

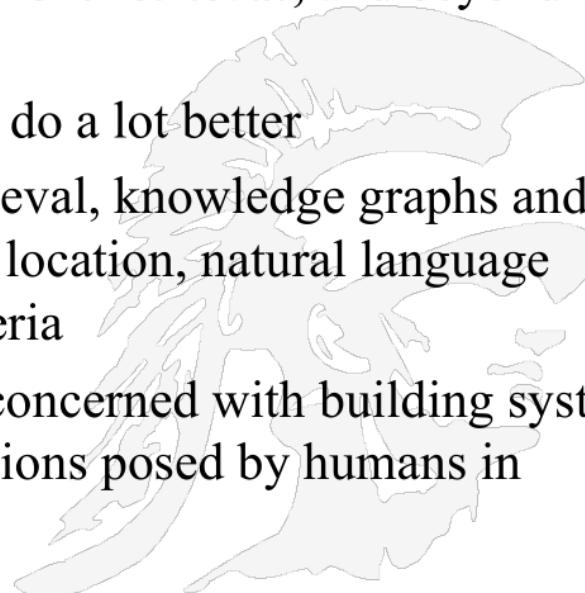
← 1/60 → *** 5:31:39

Question answering

..

Information Retrieval v. Question Answering

- The name “**information retrieval**” is standard, but as traditionally practiced, it’s not really right
- In the past all we got was ***document retrieval***, and beyond that the job is up to us
 - Modern search engines now do a lot better
- They combine information retrieval, knowledge graphs and inferencing, past query history, location, natural language processing and many other criteria
- **Question Answering (QA)** is concerned with building systems that automatically answer questions posed by humans in a natural language



..

People *want* to ask questions...

Examples from Ask.com query log

how much should I weigh

what does my name mean

how to get pregnant

where can I find pictures of hairstyles

who is the richest man in the world

what is the meaning of life

why is the sky blue

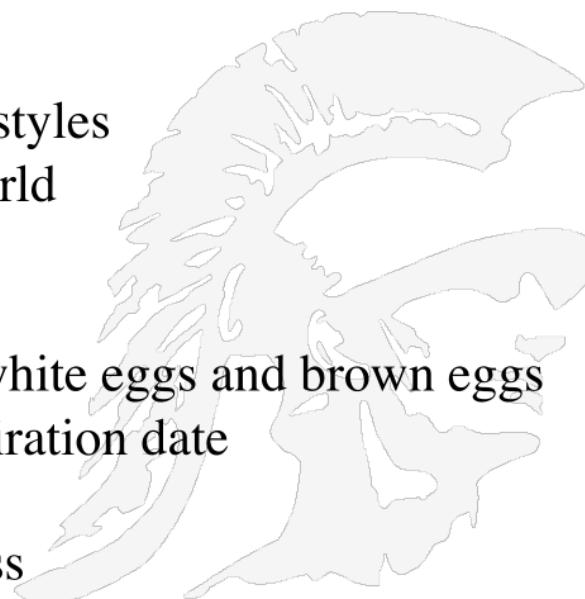
what is the difference between white eggs and brown eggs

can you drink milk after the expiration date

what is true love

what is the jonas brothers address

Around 10-20% of query logs are questions such as these



Copyright Ellis Horowitz 2011-2022

••

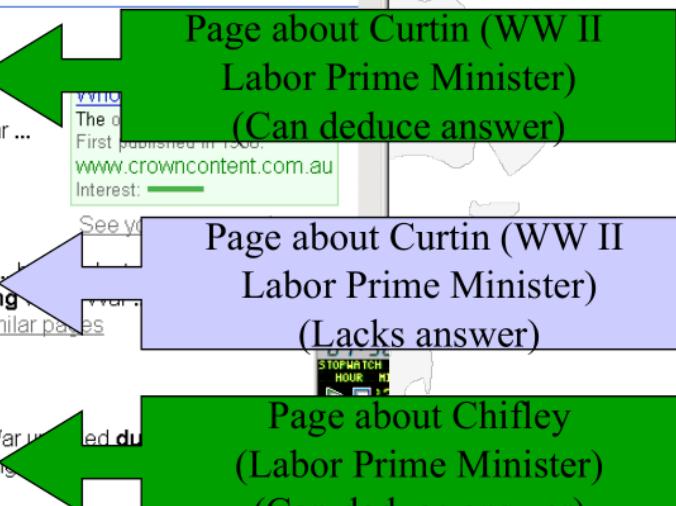
University of Southern California  **USC**

Question: *Who was the prime minister of Australia during the Great Depression?*
 Answer: James Scullin (Labor) 1929–31

Google Search: Who was the prime minister of Australia during the Great Depression? - Microsoft Internet Explorer

Address: +of+Australia+during+the+Great+Depression%3F&btnG=Google+Search | Google | Eat Depression? | Search Site | News

How Google used to respond to questions



The following words are very common and were not included in your search: **Who was the of the.** [details]

Web Images Groups Directory News
 Searched the web for **Who was the prime minister of Australia during the Great Depression?** Results 1 - 10
 Asking a question? Try out [Google Answers](#).

From Poor Boy to Prime Minister
 ... how did he come to lead **Australia during** World War ... April 1939 Menzies takes over as **Prime Minister** after the death of Lyons; Sept 3 1939 **Australia** declares war ...
john.curtin.edu.au/manofpeace/boytopm.html - 23k - Mar 1, 2003 - [Cached](#) - [Similar pages](#)

Activity: Banning of the Communist Party in World War II
 ... The **Great Depression** had brought enormous suffering to workers, ... the 'Prime Minister' and His ... the Communist Party in **Australia during** the 1930s. ... He became **Prime Minister** following Curtin's death, succeeding ...
john.curtin.edu.au/letters/activities/communism.html - 8k - [Cached](#) - [Similar pages](#)
 [More results from john.curtin.edu.au]

Prime Ministers of Australia - Chifley
 ... defying the federal United **Australia** Party government ... Second World War under led by the 1930s. ... He became **Prime Minister** following Curtin's death, succeeding ...
www.nma.gov.au/primeministers/3.htm - 30k - [Cached](#) - [Similar pages](#)

••

University of Southern California  USC

USC Viterbi School of Engineering

Question: *Who was the prime minister of Australia during the Great Depression?*

Answer: *James Scullin (Labor) 1929–31*

Google's result today

who was the prime minister of australia during the great depression

All Images News Videos Shopping More Settings Tools

About 9,390,000 results (0.72 seconds)

James Scullin became the new prime minister and Bruce lost his own seat of Flinders, the first sitting Australian prime minister to do so. However, on 24 October 1929, one week after Labor took power, the US stock market crashed.

www.nma.gov.au › defining-moments › resources › great-depression › Great Depression | National Museum of Australia

About Featured Snippets Feedback

en.wikipedia.org › wiki › Great_Depression_in_Australia - Wikipedia

Great Depression in Australia - Wikipedia

Australia suffered badly during the period of the Great Depression of the 1930s. ... The conservative Prime Minister of Australia, Stanley Bruce, wished to ... 1929–1935: Scullin and ... · Varying experiences of ... · Legacy of the Great ...

People also ask

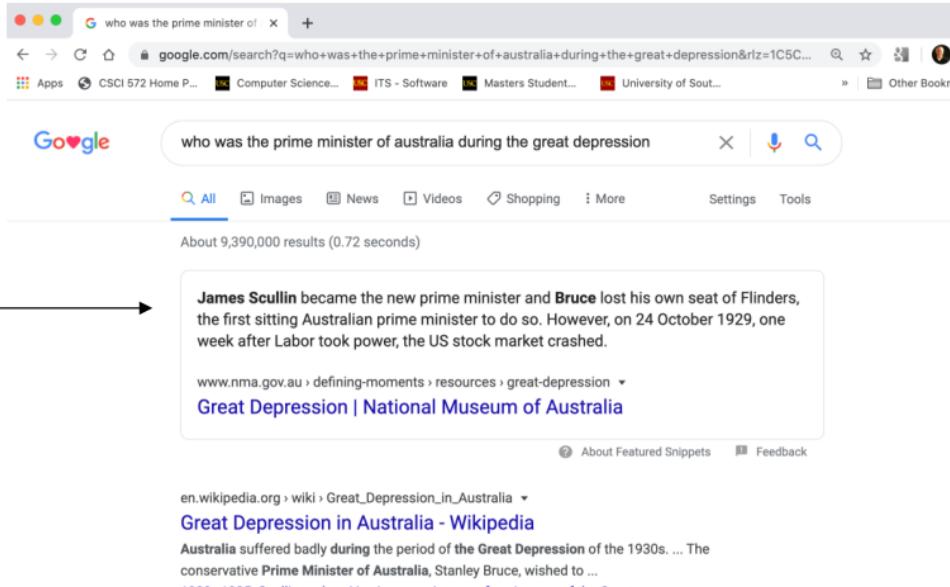
Who was the prime minister of Australia during ww2? ▾

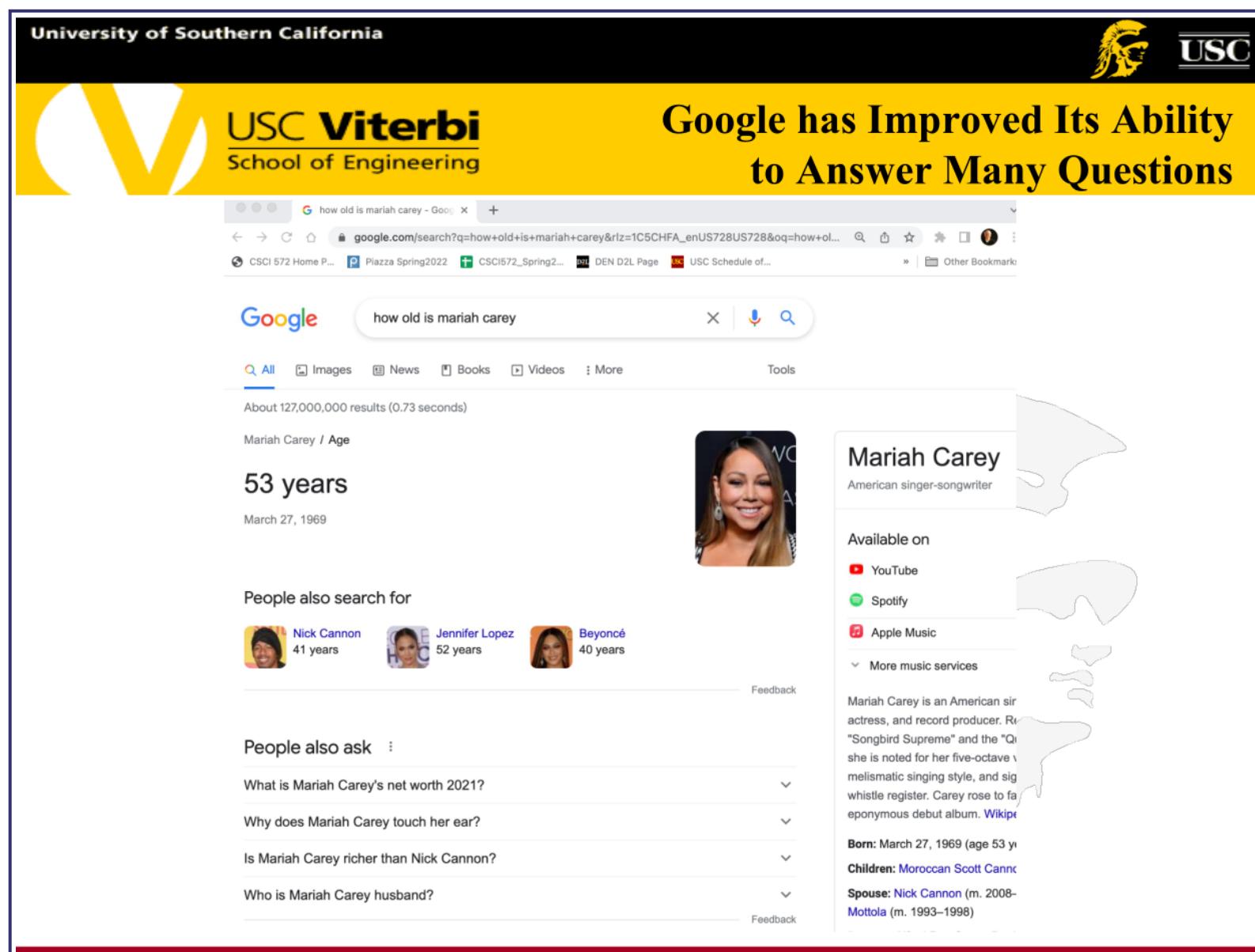
What areas of Australia were most affected by the Great Depression? ▾

