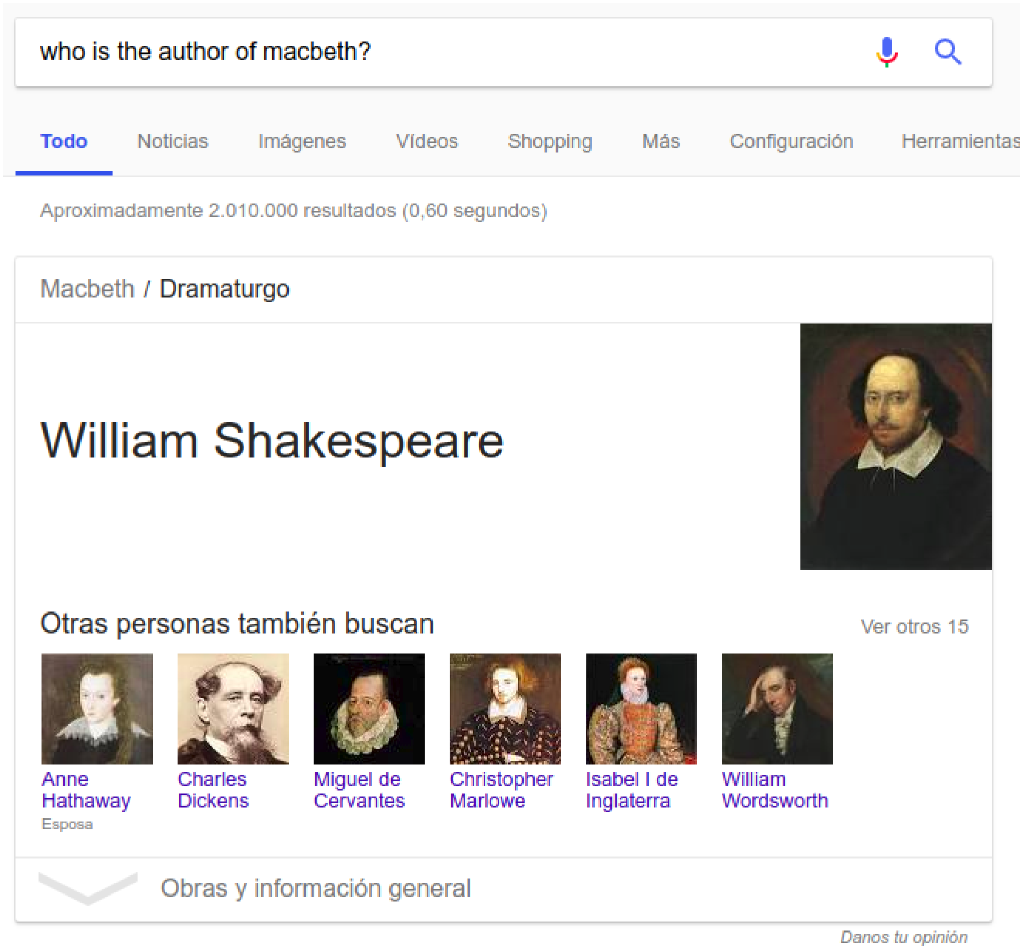
Question Answering

Natural Language Processing

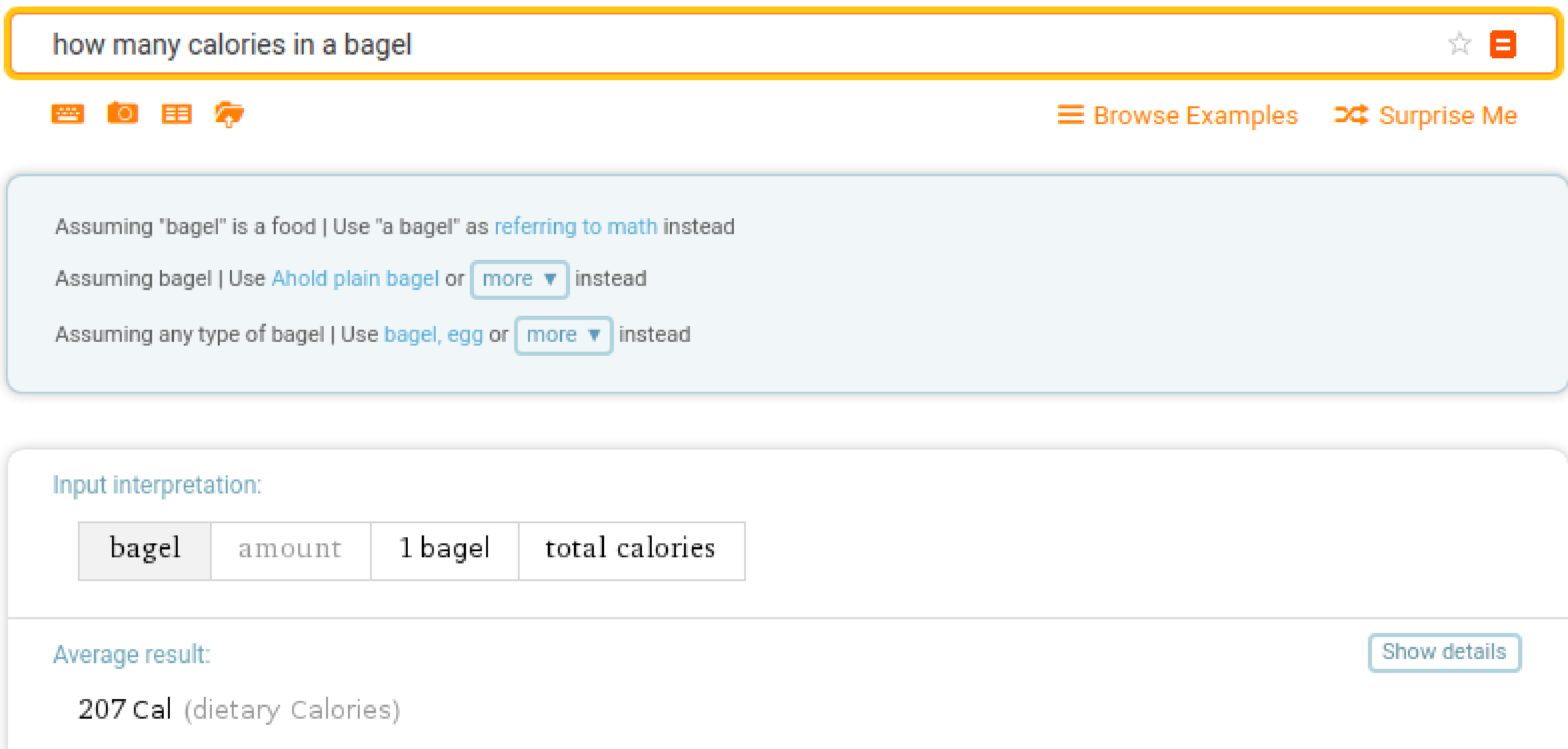
Master in Business Analytics and Big Data

acastellanos@faculty.ie.edu

QA Examples


who is the author of macbeth?
Todo
Noticias
Imágenes
Vídeos
Shopping
Más
Configuración
Herramientas
Aproximadamente 2.010.000 resultados (0,60 segundos)
Macbeth / Dramaturgo
William Shakespeare
Otras personas también buscan
Ver otros 15
Anne Hathaway
Esposa
Charles Dickens
Miguel de Cervantes
Christopher Marlowe
Isabel I de Inglaterra
William Wordsworth
Obras y información general
Danos tu opinión

