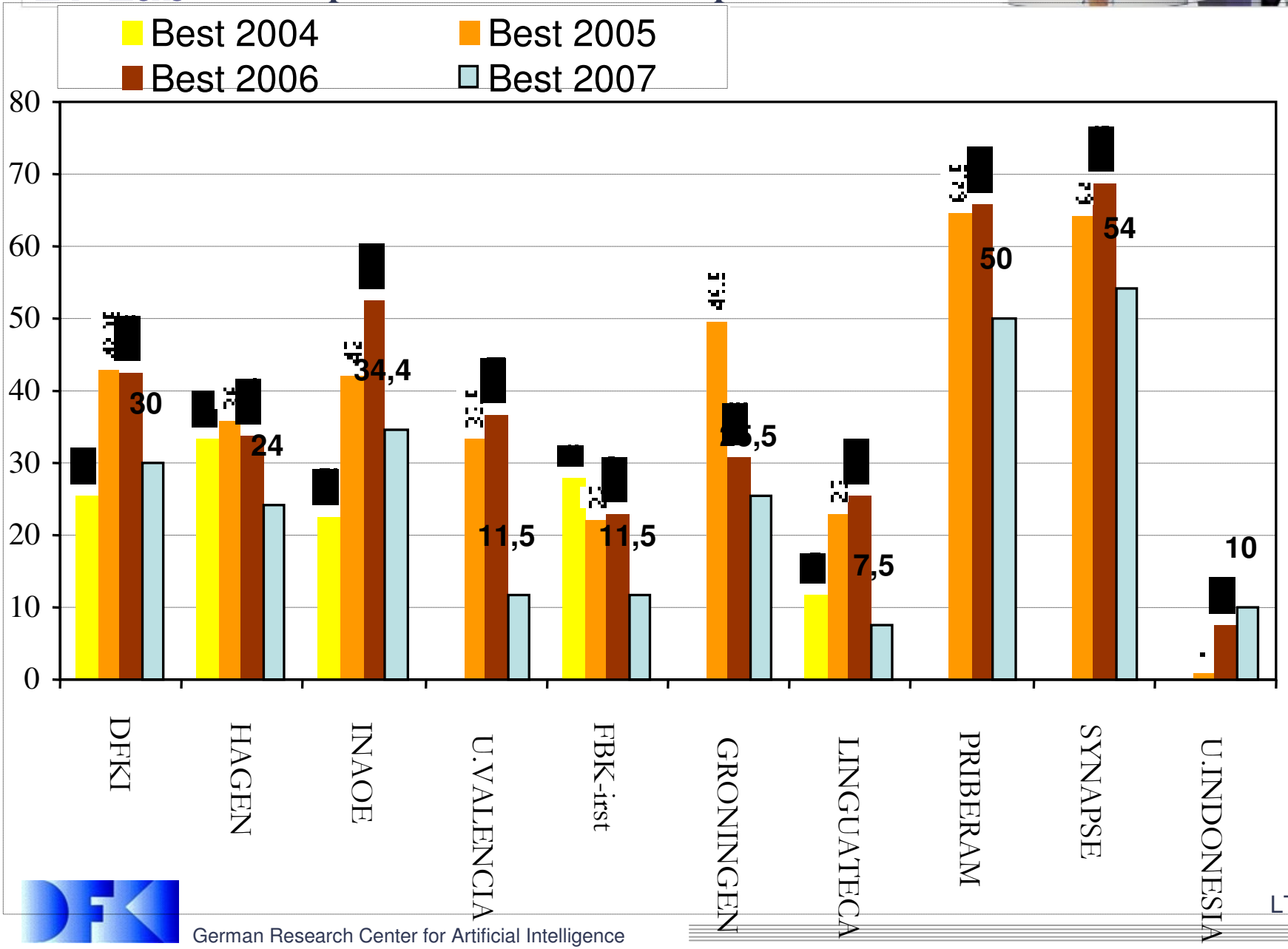


* This result is still under validation.

Best results in 2004, 2005, 2006, 2007



LT-Lab Participants in 2004 - 2007: compared best results



Lower results in 2007

- ☆ Some answers only in Wikipedia
- ☆ Closed lists
  - Almost no answers
- ☆ Temporal restrictions
  - Still very difficult
- ☆ Linked questions
  - Topic not provided
  - Fail the first, fail the rest
  - Co-reference resolution

DFKI-ODQA: Overview

- ☆ DFKI is participating since 2003
  - Focus on German monolingual QA and German/English cross-lingual QA
  - Best 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)

