



Open Domain Question Answering

Bogdan Sacaleanu

(based on slides from Bernardo Magnini, RANLP 2005)

1



Outline of the Tutorial

- I. **Introduction to QA**
- II. **QA at TREC**
- III. **System Architecture**
 - **Question Processing**
 - **Answer Extraction**
- IV. **Cross-Language QA**

2

I. Introduction to Question Answering

- What is Question Answering
- Applications
- Users
- Question Types
- Answer Types
- Evaluation
- Presentation
- Brief history

3

Query Driven vs Answer Driven Information Access

- What does LASER stand for?
- When did Hitler attack Soviet Union?
 - Using Google we find documents containing the question itself, no matter whether or not the answer is actually provided.
- Current information access is **query driven**.
- Question Answering proposes an **answer driven** approach to information access.

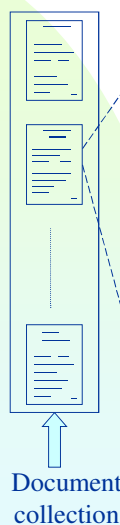
4

Question Answering

- Find the answer to a question in a large collection of documents
 - ◆ **questions** (in place of keyword-based query)
 - ◆ **answers** (in place of documents)

5

Why Question Answering?



From the Caledonian Star in the Mediterranean – September 23, 1990 (www.expeditions.com):

On a beautiful early morning the Caledonian Star approaches Naxos, situated on the east coast of Sicily. As we anchored and put the Zodiacs into the sea we enjoyed the great scenery. Under Mount Etna, the highest volcano in Europe, perches the fabulous town of Taormina. This is the goal for our morning. After a short Zodiac ride we embarked our buses with local guides and went up into the hills to reach the town of Taormina. Naxos was the first Greek settlement at Sicily. Soon a harbor was established but the town was later destroyed by invaders.[...]

Searching for: Taormina

6

Alternatives to Information Retrieval

- Document Retrieval
 - ◆ users submit queries corresponding to their information need
 - ◆ system returns (voluminous) list of full-length documents
 - ◆ it is the responsibility of the users to find their original information need, within the returned documents
- Open-Domain Question Answering (QA)
 - ◆ users ask fact-based, natural language questions

What is the highest volcano in Europe?
 - ◆ system returns list of short answers

... Under Mount Etna, the highest volcano in Europe, perches the fabulous town ...
 - ◆ more appropriate for specific information needs

7

What is QA?

- Find the answer to a question in a large collection of documents
- What is the *brightest star* visible from *Earth*?

 1. *Sirio* A is the *brightest star visible from Earth* even if it is...
 2. the planet is 12-times brighter than *Sirio*, the *brightest star* in the sky...

8

QA: a Complex Problem (1)

- **Problem: discovery implicit relations among question and answers**

*Who is the **author** of the “**Star Spangled Banner**”?*

*...**Francis Scott Key** wrote the “**Star Spangled Banner**” in 1814.*

*...comedian-actress **Roseanne Barr** sang her famous rendition of the “**Star Spangled Banner**” before ...*

9

QA: a Complex Problem (2)

- **Problem: discovery implicit relations among question and answers**

*Which is the Mozart **birth date**?*

*.... Mozart (**1751** – 1791)*

10

QA: a complex problem (3)

- **Problem: discovery implicit relations among question and answers**

Which is the distance between Naples and Ravello?

*“From the **Naples** Airport follow the sign to Autostrade (green road sign). Follow the directions to Salerno (A3). Drive for about 6 Km. Pay toll (Euros 1.20). Drive appx. 25 Km. Leave the Autostrade at Angri (Uscita Angri). Turn left, follow the sign to Ravello through Angri. Drive for about 2 Km. Turn right following the road sign "Costiera Amalfitana". Within 100m you come to traffic lights prior to narrow bridge. Watch not to miss the next Ravello sign, at appx. 1 Km from the traffic lights. Now relax and enjoy the views (follow this road for 22 Km). Once in **Ravello** ...”.*

11

QA: Applications (1)

- Information access:
 - ◆ Structured data (databases)
 - ◆ Semi-structured data (e.g. comment field in databases, XML)
 - ◆ Free text
- To search over:
 - ◆ The **Web**
 - ◆ Fixed set of **text collection** (e.g. TREC)
 - ◆ A **single text** (reading comprehension evaluation)

12

QA: Applications (2)

- Domain independent QA
- Domain specific (e.g. help systems)
- Multi-modal QA
 - ◆ Annotated images
 - ◆ Speech data

13

QA: Questions (1)

- Classification according to the **answer type**
 - ◆ **Factual questions** (*What is the larger city ...*)
 - ◆ **Opinions** (*What is the author attitude ...*)
 - ◆ **Summaries** (*What are the arguments for and against...*)
- Classification according to the **question speech act**:
 - ◆ **Yes/NO questions** (*Is it true that ...*)
 - ◆ **WH questions** (*Who was the first president ...*)
 - ◆ **Indirect Requests** (*I would like you to list ...*)
 - ◆ **Commands** (*Name all the presidents ...*)

14

QA: Questions (2)

- Difficult questions
 - ◆ **Why, How questions** require understanding causality or instrumental relations
 - ◆ **What questions** have little constraint on the answer type (e.g. *What did they do?*)

15

QA: Answers

- **Long answers**, with justification
- **Short answers** (e.g. phrases)
- **Exact answers** (named entities)
- Answer construction:
 - ◆ **Extraction**: cut and paste of snippets from the original document(s)
 - ◆ **Generation**: from multiple sentences or documents
 - ◆ QA and **summarization** (e.g. *What is this story about?*)

