第十一章：

智能问答（QA）技术

杨建武

北京大学计算机科学技术研究所

Email:yangjianwu@icst.pku.edu.cn

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.

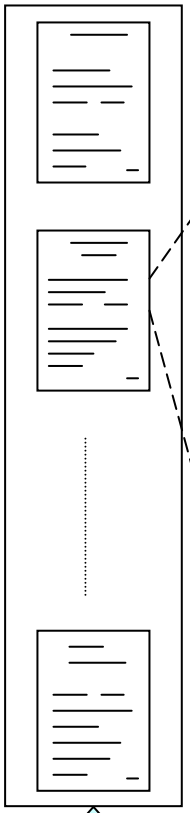
Alternatives to Information Retrieval

- 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
- Question Answering (QA)
  - users ask fact-based, natural language questions

    What is the highest volcano(火山) 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

Why Question Answering?



Document collection

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.[...]

What continent is Taormina in? → Europe

问答系统

- 输入：自然语言的提问方式
- 输出：准确的答案
- IR的扩展 + IE
  - A、更充分的表达使用者的提问意图；
  - questions (in place of keyword-based query)
  - B、直接返回答案，而不是文档列表
  - answers (in place of documents)

问答系统历史

- 1950年，A.M.Turing提出“图灵测试”
  - 如果一个人使用任意一串问题去询问两个他不能看见的对象：一个是正常思维的人；一个是机器。
  - 如果经过若干询问以后他不能得出实质的区别，则可以认为该机器已具备了人的智能。
- 1990年，Hugh Loebner设立“Loebner Prize”
  - 悬赏$100,000，奖励首次通过图灵测试的人。
  - 对每年一度“Loebner Prize”比赛的冠军，奖励$2,000。
  - “Loebner Prize”设立以来，许多程序参加了比赛，产生了许多著名的聊天机器人程序。
  - 迄今为止，没有任何一个程序通过“图灵测试”。
  - http://www.loebner.net/Prizef/loebner-prize.html

典型的聊天机器人

- ELIZA
  - 用模式及关键字匹配和置换的方法。
  - 例如，
    - 假设有：
      - 关键字: me
      - 句型模式: *you**me
      - 置换规则: what makes you think I ** you
    - 那么
      - 当输入: “Yesterday you hurt me.”时，
      - 输出为: “What makes you think I hurt you?”。
  - http://www.spaceports.com/~sjlaven/eliza.zip

典型的聊天机器人

- ALICE
  - 由宾夕法尼亚州Lehigh大学的Richard S.Wallac开发。
  - 获得2000年度、2001年度以及2002年度的“Loebner Prize”比赛冠军。
  - ALICE有40,000 多个模板，采用模式匹配的方法检索最合适的回答。
  - ALICE采用一种很好的扩充机制，AIML文件可以进行内联，许多包含特殊领域知识的AIML文件可以合并成一个更大的知识库。
  - 它遵循GNU通用公共许可协议的开放源代码。
  - http://www.alicebot.org/

典型的聊天机器人

- Jabberwock
  - Jabberwock获得2003年“Loebner Prize”冠军;
  - 用户可以通过英语或者德语，与Jabberwock进行交谈。
  - Jabberwock懂得20,000个单词，并且可以讲笑话和谜语。
  - http://www.abenteuermedien.de/jabberwock/index.php

聊天机器人共同特点

- 共同特点：
  - 在与用户的交谈过程中，都是基于谈话技巧和程序技巧，而不是根据常识。
  - 对于知道答案的问题，聊天机器人往往给出人性化的回答；
  - 对于不知道答案的问题，有三种回答方法：①猜一个答案；②老实说不知道；③用转移话题的办法回避。
  - 目前的聊天机器人，因为其知识库规模有限、甚至没有知识库，所以面对用户提出的许多专业性问题，用的就是第三种方法，也就是用转移话题的办法回避。

基于知识库的问答系统

- 拥有一个或多个知识库，并利用检索、推理等技术，来理解与求解用户问题的问答系统，称为基于知识库的问答系统。
- 一般来说，知识的数量与质量是一个基于知识库的问答系统性能是否优越的决定性因素。
- 因此，基于知识库的问答系统的主要特征是有一个或者多个知识库，其中存储一个或者多个领域的知识。
- 知识与信息不一样，知识是信息经过加工整理、解释、挑选和改造而形成的。

基于知识库的问答系统

- 共同特点：
  - 基于一个或者多个知识库（数据库），
  - 通过自然语言的形式与用户进行交流。
  - 和聊天机器人不同的是：这类系统擅长于知识问答，对于不能回答的问题，就老实回答说“不知道”，而非故意转移话题。
- 基于知识库的问答系统，主要包括几类：
  - 自然语言界面的专家系统；
  - 基于受限语言的的数据库查询系统；
  - 基于FAQ的问答系统；
  - 基于本体的问答系统；

自然语言界面的专家系统

- 1968年Feigenbaum等人于斯坦福大学建成：
  - 细菌感染诊断专家系统MYCIN
  - 探矿专家系统Prospector；
- MIT大学
  - 数学符号运算专家系统MACSYMA
- CMU大学
  - 语音识别专家系统HEARSAY
- Feigenbaum在1988年作了一次调查，当时投入运行的专家系统，约有2000个，分布在欧美和日本。

基于受限语言的数据库查询系统

- 基于受限语言的数据库查询系统，是指使用受限的自然语言，对数据库进行查询的系统。
  - 信息以关系数据库的方式存储；
  - 自然语言界面，用户用自然语言进行查询；
  - 查询结果以自然语言、表格、图形等方式返回。
- 关键步骤是要将基于自然语言的查询转换为数据库的SQL语句。

基于FAQ的问答系统

- 基于FAQ的问答系统
  - 计算用户查询和FAQ知识库中问题的相似度；
  - 找到FAQ知识库中与用户查询最为相似的问题；
  - 把此问题对应的相关答案直接提交给用户。
- 关键是一个句子相似度计算的过程；
  - 不考虑两个句子的语义信息，直接利用模式匹配技术、关键字（词）匹配技术、基于向量空间模型的TF/IDF方法等，来计算两个句子的相似度；
  - 考虑两个句子的语义信息，利用WordNet、HowNet、同义词词林等语义知识资源，计算用户查询与FAQ知识库中所有问题的语义相似度

基于本体的问答系统

- 陆汝钤院士主持开发的“盘古”知识库系统
  - 对话例子：
    - 对话者：“动物园有一头黑熊死了。”

    // 计算机：“黑熊是怎么死的？”
    - 对话者：“据说黑熊是吃塑料袋死的。”

    // 计算机：“准是哪个不文明的游客投进去的。”
    - 对话者：“是啊，太可气了。”

    // 计算机：“饲养员怎么不阻止呢？”
  - 建立一个大规模的常识知识库
    - 采用Agent和本体来表达知识
    - Agent强调包装和继承，重视事物之间的纵向联系
    - 本体主要进行常识知识的横向联想

基于本体的问答系统

- 中科院曹存根研究员率领建立NKI（国家知识基础设施）海量知识库，并在此基础上搭建NKI问答系统；
  - 庞大的、可共享的知识群体
  - 专业本体是手工获取的，专业知识的获取是半自动的方式进行的
- 在宋柔教授的率领下，北京语言文化大学计算机系语言信息处理研究所进行百科词典的知识提取工作，处理对象为行文较规范的《中国大百科全书》（光盘版），提取比较易于形式化的信息。

基于知识库的问答系统

- 优点：
  - 性能优良，对于用户提出的许多问题，回答准确;
  - 甚至可以进行一定程度的推理计算;
  - 基于知识库的，系统具有良好的可扩展性。
- 局限性：
  - 如果用户的问题落入系统的知识库范围之内，系统可以轻松的解决问题；一旦超出这个范围，系统性能很快下降为零。
  - 性能象一个窄的尖峰，适用范围非常狭窄。
  - 知识库规模不足、知识获取困难

基于自由文本的问答系统

- 自由文本，又称原始文本、非结构化文本，是指未经人工处理的文档、网页等。
- 基于自由文本的问答系统，是指
  - 接受用户以自然语言提交的问题
  - 然后利用信息检索等技术，从系统的自由文本库中检索出相关的文档、网页
  - 最后利用答案抽取等技术，从这些检索出来的自由文本中抽取出问题的答案并提交给用户

基于自由文本的问答系统

- 与其它问答系统相比，基于自由文本的问答系统：
  - 不需要建立大规模知识库，而是基于自由文本进行知识问答，节省了大量的人力物力；
  - 系统返还给用户的，是用户问题的具体答案而不只是和用户查询相关的文本或者网页。

QA: Applications

- 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)

QA: Questions

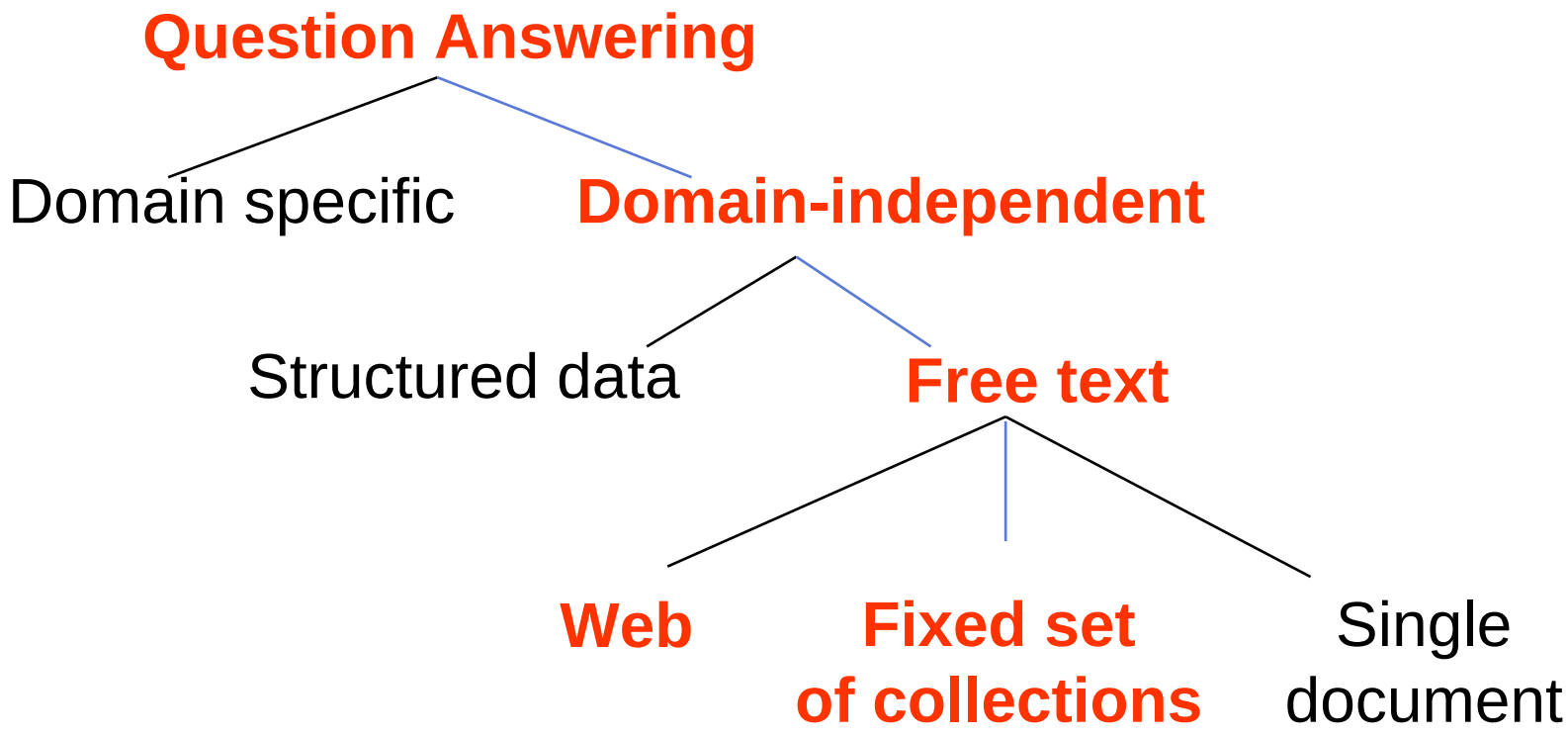
- Classification according to the **answer type**
  - **Factual questions** (*What is the larger city …*)
  - **Opinions** (*What is the author attitude …*)
  - **Summaries** (*What are the arguments for and against…*)
- Classification according to the **question speech act:**
  - **Yes/NO questions**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

