Cross Language Evaluation Forum 2004

Bath, September 16, 2004

Overview of the Multilingual Question Answering Track

Alessandro Vallin
ITC-irst, Trento - Italy

Joint work with Bernardo Magnini, Christelle Ayache, Gregor Erbach,
Anselmo Peñas, Maarten de Rijke, Paulo Rocha, Kiril Simov and Richard Sutcliffe

Outline

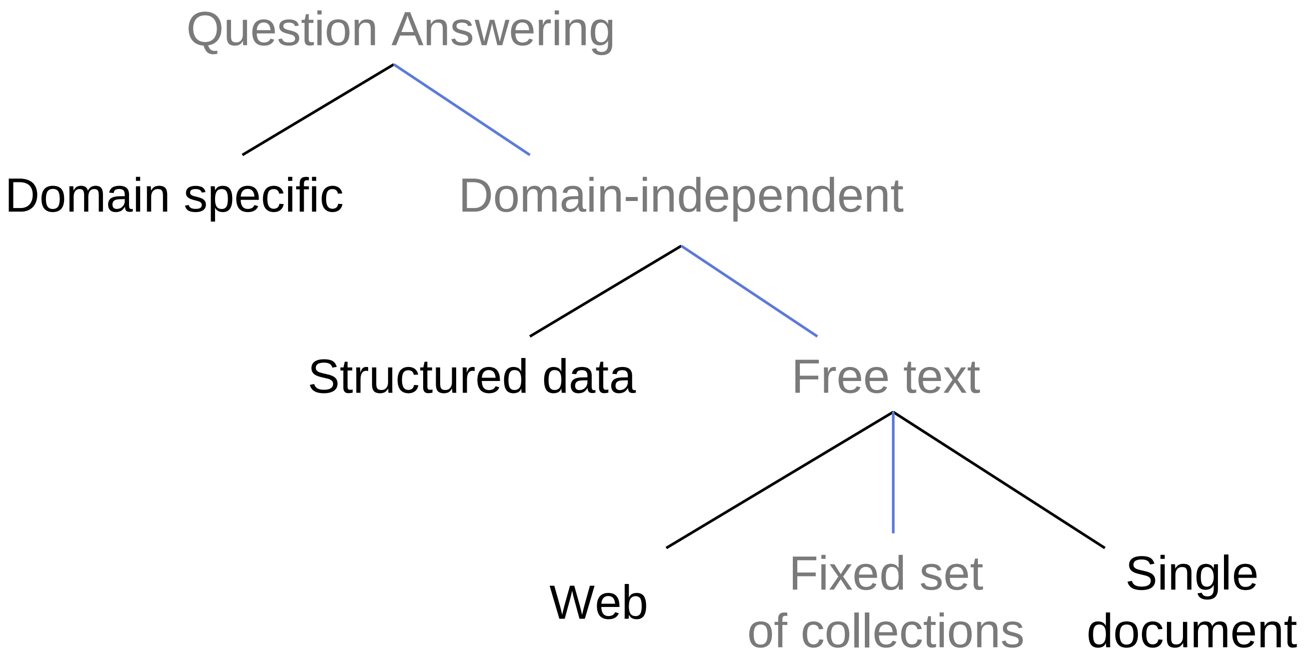
- Introduction

- QA Track Setup
  - Overview
  - Task Definition
  - Questions and Answers
  - Assessment