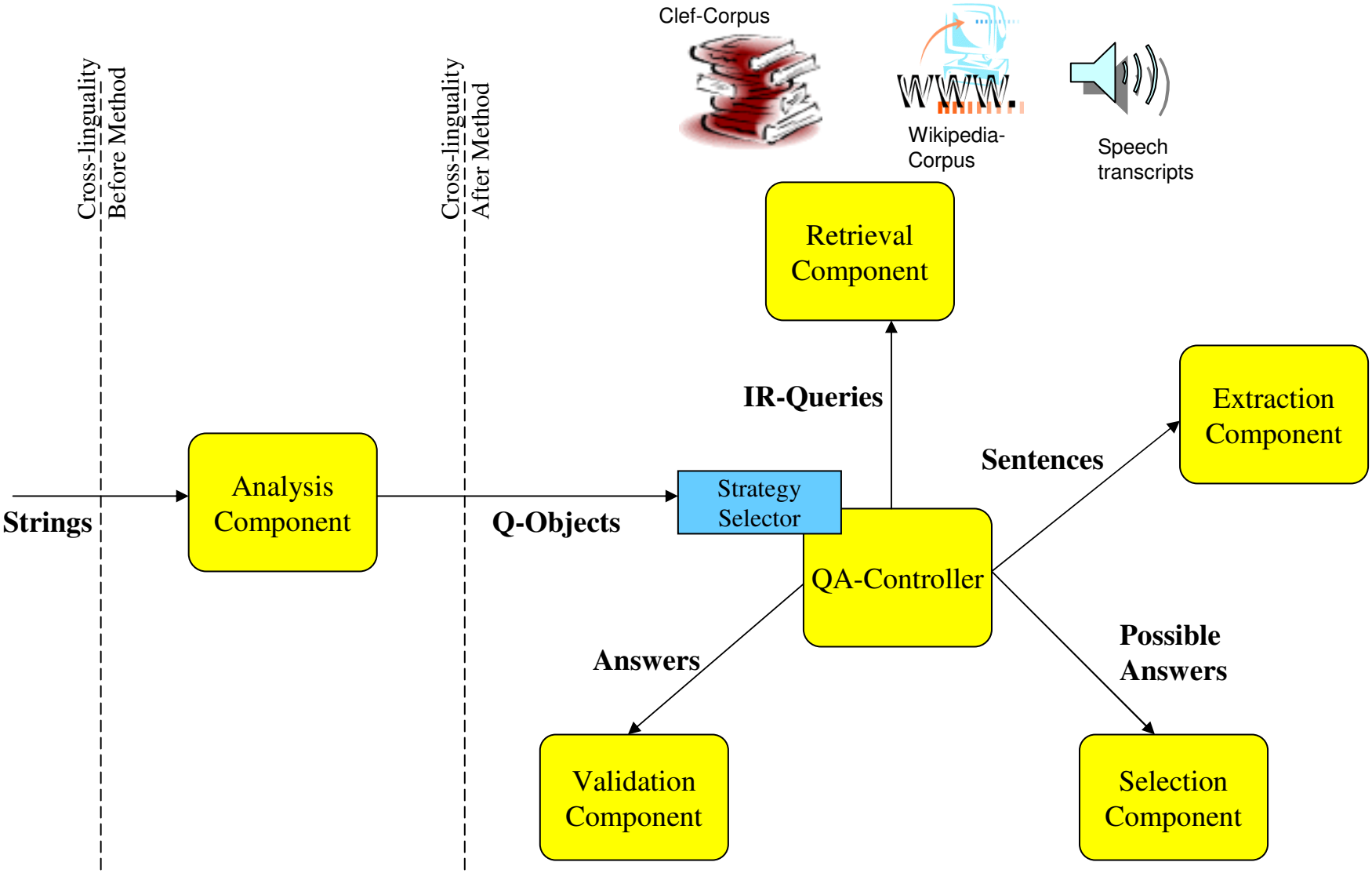
QA architecture – some design issues

- ☆ 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

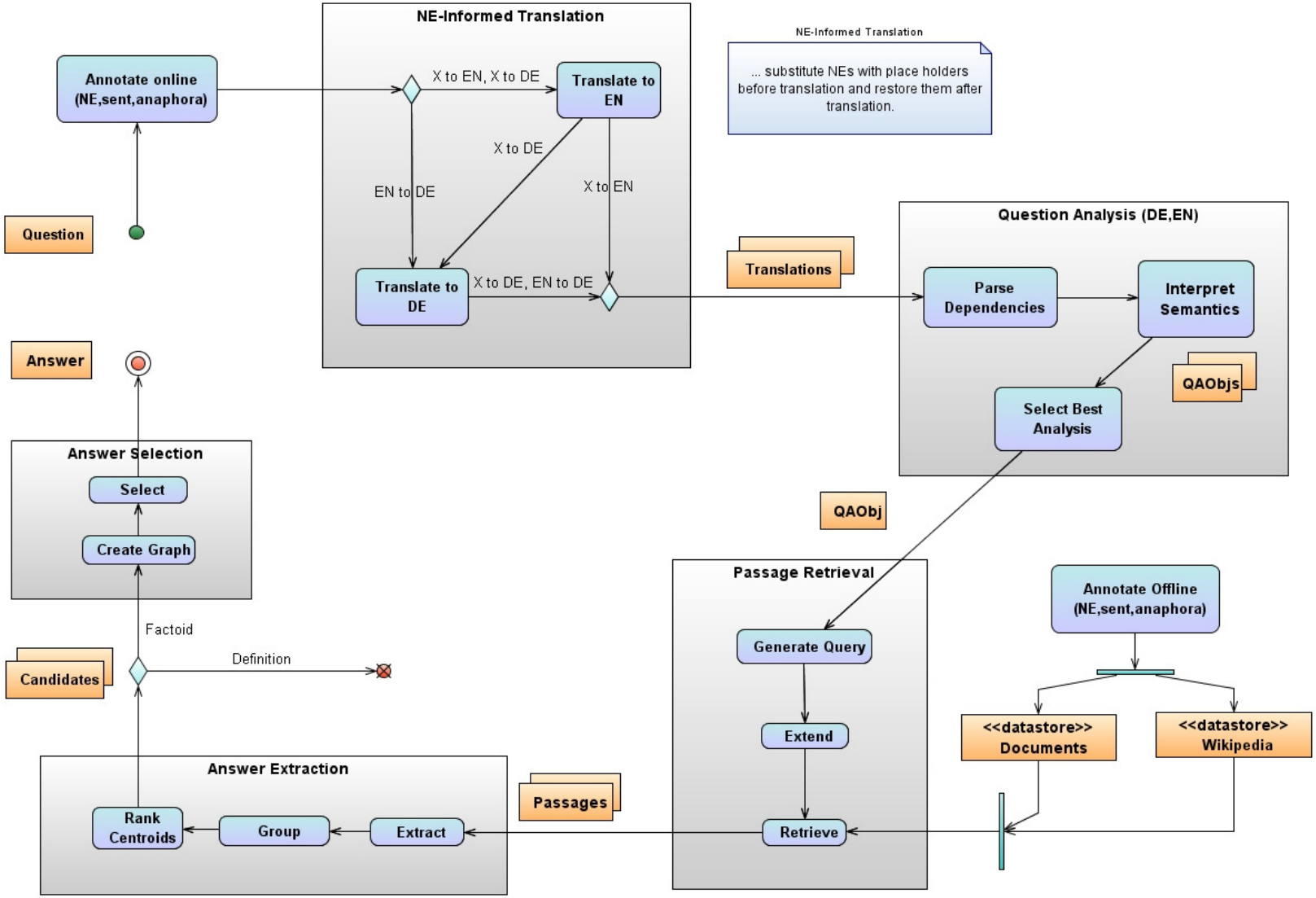
Common architecture for different answer pools

- ☆ 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)