16

QA: Information Presentation

- **Interfaces for QA**
 - ◆ Not just isolated questions, but a dialogue
 - ◆ Usability and user satisfaction
- **Critical situations**
 - ◆ Real time, single answer
- **Dialog-based interaction**
 - ◆ Speech input
 - ◆ Conversational access to the Web

17

QA: Brief History (1)

- **NLP interfaces to databases:**
 - ◆ BASEBALL (1961), LUNAR (1973), TEAM (1979), ALFRESCO (1992)
 - ◆ Limitations: structured knowledge and limited domain
- **Story comprehension:** Shank (1977), Kintsch (1998), Hirschman (1999)

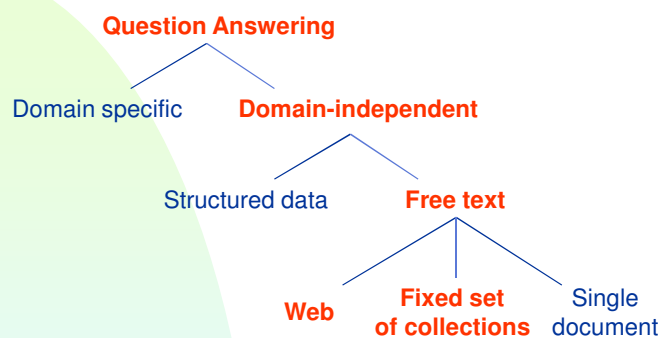
18

QA: Brief History (2)

- **Information retrieval (IR)**
 - ◆ Queries are questions
 - ◆ List of documents are answers
 - ◆ QA is close to passage retrieval
 - ◆ Well established methodologies (i.e. Text Retrieval Conferences TREC)
- **Information extraction (IE):**
 - ◆ Pre-defined templates are questions
 - ◆ Filled template are answers

19

Research Context (1)

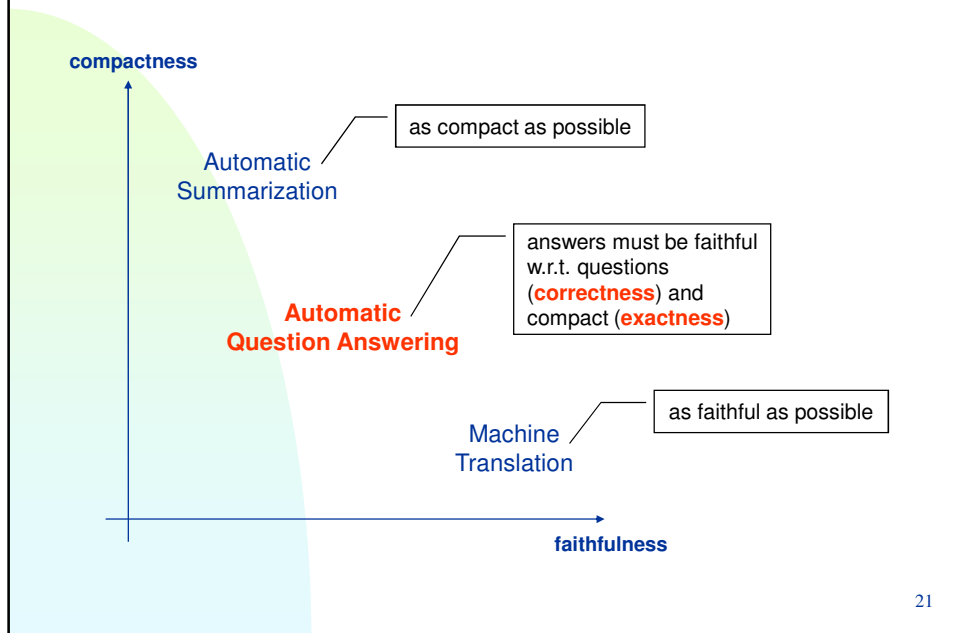


Growing interest in QA (TREC, CLEF, NT evaluation campaign).

Recent focus on **multilinguality** and **context aware QA**

20

Research Context (2)



II. Question Answering at TREC

- The problem simplified
- Questions and answers
- Evaluation metrics
- Approaches

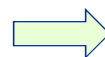
The problem simplified: The Text Retrieval Conference

- **Goal**
 - ◆ Encourage research in information retrieval based on large-scale collections
- **Sponsors**
 - ◆ NIST: National Institute of Standards and Technology
 - ◆ ARDA: Advanced Research and Development Activity
 - ◆ DARPA: Defense Advanced Research Projects Agency
- Since 1999
- Participants are research institutes, universities, industries

23

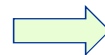
TREC Questions

Q-1391: How many feet in a mile?
Q-1057: Where is the volcano Mauna Loa?
Q-1071: When was the first stamp issued?
Q-1079: Who is the Prime Minister of Canada?
Q-1268: Name a food high in zinc.



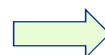
Fact-based,
short answer
questions

Q-896: Who was Galileo?
Q-897: What is an atom?



Definition
questions

Q-711: What tourist attractions are there in Reims?
Q-712: What do most tourists visit in Reims?
Q-713: What attracts tourists in Reims
Q-714: What are tourist attractions in Reims?



Reformulation
questions

24

Answer Assessment

■ Criteria for judging an answer

- ➡ ◆ **Relevance**: it should be responsive to the question
- ➡ ◆ **Correctness**: it should be factually correct
- ➡ ◆ **Conciseness**: it should not contain extraneous or irrelevant information
- ◆ **Completeness**: it should be complete, i.e. partial answer should not get full credit
- ◆ **Simplicity**: it should be simple, so that the questioner can read it easily
- ➡ ◆ **Justification**: it should be supplied with sufficient context to allow a reader to determine why this was chosen as an answer to the question

25

Questions at TREC

	Yes/ No	Entity	Definition	Opinion/ Procedure/ Explanation
Single answer	Is Berlin the capital of Germany?	What is the largest city in Germany?	Who was Galileo?	
Multiple answer		Name 9 countries that import Cuban sugar		What are the arguments for and against prayer in school?