- Evaluation Exercise
  - Participants
  - Results
  - Approaches

- Conclusions
  - Future Directions

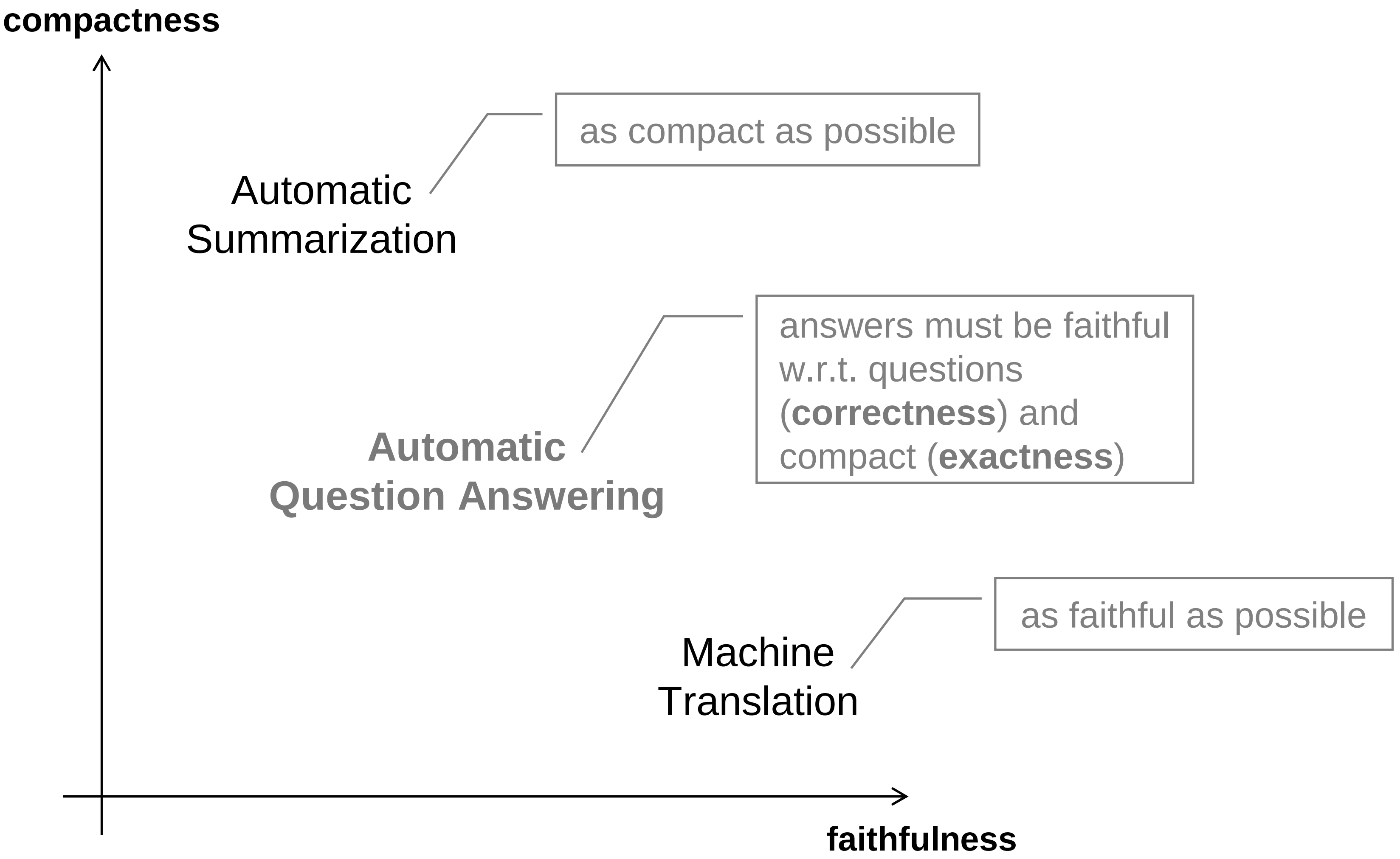
Introduction



Growing interest in QA (3 QA-related tasks at CLEF 2004).

Focus on **multilinguality** (Burger et al., 2001):
- monolingual tasks in other languages than English
- cross-language tasks

Introduction



QA @ CLEF 2004 (http://clef-qa.itc.it/2004)