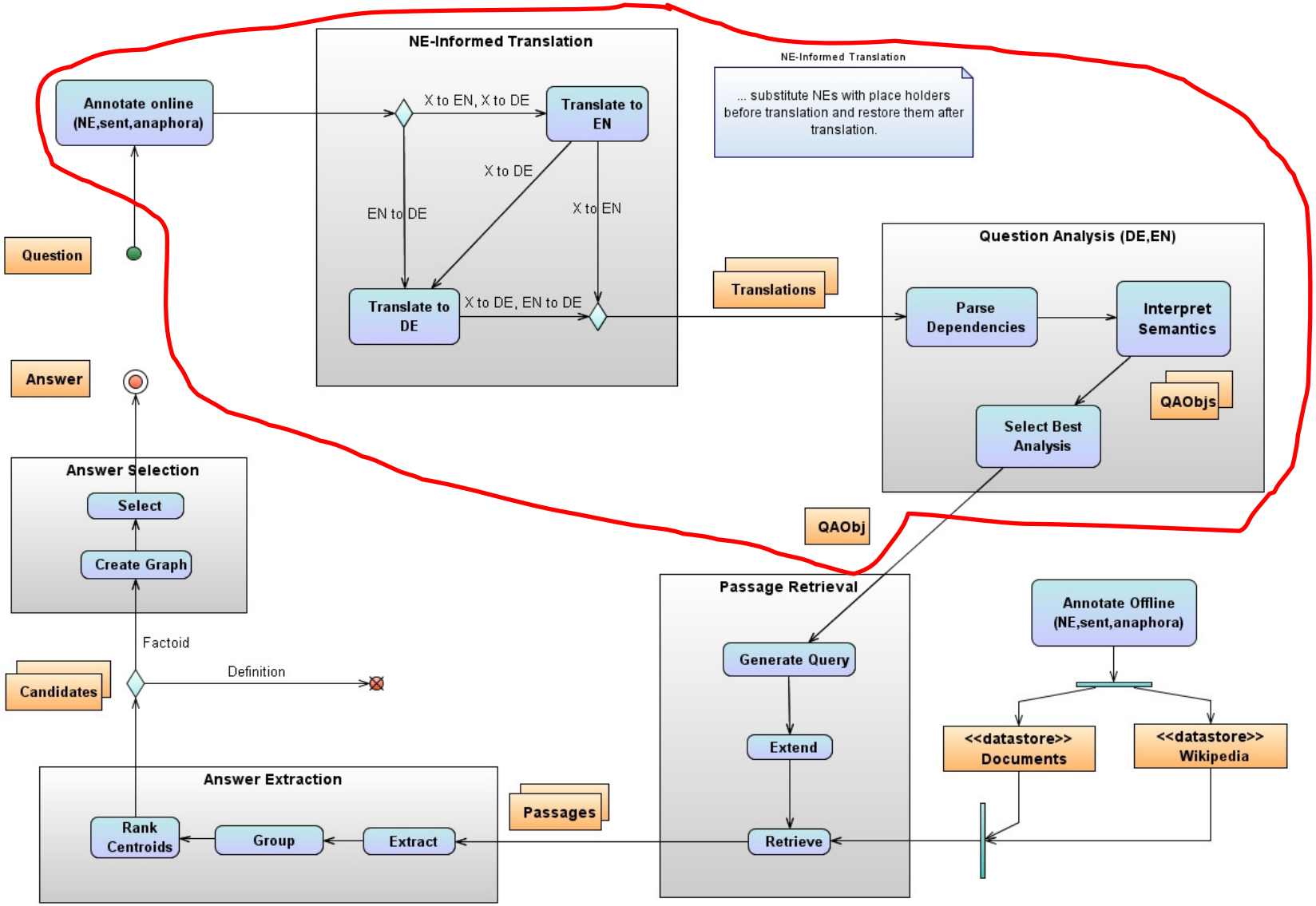
Overview QA architecture



LT-Lab System Architecture for Clef 2007



Query processing components

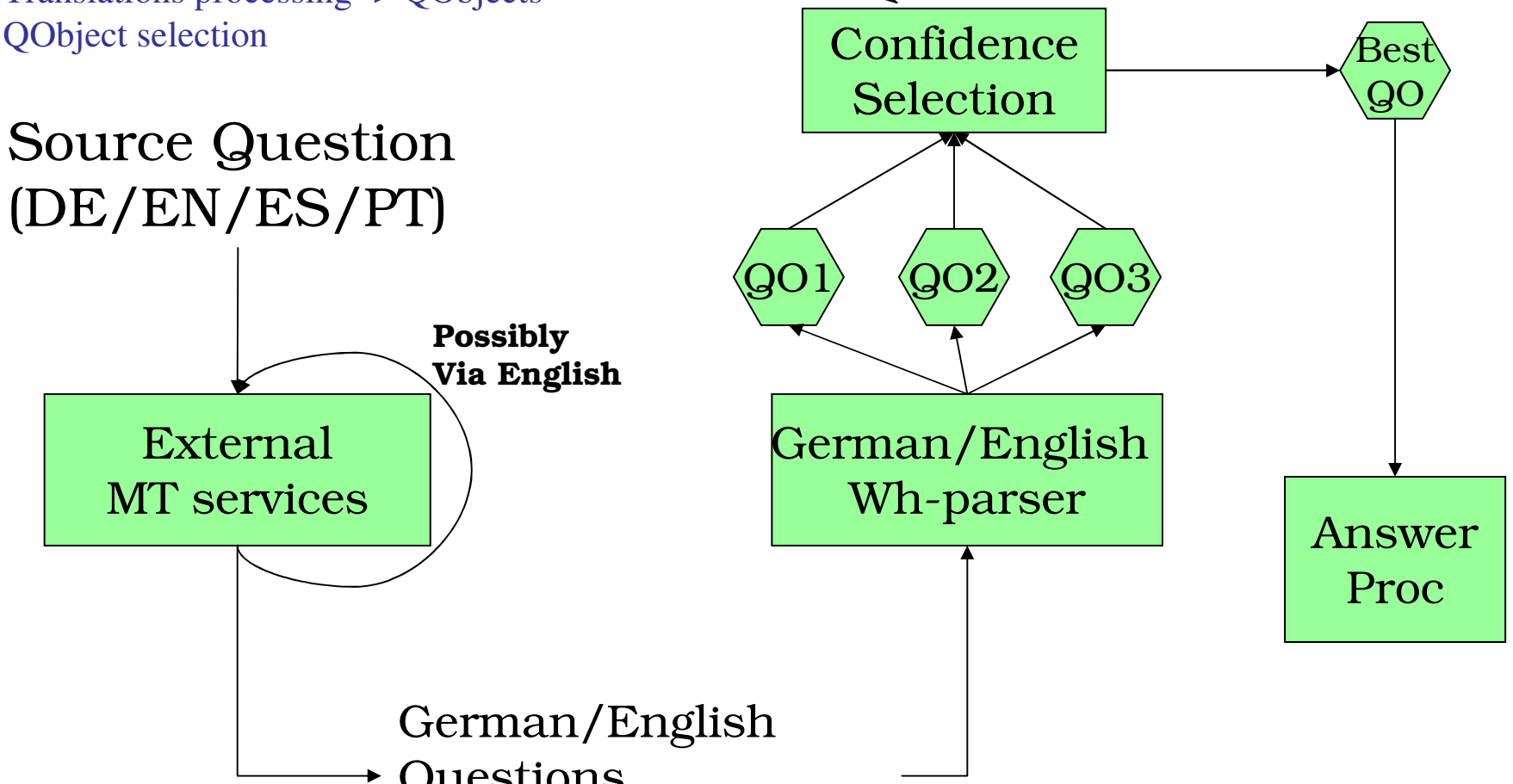


Before Method

- 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



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

Information access

Ouput example of query analysis

Which Jewish painter lived from

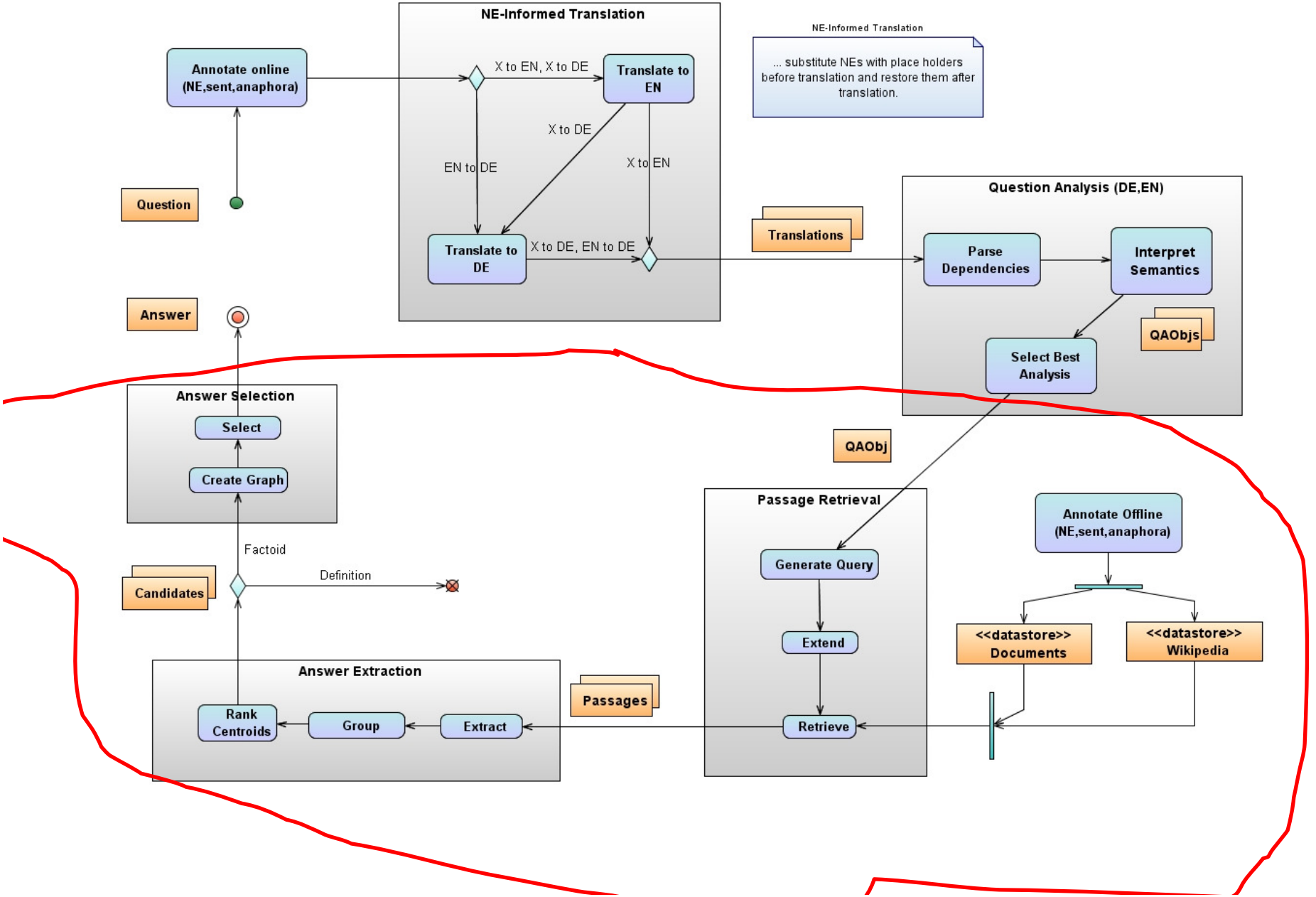
Exploiting Natural Language Generation

```
<QOBJ msg="quest" id="qId0" lang="DE" score
  <NL-STRING id="qId0">
    <SOURCE id="qId0" lang="DE">Welche juedischen
Maler lebten 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



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

Task: QAST 2007 Organization

Task jointly organized by :

- UPC, Spain (J. Turmo, P. Comas)

  Coordinator

- ELDA, France (C. Ayache, D. Mostefa)

- LIMSI-CNRS, France (S. Rosset, L. Lamel)

- ☆ 4 tasks were proposed:
  - T1 : QA in manual transcriptions of lectures
  - T2 : QA in automatic transcriptions of lectures
  - T3 : QA in manual transcriptions of meetings
  - T4 : QA in automatic transcriptions of meetings
- ☆ 2 data collections:
  - The CHIL corpus: around 25 hours (1 hour per lecture)

    Domain of lectures: *Speech and language processing*
  - The AMI corpus: around 100 hours (168 meetings)

    Domain of meetings: *Design of television remote control*

For each task, 2 sets of questions were provided:

- ☆ Development set (1 February 2007):
  - Lectures: 10 lectures (exam), 50 questions
  - Meetings: 50 meetings, 50 questions
- ☆ Evaluation set (18 June 2007):
  - Lectures: 15 lectures, 100 questions
  - Meetings: 118 meetings, 100 questions