Did the Great Depression affect Australia? ▾

Who was hit the hardest during the Great Depression? ▾

Feedback

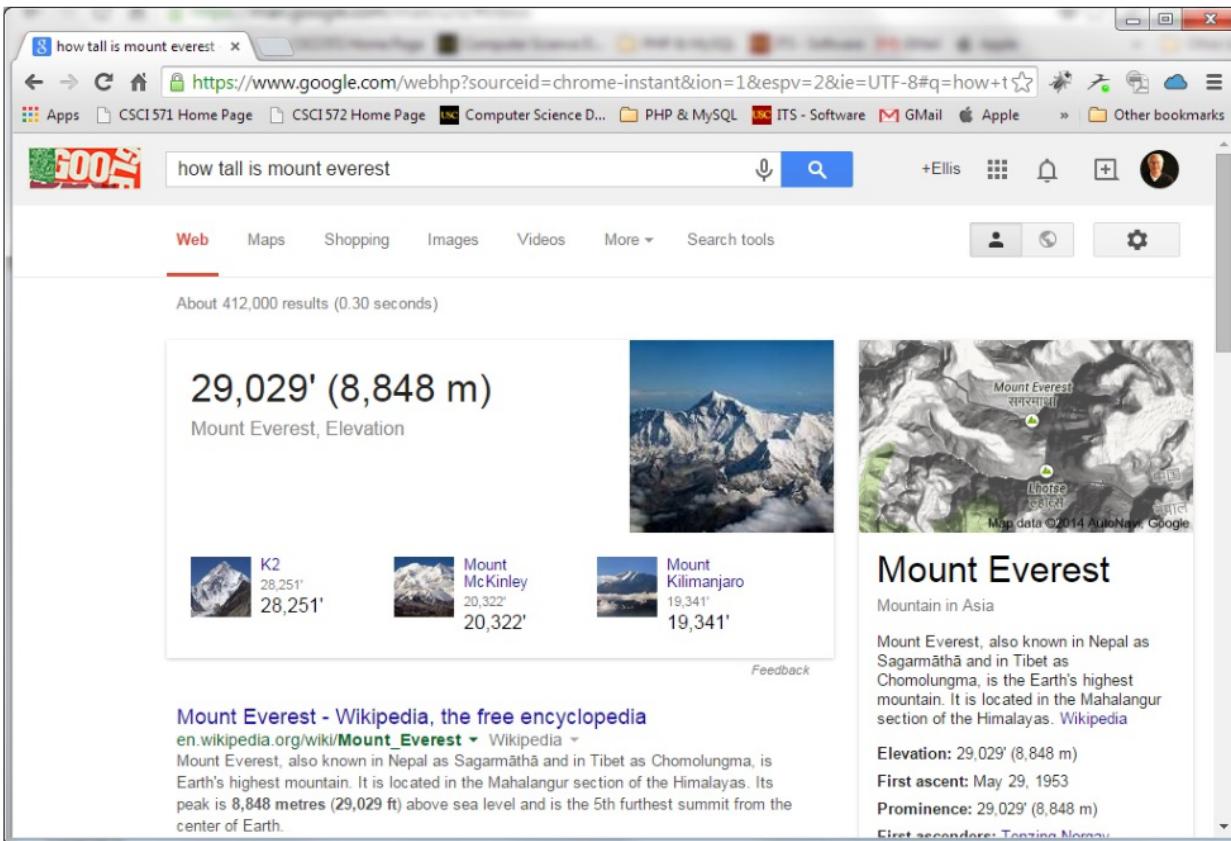




••

University of Southern California  USC

Some Questions are Easily Answered



The screenshot shows a Google search results page for the query "how tall is mount everest". The top result is a summary card for Mount Everest, stating its elevation as 29,029' (8,848 m). Below this, there are images of Mount Everest, K2, Mount McKinley, and Mount Kilimanjaro. To the right, there is a map of the Himalayas showing the location of Mount Everest. A detailed description of Mount Everest follows, including its name in various languages, its position in the Mahalangur section of the Himalayas, and its rank as the highest mountain. It also lists its elevation, first ascent date (May 29, 1953), and other notable facts.

how tall is mount everest

https://www.google.com/webhp?sourceid=chrome-instant&ion=1&espv=2&ie=UTF-8#q=how+t+mount+everest

Web Maps Shopping Images Videos More Search tools

About 412,000 results (0.30 seconds)

29,029' (8,848 m)
Mount Everest, Elevation

K2
28,251'
28,251'

Mount McKinley
20,322'
20,322'

Mount Kilimanjaro
19,341'
19,341'

Feedback

Mount Everest - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Mount_Everest • Wikipedia

Mount Everest, also known in Nepal as Sagarmāthā and in Tibet as Chomolungma, is Earth's highest mountain. It is located in the Mahalangur section of the Himalayas. Its peak is 8,848 metres (29,029 ft) above sea level and is the 5th furthest summit from the center of Earth.

Mount Everest

Mountain in Asia

Mount Everest, also known in Nepal as Sagarmāthā and in Tibet as Chomolungma, is the Earth's highest mountain. It is located in the Mahalangur section of the Himalayas. Wikipedia

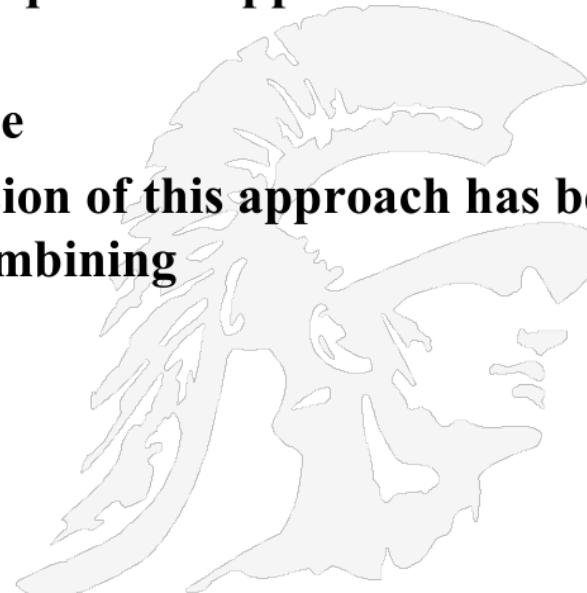
Elevation: 29,029' (8,848 m)
First ascent: May 29, 1953
Prominence: 29,029' (8,848 m)
First ascender: Tenzing Norgay

Copyright Ellis Horowitz 2011-2022

..

The Original Google Approach

- Take the question and try to find it as a string on the web
- Return the next sentence on that web page as the answer
- Works brilliantly if this exact question appears as a FAQ question, etc.
- Works poorly most of the time
- But a more sophisticated version of this approach has been introduced in recent years combining
 - Knowledge graph
 - N-grams
 - WordNet
 - NLP techniques



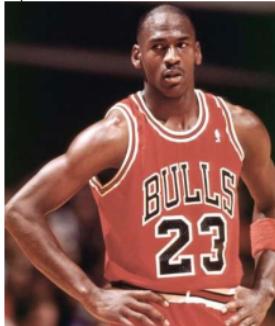
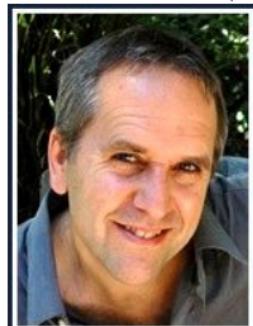
..

University of Southern California

USC Viterbi
School of Engineering

Many Questions Pose Semantic Difficulties

- Who is Michael Jordan?
 - Michael Jordan the basketball player or the Machine Learning guy?
- Key requirement is that entities get identified and disambiguated

Copyright Ellis Horowitz, 2011-2022

9

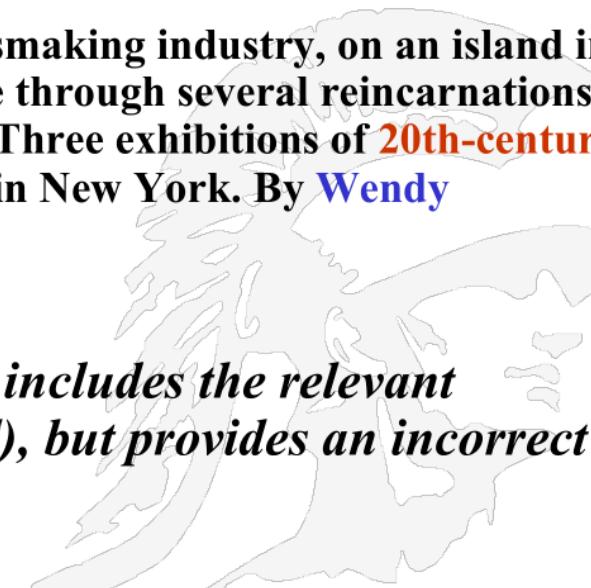
..

University of Southern California  USC

 USC **Viterbi**
School of Engineering

Why Natural Language Processing is Required

- **Question:** “When was Wendy’s founded?”
- **Passage candidate:**
 - “The renowned Murano glassmaking industry, on an island in the Venetian lagoon, has gone through several reincarnations since it was **founded** in 1291. Three exhibitions of **20th-century** Murano glass are coming up in New York. By **Wendy Moonan**.”
- **Answer:** **20th Century**
- *the candidate passage below includes the relevant keywords (Wendy's, founded), but provides an incorrect answer*



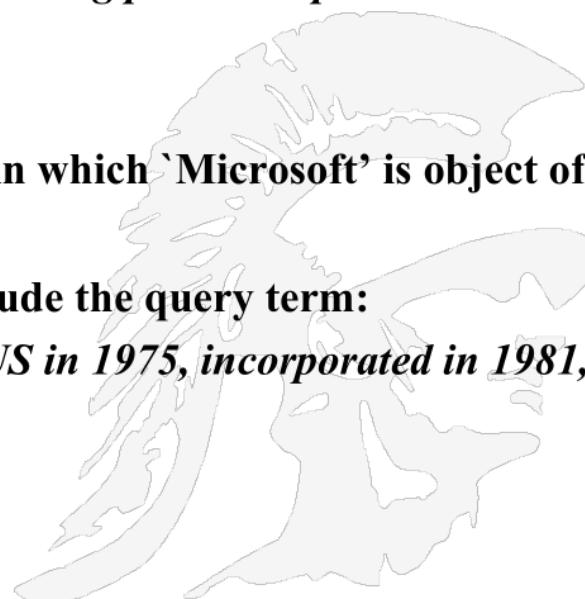
Copyright Ellis Horowitz, 2011-2022

10

..

More NLP Challenges Predicate-Argument Structure

- Q336: *When was Microsoft established?*
- Difficult because Microsoft tends to establish lots of things...
Microsoft plans to establish manufacturing partnerships in Brazil and Mexico in May.
- Need to be able to detect sentences in which 'Microsoft' is object of 'establish' or close synonym.
- A correct result might *not* even include the query term:
Microsoft Corp was founded in the US in 1975, incorporated in 1981, and established in the UK in 1982.



NLP: Natural language processing

••

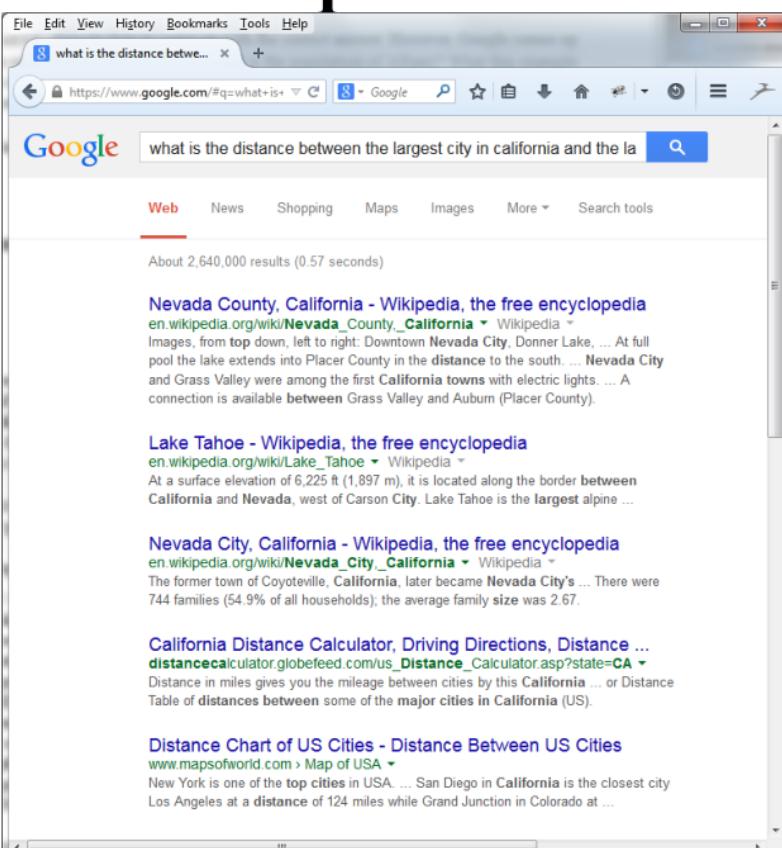
University of Southern California  USC

**Some Questions
Require Inferences**

What is the distance between the largest city in California and the largest city in Nevada?

Google does poorly on this query, misinterpreting Nevada as Nevada County, California

ps: Try the query in Google today to see if they have improved their answer



Copyright Ellis Horowitz 2011-2022

••

University of Southern California

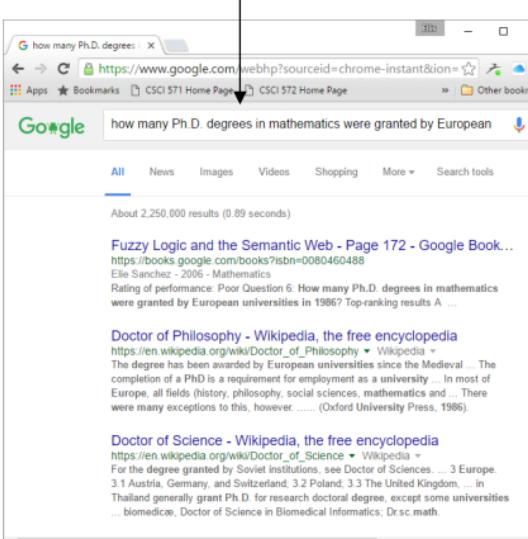
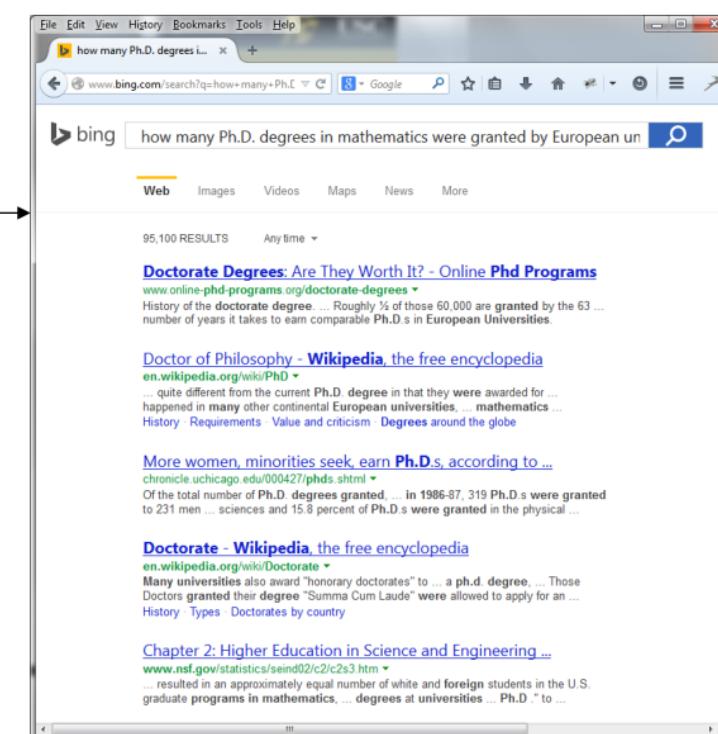
USC Viterbi
School of Engineering

**In Some Cases
the Data May Not Exist**

how many Ph.D. degrees in mathematics were granted by European universities in 1986?