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?)*

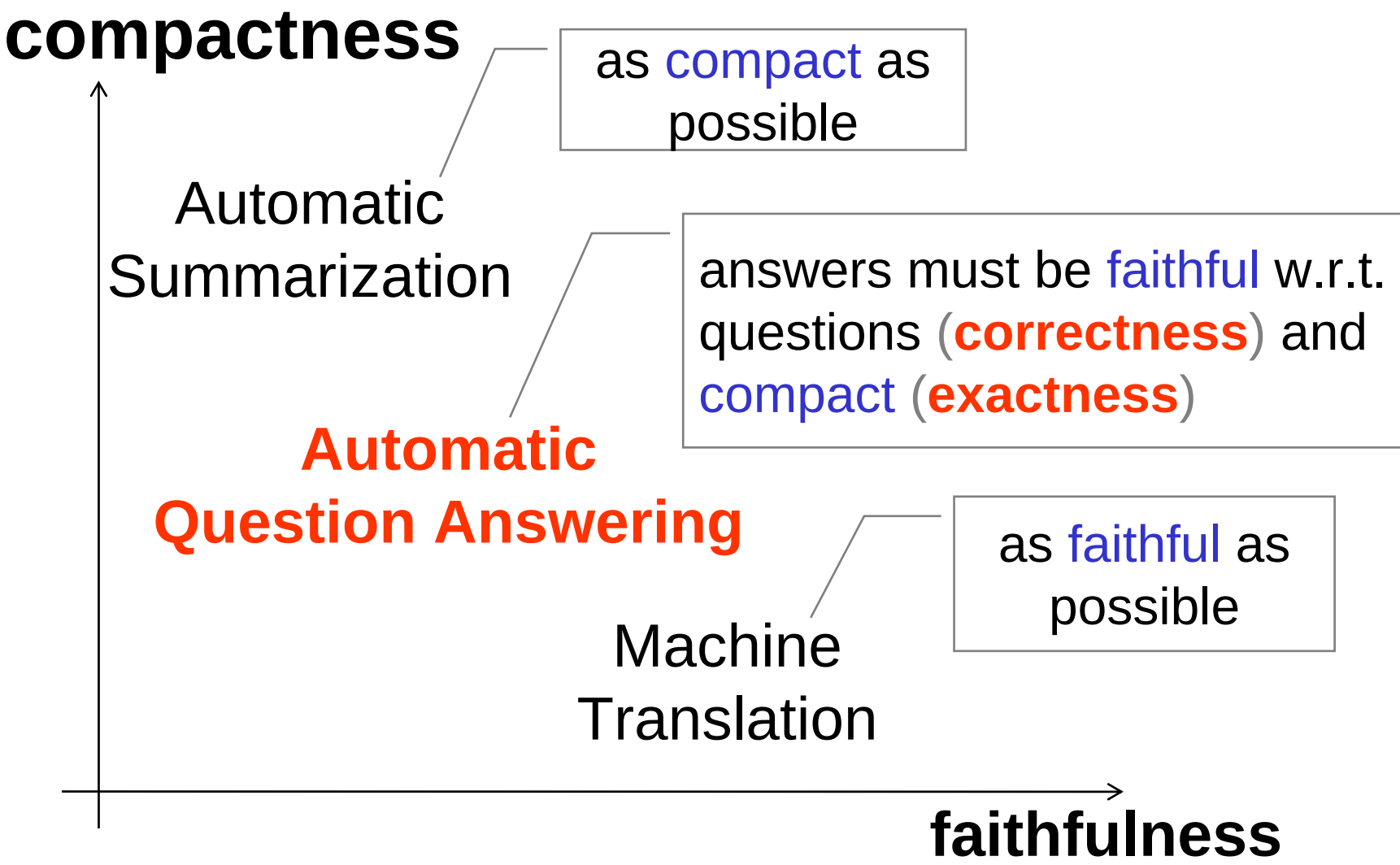
QA Research Context



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

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

QA Research Context



Question Answering at TREC

The problem simplified: The Text Retrieval Conference

- Since 1999
- **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
- Participants are research institutes, universities, industries

History of QA at TREC

- TREC 8 (Voorhees, 1999) (First QA Track)
  - 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) :
  - List questions such as *"Name 20 countries that produce coffee"*
  - Questions which don't have an answer in the collection

History of QA at TREC

- TREC 11 (Voorhees, 2002):
  - Answers had to be exact
  - Only one answer could be returned per question.
- TREC 12 (Voorhees, 2003) :
  - Introduced definition questions:
    - Define a target such as “aspirin（阿斯匹林）” 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.

History of QA at TREC

- TREC 13 (Voorhees, 2004) :
  - combines the three question types into a scenarios around targets.
  - For instance
    - Target: Hale Bopp Comet (彗星)
    - 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

History of QA at TREC

- TREC 14 (2005)
  - 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 other questions (one per target)
  - All answers had to be with reference to a document in the AQUAINT collection of newswire texts.

The TREC Document Collection

- The collection uses news articles from the following sources:
  - AP newswire, 1998-2000
  - New York Times newswire, 1998-2000
  - Xinhua News Agency newswire, 1996-2000
- In total there are 1,033,461 documents in the collection. 3GB of text

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 (旅游胜地) are there in Reims (兰斯[法国东北部城市])?
Q-712: What do most tourists visit in Reims?
Q-713: What attracts tourists in Reims

→ Reformulation questions

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 ?

Answer Assessment

- Criteria for judging an answer
  - **Relevance (相关性)**: it should be responsive to the question
  - **Correctness(正确性)**: it should be factually correct
  - **Conciseness(简明性)**: it should not contain extraneous or irrelevant information
  - **Completeness（完整性）**: it should be complete, i.e. partial answer should not get full credit
  - **Simplicity（朴素性）**: it should be simple, so that the questioner can read it easily
  - **Justification（有理有据）**: it should be supplied with sufficient context to allow a reader to determine why this was chosen as an answer to the question

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**

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;
Missipp;

Assessments

- Four possible judgments for a triple

  [Question, document, answer]

- **Right**: 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

Assessments

	[Question, document, answer]
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**

Assessments

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

DIOGENE: Enola
PARAGRAPH: NYT19991001.0143

ASSESMENT: 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.

Assessments

1402: What year did Wilt Chamberlain(张伯伦) score 100 points?

DIOGENE: 1962

ASSESMENT: 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.

Assessments

1510: Where is Anne Frank's diary?

DIOGENE: Young Girl
PARAGRAPH: NYT19980909.0338

ASSESMENT: WRONG

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

TREC Evaluation Metric

- 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]

- 评测委员会人工给出标准答案。对于每个问题，参赛系统给出5个系统运行结果，然后与标准答案进行比较。
  - 如果第一个答案就是正确的，那么系统得1分；
  - 如果第一个答案错误，而第二个答案正确，那么系统得1/2分；
  - 如果前两个答案都是错误的，而第三个答案正确，那么系统得1/3分；
  - 如果前三个答案都是错误的，而第四个答案正确，那么系统得1/4分；
  - 如果前四个答案都是错误的，而第五个答案正确，那么系统得1/5分；
  - 如果所有答案都是错误的，那么系统得0分。

- MRR: average over all questions

TREC Evaluation Metrics

- TREC Evaluation Metric: Confidence-Weighted Score (CWS)

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

System A:

1 → C 1
2 → W 0
3 → C 1
4 → C 1
5 → W 0

$$\frac{(1/1) + ((1+0)/2) + (1+0+1)/3) + ((1+0+1+1)/4) + ((1+0+1+1+0)/5)}{5}$$

Total: 0.7

System B:

1 → W 0
2 → W 0
3 → C 1
4 → C 1
5 → C 1

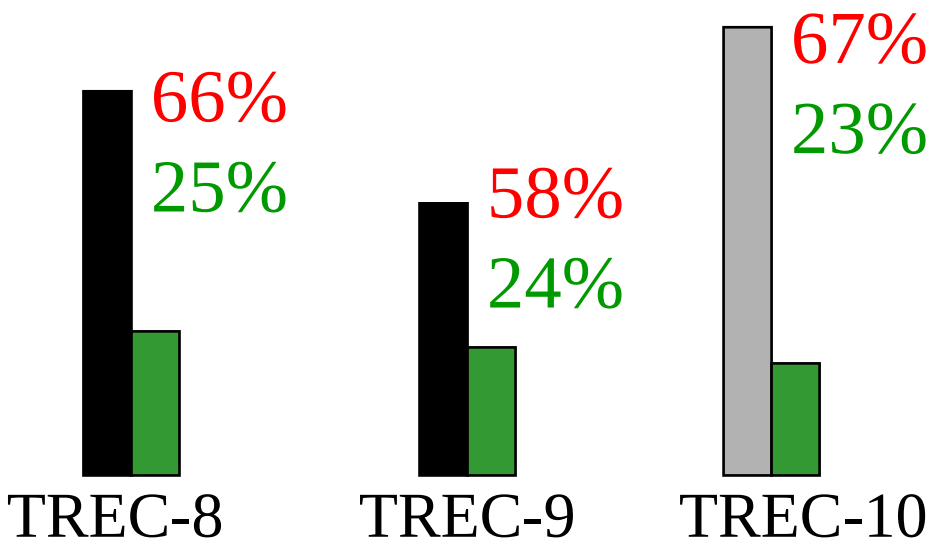
$$\frac{0 + ((0+0)/2) + (0+0+1)/3) + ((0+0+1+1)/4) + ((0+0+1+1+1)/5)}{5}$$

Total: 0.29

Score: 0 ~ 1.0， n任意个数答案

Evaluation

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



主要方法

Top Performing Systems

- Currently the best performing systems at TREC can answer approximately 60-80% of the questions
  - A pretty amazing performance!
- Approaches and successes have varied a fair deal
  - **Pattern-based**: Middle ground is to use a large collection of surface matching patterns (ISI)
  - **Knowledge-Based**: Knowledge-rich approaches, using a vast array of NLP techniques stole the show in 2000, 2001
    - Notably Harabagiu, Moldovan et al. – SMU/UTD/LCC
  - **Web-based**: AskMSR system stressed how much could be achieved by very simple methods with enough text (now has various copycats)

Pattern-Based Approach

Pattern-Based Approach -- Strategy

- ISI (USC)
- 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.
- Knowledge poor

Pattern-Based Approach -- Conditions

- 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)

QA Typology from ISI

(THING

((**AGENT (行为主体)**

(NAME (FEMALE-FIRST-NAME (EVE MARY ...))
(MALE-FIRST-NAME (LAWRENCE SAM ...))))
(COMPANY-NAME (BOEING AMERICAN-EXPRESS))
JESUS ROMANOFF ...)
(ANIMAL-HUMAN (ANIMAL (WOODCHUCK YAK ...))
PERSON)
(ORGANIZATION (SQUADRON DICTATORSHIP ...))
(GROUP-OF-PEOPLE (POSSE CHOIR ...))
(STATE-DISTRICT (TIROL MISSISSIPPI ...))
(CITY (ULAN-BATOR VIENNA ...))
(COUNTRY (SULTANATE ZIMBABWE ...))))

(**PLACE**

(STATE-DISTRICT (CITY COUNTRY...))
(GEOLOGICAL-FORMATION (STAR CANYON...))
AIRPORT COLLEGE CAPITOL ...)

(**ABSTRACT**

(LANGUAGE (LETTER-CHARACTER (A B ...)))
(QUANTITY
(NUMERICAL-QUANTITY INFORMATION-QUANTITY
MASS-QUANTITY MONETARY-QUANTITY
TEMPORAL-QUANTITY ENERGY-QUANTITY
TEMPERATURE-QUANTITY ILLUMINATION-QUANTITY
(SPATIAL-QUANTITY
(VOLUME-QUANTITY AREA-QUANTITY DISTANCE-QUANTITY)) ... PERCENTAGE)))

(**UNIT**

((INFORMATION-UNIT (BIT BYTE ... EXABYTE))
(MASS-UNIT (OUNCE ...)) (ENERGY-UNIT (BTU ...))
(CURRENCY-UNIT (ZLOTY PESO ...))
(TEMPORAL-UNIT (ATTOSECOND ... MILLENIUM))
(TEMPERATURE-UNIT (FAHRENHEIT KELVIN CELCIUS))
(ILLUMINATION-UNIT (LUX CANDELA))
(SPATIAL-UNIT
((VOLUME-UNIT (DECILITER ...))
(DISTANCE-UNIT (NANOMETER ...))))
(AREA-UNIT (ACRE)) ... PERCENT))

(**TANGIBLE-OBJECT (具体对象)**

((FOOD (HUMAN-FOOD (FISH CHEESE ...)))
(SUBSTANCE
((LIQUID (LEMONADE GASOLINE BLOOD ...))
(SOLID-SUBSTANCE (MARBLE PAPER ...))
(GAS-FORM-SUBSTANCE (GAS AIR)) ...))
(INSTRUMENT (DRUM DRILL (WEAPON (ARM GUN)) ...)
(BODY-PART (ARM HEART ...))
(MUSICAL-INSTRUMENT (PIANO)))
... *GARMENT *PLANT DISEASE)

QA Typology from ISI

(THING

((**AGENT**

(NAME (FEMALE-FIRST-NAME (EVE MARY ...))

(MALE-FIRST-NAME (LAWRENCE SAM ...))))

(COMPANY-NAME (BOEING AMERICAN-EXPRESS))

JESUS ROMANOFF ...)

(ANIMAL-HUMAN (ANIMAL (WOODCHUCK YAK ...))