Examples of speech source – manual transcripts from Chill corpus

<DOC>
<DOC_ID>ISL_20041111_B</DOC_ID>
<TOPIC>SPECTRAL ESTIMATION: NEW APPROACH FOR SPEECH RECOGNITION</TOPIC>
<DOC_TYPE>MANUAL TRANSCRIPTION</DOC_TYPE>
so yeah I just actually put the slides together so I might even surprise by myself which slide will be the next one .
so I hope we can straighten everything out and
I welcome you to my talk which I call uhm spectral estimation
new approach for speech recognition .
. so I want to start just
to give a
brief overview I want s
just start with a
first general
model for a speech recognition system .
how
hmm where we basically need the spectral estimation I will
talk about later .
so
usually we have the text generation .
then we have the the s
speech generation
and
we have a communication channel also
and then we have our speech recognition system so we get the signal with the microphone
and we do some signal processing
feature extraction before we give it to the speech detector
or the recognition system .
now I already jump to er uh the simplified filter model of speech production .
so if we talk about
speech signals it's very important
human
speech signals
it's very important to know that we have to basically separate
two different class of signals
which

Examples of speech source – automatic transcripts

so	69.400	0.440
yeah	69.840	0.250
I	70.110	0.100
just	70.210	0.360
actually	70.570	0.340
put	70.910	0.220
the	71.130	0.110
slides	71.240	0.400
together	71.640	0.550
so	72.310	0.300
I	72.610	0.120
might	72.730	0.210
even	72.940	0.220
surprise	73.160	0.490
by	73.650	0.100
myself	73.750	0.540
which	74.380	0.230
slide	74.610	0.500
will	75.110	0.110
be	75.220	0.110
