08.04 Question Answering Systems 1/2

NLP

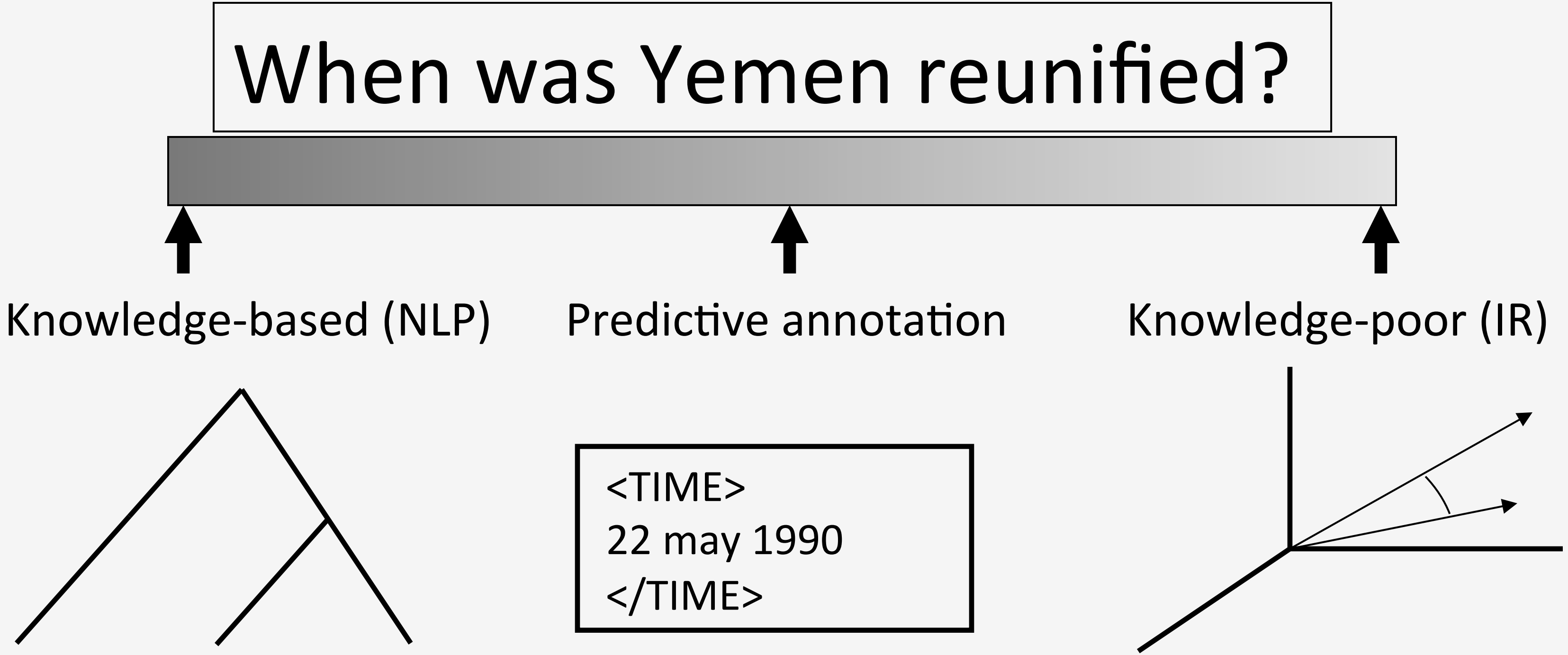
Introduction to NLP

Question Answering Systems

AnSel (Prager et al. 1999)

- IBM System
- Built for TREC
- Components
  - Predictive Annotation
  - Logistic Regression

Predictive Annotation



When was Yemen reunified?