PERSON)

(ORGANIZATION (SQUADRON DICTATORSHIP ...))

(GROUP-OF-PEOPLE (POSSE CHOIR ...))

(STATE-DISTRICT (TIROL MISSISSIPPI ...))

(CITY (ULAN-BATOR VIENNA ...))

(COUNTRY (SULTANATE ZIMBABWE ...))))

Pattern-Based Approach -- Example

- 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

Use of answer patterns

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

   How did Mahatma Gandhi(圣雄甘地) 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>

Acquisition of Answer Patterns

Relevant approaches:

- **Manually** developed surface pattern library (Soubbotin, Soubbotin, 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

   Precision = # of correct patterns with correct answer / # of total patterns

Pattern Learning

- "When was <person> born"
  - Typical answers
    - "Mozart was born in 1756."
    - "Gandhi (1869-1948)..."
  - Suggests phrases (regular expressions) like
    - "<NAME> was born in <BIRTHDATE>"
    - "<NAME> (<BIRTHDATE>-"
  - Example:
    - "The great composer Mozart (1756-1791) achieved fame at a young age"
    - "Mozart (1756-1791) was a genius"
    - "The whole world would always be indebted to the great music of Mozart (1756-1791)"
    - Longest matching substring for all 3 sentences is "Mozart (1756-1791)"

Pattern Learning (cont.)

- Repeat with different examples of same question type
  - "Gandhi 1869", "Newton 1642", etc.
- Some patterns learned for BIRTHDATE
  - a. born in <ANSWER>, <NAME>
  - b. <NAME> was born on <ANSWER> ,
  - c. <NAME> (<ANSWER> -
  - d. <NAME> (<ANSWER> -)

Experiments

- 6 different Q types
  - from Webclopedia QA Typology (Hovy et al., 2002a)
    - BIRTHDATE
    - LOCATION
    - INVENTOR
    - DISCOVERER
    - DEFINITION
    - WHY-FAMOUS

Experiments: pattern precision

- BIRTHDATE table:
  - 1.0 <NAME> (<ANSWER> -)
  - 0.85 <NAME> was born on <ANSWER>,
  - 0.6 <NAME> was born in <ANSWER>
  - 0.59 <NAME> was born <ANSWER>
  - 0.53 <ANSWER> <NAME> was born
  - 0.50 - <NAME> (<ANSWER>
  - 0.36 <NAME> (<ANSWER> -
- INVENTOR
  - 1.0 <ANSWER> invents <NAME>
  - 1.0 the <NAME> was invented by <ANSWER>
  - 1.0 <ANSWER> invented the <NAME> in

Experiments (cont.)

- DISCOVERER
  - 1.0 when <ANSWER> discovered <NAME>
  - 1.0 <ANSWER>'s discovery of <NAME>
  - 0.9 <NAME> was discovered by <ANSWER> in
- DEFINITION
  - 1.0 <NAME> and related <ANSWER>
  - 1.0 form of <ANSWER>, <NAME>
  - 0.94 as <NAME>, <ANSWER> and
- WHY-FAMOUS
  - 1.0 <ANSWER> <NAME> called
  - 1.0 laureate <ANSWER> <NAME>
  - 0.71 <NAME> is the <ANSWER> of
- LOCATION
  - 1.0 <ANSWER>'s <NAME>
  - 1.0 regional : <ANSWER> : <NAME>
  - 0.92 near <NAME> in <ANSWER>
- Depending on question type, get high MRR (0.6–0.9), with higher results from use of Web than TREC QA collection

Shortcomings & Extensions

- Long distance dependencies
  - "Where is London?"
  - "London, which has one of the most busiest airports in the world, lies on the banks of the river Thames"
  - would require pattern like: <QUESTION>, (<any_word>)*, lies on <ANSWER>
- Abundance & variety of Web data helps system to find an instance of patterns w/o losing answers to long distance dependencies

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:

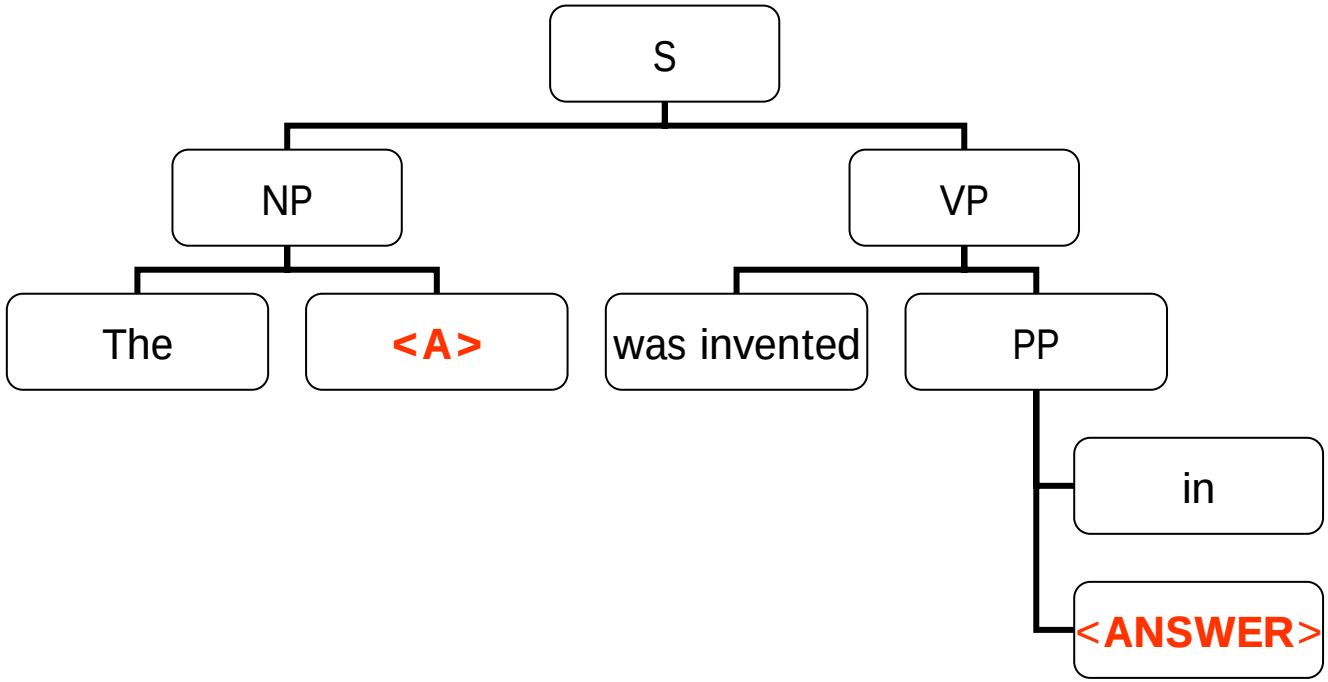
  Galileo, the famous astronomer, was born in …

- The fact that most of the QA systems use syntactic parsing (句法分析) demonstrates that the successful solution of the answer extraction problem goes beyond the surface form analysis

Syntactic answer patterns

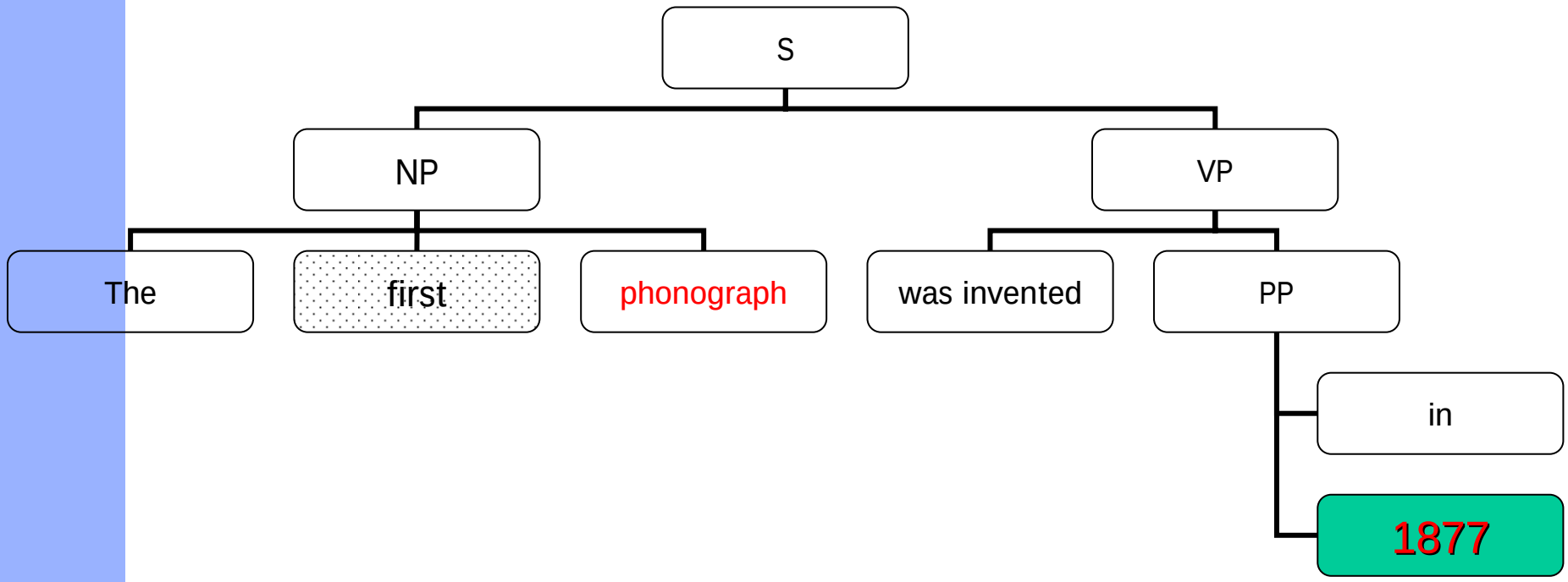
Answer patterns that capture the syntactic relations of a sentence.

When was <A> invented?



Syntactic answer patterns

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

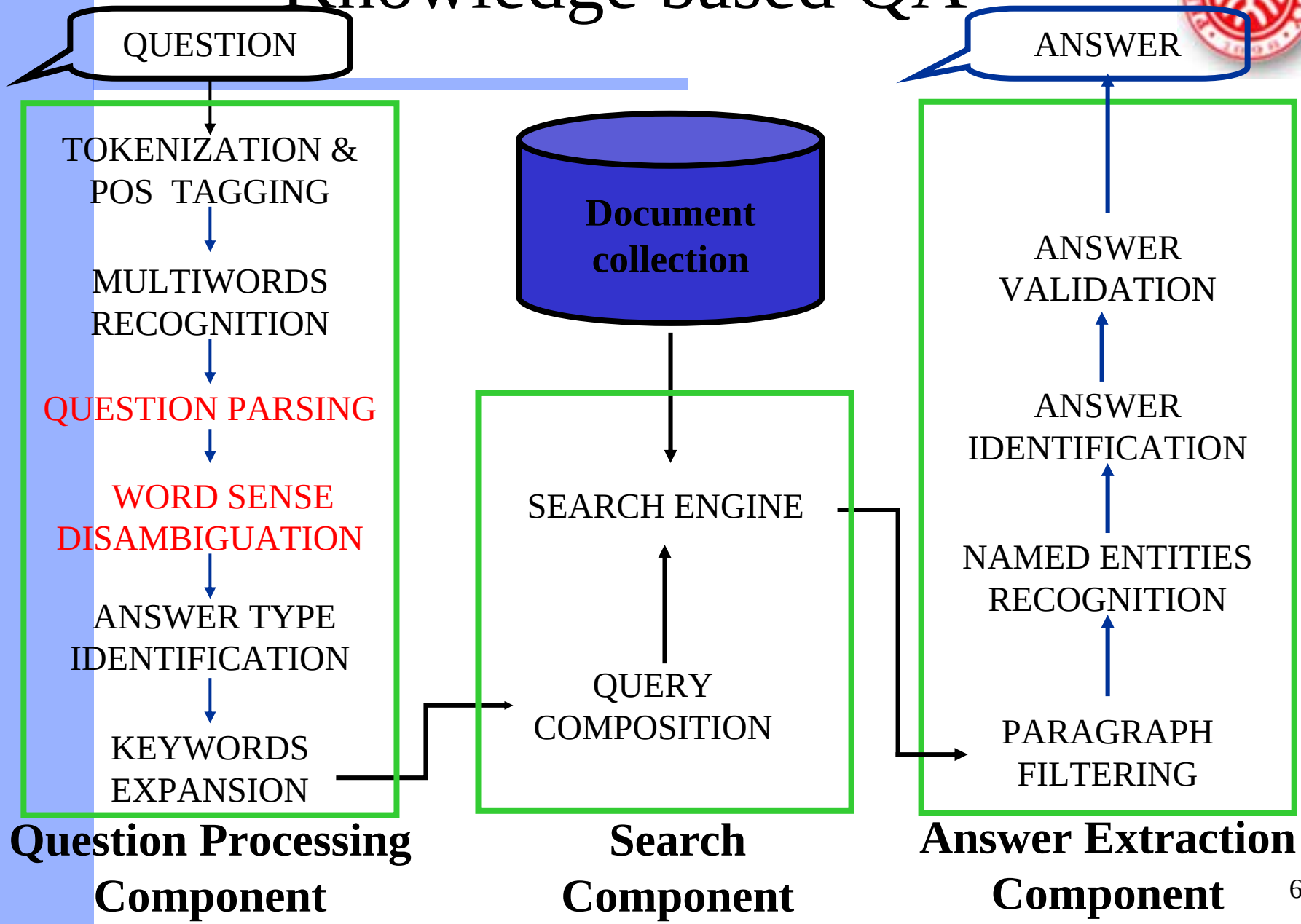


Knowledge-Based Approach

Knowledge-Based Approach

- SMU/UTD/LCC
- **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**)
  - Question Processing: To find answer type
  - Search component: To verify that a candidate answer is of the correct type
  - Answer Extraction: To get definitions
- Complex architecture...