All results are irrelevant →

a more recent result; still no relevant links

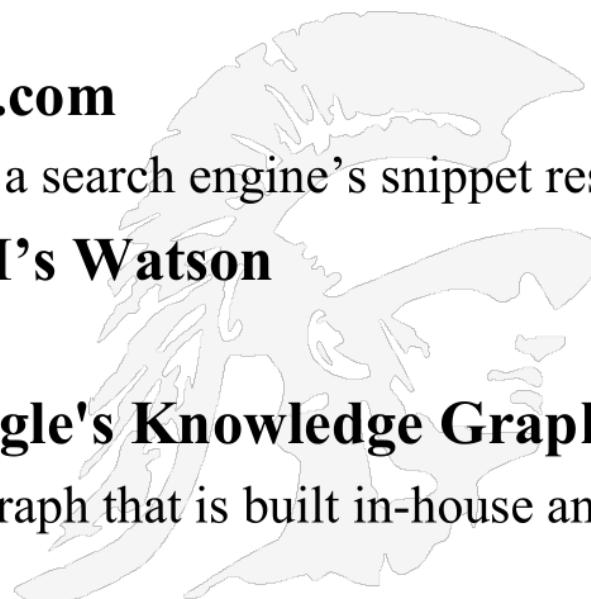



ght Ellis Horowitz 2011-2022

..

Some Popular Products Designed for Question/Answering

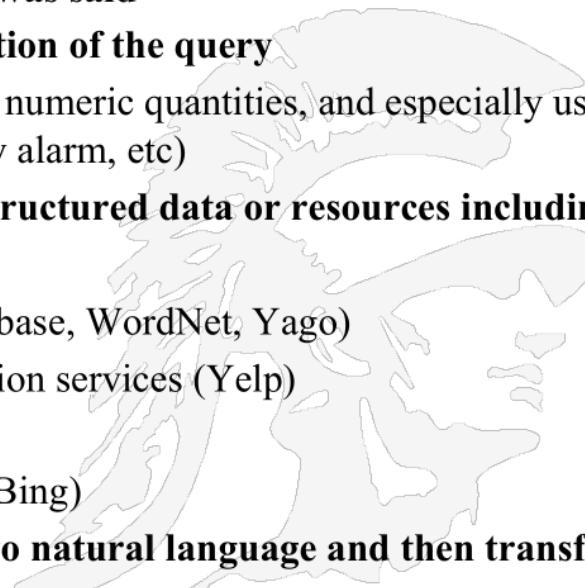
- **Approach 1: used by Siri**
 - map to known entities and use existing databases over the internet
- **Approach 2: used by Ask.com**
 - detect question type and use a search engine's snippet results
- **Approach 3: used by IBM's Watson**
 - combine approaches 1 and 2
- **Approach 4: used by Google's Knowledge Graph**
 - use an entity - relationship graph that is built in-house and infer the answer



..

Approach used by Siri: Knowledge-Based Approach

- Siri was begun as a DARPA project called CALO/PAL (Personalized Assistant that Learns)
- 1. First your voice query is put through a recognizer and a language model and Siri comes up with an interpretation of what was said
- 2. Second Siri builds a semantic representation of the query
 - Extract times, dates, locations, entities, numeric quantities, and especially user actions (e.g. schedule a meeting, set my alarm, etc)
- 3. Siri maps from this semantics to query structured data or resources including:
 - Geospatial databases
 - Ontologies (Wikipedia infoboxes, Freebase, WordNet, Yago)
 - Restaurant review sources and reservation services (Yelp)
 - Scientific databases (Wolfram Alpha)
 - Conventional search engines (Google, Bing)
- 4. Siri then transforms the output above into natural language and then transforms the text back to speech

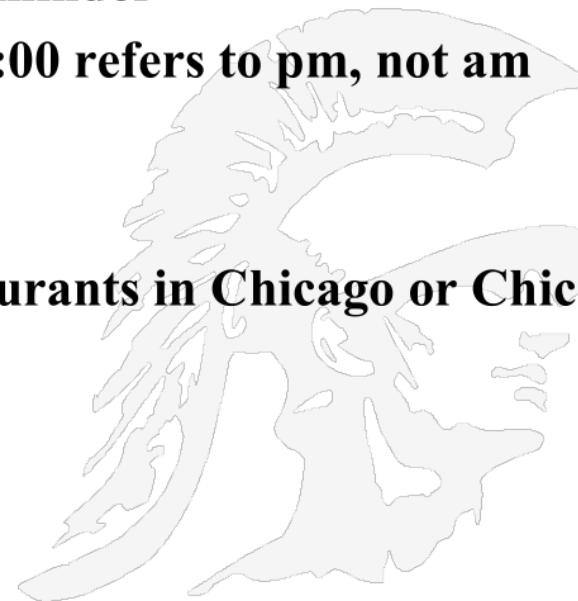


..



Context and Conversation in Virtual Assistants like Siri

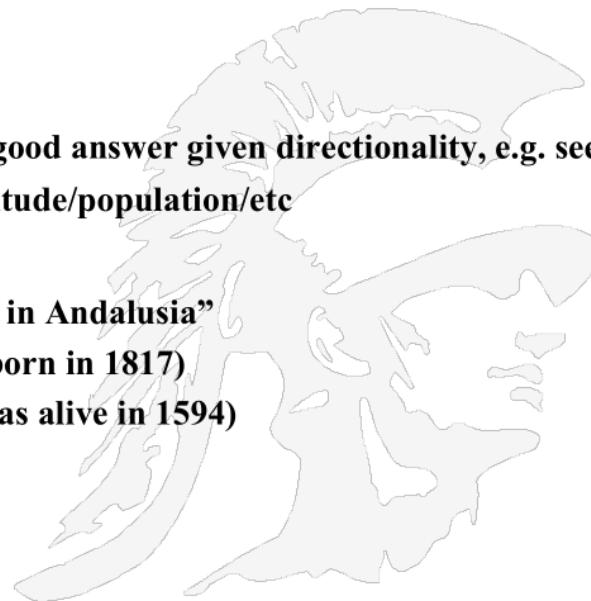
- Coreference helps resolve ambiguities
- U: “book a table at Il Fornaio at 7:00 with my mom”
- U: “also send her an email reminder”
- “her” refers to “my mom”; 7:00 refers to pm, not am
- Clarification questions:
- U: “chicago pizza”
- S: “Did you mean pizza restaurants in Chicago or Chicago-style pizza?”



..

CANDIDATE ANSWER SCORING IN IBM WATSON

- Each candidate answer gets scores from > 50 components
- From unstructured text, semi-structured text, triple stores
- Logical form (parse) match between question and candidate
- Passage source reliability
- Geospatial location
 - Denver is “southwest of Montana” is a good answer given directionality, e.g. see
 - geonames.org which gives latitude/longitude/population/etc
- Temporal relationships
 - “In 1594 he took a job as a tax collector in Andalusia”
 - Candidates: Thoreau is a bad answer (born in 1817)
 - Candidates: Cervantes is possible (he was alive in 1594)
- Taxonomic classification



••

University of Southern California

USC Viterbi
School of Engineering

AskJeeves (now Ask.com)

- **Earlier AskJeeves.com** was well-known as a search engine specializing in Questions/Answers
- Though it still exists, it performs far weaker than sites such as Google

How old is Mariah Carey
Snapshot taken 04/2022

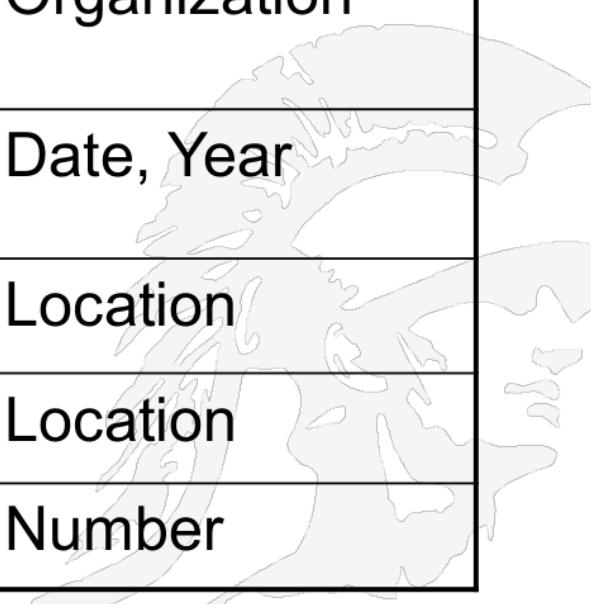
Copyright Ellis Horowitz 2011-2022

..



Question Types: Many Questions Fall into Distinct Categories

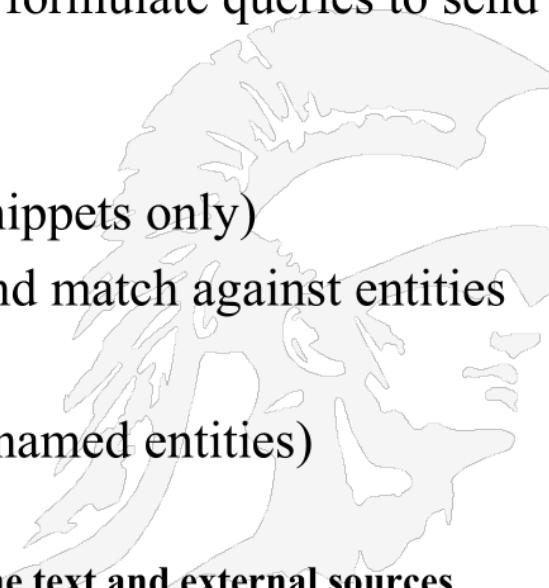
Who	Person, Organization
When	Date, Year
Where	Location
In What	Location
How many	Number



..

University of Southern California  USC

3 Main Phases for Question/Answering

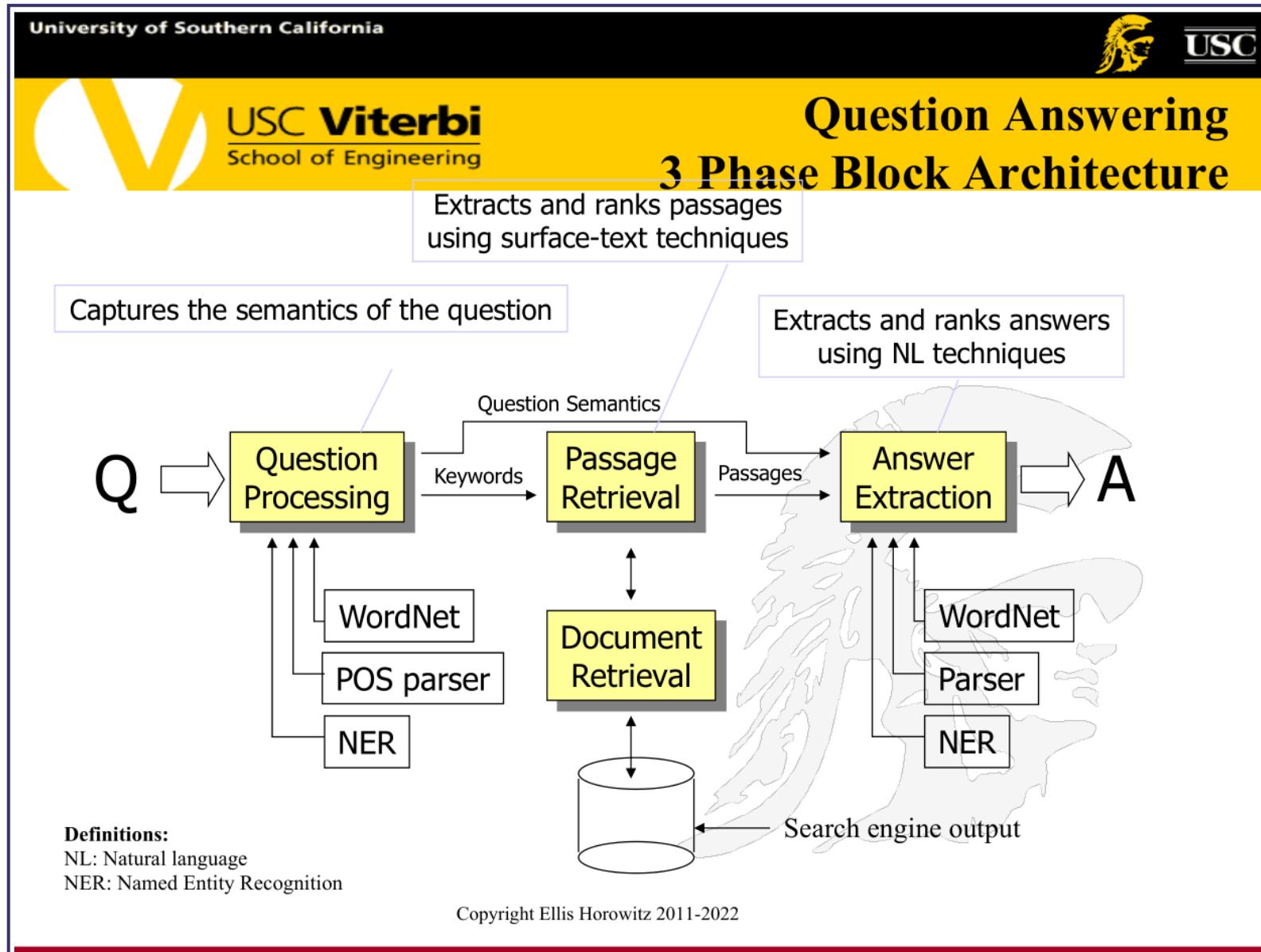


- 1. QUESTION PROCESSING**
 - Detect question type (who, what, when, where, etc)
 - Identify important entities and formulate queries to send to a search engine
- 2. PASSAGE RETRIEVAL**
 - Retrieve ranked documents (snippets only)
 - Break into suitable passages and match against entities
- 3. ANSWER PROCESSING**
 - Extract candidate answers (as named entities)
 - Rank candidates
 - using evidence from relations in the text and external sources

Copyright Ellis Horowitz, 2011-2022

20

..