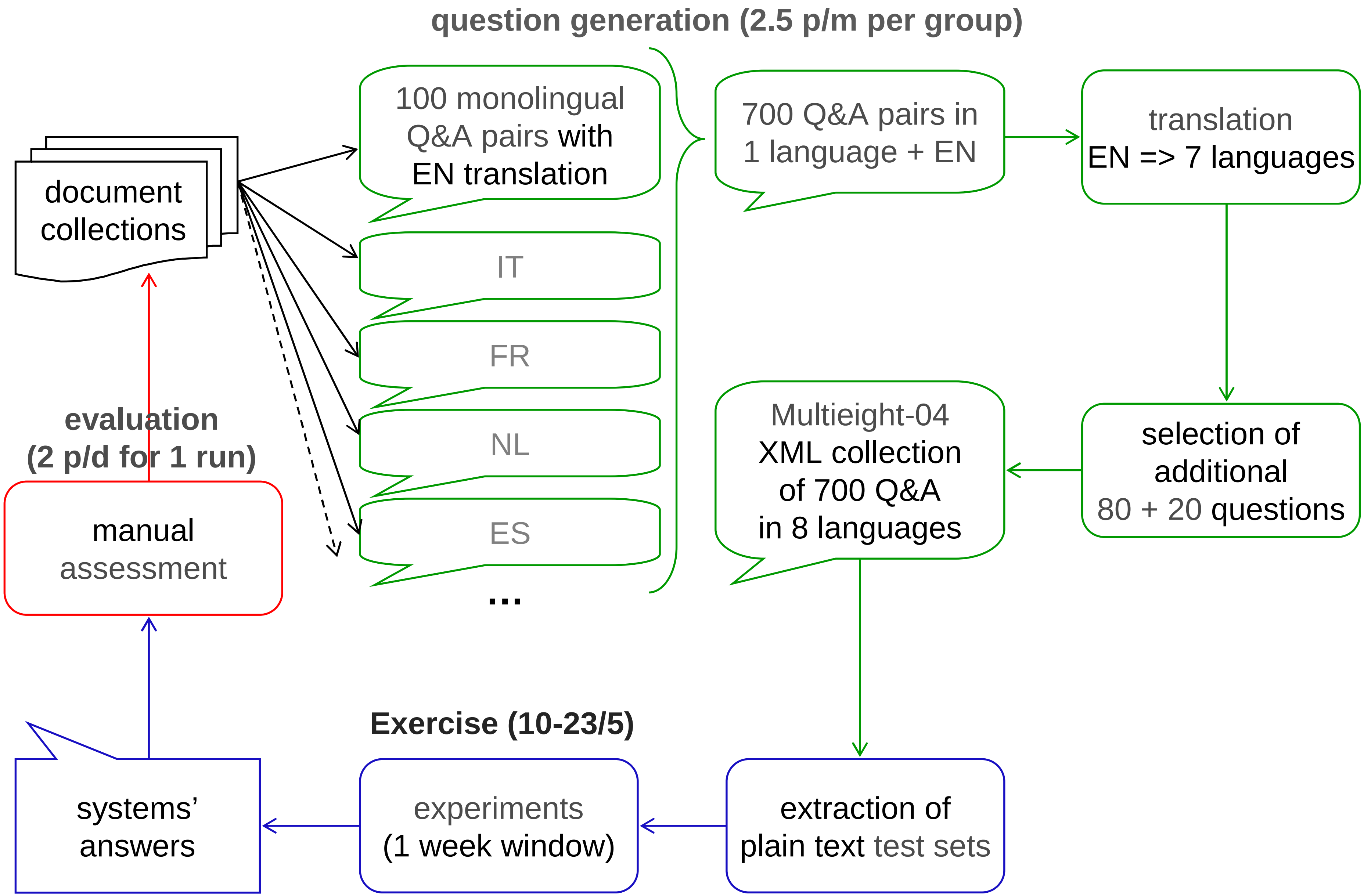
Seven groups coordinated the QA track:

- ITC-irst (IT and EN test set preparation)
- DFKI (DE)
- ELDA/ELRA (FR)
- Linguateca (PT)
- UNED (ES)
- U. Amsterdam (NL)
- U. Limerick (EN assessment)

Two more groups participated in the test set construction:

- Bulgarian Academy of Sciences (BG)
- U. Helsinki (FI)

QA Track Setup - Overview



question generation (2.5 p/m per group)

document collections

100 monolingual Q&A pairs with EN translation

IT

FR

NL

ES

...

700 Q&A pairs in 1 language + EN

translation EN => 7 languages

selection of additional 80 + 20 questions

Multieight-04 XML collection of 700 Q&A in 8 languages

extraction of plain text test sets

Exercise (10-23/5)

experiments (1 week window)

systems' answers

evaluation (2 p/d for 1 run)

manual assessment

QA Track Setup – Task Definition

Given **200 questions** in a source language, find **one exact answer** per question in a collection of documents written in a target language, and provide a justification for each retrieved answer (i.e. the **`docid`** of the unique document that supports the answer).

S \ T	DE	EN	ES	FR	IT	NL	PT
BG		🟩		🟩			
DE	🟥	🟩		🟩			
EN		⬛		🟩		🟩	
ES			🟥	🟩			
FI	⬛	🟩	⬛	⬛	⬛	⬛	⬛
FR		🟩		🟥			
IT		🟩		🟩	🟥		
NL				🟩		🟥	
PT				🟩			🟥

6 monolingual and 50 bilingual tasks.