Knowledge-based (NLP)

Predictive annotation

Knowledge-poor (IR)

<TIME>
22 may 1990
</TIME>

Predictive Annotation

```
<p><NUMBER>1</NUMBER></p>

<p><QUERY>Who is the author of the book, "The Iron Lady: A Biography of Margaret Thatcher"?
</QUERY></p>

<p><PROCESSED_QUERY>@excwin(*dynamic* @weight(2.00001001 *Ion_Lady) @weight(200
Biography_of_Margaret_Thatcher) @weight(200 Margaret) @weight(100 author) @weight(100 book)
@weight(100 iron) @weight(100 lady) @weight(100 :) @weight(100 biography) @weight(100
thatcher) @weight(400 @syn(PERSON$ ORG$ NAME$ ROLE$) ) )</PROCESSED_QUERY></p>

<p><DOC>LA090290-0118</DOC></p>

<p><SCORE>1020.8114</SCORE></p>

<TEXT><p>THE IRON LADY; A <span class=“NAME”> Biography of Margaret Thatcher </span> by
<span class=“PERSON”> Hugo Young </span> (<span class="ORG“> Farrar , Straus & Giroux </
span> ) The central riddle revealed here is why, as a woman <span class="PLACEDEF"> in a
man </span> 's world, <span class="PERSON"> Margaret Thatcher </span> evinces such an
exclusionary attitude toward women.</p></TEXT>
```

Some Observations

- In documents that contain the answers, the query terms tend to occur in close proximity to each other
- The answers to fact-seeking questions are usually phrases
- These phrases can be categorized by question type
- The phrases can be identified in text by pattern matching techniques

Feature Selection

Avgdst: the average distance in words between the beginning of the span and the words in the query that also appear in the passage. Example: given the question "Who was Johnny Mathis' high school track coach?" and the passage *"Tim O'Donohue, Woodbridge High School's varsity baseball coach, resigned Monday and will be replaced by assistant Johnny Ceballos, Athletic Director Dave Cowen said."* and the span "*Tim O'Donohue*", the value of **avgdst** is equal to 8.

Notinq: the number of words in the span that do not appear in the query. Example: **Notinq** ("*Woodbridge high school"*) = 1, because both "high" and "school" appear in the query while "Woodbridge" does not. It is set to –100 when the actual value is 0.

Frequency: number of times a given span appears in the hit list.

Sscore: passage relevance as computed by the search engine.