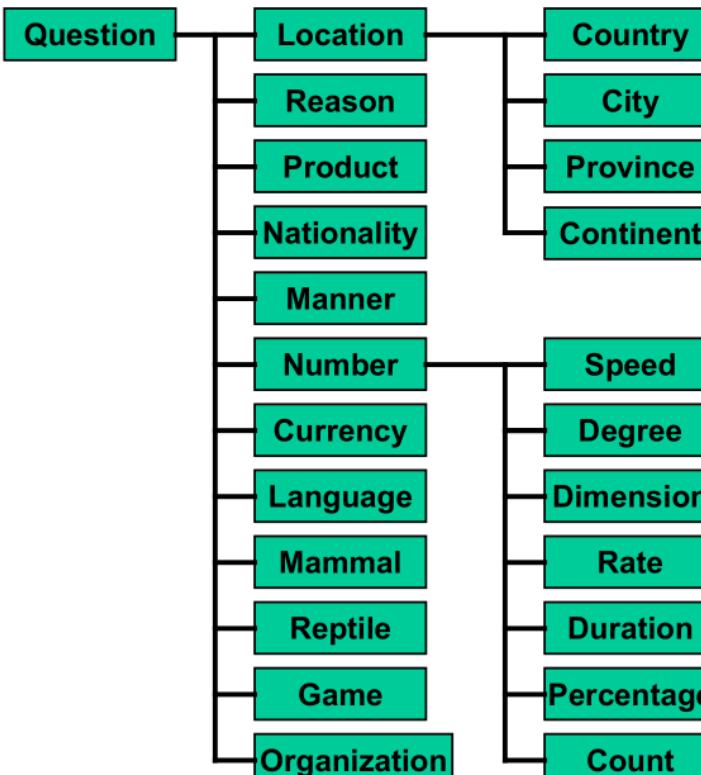


••

University of Southern California  USC

USC Viterbi
School of Engineering

Question Taxonomy



Researchers have tried to organize all question types, e.g.



Copyright Ellis Horowitz, 2011-2022

22

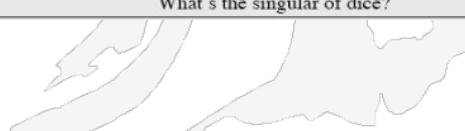
..

University of Southern California

USC Viterbi
School of Engineering

However Question Taxonomies Can Get Very Large

Tag	Example
ABBREVIATION	<p>abb exp</p> <p>What's the abbreviation for limited partnership? What does the "c" stand for in the equation E=mc2?</p>
DESCRIPTION	<p>definition description manner reason</p> <p>What are tannins? What are the words to the Canadian National anthem? How can you get rust stains out of clothing? What caused the Titanic to sink ?</p>
ENTITY	<p>animal body color creative currency disease/medicine event food instrument lang letter other plant product religion sport substance symbol technique term vehicle word</p> <p>What are the names of Odin's ravens? What part of your body contains the corpus callosum? What colors make up a rainbow ? In what book can I find the story of Aladdin? What currency is used in China? What does Salk vaccine prevent? What war involved the battle of Chapultepec? What kind of nuts are used in marzipan? What instrument does Max Roach play? What's the official language of Algeria? What letter appears on the cold-water tap in Spain? What is the name of King Arthur's sword? What are some fragrant white climbing roses? What is the fastest computer? What religion has the most members? What was the name of the ball game played by the Mayans? What fuel do airplanes use? What is the chemical symbol for nitrogen? What is the best way to remove wallpaper? How do you say " Grandma " in Irish? What was the name of Captain Bligh's ship? What's the singular of dice?</p>



Copyright Ellis Horowitz, 2011-2022

23

••

University of Southern California

USC Viterbi
School of Engineering

USC

**More Question Types
and Examples**

HUMAN	
description	Who was Confucius?
group	What are the major companies that are part of Dow Jones?
ind	Who was the first Russian astronaut to do a spacewalk?
title	What was Queen Victoria's title regarding India?
LOCATION	
city	What's the oldest capital city in the Americas?
country	What country borders the most others?
mountain	What is the highest peak in Africa?
other	What river runs through Liverpool?
state	What states do not have state income tax?
NUMERIC	
code	What is the telephone number for the University of Colorado?
count	About how many soldiers died in World War II?
date	What is the date of Boxing Day?
distance	How long was Mao's 1930s Long March?
money	How much did a McDonald's hamburger cost in 1963?
order	Where does Shanghai rank among world cities in population?
other	What is the population of Mexico?
period	What was the average life expectancy during the Stone Age?
percent	What fraction of a beaver's life is spent swimming?
speed	What is the speed of the Mississippi River?
temp	How fast must a spacecraft travel to escape Earth's gravity?
size	What is the size of Argentina?
weight	How many pounds are there in a stone?

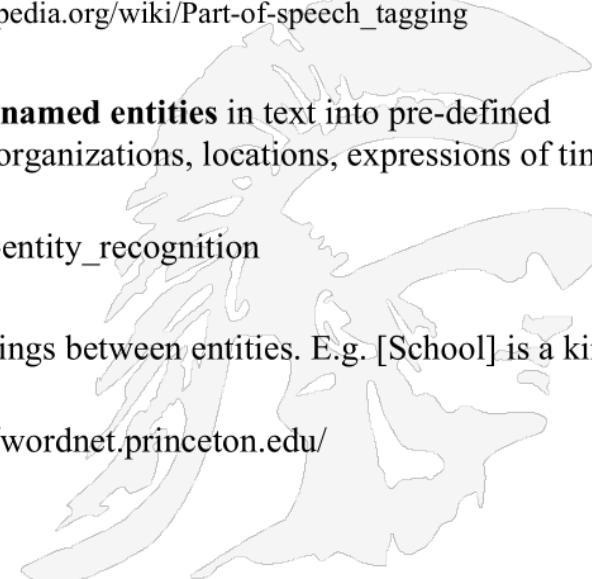
Copyright Ellis Horowitz, 2011-2022

24

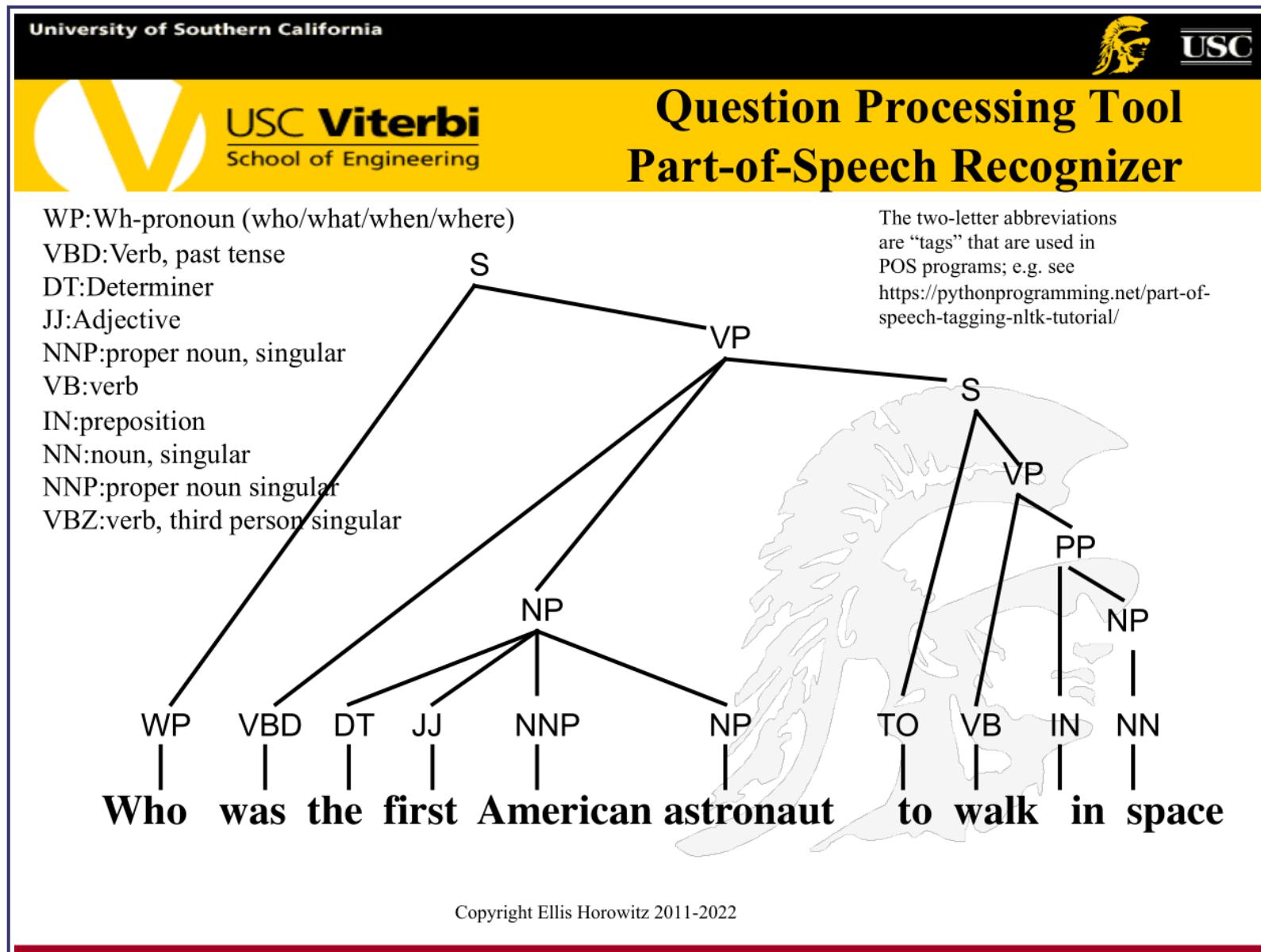
..

Some General Capabilities for Question-Answering Systems

- **Part-of-Speech Tagging**
 - a piece of software that reads text in some language and assigns **parts of speech** to each word, such as noun, verb, adjective, etc.
 - Markov Models are now the standard method for part-of-speech assignment
 - Some current major algorithms for part-of-speech tagging include the Viterbi algorithm, Brill tagger, and Baum-Welch algorithm, see https://en.wikipedia.org/wiki/Part-of-speech_tagging
- **Named Entity Extraction**
 - Software that seeks to locate and classify **named entities** in text into pre-defined categories such as the **names** of persons, organizations, locations, expressions of times, quantities ...
 - See https://en.wikipedia.org/wiki/Named-entity_recognition
- **Determining Semantic Relations**
 - **semantic relations** are concepts or meanings between entities. E.g. [School] is a kind of [educational institution]
 - Opportunity to use WordNet, <https://wordnet.princeton.edu/>
- **Dictionaries/Thesauri**



..



••

University of Southern California  USC

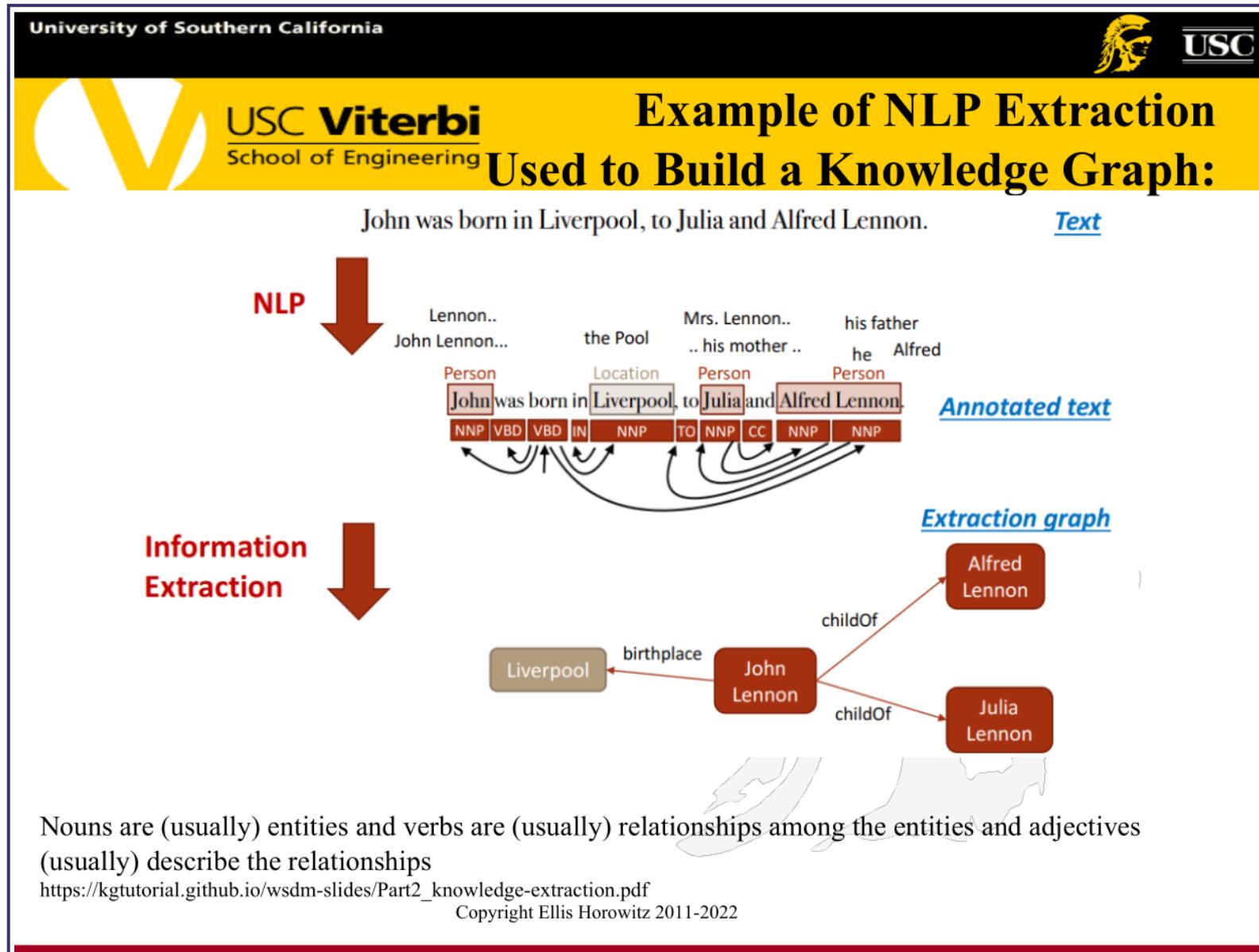
**Question Processing Tool
Named Entity Recognizer Example**