the	75.330	0.060
next	75.390	0.35
0one	75.740	0.270
<s/>	76.010	0.270
{breath}	76.280	0.460
so	76.830	0.340
I	77.200	0.100
hope	77.300	0.300
we	77.600	0.120
can	77.720	0.300

Questions from the development set

Question-id, questions string

01 Which organisation has worked with the University of Karlsruhe on the meeting transcription system?
02 Where is the IBM research centre located?
03 Who is a guru in speech recognition?
04 How many speakers were transcribed from those recorded at the Eurospeech conference?
05 Where is ICSLP?
06 How many speakers were recorded at the Eurospeech conference?
07 Most of the speakers recorded at Eurospeech were non native speakers of which language?
08 When were the IWSLT evaluations?
09 Which organisation provided a significant amount of training data?
10 Where does Florian Metze work?
11 When did KTH start working on dialog systems?
12 Who looked at different automatic methods of deriving questions?
13 What is the weight of the blue spoon headset?
14 Where did the Eurospeech conference take place?
15 Who created the “how can I help you” system?
16 Which company does the speaker for the seminar on audio visual speech for pervasive computing belong to?
17 Where was the Eurospeech conference held in ninety-five?
18 Who has performed acoustic scene analysis?
19 Where is Gales from?
20 Where did Stefan Kantak present his work?

Question-id, nickname of participant, document-id, answer string

01 elda1_t1 ISL_20041123_E Carnegie Mellon
02 elda1_t1 ISL_20050127 New York|York town
03 elda1_t1 ISL_20050420 Gales
04 elda1_t1 ISL_20041111_B thirty-one speakers|thirty-one
05 elda1_t1 ISL_20041123_A Colorado