Knowledge based QA



Question Analysis

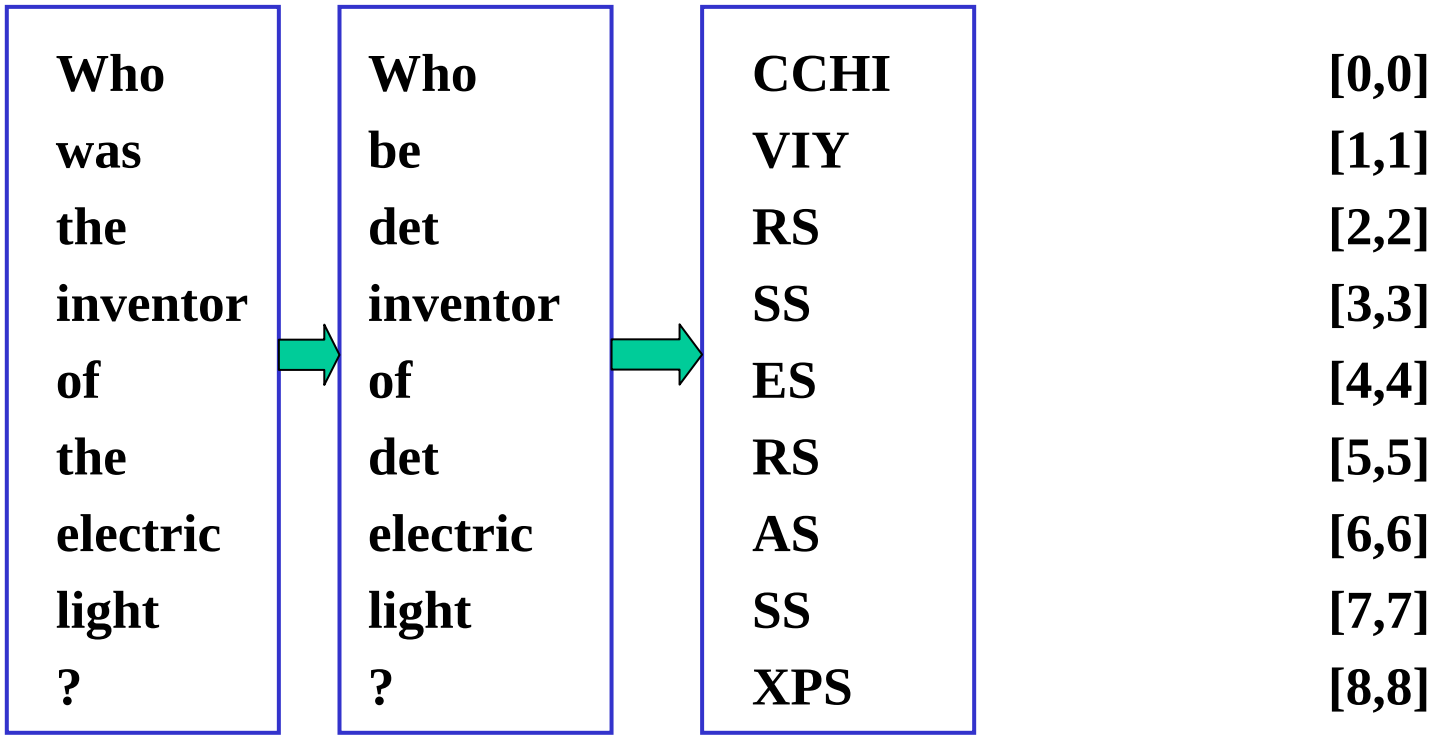
- **Input**: NLP question
- **Output**:
  - query for the search engine (i.e. a boolean composition of weighted keywords)
  - Answer type
  - Additional constraints: question focus, syntactic or semantic relations that should hold for a candidate answer entity and other entities

Question Analysis -- Steps

- **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

Tokenization and POS-tagging

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



Multi-Words recognition

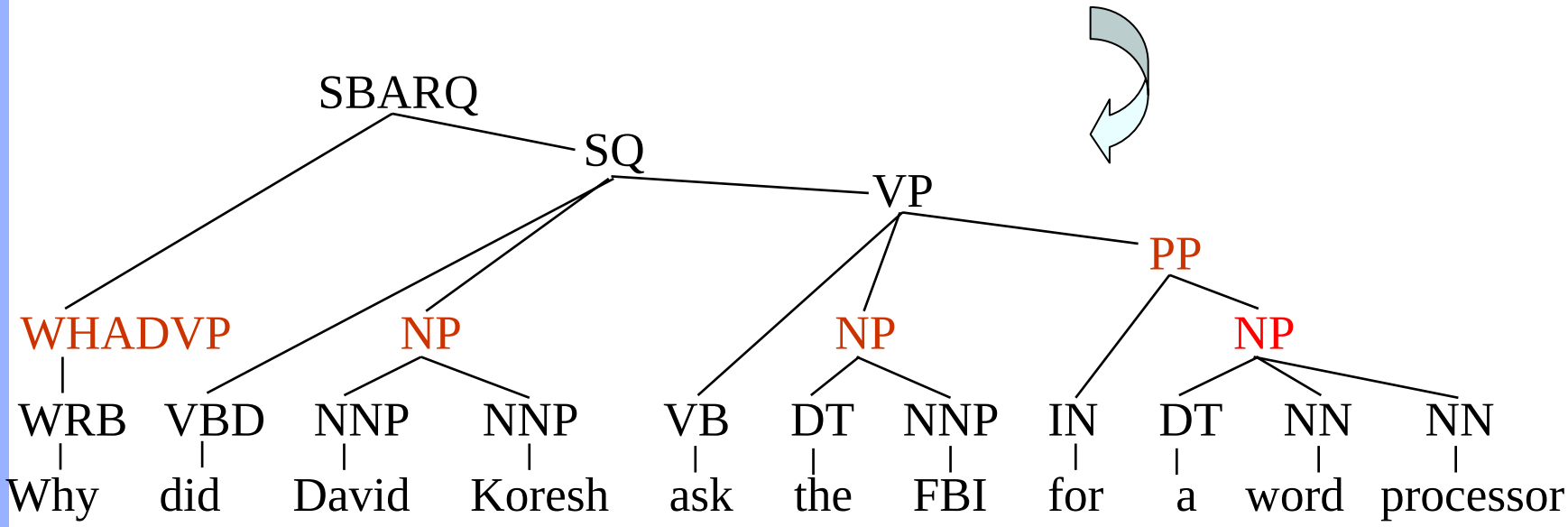
NL-QUESTION: ***Who was the inventor of the electric light?***

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]**

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?



Answer Type and Focus

- **Focus is the word that expresses the relevant entity in the question**
  - Used to select a set of relevant documents
  - ES: Where was **Mozart** born?

- **Answer Type is the category of the entity to be searched as answer**
  - PERSON, MEASURE, TIME PERIOD, DATE, ORGANIZATION, DEFINITION
  - ES: Where was **Mozart** born?
    - LOCATION

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)

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]**

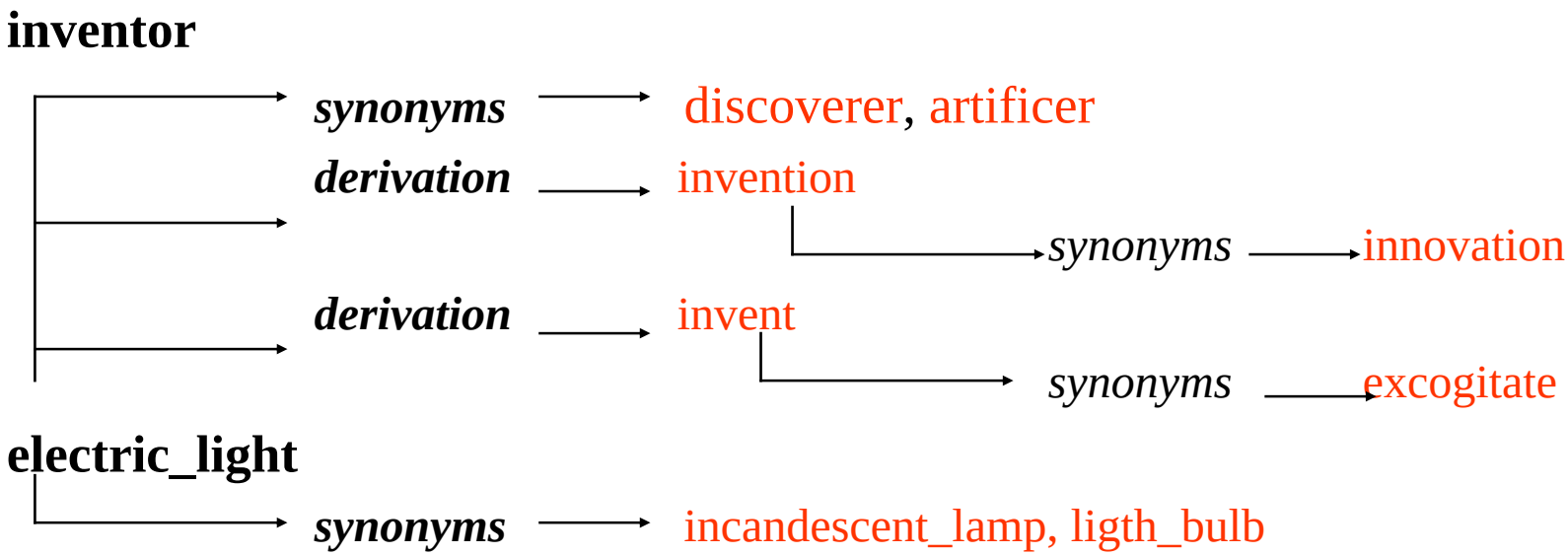
Word Sense Disambiguation

What is the brightest star visible from Earth?"