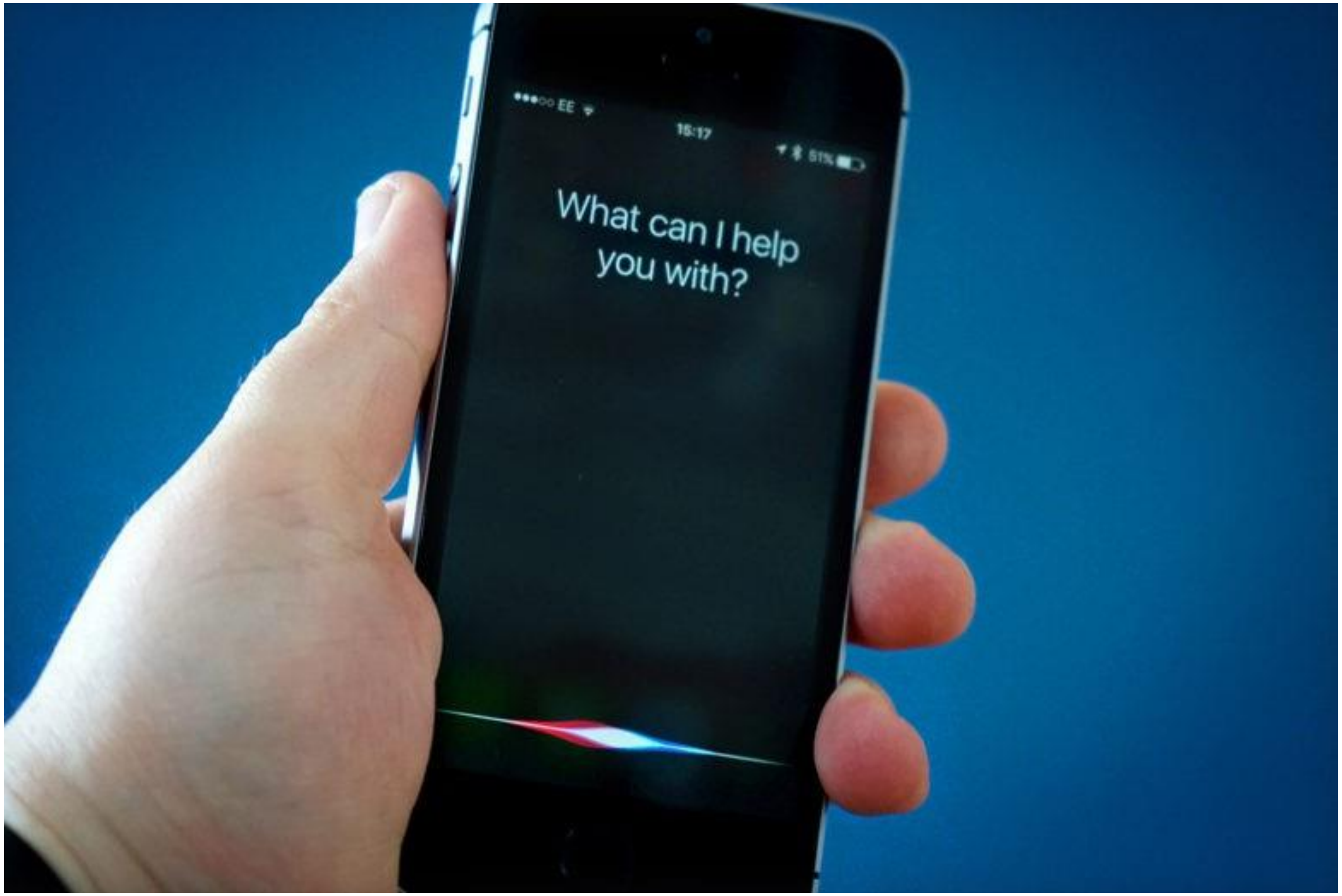
QA Examples


how many calories in a bagel
Browse Examples
Surprise Me
Assuming "bagel" is a food | Use "a bagel" as referring to math instead
Assuming bagel | Use Ahold plain bagel or more instead
Assuming any type of bagel | Use bagel, egg or more instead
Input interpretation:
bagel
amount
1 bagel
total calories
Average result:
Show details
207 Cal (dietary Calories)

QA Examples


PIENSE
THINK
ΣΚΕΨΟΥ
DENKE
सोचिए
PENSER
$300,000
$1,000,000
$200,000
Who is Stoker?
(I FOR ONE WELCOME OUR NEW COMPUTER OVERLORDS)
$ 1,000
Who is Bram Stoker?
$ 17,973
WHO IS BRAM STOKER?
$5600

QA Examples



QA Examples

Talk to Books

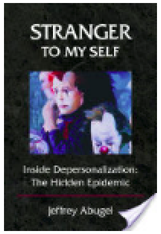
What smell brings back great memories?