06 elda1_t1 ISL_20041111_B one hundred and eighty eight speakers|one hundred and eighty eight
07 elda1_t1 ISL_20041111_B English
08 elda1_t1 ISL_20041112_A two thousand and four
09 elda1_t1 ISL_20041123_E Icsi
10 elda1_t1 ISL_20041123_E University of Karlsruhe|Karlsruhe
11 elda1_t1 NIL
12 elda1_t1 ISL_20041123_A Miriam Keller
13 elda1_t1 ISL_20041123_C ten grams
14 elda1_t1 ISL_20050127 Geneva
14 elda1_t1 ISL_20041111_B Berlin
15 elda1_t1 ISL_20041123_A AT&T
16 elda1_t1 ISL_20050127 the IBM research center|IBM research center|IBM
17 elda1_t1 NIL
18 elda1_t1 ISL_20041123_C Rob Malkin
19 elda1_t1 ISL_20050420 Cambridge
20 elda1_t1 ISL_20041123_A Colorado

Task: Questions and answer types

☆ Factual questions

Who is a guru in speech recognition?

☆ Expected answers = named entities.

List of NEs: person, location, organization, language, system/method, measure, time, color, shape, material.

☆ Examples of development set (quest, answ)

Task: Human judgment

- ☆ Assessors used QASTLE, an evaluation tool developed in Perl (by ELDA), to evaluate the data.

- ☆ Four possible judgments:
  - Correct
  - Incorrect
  - Inexact (too short or too long)
  - Unsupported (correct answers but wrong document)

Task: Scoring

- Two metrics were used:
  - Mean Reciprocal Rank (MRR):

    measures how well ranked is a right answer.
  - Accuracy:

    the fraction of correct answers ranked in the first position in the list of 5 possible answers

- Participants could submit up to 2 submissions per task and 5 answers per question.

Participants

☆ Five teams submitted results for one or more QAST tasks:

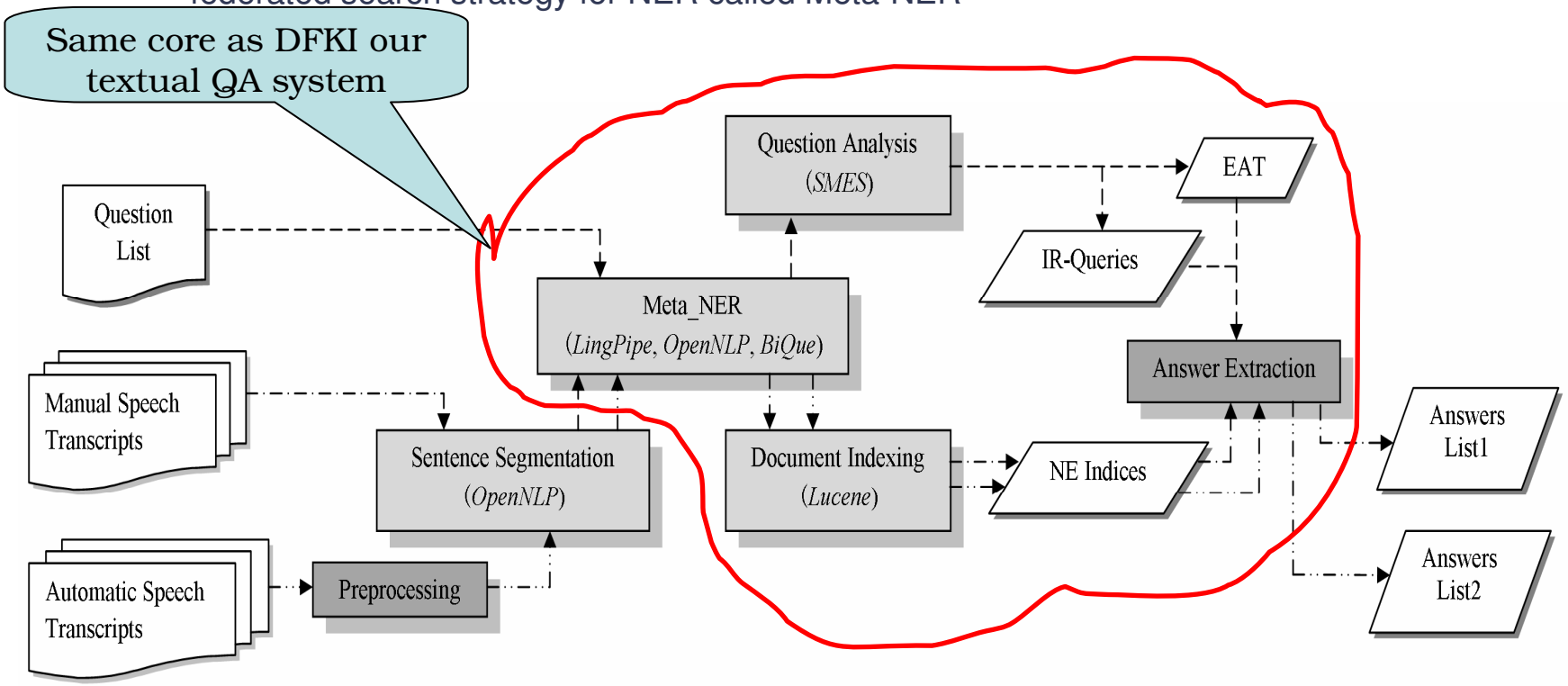
- CLT, Center for Language Technology, Australia ;
- DFKI, Germany ;