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

Expansions

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



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))
```

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

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(必需的) (e.g. the focus of the question).

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>

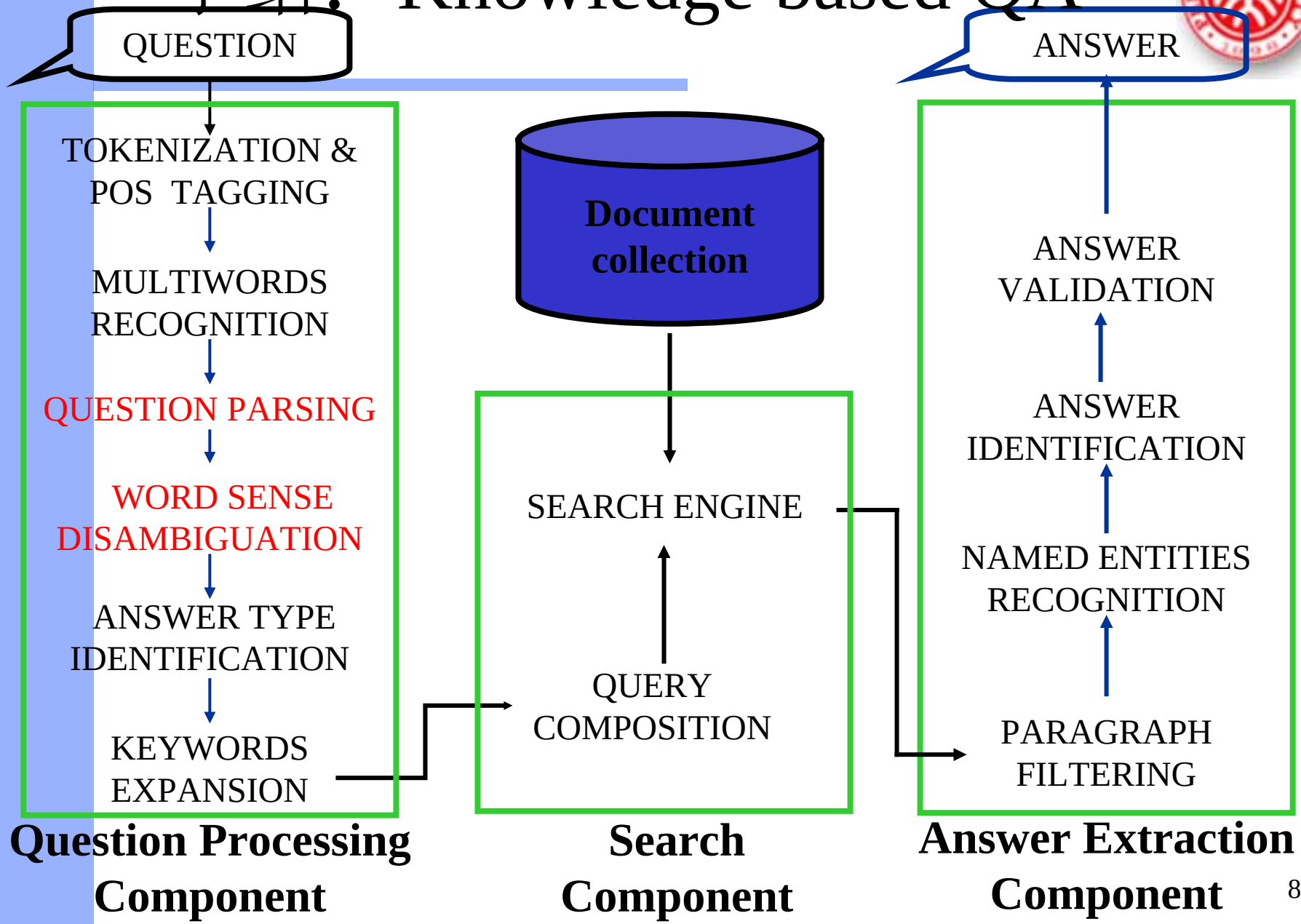
- Answer Type = **PERSON**

  Candidate Answer = **Francis Scott Key**

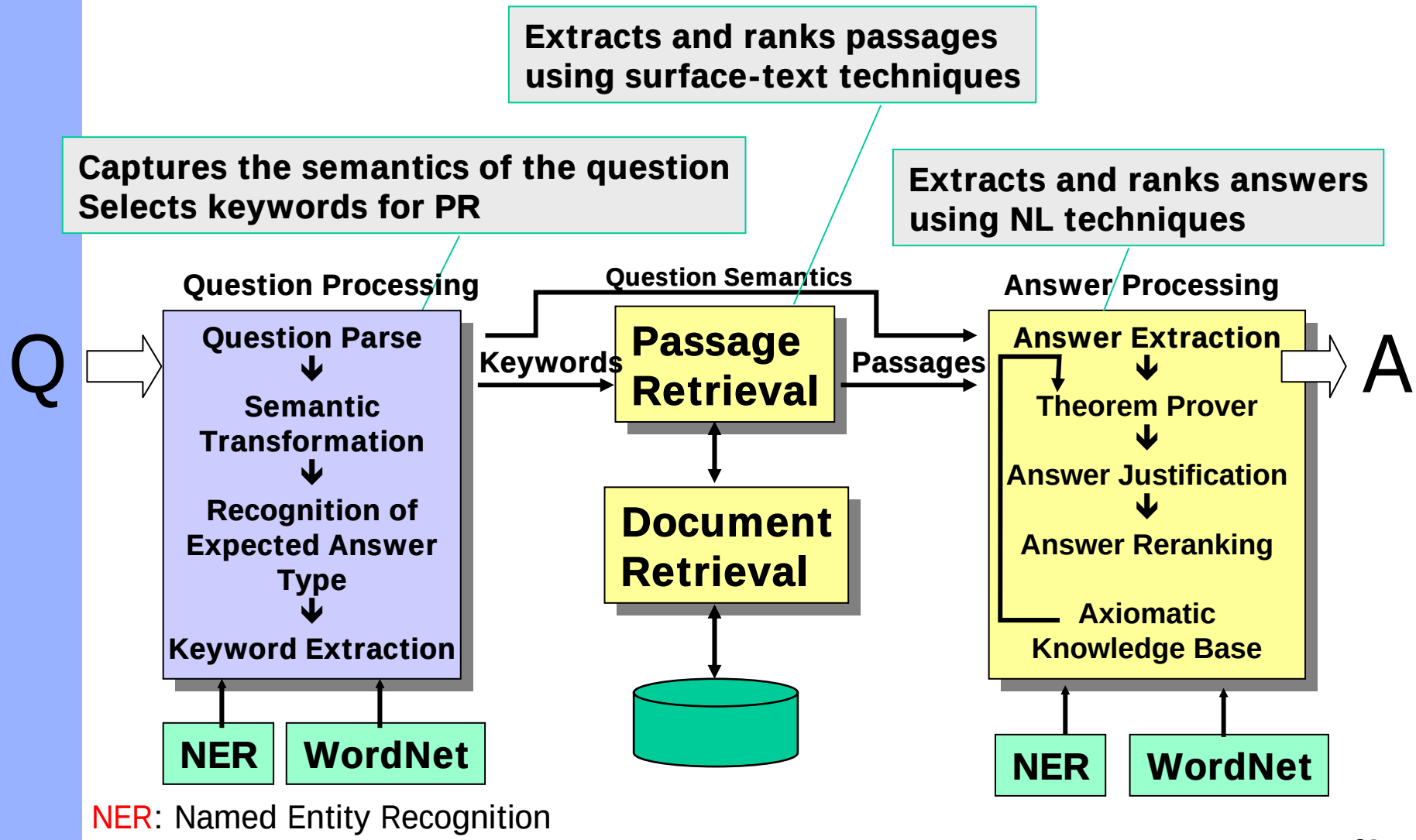
- Ranking candidate answers:

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

小结：Knowledge based QA



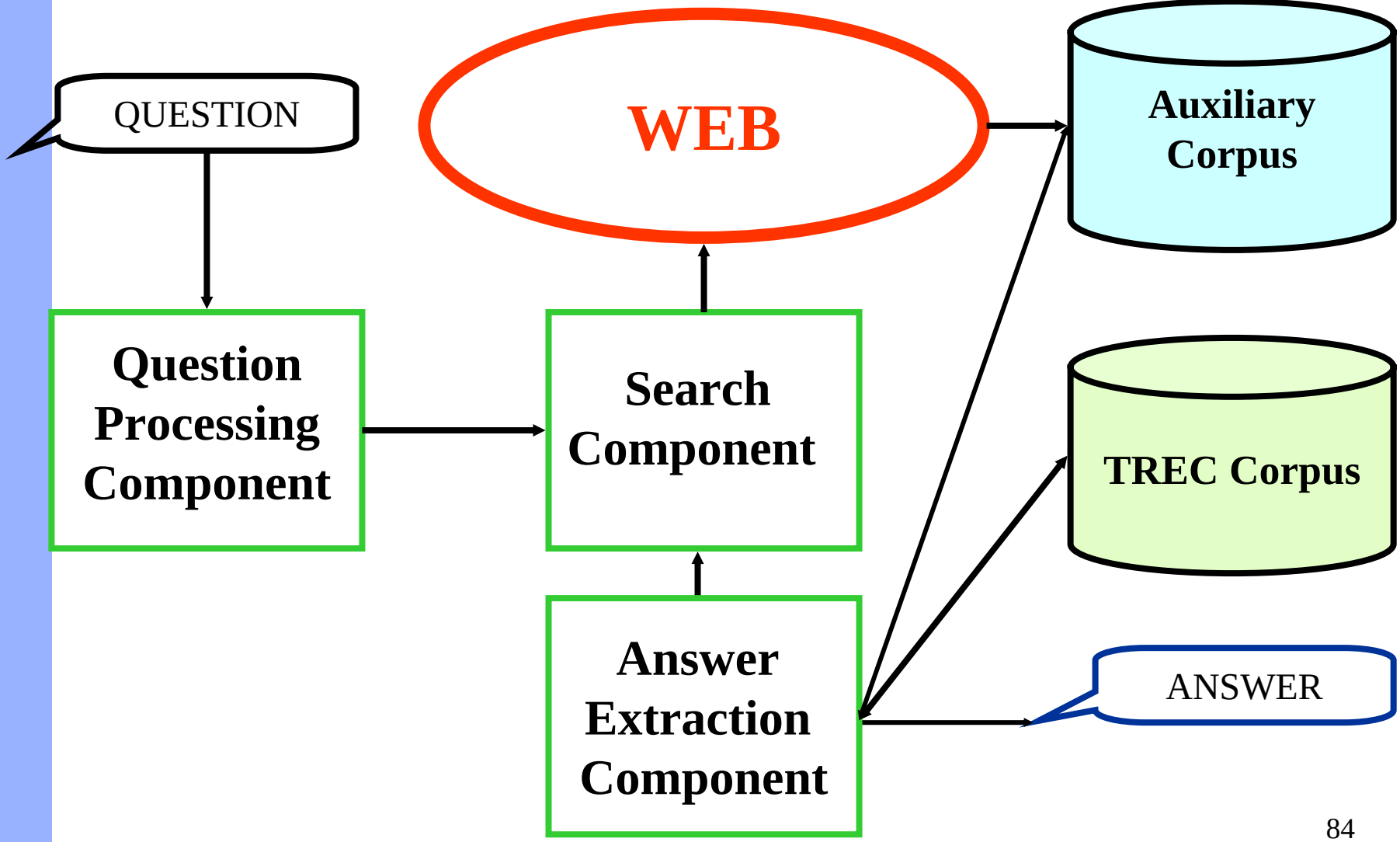
LCC Block Architecture



NER: Named Entity Recognition

Web-based Approach

Web-Based Approach

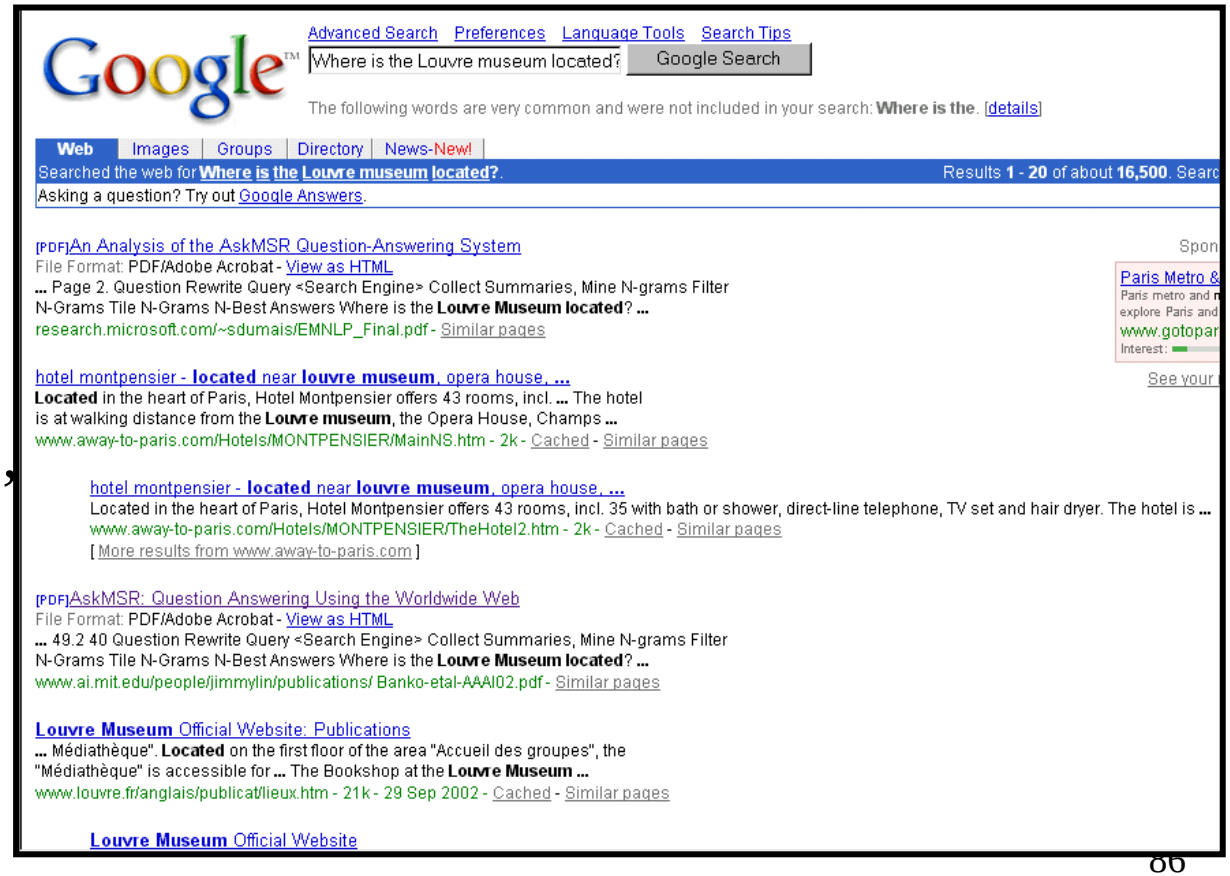


AskMSR

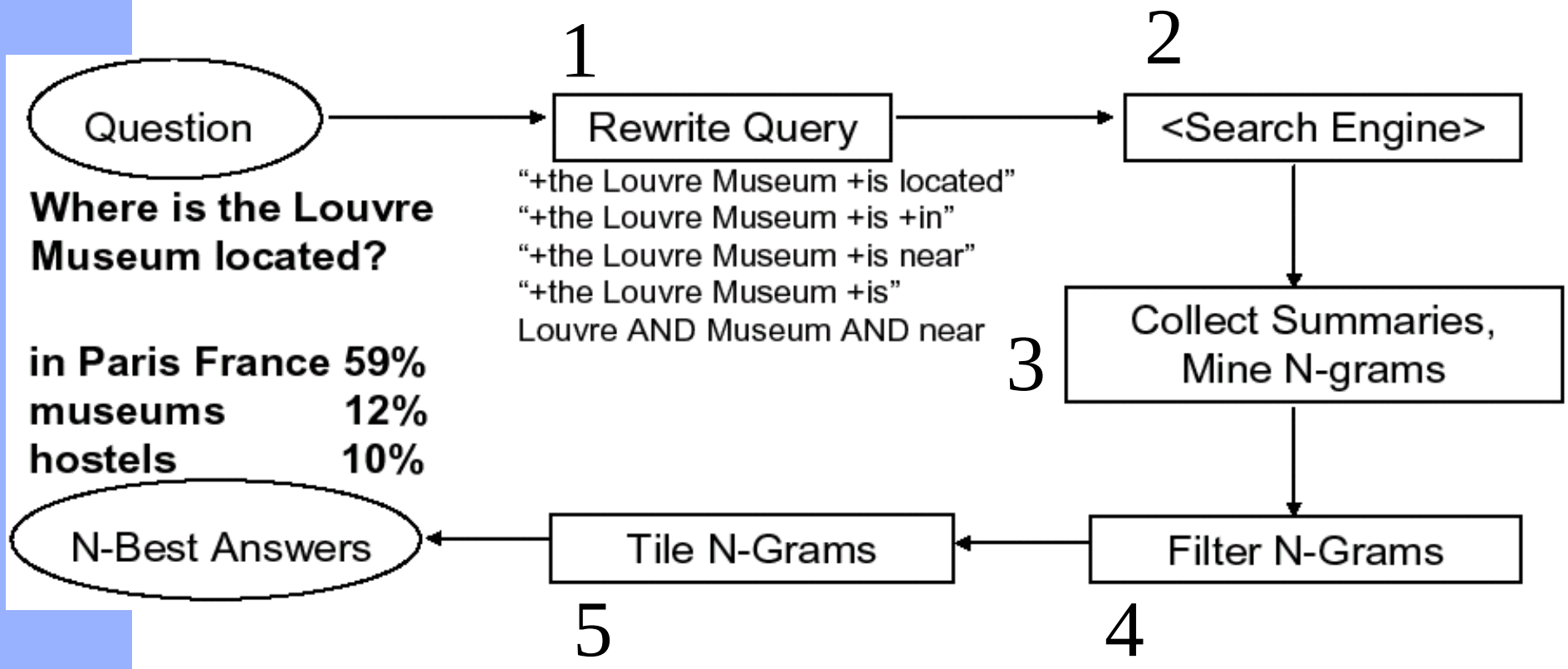
AskMSR

- ***Web Question Answering: Is More Always Better?***
  - Dumais, Banko, Brill, Lin, Ng (Microsoft, MIT, Berkeley)

- **Q: "Where is the Louvre located?"**
- Want "Paris" or "France" or "75058 Paris Cedex 01" or a map
- Don't just want URLs



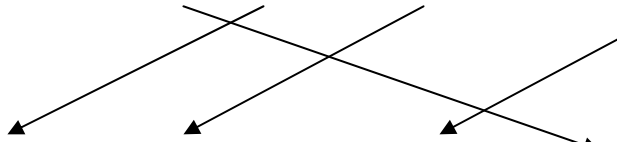
AskMSR: Details



Step 1: Rewrite queries

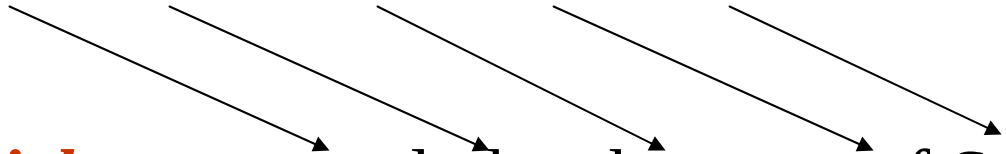
- Intuition: The user's question is often syntactically quite close to sentences that contain the answer

Where is the Louvre Museum located?



The Louvre Museum is located in ***Paris***

Who created the character of Scrooge(吝啬鬼)?



Charles Dickens created the character of Scrooge.

Query rewriting

- Classify the question into 1 of 7 categories, each mapped to a set of rewrites.
  - Rewrite sets range in size from 1 to 5 rewrites.
  - The output of a rewrite module is a 3-tuple [string, L/R/-, weight]
- **No** parser or part-of-speech tagger is used for query reformulation.
  - **But** a lexicon is used to determine part-of-speech & morphological variants of a word.
- The rewrites are **simple string-based manipulations**.
- A **final rewrite**, which is a back-off to a simple ANDing of the non stop words, is created.

Query rewriting

- Categories could be something like:
  - **Who** is/was/are/were…?