contentSkip to site indexPoliticsSubscribeLog InSubscribeLog InToday's PaperAdvertisementSupported ORG byF.B.I. Agent Peter Strzok PERSON , Who Criticized Trump PERSON in Texts, Is FiredImagePeter Strzok, a top F.B.I. GPE counterintelligence agent who was taken off the special counsel investigation after his disparaging texts about President Trump PERSON were uncovered, was fired. CreditT.J. Kirkpatrick PERSON for The New York TimesBy Adam Goldman ORG and Michael S. SchmidtAug PERSON . 13 CARDINAL , 2018WASHINGTON CARDINAL — Peter Strzok PERSON , the F.B.I. GPE senior counterintelligence agent who disparaged President Trump PERSON in inflammatory text messages and helped oversee the Hillary Clinton PERSON email and Russia GPE investigations, has been fired for violating bureau policies, Mr. Strzok PERSON 's lawyer said Monday DATE .Mr. Trump and his allies seized on the texts — exchanged during the 2016 DATE campaign with a former F.B.I. GPE lawyer, Lisa Page — in PERSON assailing the Russia GPE investigation as an illegitimate "witch hunt." Mr. Strzok PERSON , who rose over 20 years DATE at the F.B.I. GPE to become one of its most experienced counterintelligence agents, was a key figure in the early months DATE of the inquiry.Along with writing the texts, Mr. Strzok PERSON was accused of sending a highly sensitive search warrant to his personal email account.The F.B.I. GPE had been under immense political pressure by Mr. Trump PERSON to dismiss Mr. Strzok PERSON , who was removed last summer DATE from the staff of the special counsel, Robert S. Mueller III PERSON . The president has repeatedly denounced Mr. Strzok PERSON in posts on

- The process of recognizing information units like names, including persons, organizations, location names, and numeric expressions including time, date, money and percent expressions from unstructured text.
- This is an example of supervised learning as training sets are first created
- See <https://nlp.stanford.edu/software/> for a Java program and explanation

Copyright Ellis Horowitz 2011-2022

..



..

NER Example and Its Translation Jeopardy Example

The Jeopardy query:

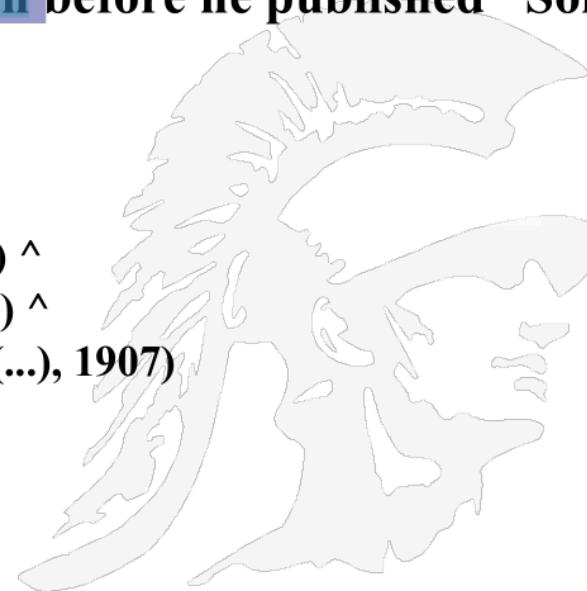
Category: Poets and Poetry: *GEO*

He was a bank clerk in the Yukon before he published “Songs of a Sourdough” in 1907.

YEAR

Produces the logic formula:

authorof(focus,“Songs of a Sourdough”) ^
publish (e1, he, “Songs of a Sourdough”) ^
in (e2, e1, 1907) ^ temporallink(publish(...), 1907)



..

University of Southern California

USC Viterbi
School of Engineering

Extracting Candidate Answers from Triple Stores

- Once we extract a relation from the question, e.g.
... he published “Songs of a sourdough”
(author-of ?x “Songs of a sourdough”)
- Many information sources support querying via a triple store
 - Wikipedia infoboxes, DBpedia, FreeBase, etc.
 - author-of(“Songs of a Sourdough”, “Robert Service”)

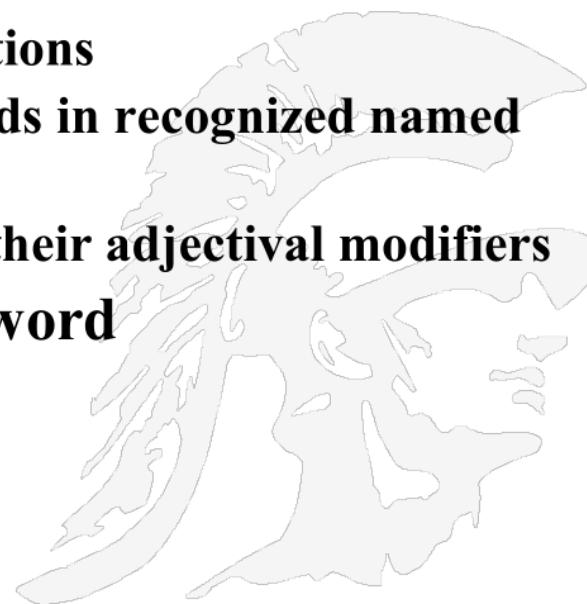


Copyright Ellis Horowitz, 2021-2022

..

General Keyword Selection Algorithm

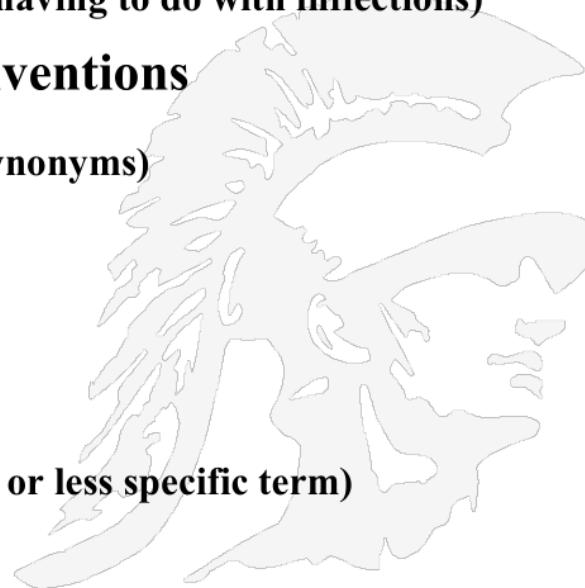
- 1. Use the part-of-speech recognizer to identify all
 - nouns
 - verbs
 - non-stopwords in quotations
 - NNP (proper noun) words in recognized named entities
 - complex nominals with their adjectival modifiers**
- 2. Select the answer type word**



..

Expanding the Keyword Set Using Variants

- There are 3 distinct ways to expand the keyword set determined by the keyword selection algorithm
- Morphological variants (having to do with inflections)
 - invented → inventor → inventions
- Lexical variants (similar to synonyms)
 - killer → assassin
 - far → distance
- Semantic variants
 - like → prefer (a more specific or less specific term)



..

How to Incorporate Lexical Variants Using Hypernins and Hyponims

Question: When was the internal combustion engine invented?

Answer: The first internal combustion engine was built in 1867.

Lexical chains:

- (1) invent:v#1 → HYPERNIM → create_by_mental_act:v#1 → HYPERNIM → create:v#1 → HYPONIM → build:v#1

Question: How many chromosomes does a human zygote have?

Answer: 46 chromosomes lie in the nucleus of every normal human cell.

Lexical chains:

- (1) zygote:n#1 → HYPERNIM → cell:n#1 → HAS.PART → nucleus:n#1

WordNet provides hypernims and hyponims

Hypernym is a superordinate grouping which includes subordinate groups

e.g. a musical instrument is a hypernym of guitar;

Hyponim is a word or phrase whose semantics is more specific than its hypernym

e.g. purple is a hyponim of color

..

University of Southern California  USC

USC Viterbi
School of Engineering

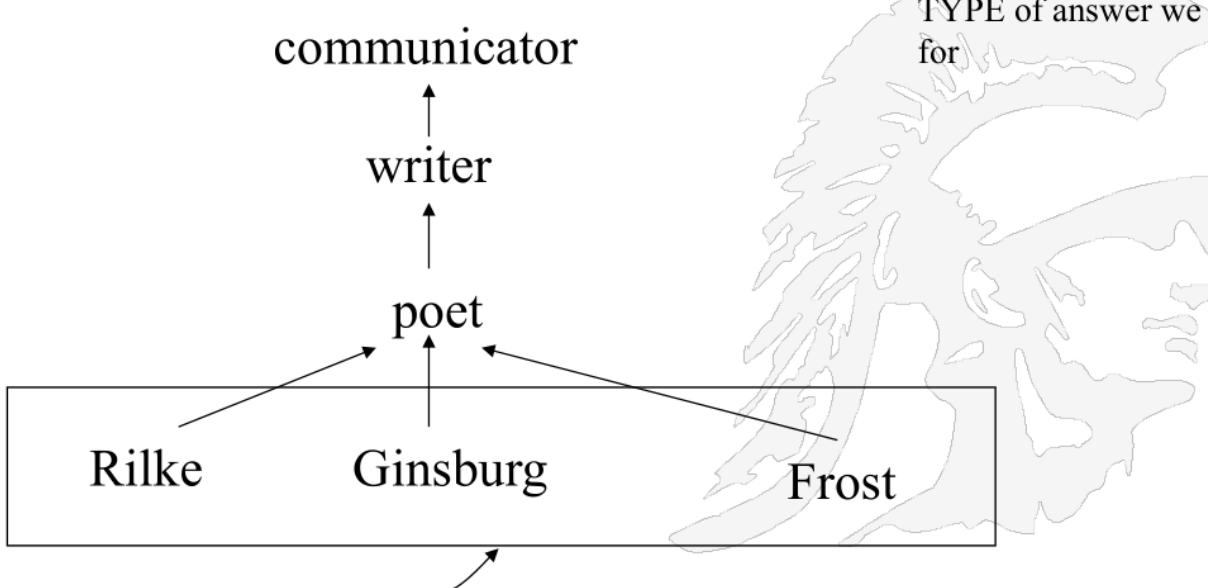
Use WordNet for Type Identification

We have already seen the use of WordNet, a lexical database of English nouns, verbs, adjectives, adverbs

“What 20th century poet wrote Howl?”

WordNet permits refinement of poet to specific instances

WordNet also helps determine the TYPE of answer we are looking for



communicator
writer
poet
Rilke Ginsburg Frost

Original keyword candidate set

Copyright Ellis Horowitz, 2011-2022

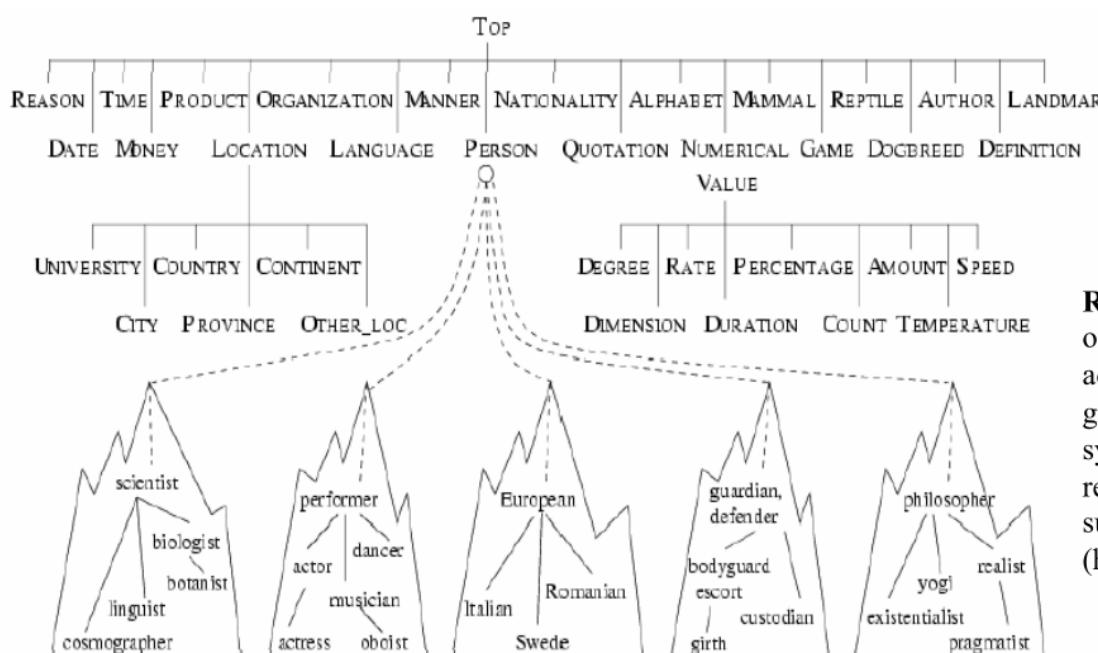
34

..