26

Exact Answers

- Basic unit of a response: [answer-string, docid] pair
- An answer string must contain a complete, exact answer and nothing else.

What is the longest river in the United States?

The following are **correct, exact answers**

Mississippi,
the Mississippi,
the Mississippi River,
Mississippi River
mississippi

while none of the following are correct exact answers

At 2,348 miles the Mississippi River is the longest river in the US.
2,348 miles; Mississippi
Missipp
Missouri

27

Assessments

- Four possible judgments for a triple
[Question, document, answer]
- **Rigth**: the answer is appropriate for the question
- **Inexact**: used for non complete answers
- **Unsupported**: answers without justification
- **Wrong**: the answer is not appropriate for the question

28

What is the capital city of New Zealand?	R 1530 XIE19990325.0298 Wellington
What is the Boston Strangler's name?	R 1490 NYT20000913.0267 Albert DeSalvo
What is the world's second largest island?	R 1503 XIE19991018.0249 New Guinea
What year did Wilt Chamberlain score 100 points?	U 1402 NYT19981017.0283 1962
Who is the governor of Tennessee?	R 1426 NYT19981030.0149 Sundquist
What's the name of King Arthur's sword?	U 1506 NYT19980618.0245 Excalibur
When did Einstein die?	R 1601 NYT19990315.0374 April 18 , 1955
What was the name of the plane that dropped the Atomic Bomb on Hiroshima?	X 1848 NYT19991001.0143 Enola
What was the name of FDR's dog?	R 1838 NYT20000412.0164 Fala
What day did Neil Armstrong land on the moon?	R 1674 APW19990717.0042 July 20 , 1969
Who was the first Triple Crown Winner?	X 1716 NYT19980605.0423 Barton
When was Lyndon B. Johnson born?	R 1473 APW19990826.0055 1908
Who was Woodrow Wilson's First Lady?	R 1622 NYT19980903.0086 Ellen
Where is Anne Frank's diary?	W 1510 NYT19980909.0338 Young Girl

R=Right, X=ineXact, U=Unsupported, W=Wrong

29

1402: What year did Wilt Chamberlain score 100 points?

DIogene: 1962

ASSESSMENT: UNSUPPORTED

PARAGRAPH: NYT19981017.0283

Petty's 200 victories, 172 of which came during a 13-year span between 1962-75, may be as unapproachable as Joe DiMaggio's 56-game hitting streak or Wilt Chamberlain's 100-point game.

30

1506: What's the name of King Arthur's sword?

ANSWER: Excalibur

PARAGRAPH: NYT19980618.0245

ASSESSMENT: UNSUPPORTED

'QUEST FOR CAMELOT,' with the voices of Andrea Carr, Gabriel Byrne, Cary Elwes, John Gielgud, Jessalyn Gilsig, Eric Idle, Gary Oldman, Bronson Pinchot, Don Rickles and Bryan White. Directed by Frederik Du Chau (G, 100 minutes). Warner Brothers' shaky entrance into the Disney-dominated sweepstakes of the musicalized animated feature wants to be a juvenile feminist ``Lion King" with a musical heart that fuses ``Riverdance" with formulaic Hollywood gush. But its characters are too wishy-washy and visually unfocused to be compelling, and the songs (by David Foster and Carole Bayer Sager) so forgettable as to be extraneous. In this variation on the Arthurian legend, a nondescript Celtic farm girl named Kayley with aspirations to be a knight wrests the magic sword **Excalibur** from the evil would-be emperor Ruber (a Hulk Hogan look-alike) and saves the kingdom (Holden).

31

1848: What was the name of the plane that dropped the Atomic Bomb on Hiroshima?

DIOGENE: Enola

PARAGRAPH: NYT19991001.0143

ASSESSMENT: INEXACT

Tibbets piloted the Boeing B-29 Superfortress **Enola Gay**, which dropped the atomic bomb on Hiroshima on Aug. 6, 1945, causing an estimated 66,000 to 240,000 deaths. He named the plane after his mother, **Enola Gay** Tibbets.

32

1716: Who was the first Triple Crown Winner?

DIOGENE: Barton

PARAGRAPH: NYT19980605.0423

ASSESSMENT: INEXACT

Not all of the Triple Crown winners were immortals. The first, Sir Barton, lost six races in 1918 before his first victory, just as Real Quiet lost six in a row last year. Try to find Omaha and Whirlaway on anybody's list of all-time greats.

33

1510: Where is Anne Frank's diary?

DIOGENE: Young Girl

PARAGRAPH: NYT19980909.0338

ASSESSMENT: WRONG

Otto Frank released a heavily edited version of “B” for its first publication as “Anne Frank: Diary of a Young Girl” in 1947.

34

TREC Evaluation Metric: Mean Reciprocal Rank (MRR)

- **Reciprocal Rank** = inverse of rank at which first correct answer was found:
[1, 0.5, 0.33, 0.25, 0.2, 0]
- **MRR**: average over all questions
- **Strict score**: unsupported count as incorrect
- **Lenient score**: unsupported count as correct

35

TREC Evaluation Metrics: Confidence-Weighted Score (CWS)

$$\frac{\text{Sum for } i = 1 \text{ to } 500 (\# \text{-correct-up-to-question } i / i)}{500}$$

System A:

$$\begin{array}{l} 1 \rightarrow C \\ 2 \rightarrow W \\ 3 \rightarrow C \\ 4 \rightarrow C \\ 5 \rightarrow W \end{array} \quad \begin{array}{l} (1/1) + ((1+0)/2) + (1+0+1)/3 + ((1+0+1+1)/4) + ((1+0+1+1+0)/5) \\ \hline 5 \end{array}$$