Teams participated in 19 tasks,

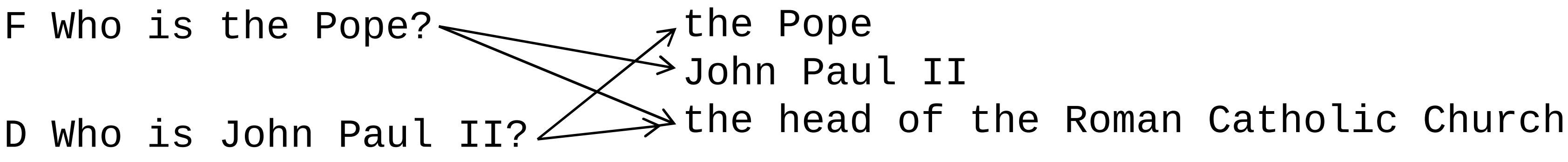
QA Track Setup - Questions

All the test sets were made up of 200 questions:

- ~90% **factoid** questions
- ~10% **definition** questions
- ~10% of the questions did not have any answer in the corpora (right answer-string was “NIL”)

Problems in introducing definition questions:

- What’s the right answer? (it depends on the user’s model)
- What’s the easiest and more efficient way to assess their answers?
- Overlap with factoid questions:



F Who is the Pope?

D Who is John Paul II?

the Pope

John Paul II

the head of the Roman Catholic Church

QA Track Setup - Answers

One exact answer per question was required.

Exactness subsumes **conciseness** (the answer should not contain extraneous or irrelevant information) and **completeness** (the answer should be complete).

✓ Exactness is relatively easy to assess for factoids:

```
F Where is Heathrow airport?
R F London

F What racing team is Flavio Briatore the manager of?
X F I am happy for this thing in particular said Benetton Flavio Briatore
```

✓ Definition questions and How- questions could have a passage as answer (depending on the user's model).

```
D Who is Jorge Amado?
R D Bahian novelist

R D American authors such as Stephen King and Sidney Sheldon are perennial
best sellers in Latin American countries, while Brazilian Jorge Amado,
Colombian Gabriel Garcia Marquez and Mexican Carlos Fuentes are renowned in
U.S. literary circles.
```

QA Track Setup – Multieight

```
<q cnt="0675" category="F" answer_type="MANNER">
     <language val="BG" original="FALSE">
          <question group="BTB">Как умира Пазолини?</question>
          <answer n="1" docid="">TRANSLATION[убит]</answer>
     </language>
     <language val="DE" original="FALSE">
          <question group="DFKI">Auf welche Art starb Pasolini?</question>
          <answer n="1" docid="">TRANSLATION[ermordet]</answer>
          <answer n="2" docid="SDA.951005.0154">ermordet</answer>
     </language>
     <language val="EN" original="FALSE">
          <question group="LING">How did Pasolini die?</question>
          <answer n="1" docid="">TRANSLATION[murdered]</answer>
          <answer n="2" docid="LA112794-0003">murdered</answer>
     </language>
     <language val="ES" original="FALSE">
          <question group="UNED">¿Cómo murió Pasolini?</question>
          <answer n="1" docid="">TRANSLATION[Asesinado]</answer>
          <answer n="2" docid="EFE19950724-14869">Brutalmente asesinado en los arrabales de Ostia</answer>
     </language>
     <language val="FR" original="FALSE">
          <question group="ELDA">Comment est mort Pasolini ?</question>
          <answer n="1" docid="">TRANSLATION[assassiné]</answer>
          <answer n="2" docid="ATS.951101.0082">assassiné</answer>
          <answer n="3" docid="ATS.950904.0066">assassiné en novembre 1975 dans des circonstances mystérieuses</answer>
          <answer n="4" docid="ATS.951031.0099">assassiné il y a 20 ans</answer>
     </language>
     <language val="IT" original="FALSE">
          <question group="IRST">Come è morto Pasolini?</question>
          <answer n="1" docid="">TRANSLATION[assassinato]</answer>
          <answer n="2" docid="AGZ.951102.0145">massacrato e abbandonato sulla spiaggia di Ostia</answer>
     </language>
     <language val="NL" original="FALSE">
          <question group="UoA">Hoe stierf Pasolini?</question>
          <answer n="1" docid="">TRANSLATION[vermoord]</answer>
          <answer n="2" docid="NH19951102-0080">vermoord</answer>
     </language>
     <language val="PT" original="TRUE">
          <question group="LING">Como morreu Pasolini?</question>
          <answer n="1" docid="LING-951120-088">assassinado</answer>
     </language>
</q>
```

QA Track Setup - Assessment

Judgments taken from the TREC QA tracks:

- Right
- Wrong
- ineXact
- Unsupported

Other criteria, such as the length of the answer-strings (instead of X, which is underspecified) or the usefulness of responses for a potential user, have not been considered.