University of Southern California  USC

Answer Type Taxonomy

- Use WordNet to merge named entities with the WordNet hierarchy



Recall: WordNet is a database of English nouns, verbs, adjectives and adverbs grouped into sets of synonyms (synsets) and relations showing super/subordinate relations (hyperonymy/hyponymy)

If you know the answer should be a person
WordNet helps determine what sort of person

Copyright Ellis Horowitz, 2011-2022

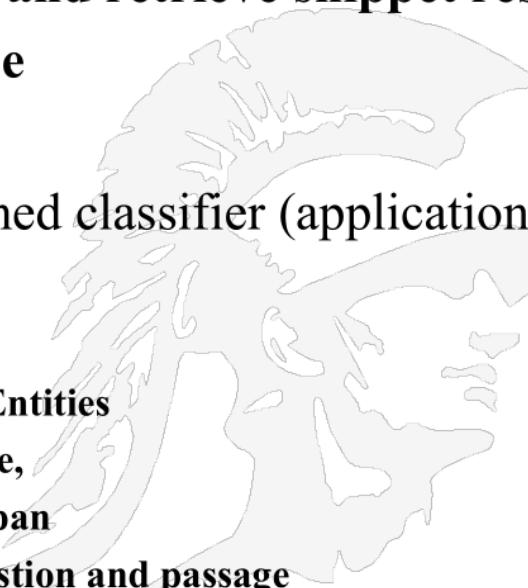
35

..



Part 2: Passage Retrieval

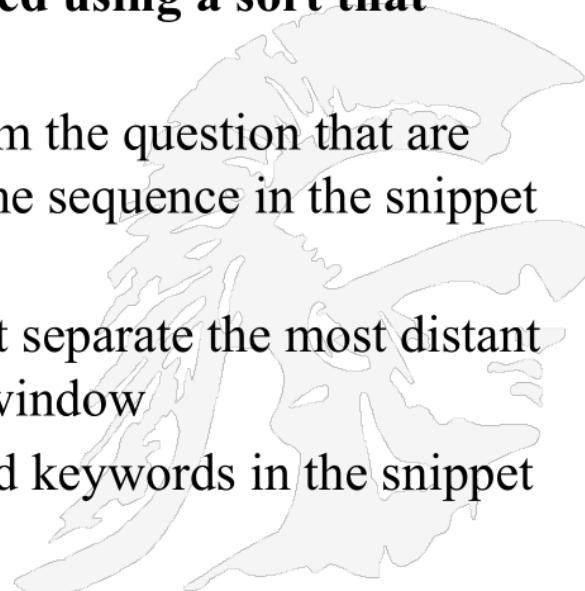
- Once we have formulated queries using tools like NER, POS, variant expansion and WordNet results
- Send queries to a search engine and retrieve snippet results
- Filter the results for correct type
 - use answer type classification
 - Rank passages based on a trained classifier (application of machine learning)
 - Features:
 - Question keywords, Named Entities
 - Longest overlapping sequence,
 - Shortest keyword-covering span
 - N-gram overlap between question and passage



..

Passage Scoring Method

- Focus on the snippets that are returned, the answers must be extracted from them
- Passage ordering is performed using a sort that involves three scores:
 1. The number of words from the question that are recognized and in the same sequence in the snippet window
 2. The number of words that separate the most distant keywords in the snippet window
 3. The number of unmatched keywords in the snippet window



..

University of Southern California  USC

Ranking Candidate Answers

Q066: Name the first private citizen to fly in space.

- Answer type: Person
 - Text passage:
"Among them was Christa McAuliffe, the first private citizen to fly in space. Karen Allen, best known for her starring role in "Raiders of the Lost Ark", plays McAuliffe. Brian Kerwin is featured as shuttle pilot Mike Smith..."

Scoring
There are five words from the question "the first private citizen space"
The answer is adjacent to "the first private citizen. . ." "
There are no unmatched keywords in "the first private citizen. . ." "



Copyright Ellis Horowitz, 2011-2022

38

..

Ranking Candidate Answers

Q066: Name the first private citizen to fly in space.

■ Answer type: Person

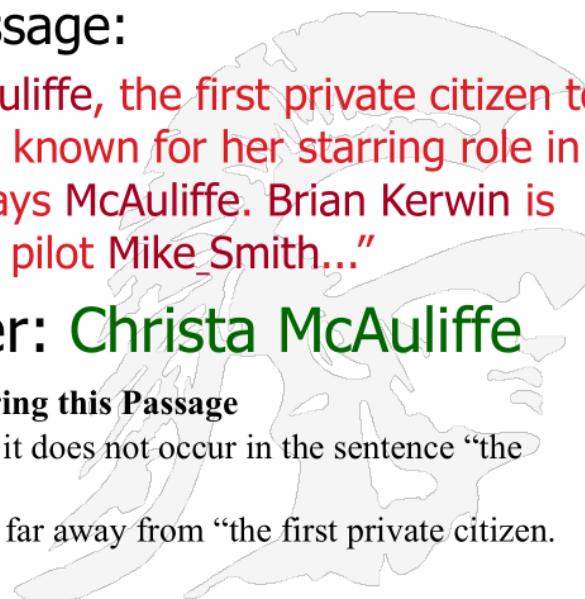
■ Text passage:

“Among them was Christa McAuliffe, the first private citizen to fly in space. Karen Allen, best known for her starring role in “Raiders of the Lost Ark”, plays McAuliffe. Brian Kerwin is featured as shuttle pilot Mike Smith...”

■ Best candidate answer: Christa McAuliffe

Comments on Scoring this Passage

- Karen Allen is rejected as an answer as it does not occur in the sentence “the first private citizen...”
- Brian Kerwin is rejected as the name is far away from “the first private citizen. . .”

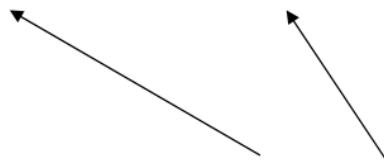


..

Local Alignment Example (1 of 7)

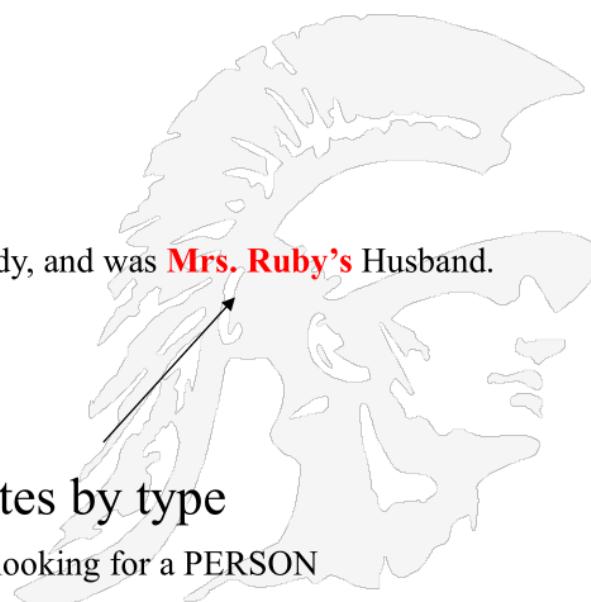
Who shot Kennedy?

Jack assassinated **Oswald**, the man who shot Kennedy, and was **Mrs. Ruby's** Husband.



Three Potential Candidates by type

WHO indicates we are looking for a PERSON



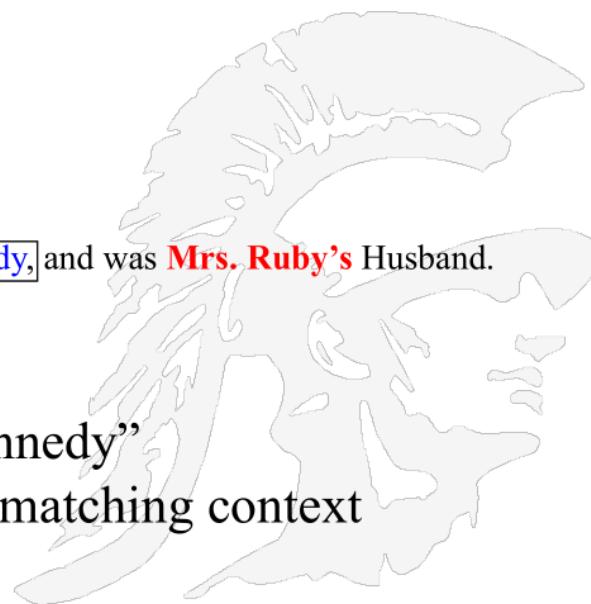
..

Local Alignment Example (2 of 7)

Question

Head → ***Who shot Kennedy?***
word

Jack assassinated **Oswald**, the man who **shot Kennedy**, and was **Mrs. Ruby's** Husband.



“shot Kennedy”
gives us a verb and matching context

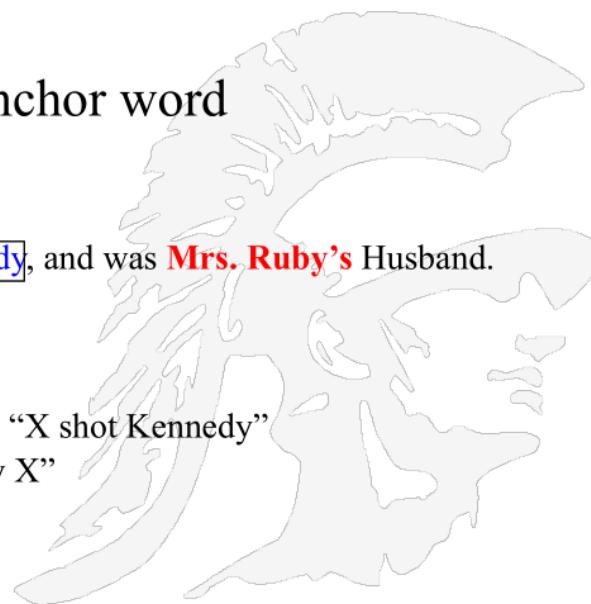
••

Local Alignment Example (3 of 7)

Who shot Kennedy?

Anchor word

Jack assassinated **Oswald**, the man who shot **Kennedy**, and was **Mrs. Ruby's** Husband.



Look for phrases such as “X shot Kennedy”
or “Kennedy was shot by X”

••

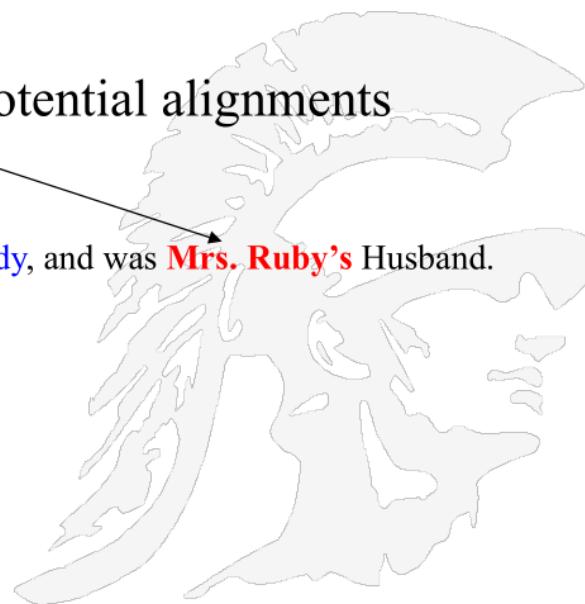
Local Alignment (4 of 7)

In principle it can be anyone of the three people identified

Who shot Kennedy?

Potential alignments

Jack assassinated **Oswald**, the man who shot **Kennedy**, and was **Mrs. Ruby's** Husband.



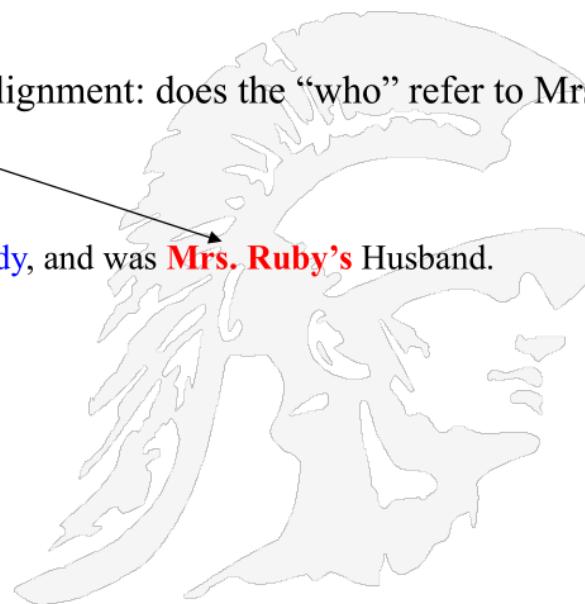
..

Local Alignment Example (5 of 7)

Who shot Kennedy?

One Alignment: does the “who” refer to Mrs. Ruby?

Jack assassinated **Oswald**, the man who shot **Kennedy**, and was **Mrs. Ruby’s** Husband.



Three Alignment Features :

••

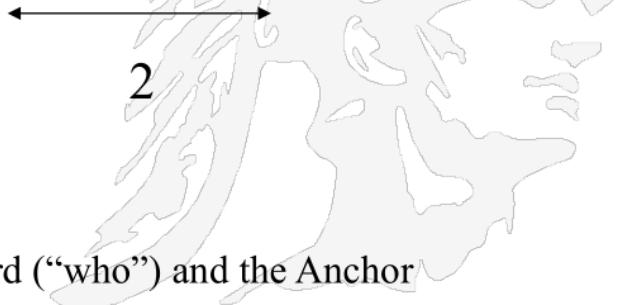
Local Alignment Example (6 of 7)

1
↔
Who shot Kennedy?

The distance between the question head word “who” and the anchor word Kennedy is 1

One Alignment : does the “who” refer to Mrs. Ruby?
The distance from Kennedy to Mrs. Ruby

Jack assassinated Oswald, the man who shot Kennedy, and was Mrs. Ruby's Husband.