Total: 0.7

System B:

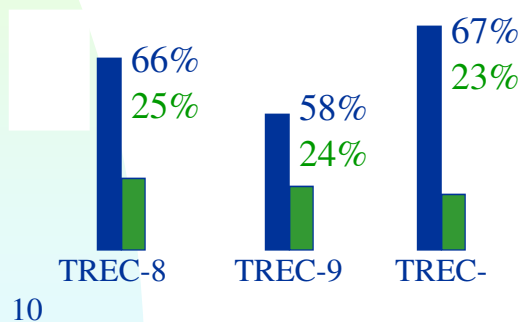
$$\begin{array}{l} 1 \rightarrow W \\ 2 \rightarrow W \\ 3 \rightarrow C \\ 4 \rightarrow C \\ 5 \rightarrow C \end{array} \quad \begin{array}{l} 0 + ((0+0)/2) + (0+0+1)/3 + ((0+0+1+1)/4) + ((0+0+1+1+1)/5) \\ \hline 5 \end{array}$$

Total: 0.29

36

Evaluation

- Best result: 67%
- Average over 67 runs: 23%



37

Main Approaches at TREC

- Knowledge-Based
- Web-based
- Pattern-based

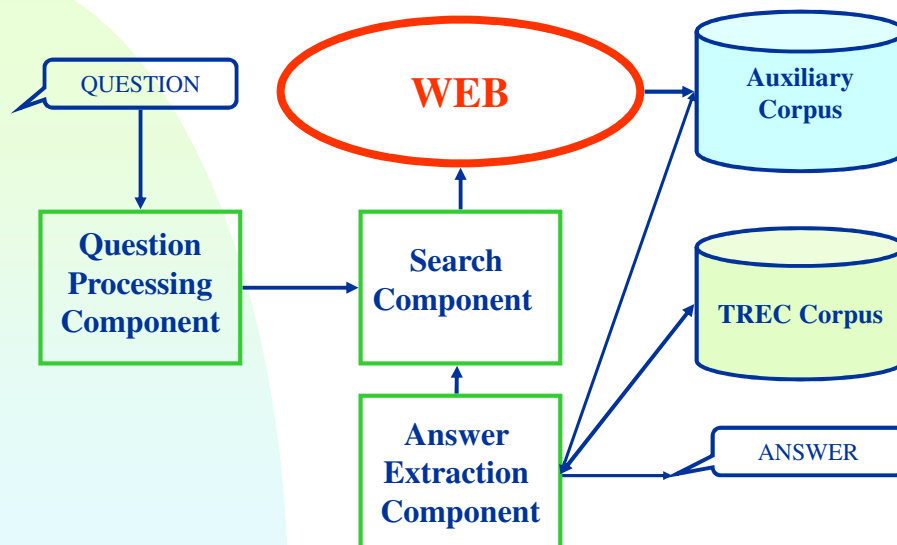
38

Knowledge-Based Approach

- **Linguistic-oriented** methodology
 - ◆ Determine the answer type from question form
 - ◆ Retrieve small portions of documents
 - ◆ Find entities matching the answer type category in text snippets
- Majority of systems use a lexicon (usually **WordNet**)
 - ◆ To find answer type
 - ◆ To verify that a candidate answer is of the correct type
 - ◆ To get definitions
- Complex architecture...

39

Web-Based Approach



40

Pattern-Based Approach (1/3)

- Knowledge poor
- Strategy
 - ◆ **Search for predefined patterns** of textual expressions that may be interpreted as answers to certain question types.
 - ◆ The presence of such patterns in answer string candidates may provide evidence of the right answer.

41

Pattern-Based Approach (2/3)

- Conditions
 - ◆ Detailed categorization of **question types**
 - ↳ Up to 9 types of the “Who” question; 35 categories in total
 - ◆ Significant number of **patterns** corresponding to each question type
 - ↳ Up to 23 patterns for the “Who-Author” type, average of 15
 - ◆ Find multiple **candidate snippets** and check for the presence of patterns (emphasis on recall)

42

Pattern-based approach (3/3)

- Example: patterns for definition questions
- Question: **What is A?**

1. <A; is/are; [a/an/the]; X> ...23 correct answers
2. <A; comma; [a/an/the]; X; [comma/period]> ...26 correct answers
3. <A; [comma]; or; X; [comma]> ...12 correct answers
4. <A; dash; X; [dash]> ...9 correct answers
5. <A; parenthesis; X; parenthesis> ...8 correct answers
6. <A; comma; [also] called; X [comma]> ...7 correct answers
7. <A; is called; X> ...3 correct answers

total: 88 correct answers

43

Use of answer patterns

1. **For generating queries to the search engine.**

How did Mahatma Gandhi die?

Mahatma Gandhi die <HOW>

Mahatma Gandhi die of <HOW>

Mahatma Gandhi lost his life in <WHAT>

The TEXTMAP system (ISI) uses 550 patterns, grouped in 105 equivalence blocks. On TREC-2003 questions, the system produced, on average, 5 reformulations for each question.

2. **For answer extraction**

When was Mozart born?

P=1 <PERSON> (<BIRTHDATE> - DATE)

P=.69 <PERSON> was born on <BIRTHDATE>

44

Acquisition of Answer Patterns

Relevant approaches:

- ◆ Manually developed surface pattern library (Soubotin, Soubotin, 2001)
- ◆ Automatically extracted surface patterns (Ravichandran, Hovy 2002)

Pattern learning:

1. Start with a seed, e.g. (Mozart, 1756)
2. Download Web documents using a search engine
3. Retain sentences that contain both question and answer terms
4. Construct a suffix tree for extracting the longest matching substring that spans <Question> and <Answer>
5. Calculate precision of patterns
$$\text{Precision} = \# \text{ of correct patterns with correct answer} / \# \text{ of total patterns}$$

45

Capturing variability with patterns

- Pattern based QA is more effective when supported by **variable typing** obtained using NLP techniques and resources.