Main evaluation measure was **accuracy** (fraction of Right responses).

Whenever possible, a **Confidence-Weighted Score** was calculated:

$$\text{CWS} = \frac{1}{Q} \sum_{i=1}^{Q} \frac{\text{number of correct responses in first i ranks}}{i}$$

Evaluation Exercise - Participants

Distribution of participating groups in different QA evaluation campaigns.

	America	Europe	Asia	Australia	TOTAL	submitted runs
TREC-8	13	3	3	1	20	46
TREC-9	14	7	6	-	27	75
TREC-10	19	8	8	-	35	67
TREC-11	16	10	6	-	32	67
TREC-12	13	8	4	-	25	54
NTCIR-3 (QAC-1)	1	-	15	-	16	36
CLEF 2003	3	5	-	-	8	17
CLEF 2004	**1**	**17**	**-**	**-**	**18**	**48**

Evaluation Exercise - Participants

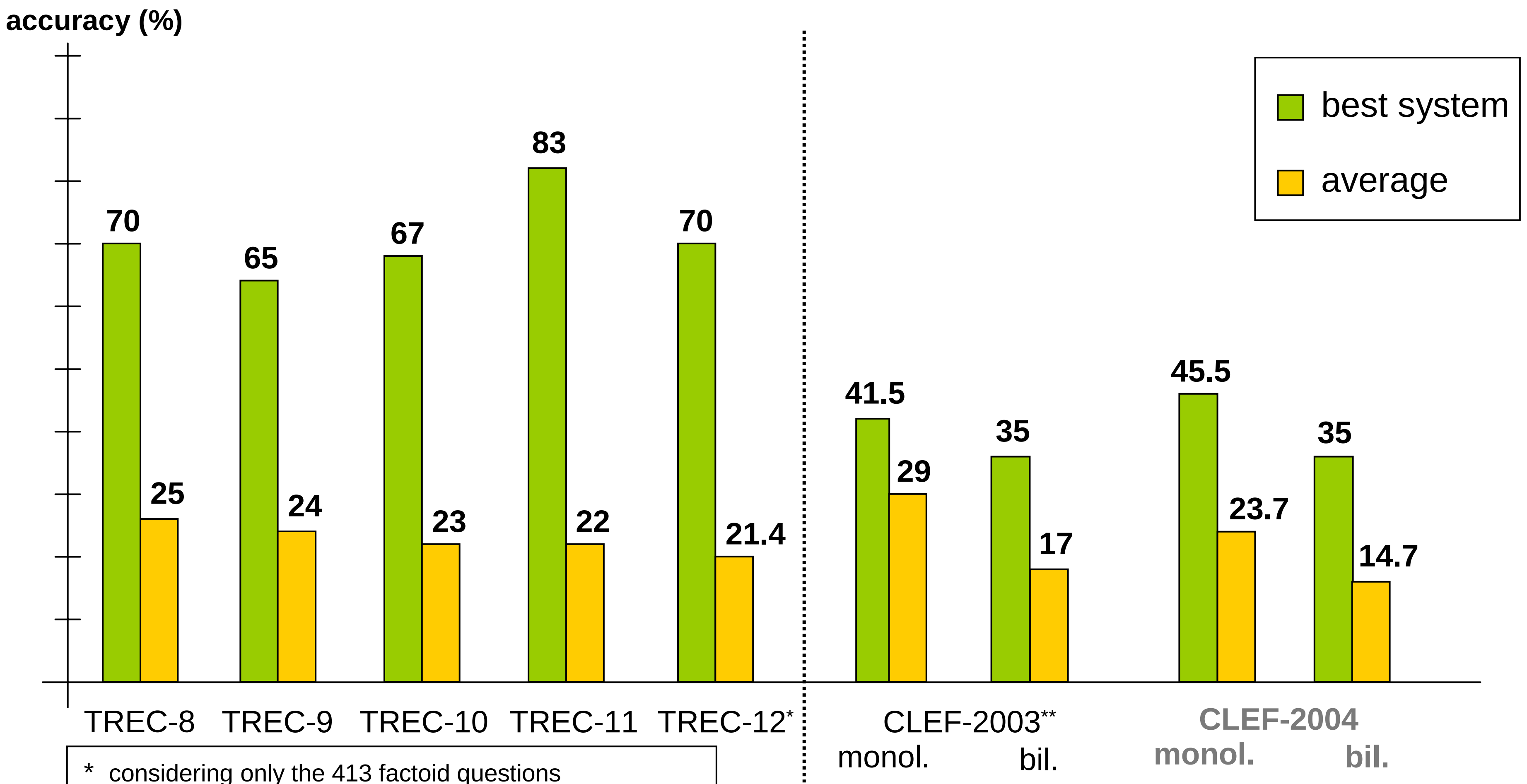
Number of participating teams-number of submitted runs at CLEF 2004.