2
↔

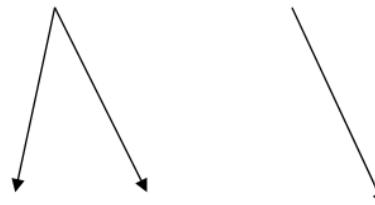
Three Alignment Features :

1. Distance between Question Head word (“who”) and the Anchor word (“Kennedy”) in the sentence

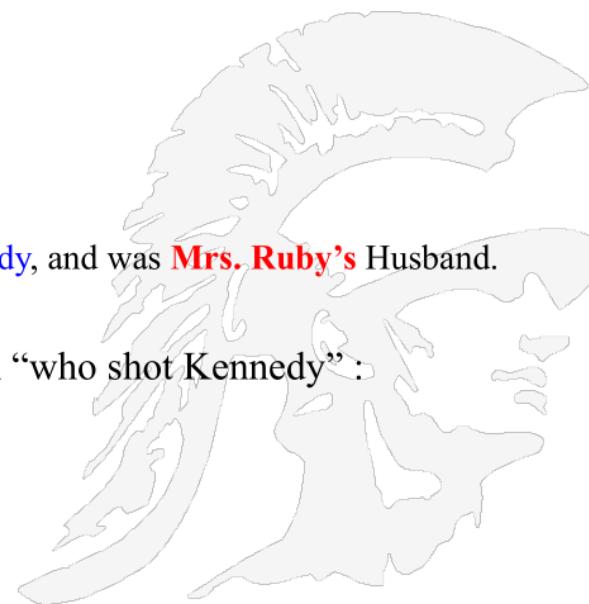
••

Local Alignment Example (7 of 7)

Who shot Kennedy?



Jack assassinated **Oswald**, the man who shot **Kennedy**, and was **Mrs. Ruby's** Husband.

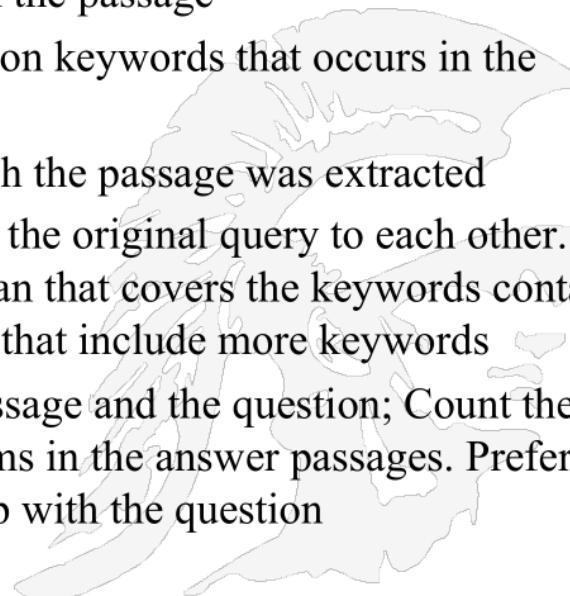


Oswald is properly aligned with “who shot Kennedy” :

..

A Refined Ranking Scheme

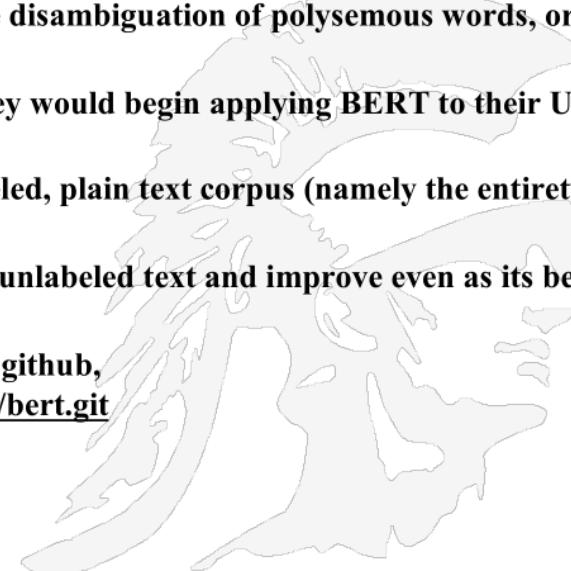
- **Refining the Passage Scoring Method, we can use supervised machine learning to rank the candidate passages according to six criteria**
 1. The number of named entities of the right type in the passage
 2. The number of question keywords in the passage
 3. The longest exact sequence of question keywords that occurs in the passage
 4. The rank of the document from which the passage was extracted
 5. The proximity of the keywords from the original query to each other. For each passage identify the shortest span that covers the keywords contained in that passage. Prefer smaller spans that include more keywords
 6. The N-gram overlap between the passage and the question; Count the N-grams in the question and the N-grams in the answer passages. Prefer the passages with higher N-gram overlap with the question



..

What is BERT

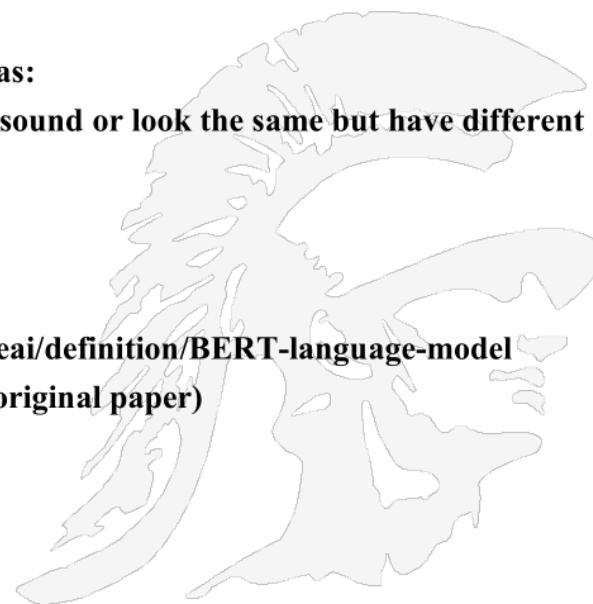
- **Bidirectional Encoder Representations from Transformers**
- In 2018, Google introduced and open-sourced BERT
 - BERT is designed to help computers understand the meaning of ambiguous language in text by using surrounding text to establish context
 - It achieved strong results on problems such as sentiment analysis, semantic role labeling, sentence classification and the disambiguation of polysemous words, or words with multiple meanings
- In October 2019, Google announced that they would begin applying BERT to their United States based production search algorithms.
- BERT was pre-trained using only an unlabeled, plain text corpus (namely the entirety of the English Wikipedia).
- It continues to learn unsupervised from the unlabeled text and improve even as its being used in practical applications
- the BERT Github repository is available on [github](https://github.com/google-research/bert.git),
`git clone https://github.com/google-research/bert.git`



••

What is BERT used for

- Sequence-to-sequence based language generation tasks such as:
 - Question answering
 - Abstract summarization
 - Sentence prediction
 - Conversational response generation
- Natural language understanding tasks such as:
 - Polysemy and Coreference (words that sound or look the same but have different meanings) resolution
 - Word sense disambiguation
 - Natural language inference
 - Sentiment classification
- <https://www.techtarget.com/searchenterpriseai/definition/BERT-language-model>
- <https://csci572.com/papers/BERT.pdf> (the original paper)



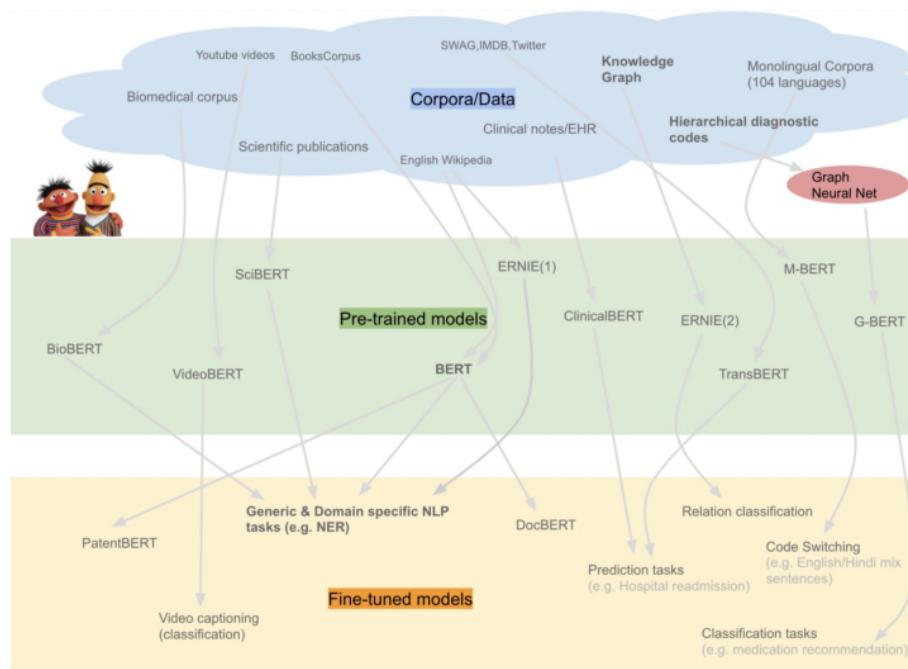
••

University of Southern California  **USC**

USC Viterbi
School of Engineering

BERT Has Many Pre-Trained Models

- BERT encoders have larger feedforward networks (768 and 1024 nodes in Base and Large respectively) and more attention heads (12 and 16 respectively).
- BERT was trained on Wikipedia and other datasets
- To the right you can see a diagram of additional variants of BERT pre-trained on specialized corpora



<https://towardsdatascience.com/bert-for-dummies-step-by-step-tutorial-fb90890ffe03>

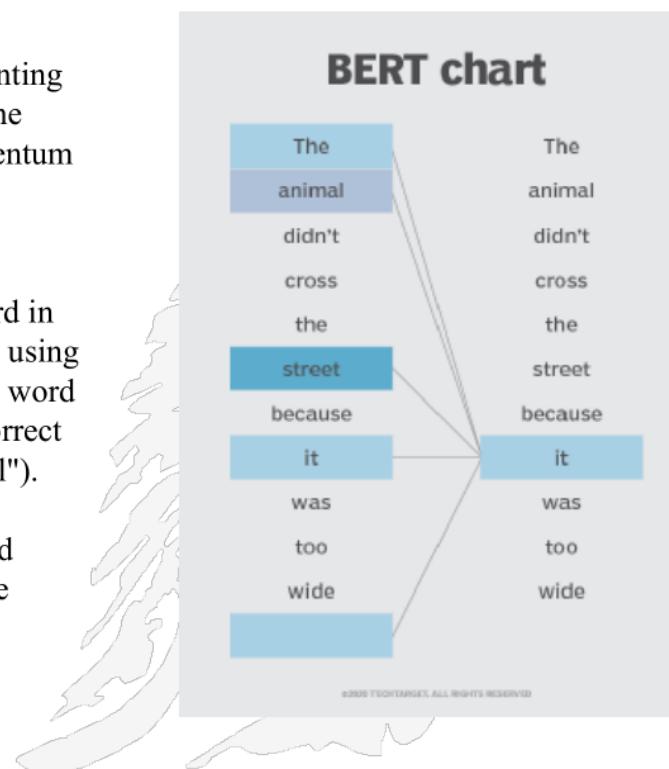
Copyright Ellis Horowitz, 2011-2022

50

••

How BERT works

- BERT functions by reading bidirectionally, accounting for the effect of all other words in a sentence on the focus word and eliminating the left-to-right momentum that biases words towards a certain meaning as a sentence progresses.
- At the right BERT is determining which prior word in the sentence the word "it" is referring to, and then using its attention mechanism to weigh the options. The word with the highest calculated score is deemed the correct association (i.e., "it" refers to "street", not "animal").
- If this phrase was a search query, the results would reflect this subtler, more precise understanding the BERT reached.



Copyright Ellis Horowitz 2011-2022

••

University of Southern California  USC

 USC **Viterbi**
School of Engineering

BERT for Question-Answering

- question answering is just a prediction task
- on receiving a question as input, the goal of the application is to identify the right answer from some corpus
- given a question and a context paragraph, the model predicts a start and an end token from the paragraph that most likely answers the question.