When was <A> born?

<A:PERSON> (<ANSWER:DATE> -

<A :PERSON > was born in <ANSWER :DATE >

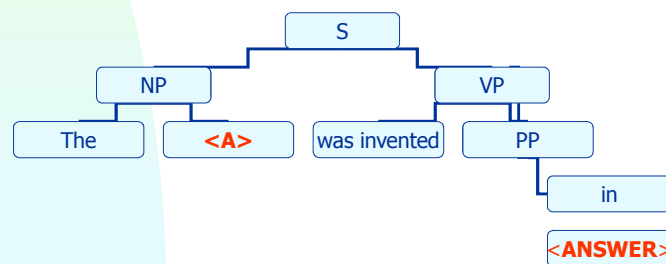
- Surface patterns can not deal with word reordering and apposition phrases:
Galileo, the famous astronomer, was born in ...
- The fact that most of the QA systems use syntactic parsing demonstrates that the successful solution of the answer extraction problem goes beyond the surface form analysis

46

Syntactic answer patterns (1)

Answer patterns that capture the syntactic relations of a sentence.

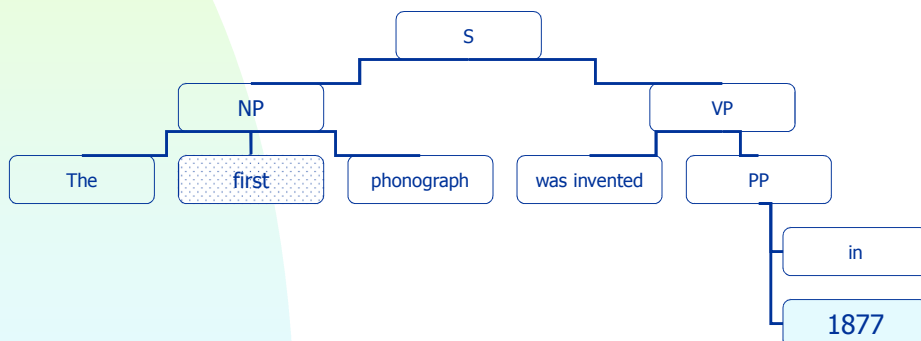
When was <A> invented?



47

Syntactic answer patterns (2)

The matching phase turns out to be a problem of partial match among syntactic trees.

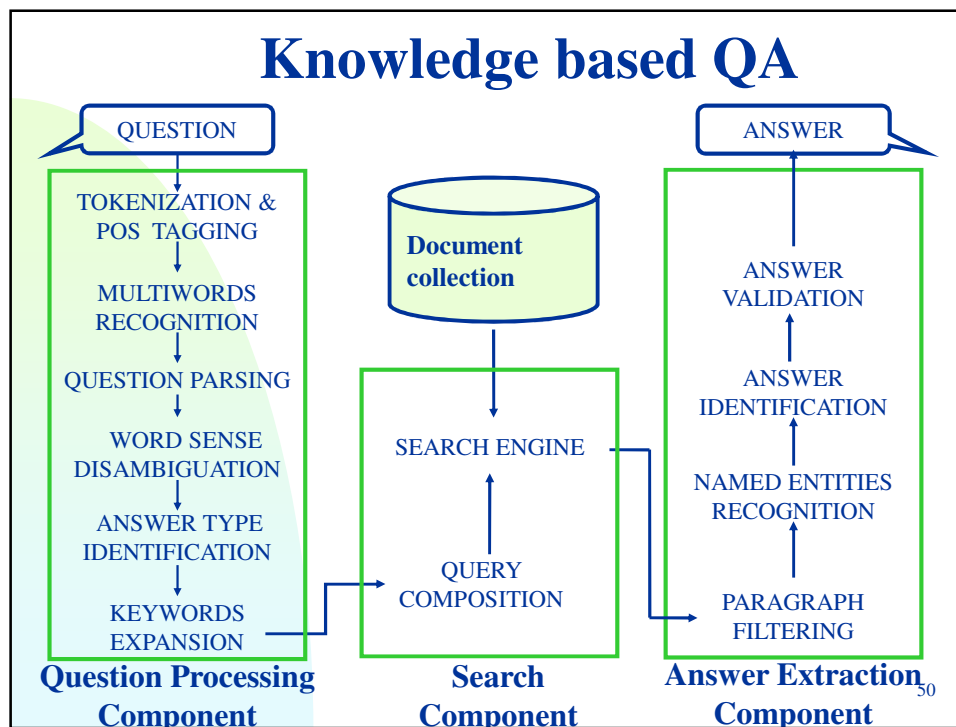


48

III. System Architecture

- **Knowledge Based** approach
 - ◆ Question Processing
 - ◆ Search component
 - ◆ Answer Extraction

49



Question Analysis (1)

- **Input:** NLP question
- **Output:**
 - ◆ query for the search engine (i.e. a boolean composition of weighted keywords)
 - ◆ Answer type
 - ◆ Additional constraints: question focus, syntactic or semantic relations that should hold for a candidate answer entity and other entities

51

Question Analysis (2)

- **Steps:**
 1. Tokenization
 2. POS-tagging
 3. Multi-words recognition
 4. Parsing
 5. Answer type and focus identification
 6. Keyword extraction
 7. Word Sense Disambiguation
 8. Expansions

52

Tokenization and POS-tagging

NL-QUESTION: *Who was the inventor of the electric light?*

Who	Who	CCHI	[0,0]
was	be	VIY	[1,1]
the	det	RS	[2,2]
inventor	inventor	SS	[3,3]
of	of	ES	[4,4]
the	det	RS	[5,5]
electric	electric	AS	[6,6]
light	light	SS	[7,7]
?	?	XPS	[8,8]