Go!

Passages from books

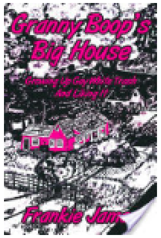
Part of the human experience is sudden recollection, spawned by new information coming in through the senses. **The aroma of frying chicken take me back to summers in Alabama at my grandmother's house.** Certain blossoms in Spring evoke memories of people and places from the past. (view in book)

from Stranger to My Self: Inside Depersonalization : the Hidden Epidemic
by Jeffrey Abugel



I worked the tissues. **The smell of this old place, old people, alcohol and cigarettes from years of partying, brought back so many memories.** Miss Barbara, Miss Dobbs, Miss Lang, Miss Jackson and all of them who I knew from growing up in this Big House. (view in book)

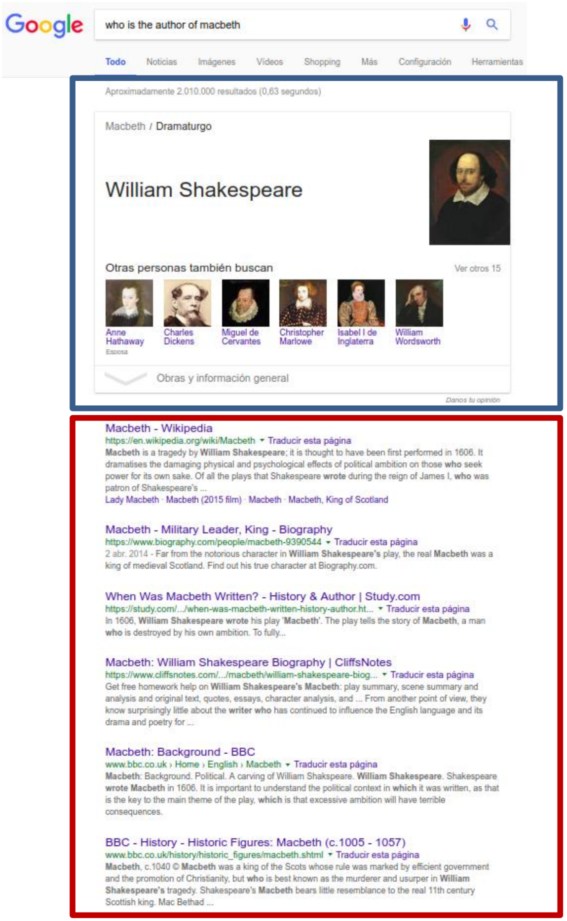
from Granny Boop's Big House: Growing up Gay White Trash and Liking It
by Frankie James



QA vs. IR

- **Information Retrieval**
  - **Retrieve documents** related to the query (and the snippet where the query appears)
  - Boolean like
  - Query Driven
- **Question Answering**
  - **Answer** the query
  - NLP like
  - Answer Driven
- IR is usually **included** in the QA Pipeline
  - IR to retrieve relevant documents --> QA to find the answer in the documents