Comparability issue.

T S	DE	EN	ES	FR	IT	NL	PT
BG		1-1		1-2			
DE	2−2	2-3		1-2			
EN				1-2		1-1	
ES			5-8	1-2			
FI		1-1					
FR		3-6		1-2			
IT		1-2		1-2	2-3		
NL				1-2		1-2	
PT				1-2			2-3

Evaluation Exercise - Results

Systems' performance at the TREC and CLEF QA tracks.



* considering only the 413 factoid questions

** considering only the answers returned at the first rank

Evaluation Exercise – Results (DE)

Results of the runs with German as target language.

Run Name	R	W	X	U	Overall Accuracy (%)	Accuracy over F (%)	Accuracy over D (%)	NIL Accuracy		CWS
								Precision	Recall	
dfki041dede	50	143	1	3	25.38	28.25	0.00	0.14	0.85	-
FUHA041-dede	67	128	2	0	**34.01**	31.64	55.00	0.14	1.00	0.333

Evaluation Exercise – Results (EN)

Results of the runs with English as target language.

Run Name	R	W	X	U	Overall Accuracy (%)	Accuracy over F (%)	Accuracy over D (%)	NIL Accuracy		CWS
								Precision	Recall	
bgas041bgen	26	168	5	1	13.00	11.67	25.00	0.13	0.40	0.056
dfki041deen	47	151	0	2	**23.50**	23.89	20.00	0.10	0.75	0.177
dltg041fren	38	155	7	0	19.00	17.78	30.00	0.17	0.55	-
dltg042fren	29	164	7	0	14.50	12.78	30.00	0.14	0.45	-
edin041deen	28	166	5	1	14.00	13.33	20.00	0.14	0.35	0.049
edin041fren	33	161	6	0	16.50	17.78	5.00	0.15	0.55	0.056
edin042deen	34	159	7	0	17.00	16.11	25.00	0.14	0.35	0.052
edin042fren	40	153	7	0	**20.00**	20.56	15.00	0.15	0.55	0.058
hels041fien	21	171	1	0	10.88	11.56	5.00	0.10	0.85	0.046
irst041iten	45	146	6	3	**22.50**	22.22	25.00	0.24	0.30	0.121
irst042iten	35	158	5	2	17.50	16.67	25.00	0.24	0.30	0.075
lire041fren	22	172	6	0	11.00	10.00	20.00	0.05	0.05	0.032
lire042fren	39	155	6	0	19.50	20.00	15.00	0.00	0.00	0.075

Evaluation Exercise – Results (ES)

Results of the runs with Spanish as target language.

Run Name	R	W	X	U	Overall Accuracy (%)	Accuracy over F (%)	Accuracy over D (%)	NIL Accuracy		CWS
								Precision	Recall	
aliv041eses	63	130	5	2	31.50	30.56	40.00	0.17	0.35	0.121
aliv042eses	65	129	4	2	**32.50**	31.11	45.00	0.17	0.35	0.144
cole041eses	22	178	0	0	11.00	11.67	5.00	0.10	1.00	-
inao041eses	45	145	5	5	22.50	19.44	50.00	0.19	0.50	-
inao042eses	37	152	6	5	18.50	17.78	25.00	0.21	0.50	-
mira041eses	18	174	7	1	9.00	10.00	0.00	0.14	0.55	-
talp041eses	48	150	1	1	24.00	18.89	70.00	0.19	0.50	0.087
talp042eses	52	143	3	2	26.00	21.11	70.00	0.20	0.55	0.102

Evaluation Exercise – Results (FR)

Results of the runs with French as target language.