53

Multi-Words recognition

NL-QUESTION: *Who was the inventor of the electric light?*

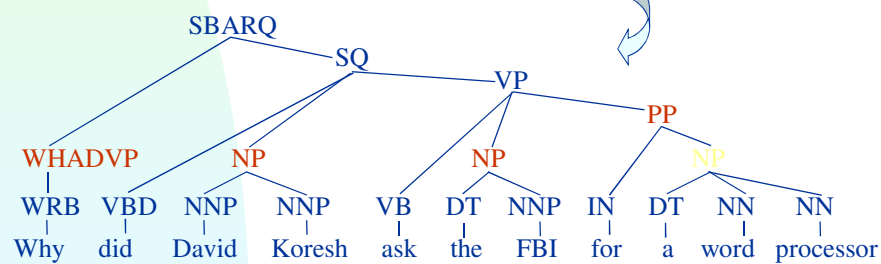
Who	Who	CCHI	[0,0]
was	be	VIY	[1,1]
the	det	RS	[2,2]
inventor	inventor	SS	[3,3]
of	of	ES	[4,4]
the	det	RS	[5,5]
electric_light	electric_light	SS	[6,7]
?	?	XPS	[8,8]

54

Syntactic Parsing

- Identify syntactic structure of a sentence
 - noun phrases (NP), verb phrases (VP), prepositional phrases (PP) etc.

Why did David Koresh ask the FBI for a word processor?



55

Answer Type and Focus

- Focus** is the word that characterises the correct answer to the question
 - Used to narrow down a potential set of relevant answer candidates
 - EX: Who is the **president** of the USA?
 - EX: What is the **distance** between A and B?
- Answer Type** is the category of the entity to be searched as answer
 - PERSON, MEASURE, TIME PERIOD, DATE, ORGANIZATION, DEFINITION
 - EX: Where was **Mozart** born?
 - LOCATION

56

Answer Type and Focus

What famous communist leader died in Mexico City?

RULENAME: WHAT-WHO
TEST: ["what" [¬ NOUN]* [NOUN:person-p]_J +]
OUTPUT: ["PERSON" J]

Answer type: **PERSON**

Focus: **leader**

This rule matches any question starting with *what*, whose first noun, if any, is a person (i.e. satisfies the *person-p* predicate)

57

Keywords Extraction

NL-QUESTION: *Who was the inventor of the electric light?*

Who	Who	CCHI	[0,0]
was	be	VIY	[1,1]
the	det	RS	[2,2]
inventor	inventor	SS	[3,3]
of	of	ES	[4,4]
the	det	RS	[5,5]
electric_light	electric_light	SS	[6,7]
?	?	XPS	[8,8]

58

Word Sense Disambiguation

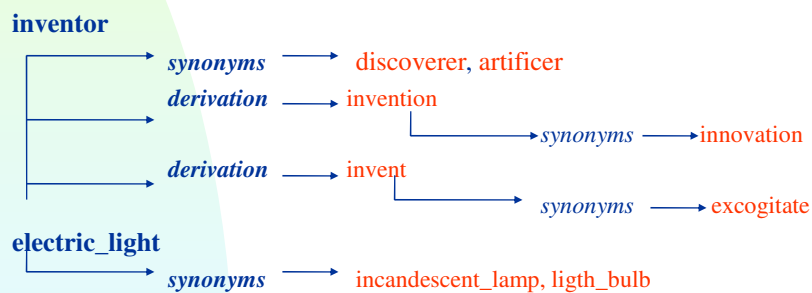
What is the brightest *star* visible from Earth?"

STAR	star#1: celestial body star#2: an actor who play ...	ASTRONOMY ART
BRIGHT	bright #1: bright brilliant shining bright #2: popular glorious bright #3: promising auspicious	PHYSICS GENERIC GENERIC
VISIBLE	visible#1: conspicuous obvious visible#2: visible seeable	PHYSICS ASTRONOMY
EARTH	earth#1: Earth world globe earth #2: estate land landed_estate acres earth #3: clay earth #4: dry_land earth solid_ground earth #5: land ground soil earth #6: earth ground	ASTRONOMY ECONOMY GEOLOGY GEOGRAPHY GEOGRAPHY GEOLOGY

59

Expansions

- NL-QUESTION: *Who was the inventor of the electric light?*
- BASIC-KEYWORDS: *inventor electric-light*



60

Keyword Composition

- Keywords and expansions are composed in a boolean expression with AND/OR operators
- Several possibilities:
 - ◆ AND composition
 - ◆ Cartesian composition

```
(OR (inventor AND electric_light)
OR (inventor AND incandescent_lamp)
OR (discoverer AND electric_light)
.....
OR inventor OR electric_light))
```

61

Document Collection Pre-processing

- For real time QA applications off-line pre-processing of the text is necessary
 - ◆ Term indexing
 - ◆ POS-tagging
 - ◆ Named Entities Recognition

62

Candidate Answer Document Selection

- **Passage Selection:** Individuate relevant, small, text portions
- Given a document and a list of keywords:
 - ◆ Paragraph length (e.g. 200 words)
 - ◆ Consider the percentage of keywords present in the passage
 - ◆ Consider if some keyword is obligatory (e.g. the focus of the question).

63

Candidate Answer Document Analysis

- Passage text tagging
- **Named Entity Recognition**

Who is the author of the “Star Spangled Banner”?

...<PERSON>Francis Scott Key </PERSON> wrote the “Star Spangled Banner” in <DATE>1814</DATE>

- Some systems:
 - ◆ passages parsing (Harabagiu, 2001)
 - ◆ Logical form (Zajac, 2001)

64

Answer Extraction (1)

- Who is the author of the “Star Spangled Banner”?