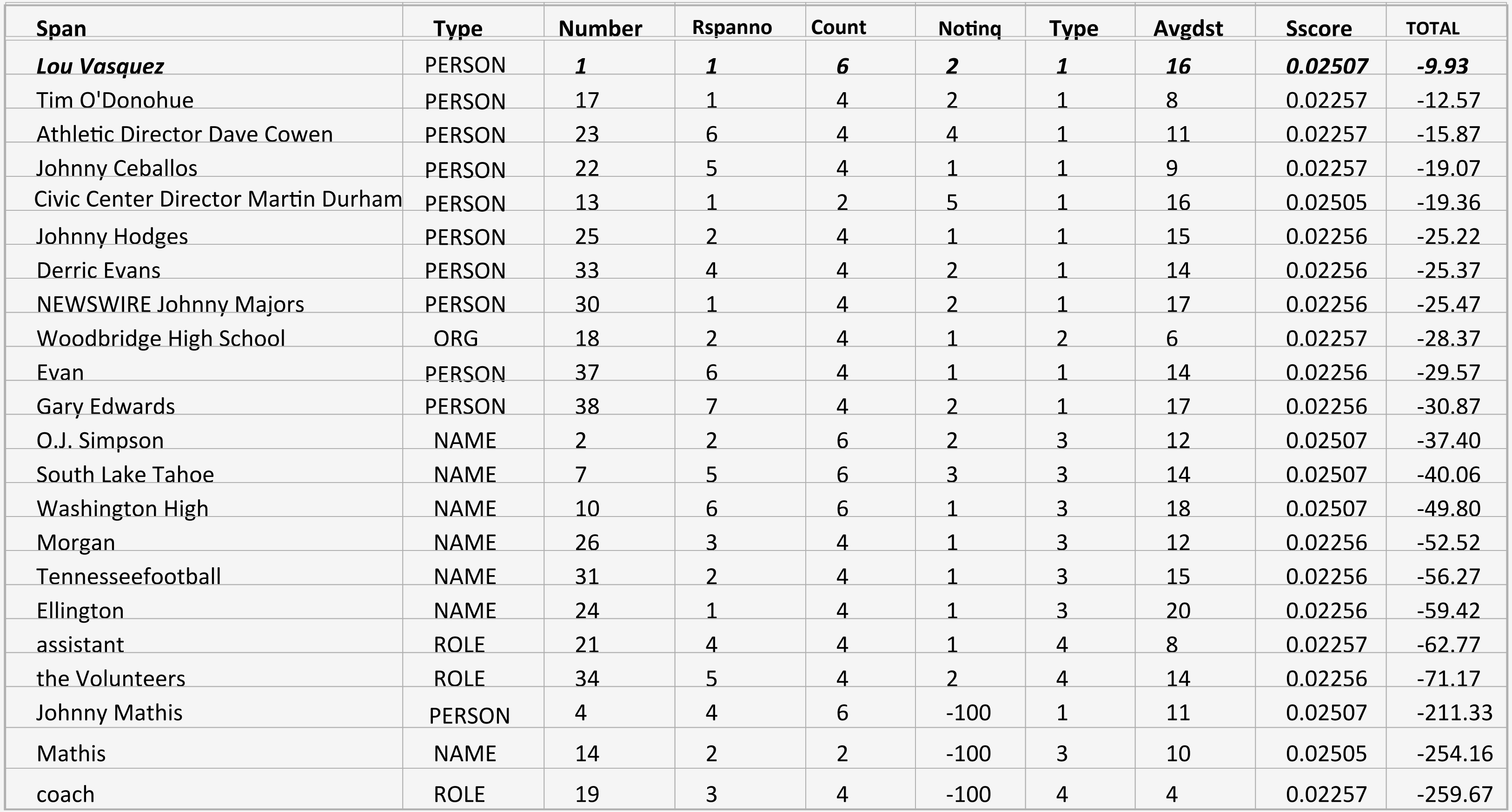
Number: position of the span among all spans returned. Example: "*Lou Vasquez*" was the first span returned by GuruQA on the sample question.

Rspanno: position of the span among all spans returned within the current passage.

Count: number of spans of any span class retrieved within the current passage.

Type: the position of the span type in the list of potential span types. Example: **Type** ("*Lou Vasquez*") = 1, because the span type of "*Lou Vasquez*", namely "PERSON" appears first in the list of potential span types, "PERSON ORG NAME ROLE".

08.04 Question Answering Systems 1/2



Span	Type	Number	Rspanno	Count	Notinq	Type	Avgdst	Sscore	TOTAL
Lou Vasquez	PERSON	***1***	***1***	***6***	***2***	***1***	***16***	***0.02507***	***-9.93***
Tim O'Donohue	PERSON	17	1	4	2	1	8	0.02257	-12.57
Athletic Director Dave Cowen	PERSON	23	6	4	4	1	11	0.02257	-15.87
Johnny Ceballos	PERSON	22	5	4	1	1	9	0.02257	-19.07
Civic Center Director Martin Durham	PERSON	13	1	2	5	1	16	0.02505	-19.36
Johnny Hodges	PERSON	25	2	4	1	1	15	0.02256	-25.22
Derric Evans	PERSON	33	4	4	2	1	14	0.02256	-25.37
NEWSWIRE Johnny Majors	PERSON	30	1	4	2	1	17	0.02256	-25.47
Woodbridge High School	ORG	18	2	4	1	2	6	0.02257	-28.37
Evan	PERSON	37	6	4	1	1	14	0.02256	-29.57
Gary Edwards	PERSON	38	7	4	2	1	17	0.02256	-30.87
O.J. Simpson	NAME	2	2	6	2	3	12	0.02507	-37.40
South Lake Tahoe	NAME	7	5	6	3	3	14	0.02507	-40.06
Washington High	NAME	10	6	6	1	3	18	0.02507	-49.80
Morgan	NAME	26	3	4	1	3	12	0.02256	-52.52
Tennesseefootball	NAME	31	2	4	1	3	15	0.02256	-56.27
Ellington	NAME	24	1	4	1	3	20	0.02256	-59.42
assistant	ROLE	21	4	4	1	4	8	0.02257	-62.77
the Volunteers	ROLE	34	5	4	2	4	14	0.02256	-71.17
Johnny Mathis	PERSON	4	4	6	-100	1	11	0.02507	-211.33
Mathis	NAME	14	2	2	-100	3	10	0.02505	-254.16
coach	ROLE	19	3	4	-100	4	4	0.02257	-259.67