Run Name	R	W	X	U	Overall Accuracy (%)	Accuracy over F (%)	Accuracy over D (%)	NIL Accuracy		CWS
								Precision	Recall	
gine041bgfr	13	182	5	0	6.50	6.67	5.00	0.10	0.50	0.051
gine041defr	29	161	10	0	14.50	14.44	15.00	0.15	0.20	0.079
gine041enfr	18	170	12	0	9.00	8.89	10.00	0.05	0.10	0.033
gine041esfr	27	165	8	0	13.50	14.44	5.00	0.12	0.15	0.056
gine041frfr	27	160	13	0	13.50	13.89	10.00	0.00	0.00	0.048
gine041itfr	25	165	10	0	12.50	13.33	5.00	0.15	0.30	0.049
gine041nlfr	20	169	11	0	10.00	10.00	10.00	0.12	0.20	0.044
gine041ptfr	25	169	6	0	12.50	12.22	15.00	0.11	0.15	0.044
gine042bgfr	13	180	7	0	6.50	6.11	10.00	0.10	0.35	0.038
gine042defr	34	154	12	0	17.00	15.56	30.00	0.23	0.20	0.097
gine042enfr	27	164	9	0	13.50	12.22	25.00	0.06	0.10	0.051
gine042esfr	34	162	4	0	17.00	17.22	15.00	0.11	0.10	0.075
gine042frfr	49	145	6	0	**24.50**	23.89	30.00	0.09	0.05	0.114
gine042itfr	29	164	7	0	14.50	15.56	5.00	0.14	0.30	0.054
gine042nlfr	29	156	15	0	14.50	13.33	25.00	0.14	0.20	0.065
gine042ptfr	29	164	7	0	14.50	13.33	25.00	0.10	0.15	0.056

Evaluation Exercise – Results (IT)

Results of the runs with Italian as target language.

Run Name	R	W	X	U	Overall Accuracy (%)	Accuracy over F (%)	Accuracy over D (%)	NIL Accuracy		CWS
								Precision	Recall	
ILCP-QA-ITIT	51	117	29	3	25.50	22.78	50.00	0.62	0.50	-
irst041itit	56	131	11	2	**28.00**	26.67	40.00	0.27	0.30	0.155
irst042itit	44	147	9	0	22.00	20.00	40.00	0.66	0.20	0.107

Evaluation Exercise – Results (NL)

Results of the runs with Dutch as target language.