  - **When** is/did/will/are/were …?
  - **Where** is/are/were …?

a. Category-specific transformation rules

e.g. "For Where questions, move 'is' to all possible locations"

"Where is the Louvre Museum located"

→ "is the Louvre Museum located"

→ "the is Louvre Museum located"

→ "the Louvre is Museum located"

→ "the Louvre Museum is located"

→ "the Louvre Museum located is"

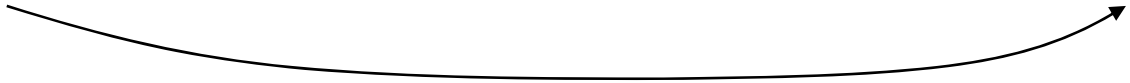
Nonsense, but who cares? It's only a few more queries to Google.

Query rewriting

b. Expected answer "Datatype"

(e.g., Date, Person, Location, …)

When was the French Revolution? → DATE

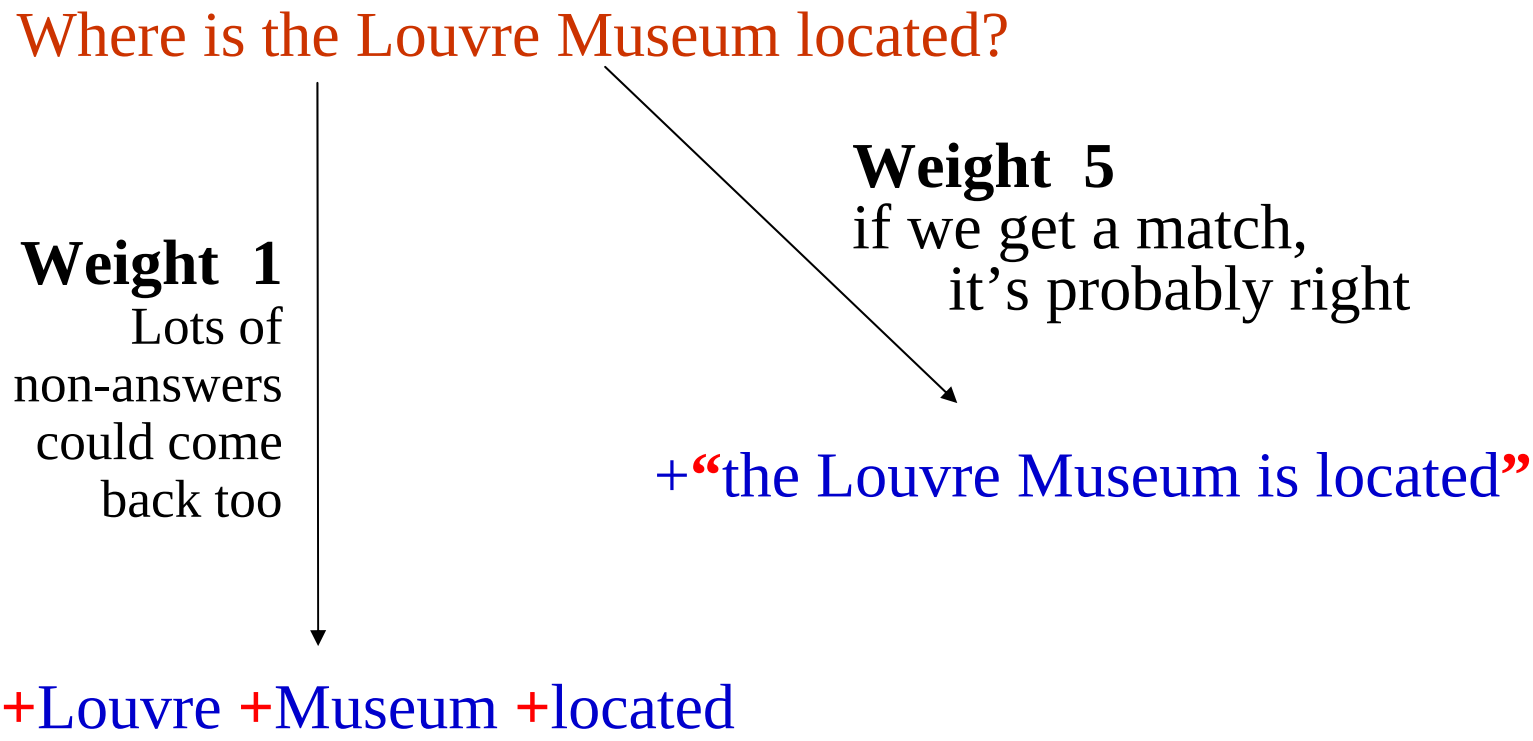


- Hand-crafted classification/rewrite/datatype rules

  (Could they be automatically learned?)

Query Rewriting - weights

- Some query rewrites are more reliable than others



Step 2: Query search engine

- Throw all rewrites to a Web-wide search engine
- Retrieve top N answers (100?)
- For speed, rely just on search engine's "snippets", not the full text of the actual document.
  - Truncation might occur, since the summary contains the query terms with a few surrounding words.

Step 3: Mining N-Grams

- Unigram, bigram, trigram, … N-gram: list of N adjacent terms in a sequence

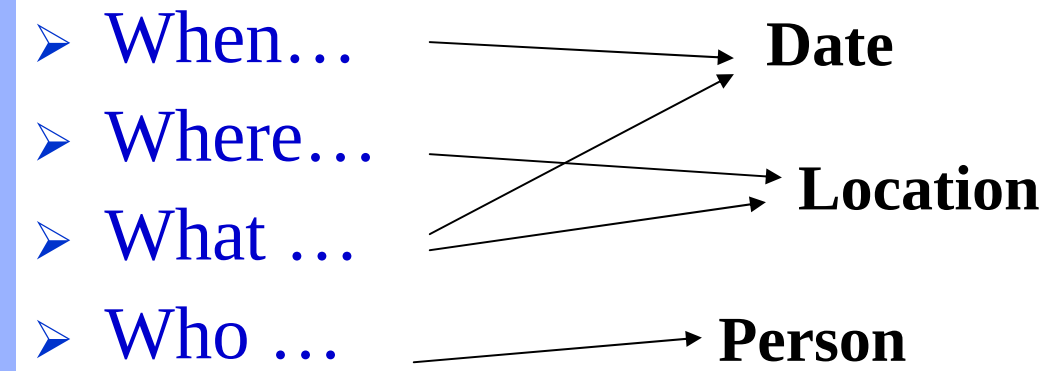
- E.g., "Web Question Answering: Is More Always Better"
  - **Unigrams**: Web, Question, Answering, Is, More, Always, Better
  - **Bigrams**: Web Question, Question Answering, Answering Is, Is More, More Always, Always Better
  - **Trigrams**: Web Question Answering, Question Answering Is, Answering Is More, Is More Always, More Always Betters

Mining N-Grams

- Enumerate all N-grams (N=1,2,3) in all retrieved snippets.
- An N-gram is scored according to the weight of the query rewrite that retrieved it.
- The scores are summed across the summaries that contain the N-gram (opposite of the usual idf ranking scheme)
- Frequency of occurrence within the summary isn't counted (the tf component in ranking schemes).
- Example: "Who created the character of Scrooge?"
  - Dickens – 117
  - Charles Dickens – 75
  - Carl Banks – 54
  - Uncle - 31
  - Christmas Carol - 78
  - Disney - 72
  - A Christmas - 41