QA vs. IR



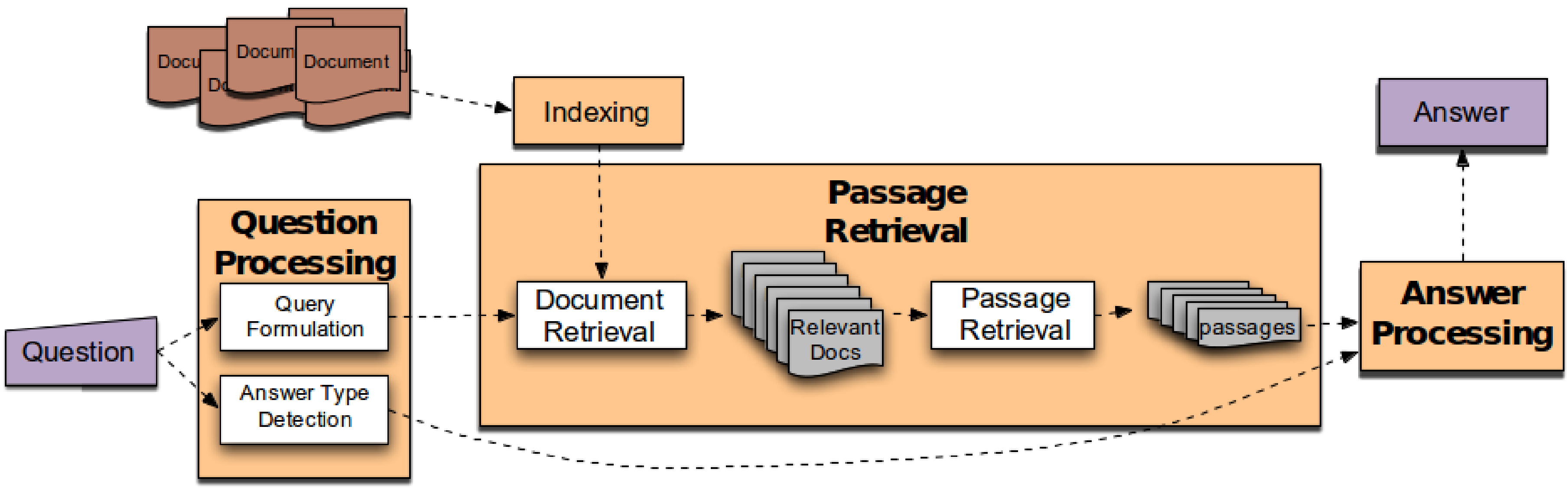
Question Answering

Information Retrieval

QA Paradigms

- **IR-based approaches**
  - Google
- **Knowledge-based and Hybrid approaches**
  - IBM Watson
  - Apple Siri
  - Wolfram Alpha

