CS345a Data Mining Project

42

A Web Based Question Answering System

Vincenzo Di Nicola
Jyotika Prasad

The Ultimate question answering system

- What is the meaning of life?
- Who are we?
- Why are we doing CS?

Or, less philosophically,

- What questions will the CS345 final contain?
- Who will win the next World Cup?
  (that's an easy one, though)

Project Aim

Well, our system has a humbler aim:

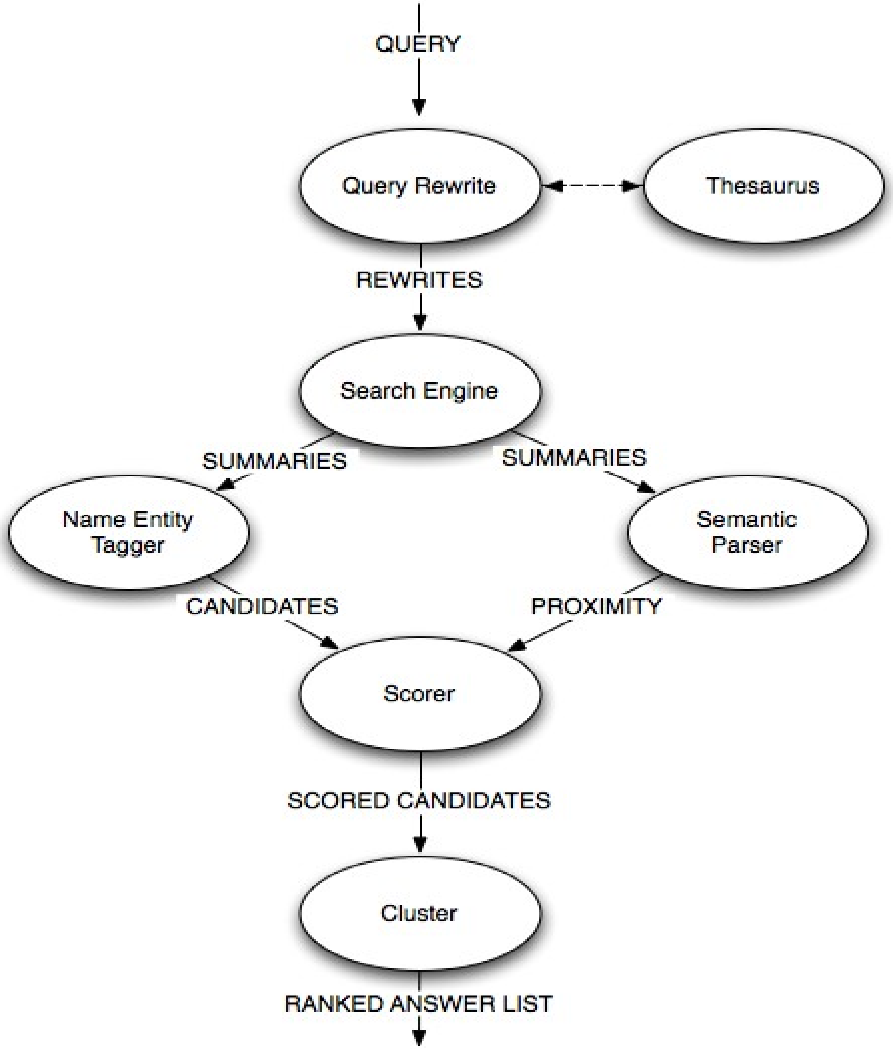
- **To find the answer to certain categories of factoid questions by exploiting the redundancy of the data available on the Internet**
- E.g. : "*Who teaches Data Mining at Stanford?*"

Question types:

- *Who*
- *Where*
- *When*

Also, *What time, How long, How much, How many ...*

System Overview



Previous Work - AskMSR

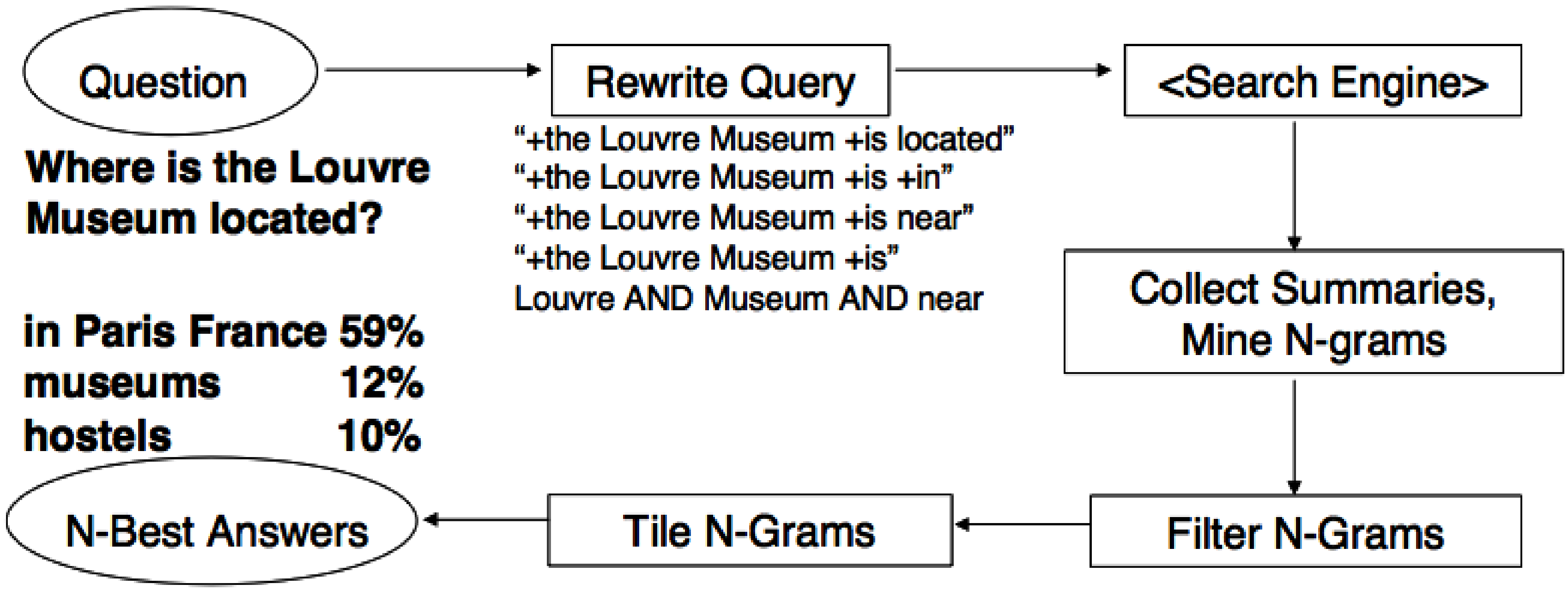

Question
Where is the Louvre Museum located?
Rewrite Query
"+the Louvre Museum +is located"
"+the Louvre Museum +is +in"
"+the Louvre Museum +is near"
"+the Louvre Museum +is"
Louvre AND Museum AND near
<Search Engine>
Collect Summaries, Mine N-grams
Filter N-Grams
Tile N-Grams
N-Best Answers
in Paris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

Figure 1. System Architecture

42 : New Features

- Semantic query rewriting
- Name Entity tagging to generate candidate answers
- Semantic distance metric
- Clustering of candidates rather than tiling
- Scoring Module
- Returning straight answers instead of paragraphs
- Multi-language leap ahead scenario

Semantic Distance

- Jaccard distance
  - A possible choice

- “Ad hoc” semantic distance (or, better, “proximity”)
  - Analyze the semantic structure of the question and the snippet answers
  - Discover the semantic part to retrieve (e.g. subject, passive complement, predicate, etc...)
  - Compute the semantic distance
  - Finer results

Semantic Distance

"Who killed John Lennon?"

- "John Lennon was brutally killed by **Mark Chapman**"
  Chapman's Proximity: 10
- "**Mark Chapman** killed the famous John Lennon…"
  Chapman's Proximity: 10
- "Mark Chapman, **who** killed John Lennon…"
  Chapman's Proximity: 7
- "Mark Chapman, the murder **who** killed John Lennon…".
  Chapman's Proximity: 6
- "While John Lennon was leaving his residence, Mark Chapman killed **him**…"
  Chapman's Proximity: 5

What else we tried

- Using rank of the page where the candidate came from in scoring.
- Averaging the score over all candidates in an answer
- Using a euclidean distance metric.

Results - Scores

	Who killed John Lennon?	Who was the second president of the USA	Who wrote Wuthering Heights?	Who discovered the New World?	Where is the Taj Mahal?	Where is the next World Cup?	Who teaches Data Mining at Stanford?
Rank1	mark david chapman	john quincy adams	emily bronte	john cabot	agra india	south africa	anand rajaraman
Score	129	55	155	16	226	138	36
Rank2	fenton bresler	michael bond	charlotte bronte	christopher columbus	chauk india	west germany	jeff ullman &amp; wei li
Score	12	4	121	14	165	34	24
Rank3	stephen king	thomas jefferson	jane bronte	amerigo vespucci	northern india	france	doug brutlag
Score	10	2	115	11	192	32	10

Results - Comparison

	Who killed John Lennon?	Who was the second president of the USA?	Who wrote Wuthering Heights?	Where is the Taj Mahal?	Where is the next World Cup?	Who teaches Data Mining at Stanford?
Returning Summaries						
LCC	Thomas Johnson	John Adams	Currer Bell	agra, India	Germany	Andreas Weigend
Ask	mark chapman	dont know	Emily Bront	Agra, India	Europe	Jimison
AnswerBus	mark david chapman	John Adams	Emily	India	dont know	dont know
Returning	Straight Answers					
42	mark david chapman	john quincy adams	emily bronte	agra india	South Africa	Anand Rajaraman
Start	dont know	John Adams	Bronte	India	dont know	dont know

Demo

Reference

- S. Dumais, M. Banko, E. Brill, J. Lin and A. Ng (2002). P. Bennett, S. Dumais and E. Horvitz (2002).

  Web question answering: Is more always better? *In Proceedings of SIGIR'02, Aug 2002, pp. 291-298.*

- E. Brill, S. Dumais and M. Banko (2002).

  An analysis of the AskMSR question-answering system. *In Proceedings of 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002).*