Step 4: Filtering N-Grams

- The query is analyzed & assigned to one of seven questions types:
  - *who-question, what-question, how-many-question, etc …*
- Each question type is associated with one or more "**data-type filters**" = regular expression

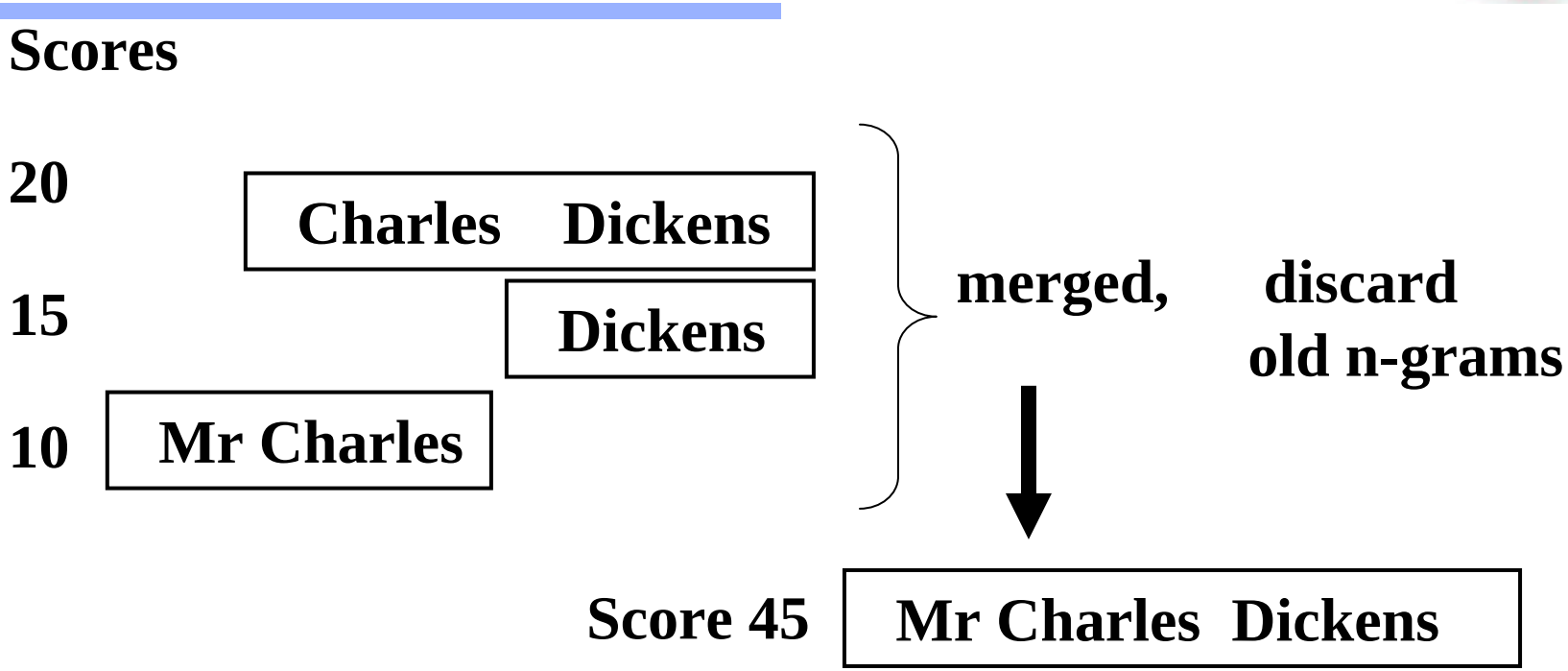


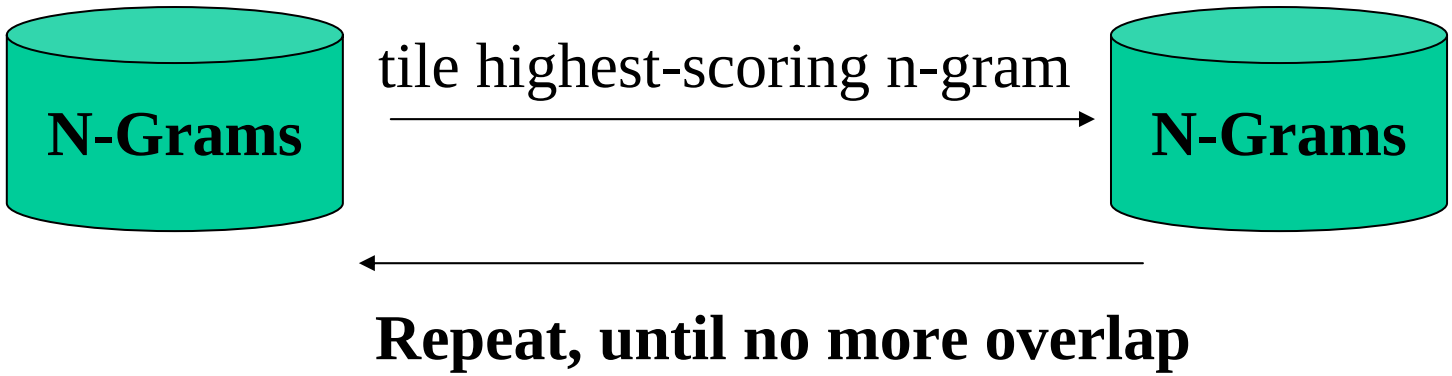
Step 4: Filtering N-Grams

- A collection of about 15 **handwritten** filters was created based on human knowledge

- The determined filters are applied to each candidate string in order to:
  - Boost score of N-grams that do match “regular expression”
  - Lower score of N-grams that don’t match “regular expression”

Step 5: Tiling the Answers





Experiments

- Standard TREC-9 contest test-bed:
  ~1 Million documents; 900 questions
- Systems receive query and generate "top 5 candidate answers"
- Standard performance metric: MRR (Mean Reciprocal Rank)
- Score = 1/R, where R is rank of the correct answer
  - 1 : 1;
  - 2 : 0.5;
  - 3 : 0.33;
  - 4 : 0.25;
  - 5 : 0.2;
  - 6+: 0.

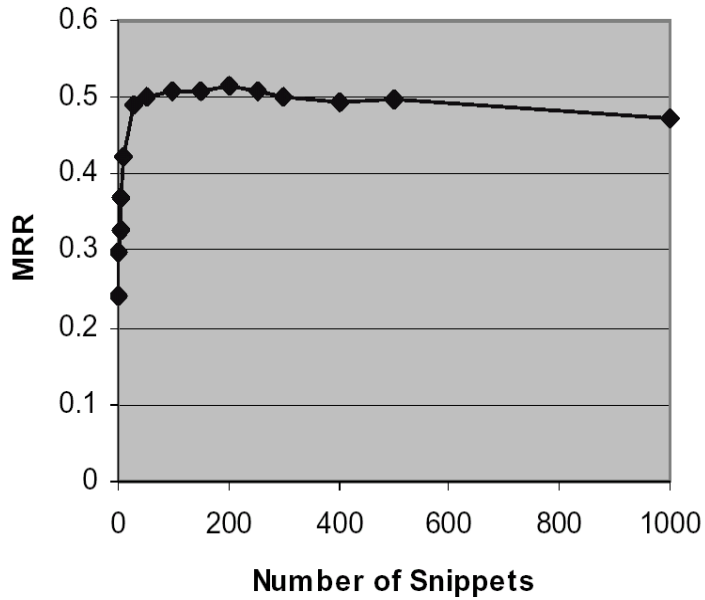
$$MRR = \frac{1}{\#questions} \sum_{\#questions} \frac{1}{R}$$

Results [summary]

- MRR=0.507 (right answer ranked about #2).
- About 61% of the questions are answered.
- The average answer length is 12 bytes.
- The system returns short answers, and not passages.
- Information redundancy,
  - Helpful ?
  - Worthless ?

Is more always better?

- What is the influence of the number of snippets the search engine returns on the quality of the answers?
- Performance improves sharply as the number of snippets increases to 50.
- Increases slowly after that (peaking at 200 snippets)
- Descends as more snippets are included for N-gram analysis.



Web-based Answer Validation

The problem: Answer Validation

*Given a question **q** and a candidate answer **a**, decide if **a** is a correct answer for **q***

What is the capital of the USA?

Washington D.C.	correct
San Francisco	wrong
Rome	wrong

Requirements for Automatic AV

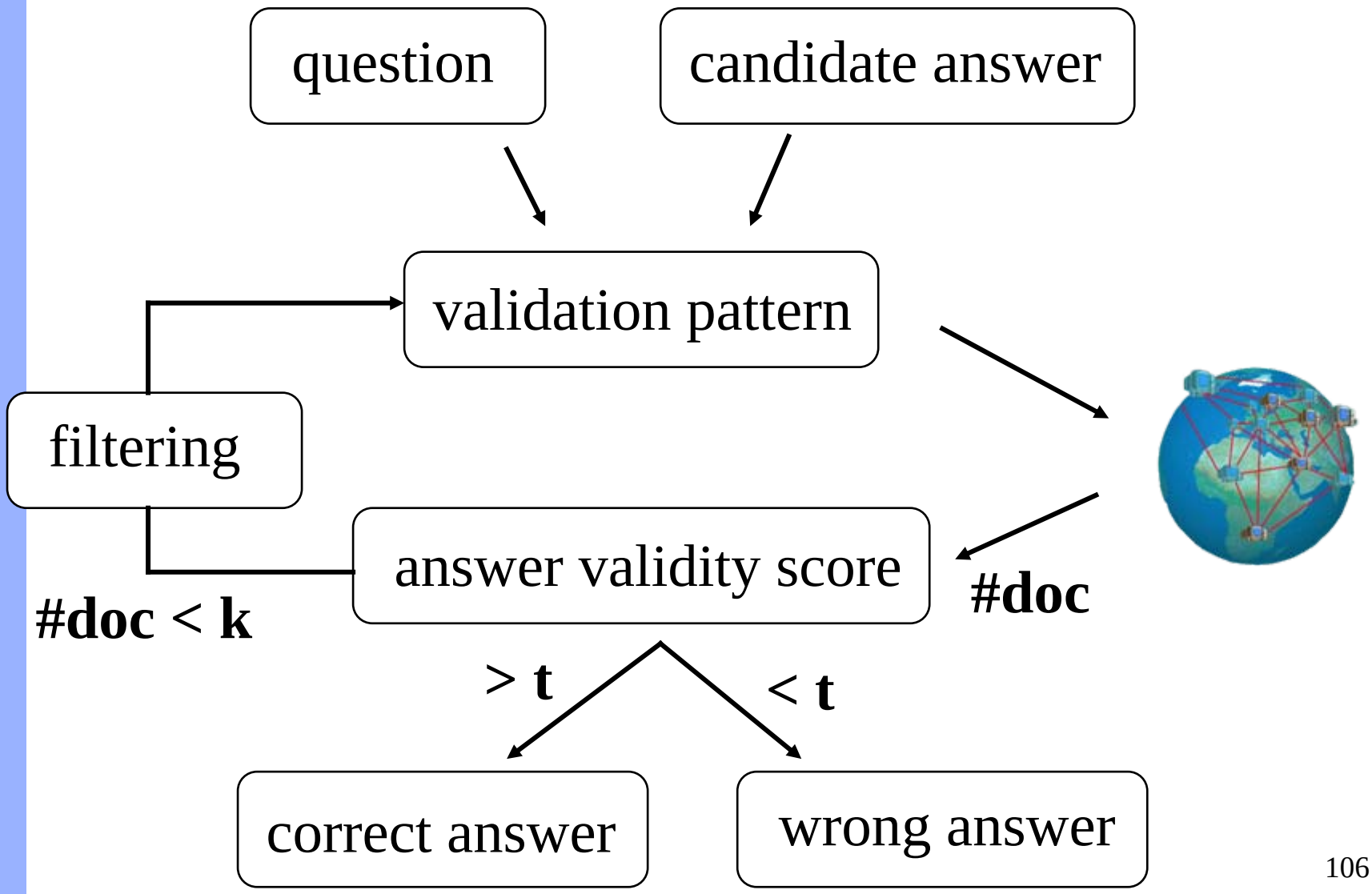
- **Accuracy**: it has to compare well with respect to human judgments
- **Efficiency**: large scale (Web), real time scenarios