...<PERSON>Francis Scott Key </PERSON> wrote the “Star Spangled Banner” in <DATE>1814</DATE>

Answer Type = PERSON

Candidate Answer = Francis Scott Key

Ranking candidate answers: keyword density in the passage, apply additional constraints (e.g. syntax, semantics), rank candidates using the Web

65

Answer Identification



Thomas E.
Edison

66

V. Cross-Language QA

- Motivations
- [QA@CLEF](#)
- Performances
- Approaches

67

Motivations

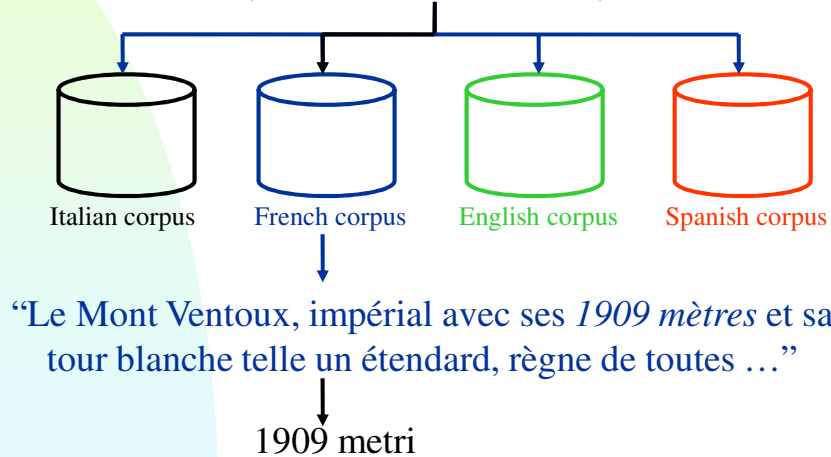
- Answers may be found in languages different from the language of the question.
- Interest in QA systems for languages other than English.
- Force the QA community to design real multilingual systems.
- Check/improve the portability of the technologies implemented in current English QA systems.

68

Cross-Language QA

Quanto è alto il Mont Ventoux?

(How tall is Mont Ventoux?)



69

CL-QA at CLEF

- Adopt the same rules used at TREC QA
 - ◆ Factoid questions (i.e. no definition questions)
 - ◆ Exact answers + document id
- Use the CLEF corpora (news, 1994 -1995)
- Return the answer in the language of the text collection in which it has been found (i.e. no translation of the answer)
- QA-CLEF-2003 was an initial step toward a more complex task organized at CLEF-2004 and 2005.

70

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

Seven groups coordinated the QA track:

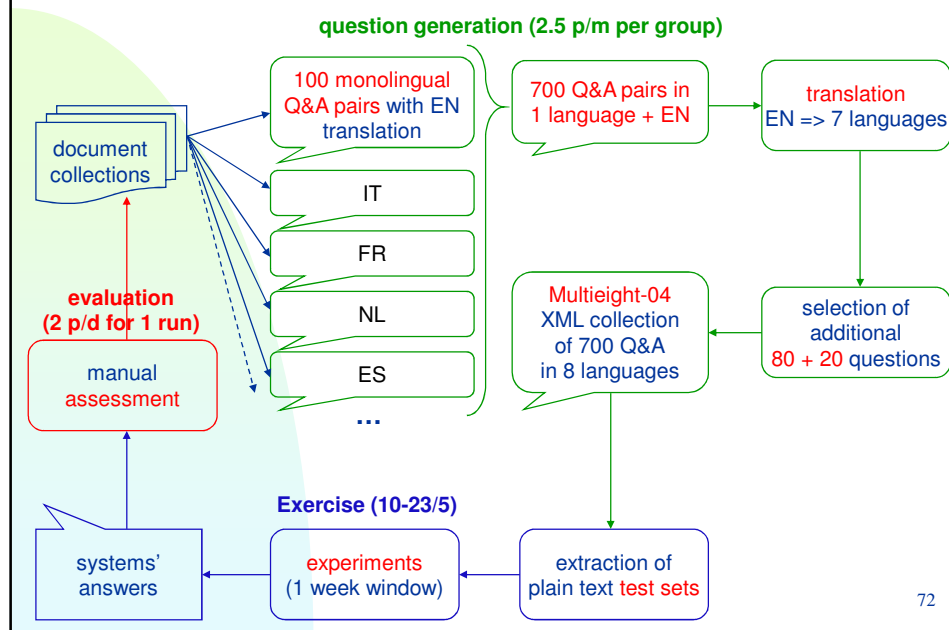
- ITC-irst (IT and EN test set preparation)
- DFKI (DE)
- ELDA/ELRA (FR)
- Linguatca (PT)
- UNED (ES)
- U. Amsterdam (NL)
- U. Limerick (EN assessment)

Two more groups participated in the test set construction:

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

71

CLEF QA - Overview



72

CLEF QA – 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,

73

CLEF QA - 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? → the Pope
 → John Paul II
 D Who is John Paul II? → the head of the Roman Catholic Church

74

CLEF QA – Multieight

```
<q cnt="0675" category="F" answer_type="MANNER">
  <language val="BG" original="FALSE">
    <question group="BTB">Как умрѣ 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-951120-089">assassinado</answer>
  </language>
</q>
```

75

CLEF QA - 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:

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

76

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

77

Evaluation Exercise - Participants

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

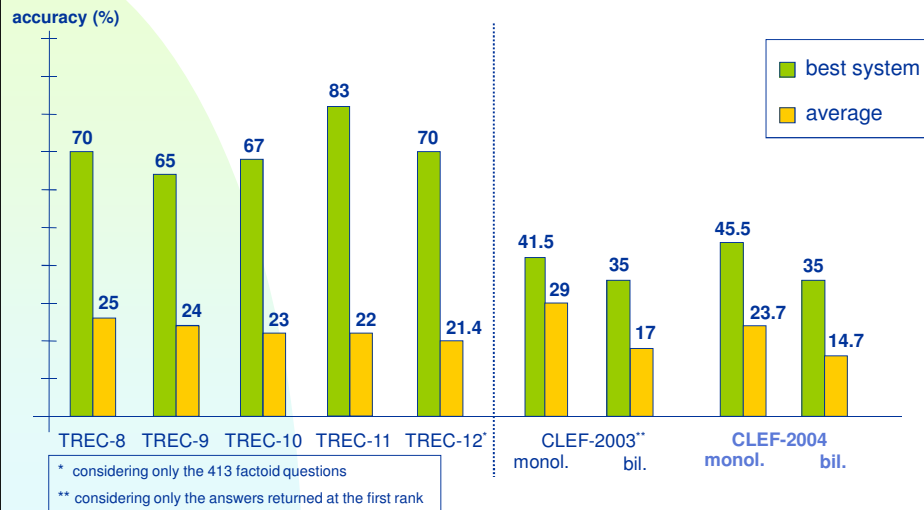
Comparability issue.

S \ T	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

78

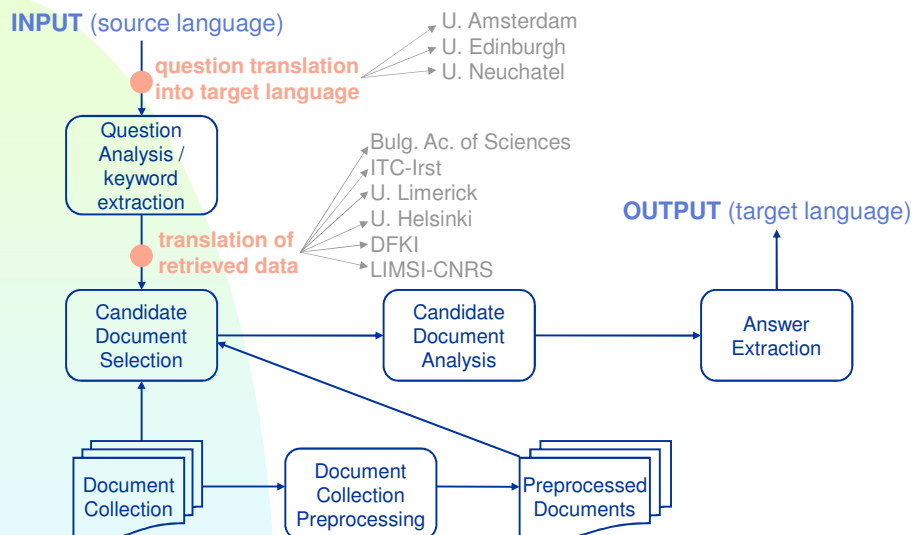
Evaluation Exercise - Results

Systems' performance at the TREC and CLEF QA tracks.



79

Evaluation Exercise – CL Approaches



80

Discussion on Cross-Language QA

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.

81

References

- **Books**
- Pasca, Marius, Open Domain Question Answering from Large Text Collections, CSLI, 2003.
- Maybury, Mark (Ed.), New Directions in Question Answering, AAAI Press, 2004.
- **Journals**
- Hirshman, Gaizauskas. *Natural Language question answering: the view from here*. JNLE, 7 (4), 2001.
- **TREC**
- E. Voorhees. *Overview of the TREC 2001 Question Answering Track*.
- M.M. Soubbotin, S.M. Soubbotin. *Patterns of Potential Answer Expressions as Clues to the Right Answers*.
- S. Harabagiu, D. Moldovan, M. Pasca, M. Surdeanu, R. Mihalcea, R. Girju, V. Rus, F. Lacatusu, P. Morarescu, R. Brunesco. *Answering Complex, List and Context questions with LCC's Question-Answering Server*.
- C.L.A. Clarke, G.V. Cormack, T.R. Lynam, C.M. Li, G.L. McLearn. *Web Reinforced Question Answering (MultiText Experiments for TREC 2001)*.
- E. Brill, J. Lin, M. Banko, S. Dumais, A. Ng. *Data-Intensive Question Answering*.

82

References

- **Workshop Proceedings**
- H. Chen and C.-Y. Lin, editors. 2002. *Proceedings of the Workshop on Multilingual Summarization and Question Answering at COLING-02*, Taipei, Taiwan.
- M. de Rijke and B. Webber, editors. 2003. *Proceedings of the Workshop on Natural Language Processing for Question Answering at EACL-03*, Budapest, Hungary.
- R. Gaizauskas, M. Hepple, and M. Greenwood, editors. 2004. *Proceedings of the Workshop on Information Retrieval for Question Answering at SIGIR-04*, Sheffield, United Kingdom.

83

References

- N. Kando and H. Ishikawa, editors. 2004. *Working Notes of the 4th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Summarization (NTCIR-04)*, Tokyo, Japan.
- M. Maybury, editor. 2003. *Proceedings of the AAAI Spring Symposium on New Directions in Question Answering*, Stanford, California.
- C. Peters and F. Borri, editors. 2004. *Working Notes of the 5th Cross-Language Evaluation Forum (CLEF-04)*, Bath, United Kingdom.
- J. Pustejovsky, editor. 2002. *Final Report of the Workshop on TERQAS: Time and Event Recognition in Question Answering Systems*, Bedford, Massachusetts.
- Y. Ravin, J. Prager and S. Harabagiu, editors. 2001. *Proceedings of the Workshop on Open-Domain Question Answering at ACL-01*, Toulouse, France.

84