IONAUT (Abney et al. 2000)

- Passage retrieval
  - Uses START (Salton, Buckley)
- Entity recognition
  - Uses Cass (Abney) – partial parser
- Entity classification
  - Simple patterns for 8 question types

Mulder (Kwok et al. 2001)

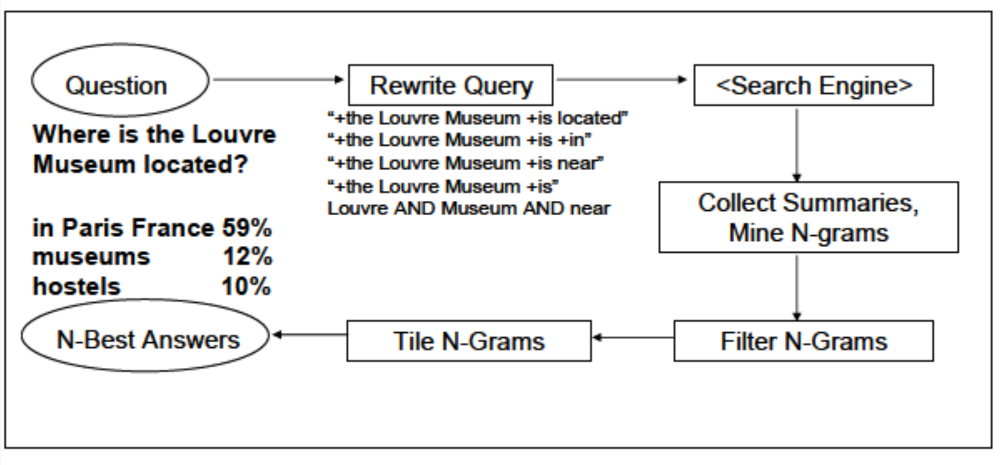
- First large-scale Web QA system
- Components
  - Maximum entropy parser (Charniak)
  - PC-Kimmo for unknown words
  - Link parser (Sleator and Temperley)
  - Google
- Tokenization
  - phrases in quotes
- Query transformations
  - “When did Nixon visit China” -> “Nixon visited China”

NSIR (Radev et al. 2002)

- Probabilistic phrase reranking
  - P(qtype|signature)
  - Signature = POS sequence (e.g., “NNP NNP” for “Bill Gates”)
- Search engines
  - AlltheWeb, NorthernLight, Altavista, Google

AskMSR (Banko et al. 2002)

- Assumption
  - Someone has already answered this question on the Web
- Components
  - Query rewriting
  - Snippet retrieval
  - N-gram ranking
- Tiling matches
  - Combining A B C and B C D into A B C D
  - E.g., “Mr. Charles” and “Charles Dickens” into “Mr. Charles Dickens”


Question
Where is the Louvre Museum located?
Rewrite Query
"+the Louvre Museum +is located"
"+the Louvre Museum +is +in"
"+the Louvre Museum +is near"
"+the Louvre Museum +is"
Louvre AND Museum AND near
<Search Engine>
Collect Summaries, Mine N-grams
Filter N-Grams
Tile N-Grams
N-Best Answers
in Paris France 59%
museums 12%
hostels 10%

Echihabi and Marcu

- Based on the noisy-channel model
- Find the sentence S that maximizes

  $p(q|S)$
- Requires simplifying the sentences

NLP