- **Simplicity**: avoid the complexity of QA systems

Web Redundancy

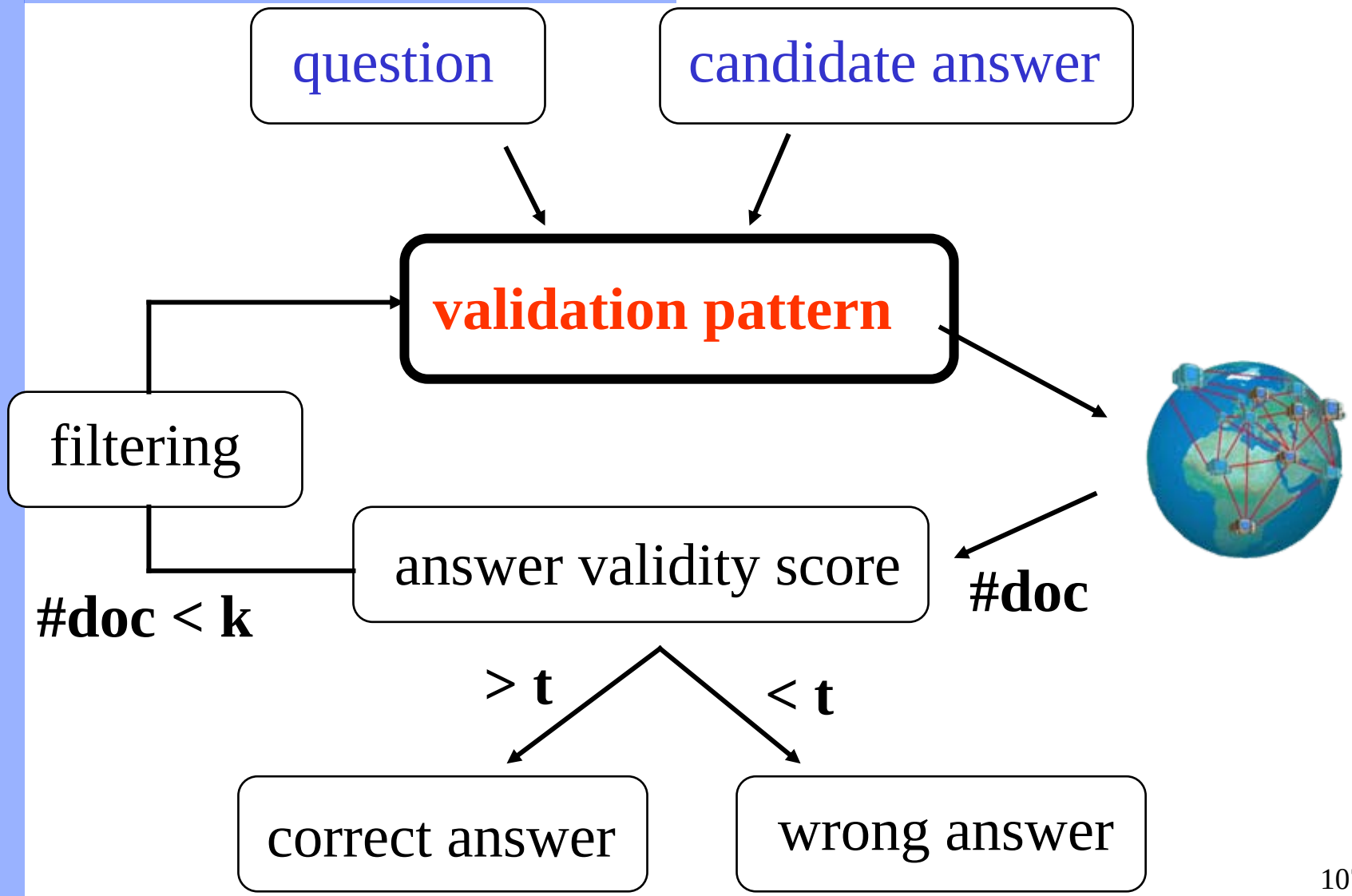
What is the capital of the USA? Washington

Capital Region USA: Fly-Drive Holidays in and Around Washington D.C.
the Insider's Guide to the Capital Area Music Scene (Washington D.C., USA).
The Capital Tangueros (Washington DC Area, USA)
I live in the Nations's Capital, Washington Metropolitan Area (USA)
In 1790 Capital (also USA's capital): Washington D.C. Area: 179 square km

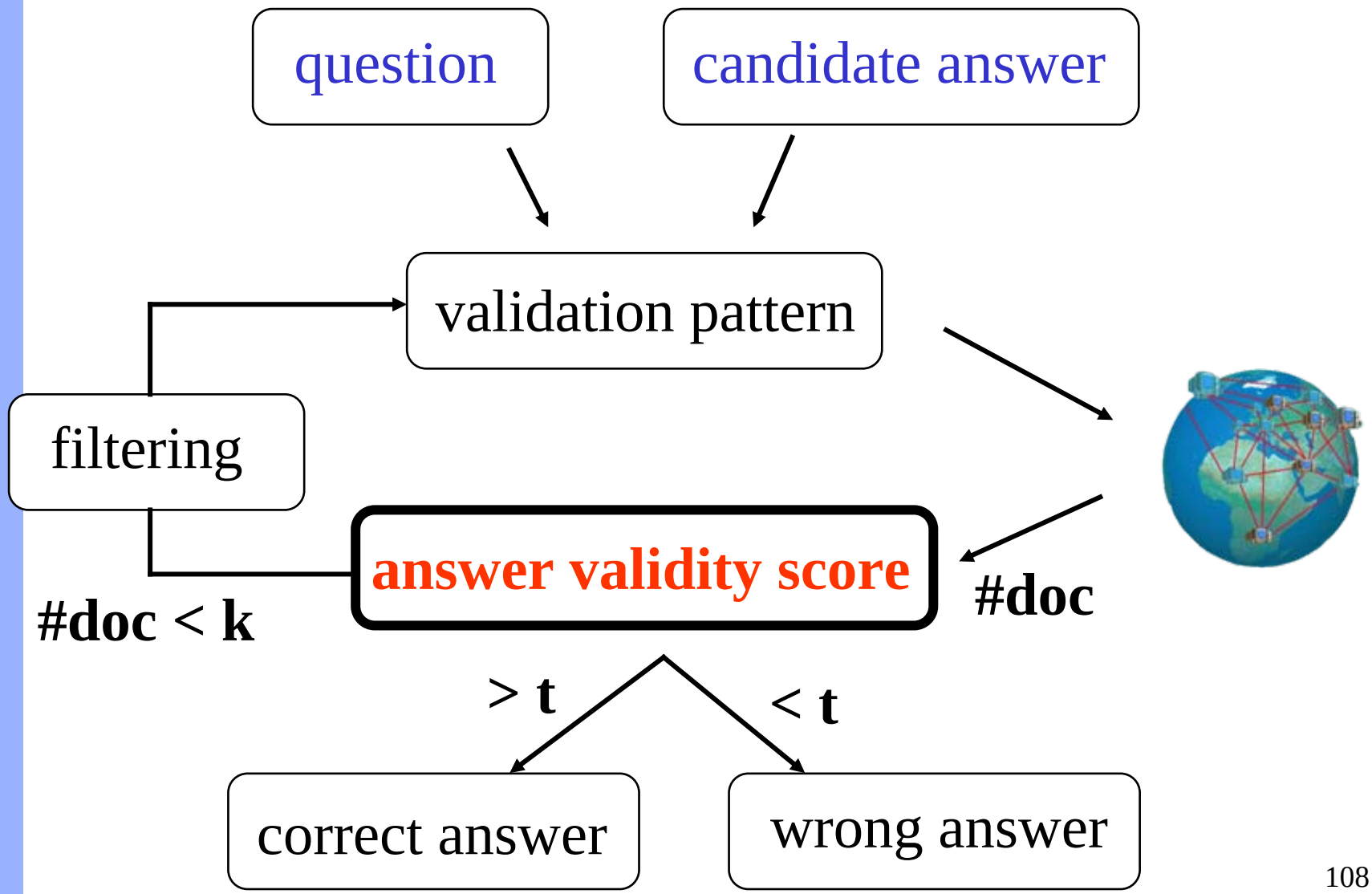
Architecture



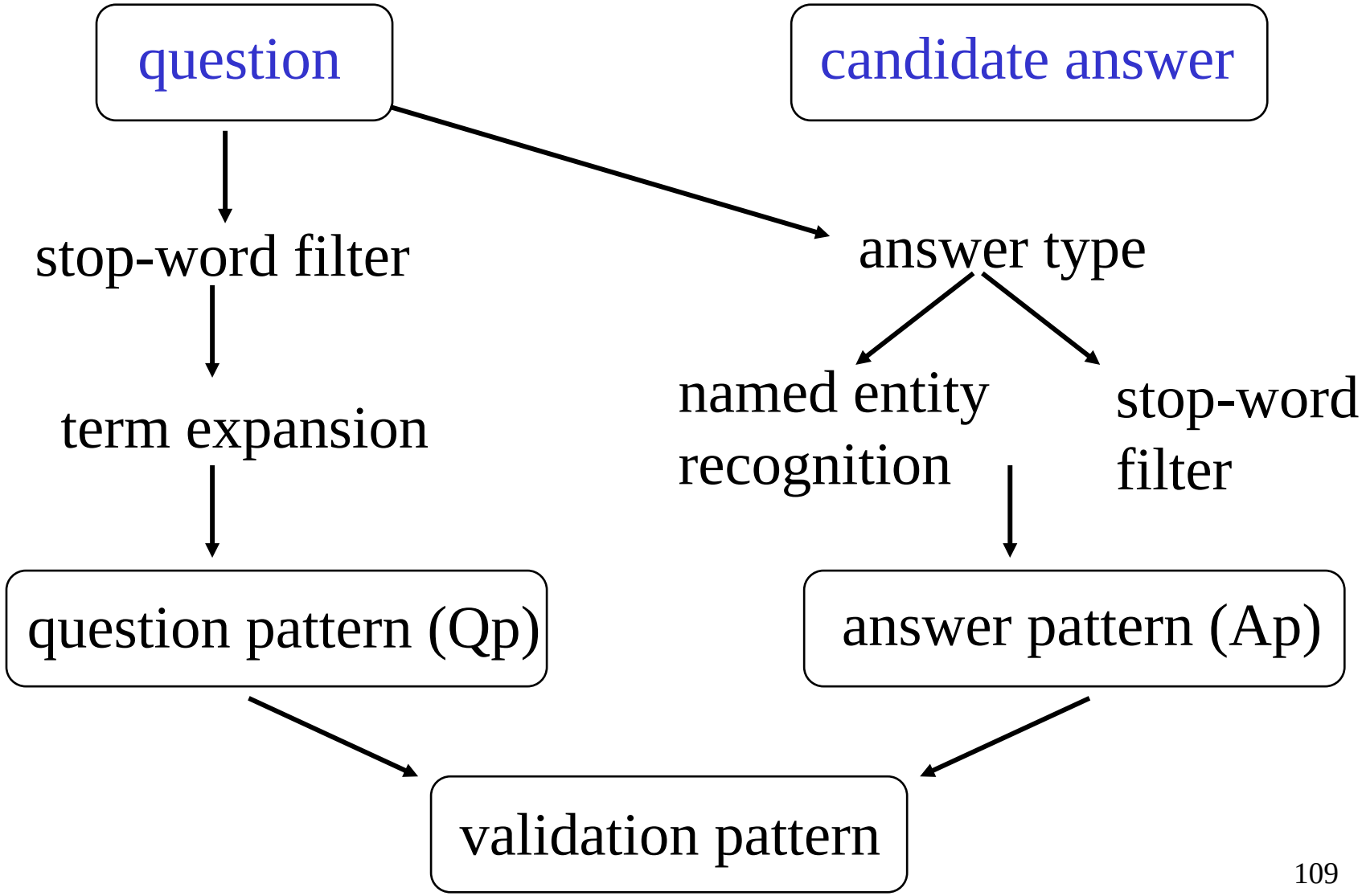
Architecture



Architecture



Extracting Validation Patterns



Extracting Validation Patterns

Capital Region USA: Fly-Drive Holidays in and Around Washington D.C.
the Insider's Guide to the Capital Area Music Scene (Washington D.C., USA).
The Capital Tangueros (Washington DC Area, USA)
I live in the Nations's Capital, Washington Metropolitan Area (USA)
In 1790 Capital (also USA's capital): Washington D.C. Area: 179 square km

[Capital **NEAR** USA **NEAR** Washington]

Answer Validity Score

- **PMI-IR** algorithm (Turney, 2001)

$$\text{PMI}(Qp, Ap) = \frac{\text{P}(Qp, Ap)}{\text{P}(Qp) * \text{P}(Ap)}$$

- The result is interpreted as evidence that the validation pattern is consistent, which imply answer accuracy

Answer Validity Score

$$\text{PMI}(Qp, Ap) = \frac{\text{hits}(Qp \texttt{ NEAR } Ap)}{\text{hits}(Qp) * \text{hits}(Ap)}$$

- Three searches are submitted to the Web:

  hits(*Qp*)
  hits(*Ap*)
  hits(*Qp* **NEAR** *Ap*)

Answer Validity Score

- Asymmetric Conditional Probability (ACP)

$$\mathrm{ACP}(Qsp, Asp) = \frac{\mathrm{P}(Qsp \mid Asp)}{\mathrm{P}(Qsp) * \mathrm{P}(Asp)^{2/3}}$$

$$= \frac{\mathrm{hits}(Qsp \ \mathrm{NEAR} \ Asp)}{\mathrm{hits}(Qsp) * \mathrm{hits}(Asp)^{2/3}}$$

Comparing PMI and ACP

$$\frac{PMI(Great\ Lakes,\ five) = 0.036}{PMI(Great\ Lakes, 19.2) = 0.02} = 1.8$$

$$\frac{ACP(Great\ Lakes,\ five) = 0.015}{ACP(Great\ Lakes, 19.2) = 0.0029} = 5.17$$

ACP increases the difference between the right and the wrong answer.

小结

- 问答系统的概念与历史
- QA@TREC
- 主要方法
  - Pattern-based Approach
  - Knowledge-Based Approach
  - Web-based Approach

Any Question?