- LIMSI-CNRS, Laboratoire d'Informatique et de Mécanique des Sciences de l'Ingénieur, France ;
- Tokyo Institute of Technology, Japan ;
- UPC, Universitat Politècnica de Catalunya, Spain.

☆ In total, 28 submission files were evaluated:

CHIL Corpus (lectures)		AMI Corpus (meetings)	
T1 (manual)	T2 (ASR)	T3 (manual)	T4 (ASR)
8 submissions	9 submissions	5 submissions	6 submissions

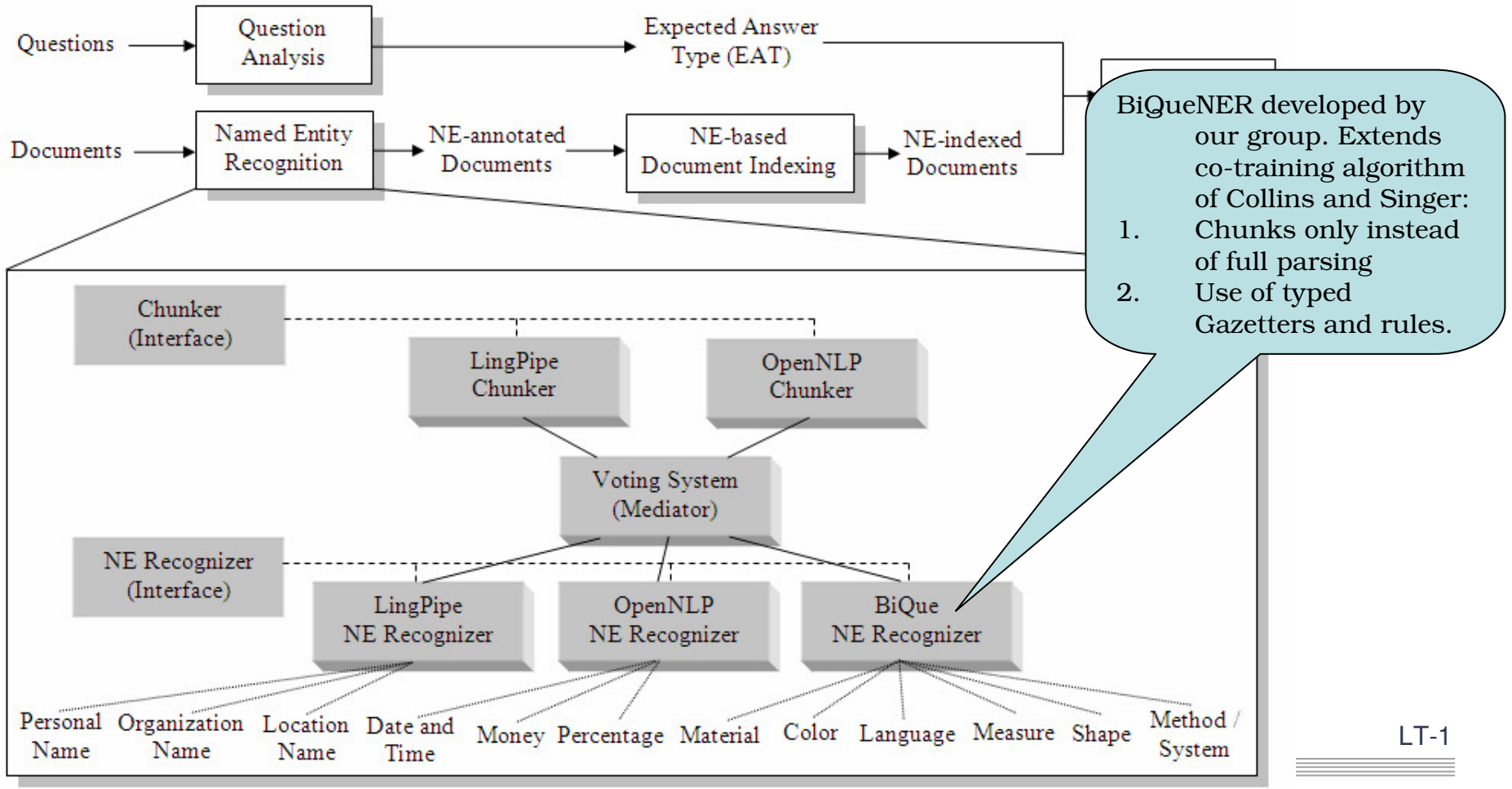
DFKI at QAST pilot task

- 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



Results for T1

☆ QA on CHIL manual transcriptions:

System	# Questions Returned	# Correct Answers	MRR	Accuracy
clt1_t1	98	16	0.09	0.06
clt2_t1	98	16	0.09	0.05
dfki1_t1	98	19	0.17	0.15
limsi1_t1	98	43	0.37	0.32
limsi2_t1	98	**56**	0.46	0.39
tokyo1_t1	98	32	0.19	0.14
tokyo2_t1	98	34	0.20	0.14
upc1_t1	**98**	54	**0.53**	**0.51**

Results for T2

☆ QA on CHIL automatic transcriptions:

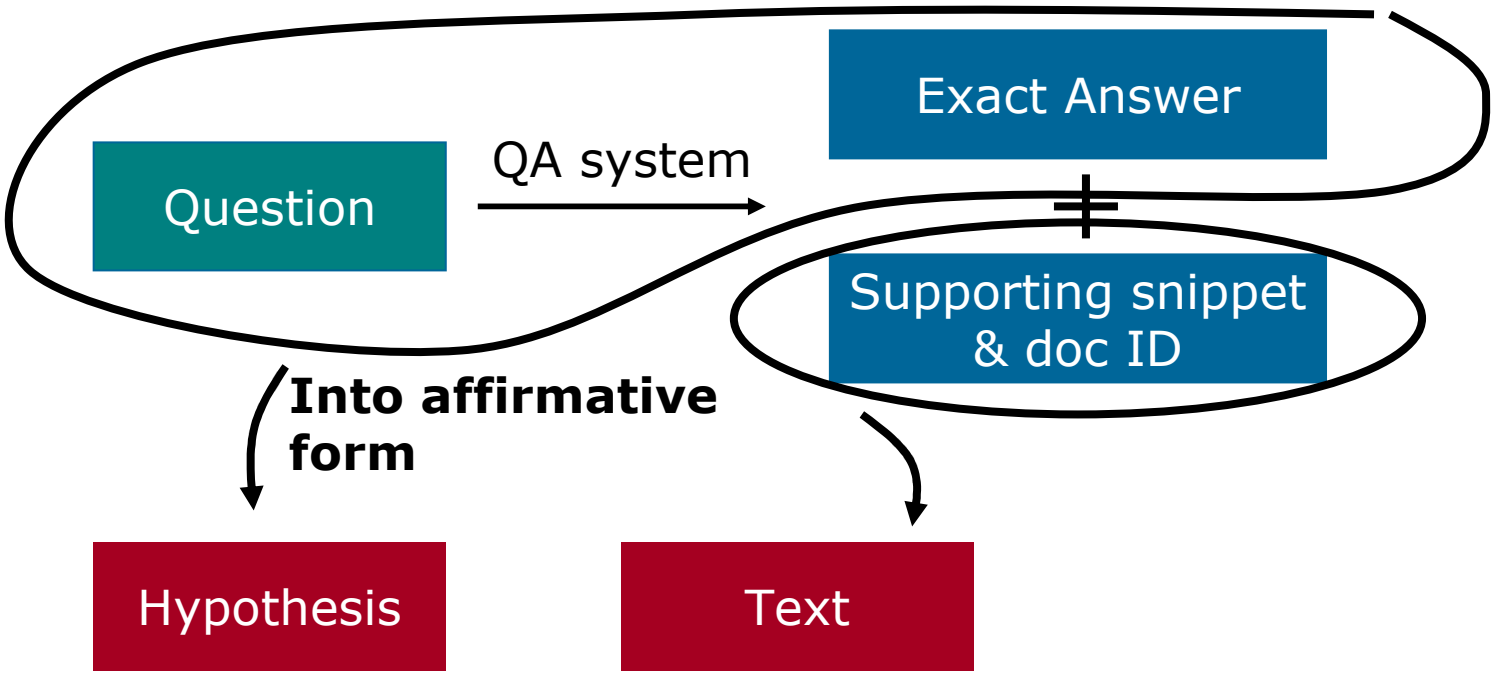
System	# Questions Returned	# Correct Answers	MRR	Accuracy
clt1_t2	98	13	0.06	0.03
clt2_t2	98	12	0.05	0.02
dfki1_t2	98	9	0.09	0.09
limsi1_t2	98	28	0.23	0.20
limsi2_t2	98	28	0.24	0.21
tokyo1_t2	98	17	0.12	0.08
tokyo2_t2	98	18	0.12	0.08
upc1_t2	**96**	**37**	**0.37**	**0.36**
upc2_t2	97	29	0.25	0.24

Validate the correctness of the answers…

... given by the participants

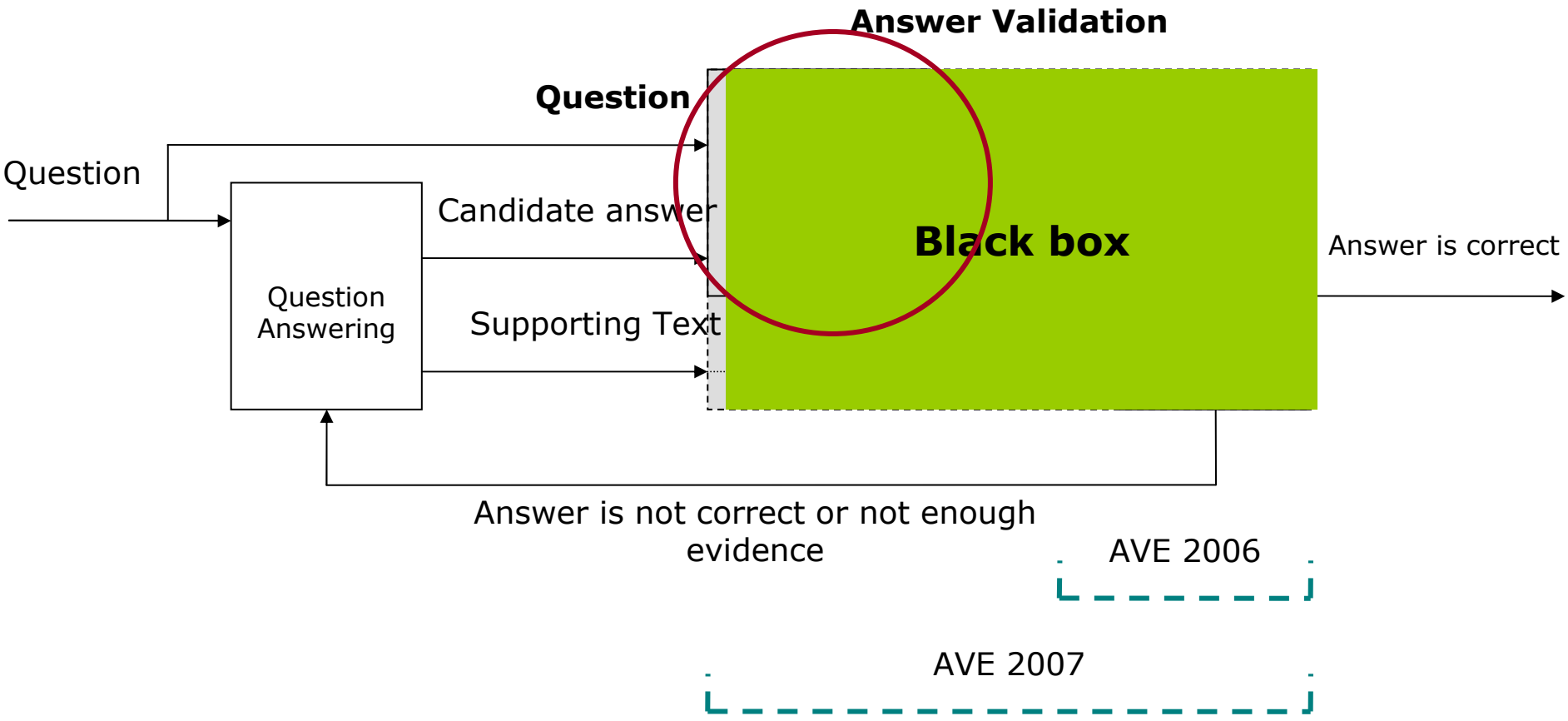
at CLEF QA 2007

AVE 2006: an RTE exercise



If the text semantically entails the hypothesis, then the answer is expected to be correct.

Answer Validation Exercise



- ☆ AVE 2006 😐
  - Not possible to quantify the potential gain that AV modules give to QA systems
- ☆ Change in AVE 2007 methodology
  - Group answers by question
  - Systems must validate all
  - But select one