- Input Question:
Where do water droplets collide with ice crystals to form precipitation?
- Input Paragraph:
... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. ...
- Output Answer:
within a cloud



Copyright Ellis Horowitz, 2011-2022

52

••

University of Southern California

USC Viterbi
School of Engineering

Microsoft's AskMSR Answering System

Microsoft Store Products Support Search Microsoft Research

Research Research areas Products & Downloads Programs & Events People Careers About

An Analysis of the AskMSR Question-Answering System
January 1, 2002

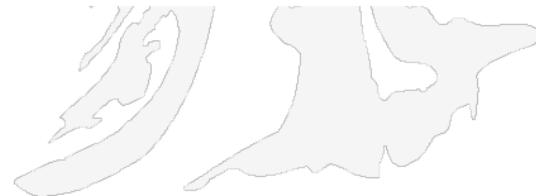
Download PDF Abstract Related Info

BibTeX

Authors
Eric Brill
Susan Dumais
Michele Banko

Published In
Proceedings of EMNLP 2002

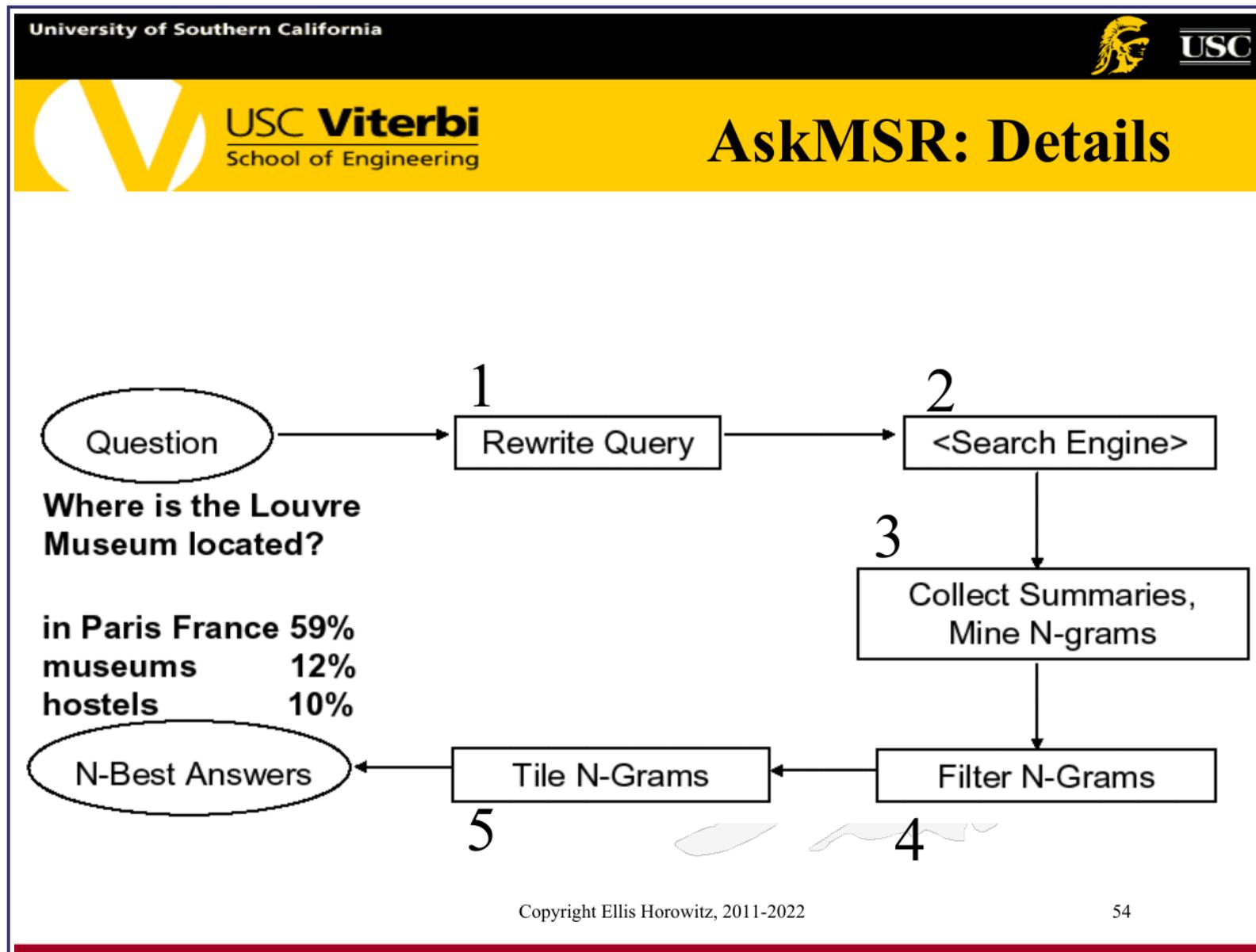
We describe the architecture of the AskMSR question answering system and systematically evaluate contributions of different system components to accuracy. The system differs from most question answering systems in its dependency on data redundancy rather than sophisticated linguistic analyses of either questions or candidate answers. Because a wrong answer is often worse than no answer, we also explore strategies for predicting when the question answering system is likely to give an incorrect answer.



Copyright Ellis Horowitz, 2011-2022

53

..



..

AskMSR: Step 1:Query Rewriting

- Classify question into categories
 - Who is/was/are/were...?
 - When is/did/will/are/were ...?
 - Where is/are/were ...?

a. Category-specific transformation rules

eg “For Where questions, move ‘is’ to all possible locations”

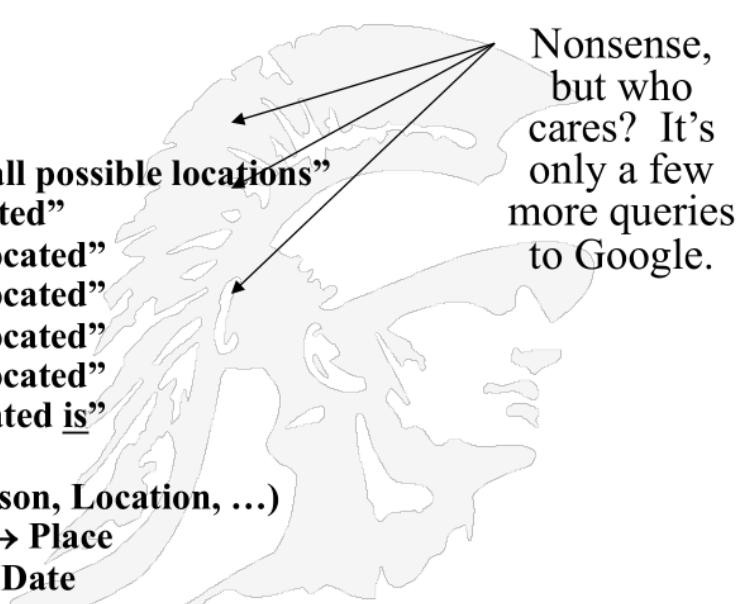
“Where is the Louvre Museum located”

- “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”

b. Expected answer “Datatype” (eg, Date, Person, Location, ...)

Where is the Louvre Museum located → Place

When was the French Revolution? → Date



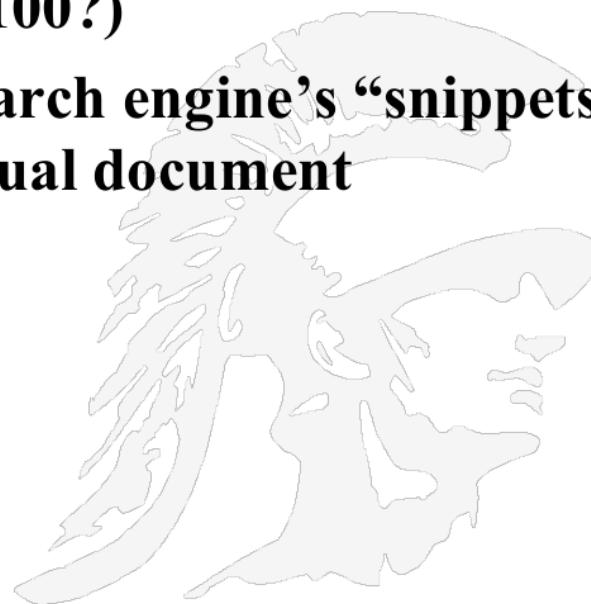
..

University of Southern California  USC

 USC **Viterbi**
School of Engineering

AskMSR:
Step 2: Query Search Engine

- Send all rewrites to a Web search engine
- Retrieve top N answers (100?)
- For speed, rely just on search engine's "snippets", not the full text of the actual document



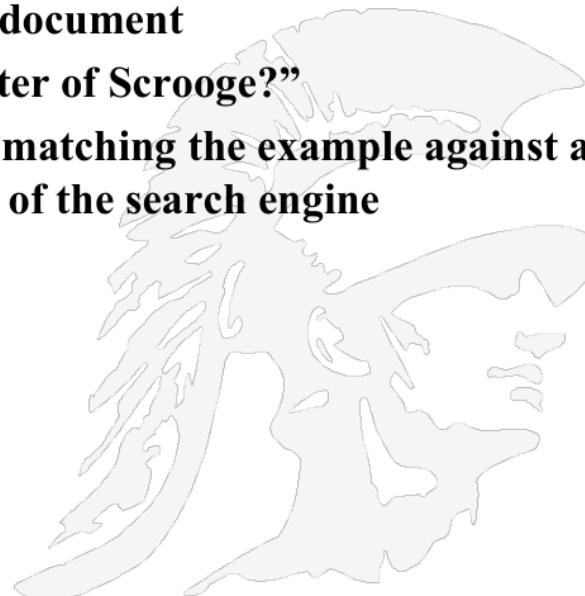
Copyright Ellis Horowitz, 2011-2022

56

..

AskMSR: Step 3: Mining N-Grams

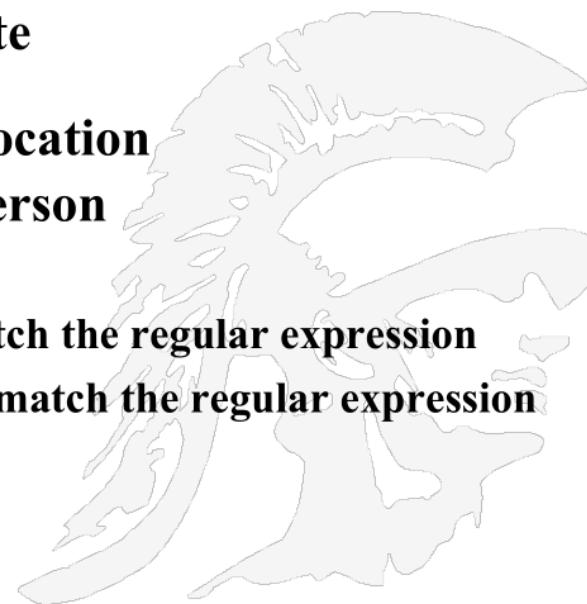
- **Simple:** Enumerate all N-grams (N=1,2,3 say) in all retrieved snippets
 - Use hash table and other data structures to make this efficient
- **Weight of an n-gram:** occurrence count, each weighted by “reliability” (weight) of rewrite that fetched the document
- **Example:** “Who created the character of Scrooge?”
- **Below are the weights produced by matching the example against a set of N-grams in the N-gram database of the search engine**
 - Dickens - 117
 - Christmas Carol - 78
 - Charles Dickens - 75
 - Disney - 72
 - Carl Banks - 54
 - A Christmas - 41
 - Christmas Carol - 45
 - Uncle - 31



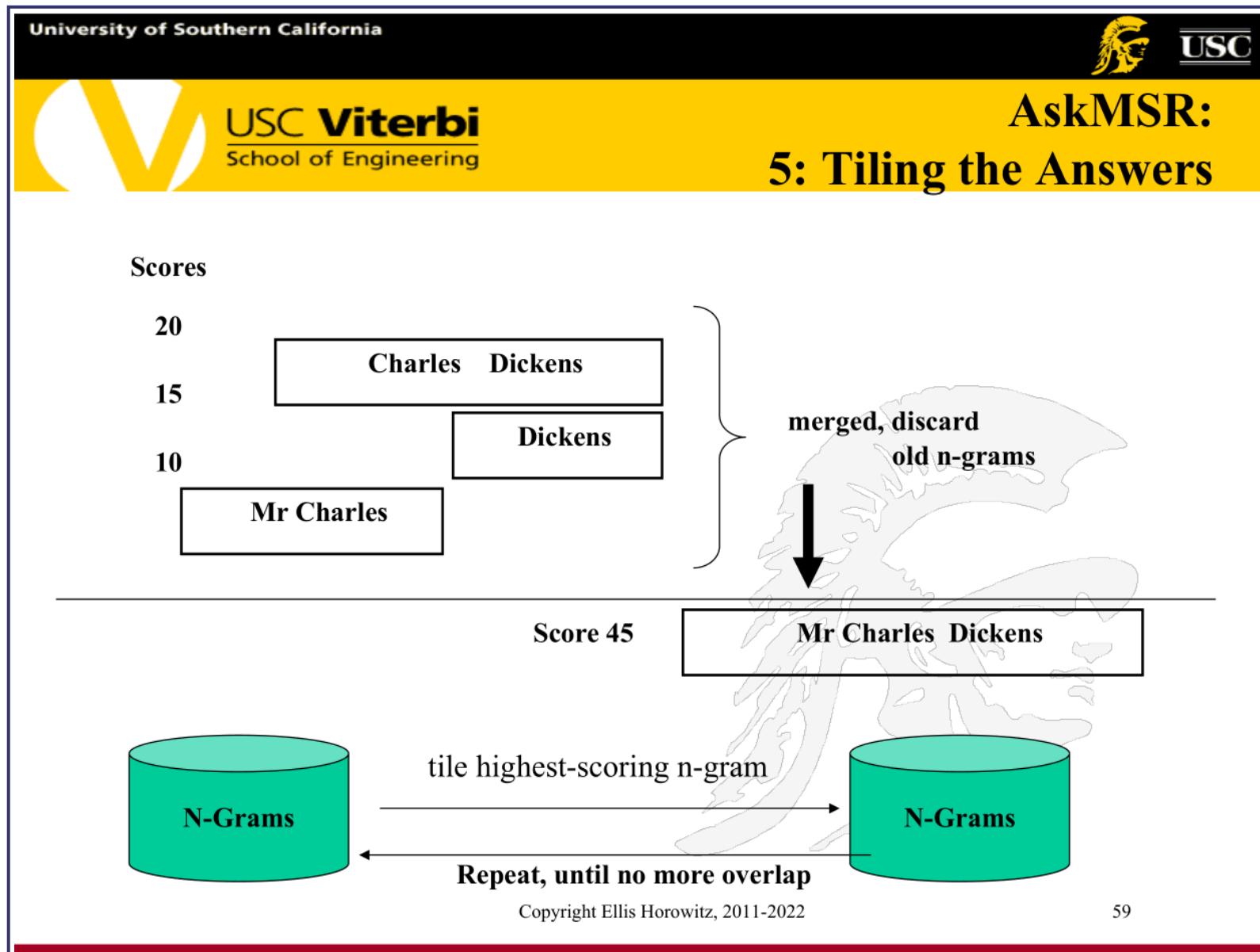
..

Step 4: Filtering N-Grams

- Each question type is associated with one or more “data-type filters” = regular expression
- When... → Date
- Where... → Location
- What ... → Person
- Who ... → Person
- Boost score of n-grams that do match the regular expression
- Lower score of n-grams that don’t match the regular expression



••



..

Common Evaluation Metric

- Accuracy (does answer match gold-labeled answer?)
- Mean Reciprocal Rank
 - For each query return a ranked list of M candidate answers
 - Its score is 1/Rank of the first right answer (or 0 if no answers are correct)
 - Take the mean over all N queries
 - $$\text{MRR} = \sum_{i=1}^{i=N} (1/\text{rank}_i) / N$$