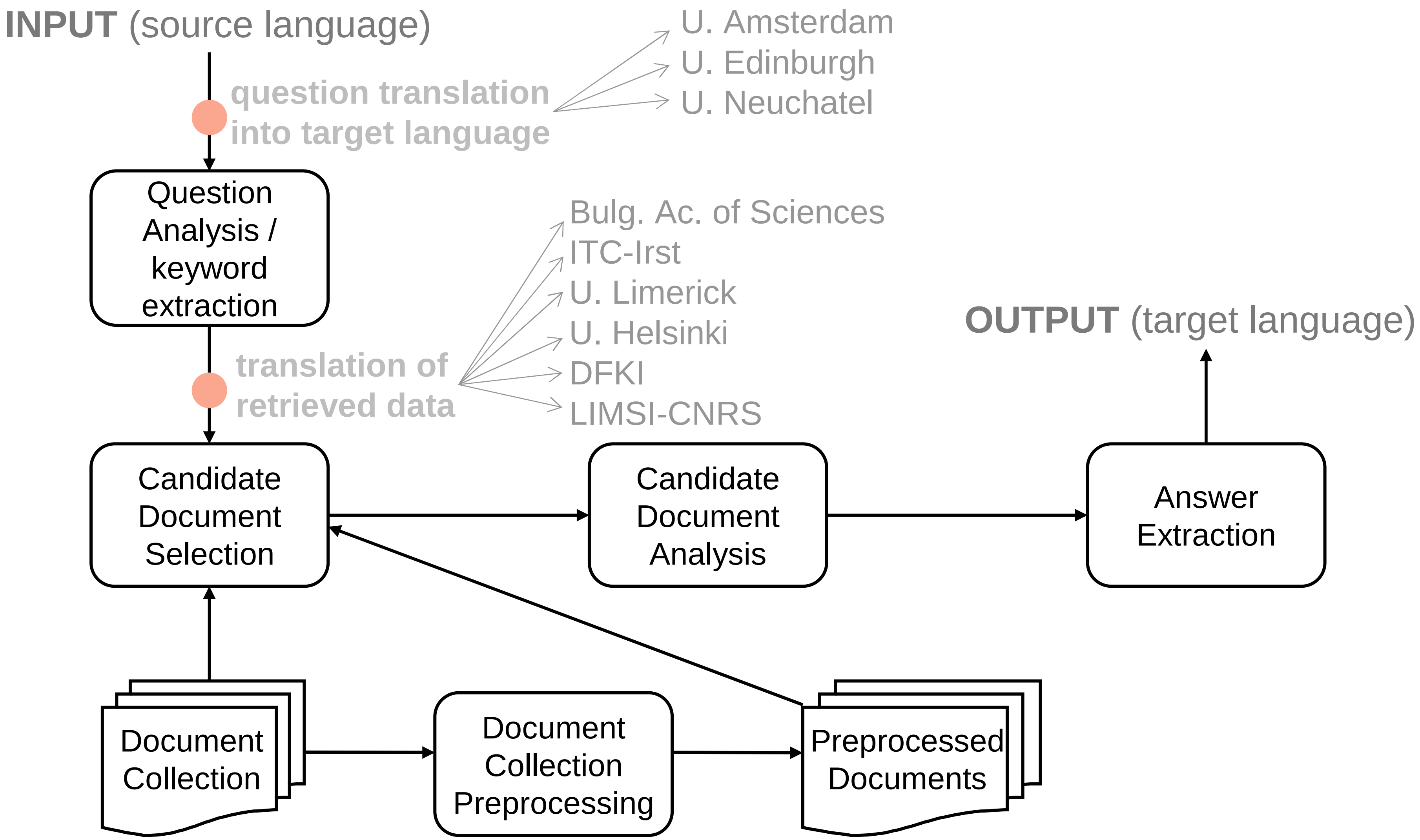
Run Name	R	W	X	U	Overall Accuracy (%)	Accuracy over F (%)	Accuracy over D (%)	NIL Accuracy		CWS
								Precision	Recall	
uams041ennl	70	122	7	1	35.00	31.07	65.22	0.00	0.00	-
uams041nlnl	88	98	10	4	44.00	42.37	56.52	0.00	0.00	-
uams042nlnl	91	97	10	2	**45.50**	45.20	47.83	0.56	0.25	-

Evaluation Exercise – Results (PT)

Results of the runs with Portuguese as target language.

Run Name	R	W	X	U	Overall Accuracy (%)	Accuracy over F (%)	Accuracy over D (%)	NIL Accuracy		CWS
								Precision	Recall	
PTUE041ptpt	57	125	18	0	**28.64**	29.17	25.81	0.14	0.90	-
sfnx041ptpt	22	166	8	4	11.06	11.90	6.45	0.13	0.75	-
sfnx042ptpt	30	155	10	5	15.08	16.07	9.68	0.16	0.55	-

Evaluation Exercise – CL Approaches



INPUT (source language)

question translation into target language

U. Amsterdam
U. Edinburgh
U. Neuchatel

Question Analysis / keyword extraction

translation of retrieved data

Bulg. Ac. of Sciences
ITC-Irst
U. Limerick
U. Helsinki
DFKI
LIMSI-CNRS

OUTPUT (target language)

Candidate Document Selection

Candidate Document Analysis

Answer Extraction

Document Collection

Document Collection Preprocessing

Preprocessed Documents

Conclusions

CLEF multilingual QA track (like TREC QA) represents a formal evaluation, designed with an eye to replicability. As an exercise, it is an **abstraction** of the real problems.

Future challenges:

- investigate QA in combination with other applications (for instance summarization)
- access not only free text, but also different sources of data (multimedia, spoken language, imagery)
- introduce automated evaluation along with judgments given by humans
- focus on user's need: develop real-time interactive systems, which means modeling a potential user and defining suitable answer types.

Conclusions

Possible improvements of the CLEF QA track:

Questions

- closer interaction with IR track (use topics)
- individuate common areas of interest in the collections (minimize translation efforts)
- more realistic factoid questions (focus on real newspapers readers, introduce yes/no questions)

Answers

- accept short snippets in response to definition and How- questions

Evaluation

- introduce new judgments (consider usefulness)
- introduce new evaluation measures (study measures that reward self scoring, introduce automated evaluation)