```
<q id="116" lang="EN">
    <q_str>What is Zanussi?</q_str>
    <a id="116_1" value="">
        <a_str>was an Italian producer of home appliances</a_str>
        <t_str doc="Zanussi">Zanussi For the Polish film director, see
        Krzysztof Zanussi. For the hot-air balloon, see Zanussi (balloon). Zanussi
        was an Italian producer of home appliances that in 1984 was
        bought</t_str>
    </a>
    <a id="116_2" value="">
        <a_str>who had also been in Cassibile since August 31</a_str>
        <t_str doc="en/p29/2998260.xml">Only after the signing had taken
        place was Giuseppe Castellano informed of the additional clauses that had
        been presented by general Ronald Campbell to another Italian general,
        Zanussi, who had also been in Cassibile since August 31.</t_str>
    </a>
    <a id="116_4" value="">
        <a_str>3</a_str>
        <t_str doc="1618911.xml">(1985) 3 Out of 5 Live (1985)      What Is
        This?</t_str>
    </a>
</q>
```

Collections

- ☆ Remove duplicated answers inside the same question group
- ☆ Discard NIL answers, void answers and answers with too long supporting snippet
- ☆ This processing lead to a reduction in the number of answers to be validated

Collections (# answers to validate)

	Testing	Development
English	202	1121
Spanish	564	1817
German	282	504
French	187	1503
Italian	103	476
Dutch	202	528
Portuguese	367	817
Bulgarian	-	70
Romanian	127	-

Available for CLEF participants at **nlp.uned.es/QA/ave/**

Evaluation

- ☆ Not balanced collections
- ☆ Approach: Detect if there is enough evidence to accept an answer
- ☆ Measures: Precision, recall and F over ACCEPTED answers
- ☆ Baseline system: Accept all answers

Evaluation

Precision, Recall and F measure over correct answers for **English**

Group	System	F	Precision	Recall
DFKI	ltqa_2	0.55	0.44	0.71
DFKI	ltqa_1	0.46	0.37	0.62
U. Alicante	ofe_1	0.39	0.25	0.81
Text-Mess Project	Text-Mess_1	0.36	0.25	0.62
Iasi	adiftene	0.34	0.21	0.81
UNED	rodrigo	0.34	0.22	0.71
Text-Mess Project	Text-Mess_2	0.34	0.25	0.52
U. Alicante	ofe_2	0.29	0.18	0.81
100% VALIDATED		0.19	0.11	1
50% VALIDATED		0.18	0.11	0.5

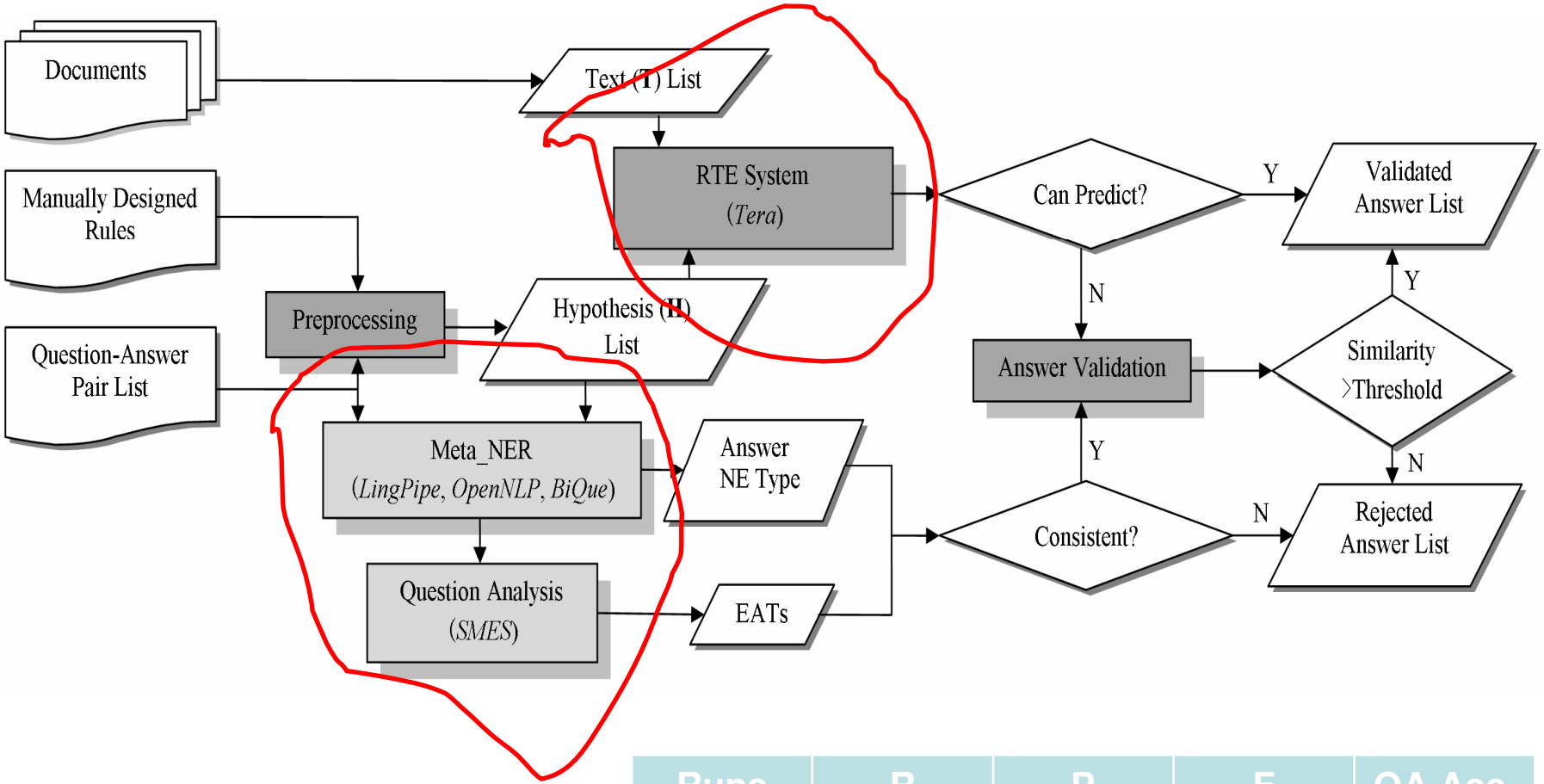
DFKI's AVE System

- ☆ 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

☆ Details about our core RTE method

- System Called TERA
- Implemented and evaluated by Rui Wang as part of his Master Thesis
- References
  - R. Wang and G. Neumann
    Recognizing Textual Entailment Using a Subsequence Kernel Method.
    AAAI-2007, Vancoucer.
  - R. Wang and G. Neumann
    Recognizing Textual Entailment Using Sentence Similarity based on Dependency Tree Skeletons
    Workshop proceedings of the RTE-3 challenge, 2007, Association for Computational Linguistics.

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!