IR-based QA

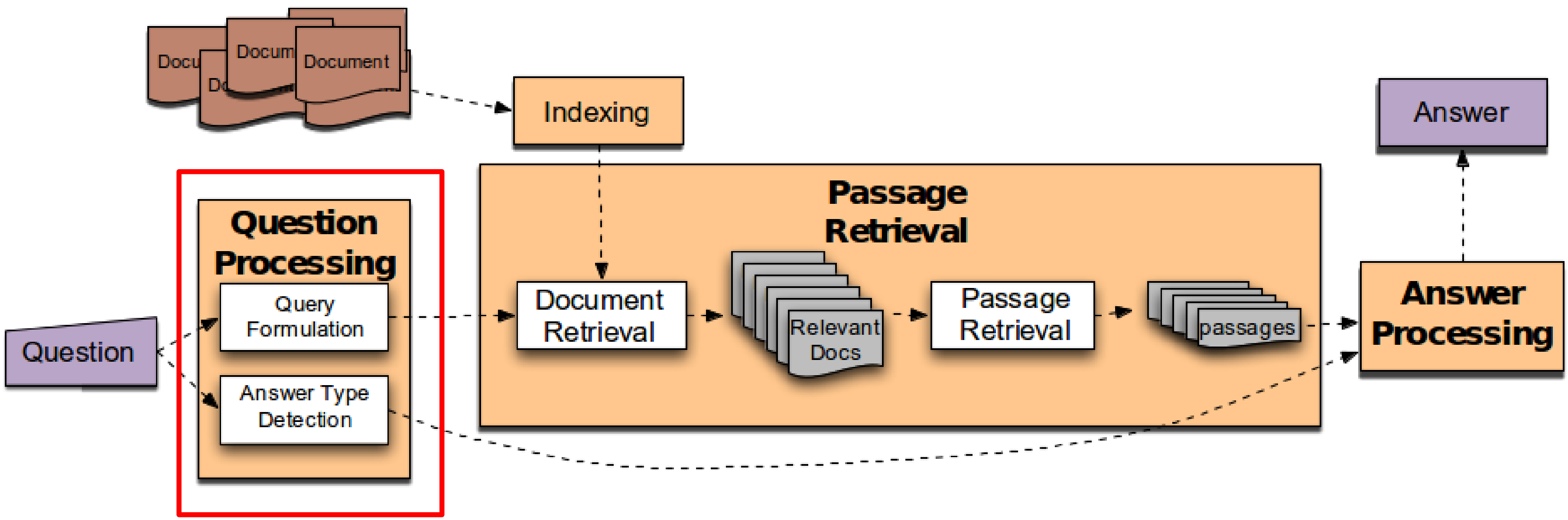


Question Answering [Dan Jurafsky, Stanford]

IR-based QA

- **Question Processing**
  - Detect **question type**: (Who?, Where? How much?...)
  - **Formulate** (i.e. construct) IR-based queries
- **Passage Retrieval**
  - Retrieve **related documents**
  - Extract **suitable passages** and re-rank
- **Answer Processing**
  - Extract **candidate answers**
  - **Rank candidates**
  - Using **evidence from the text and external sources**

IR-based QA



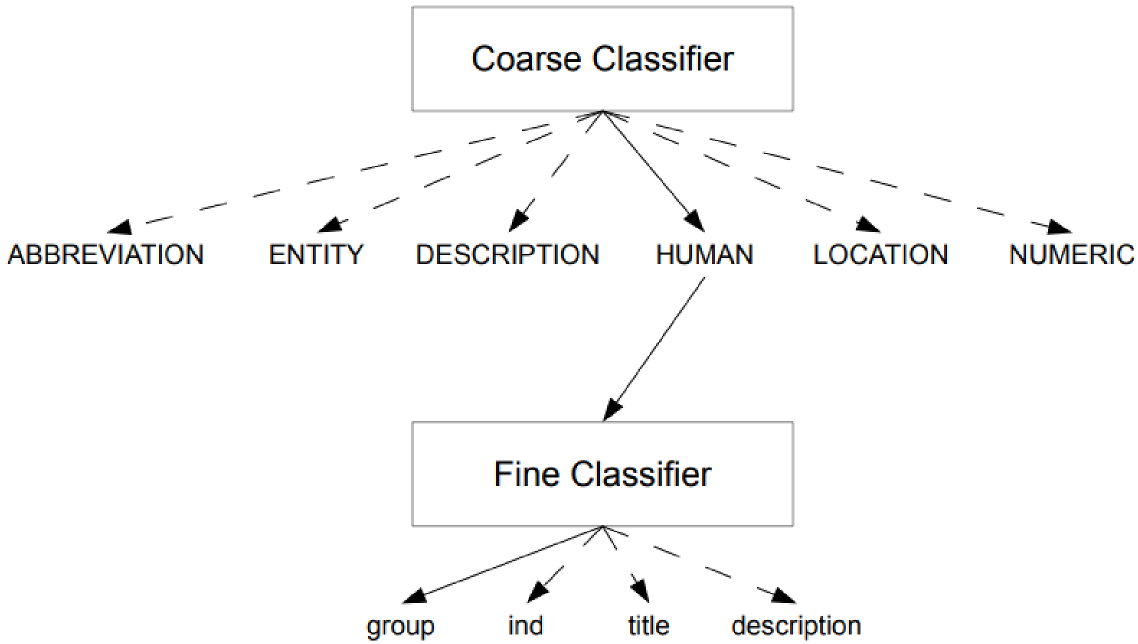
Question Answering [Dan Jurafsky, Stanford]

Answer Type Detection

- Who founded Tesla?
  - PERSON
- What Spanish city has the largest population?
  - CITY
- Where is the next world cup going to be held?
  - LOCATION

Answer Type Detection

- Answer Type Taxonomy
  - Xin Li, Dan Roth. 2002. Learning Question Classifiers. COLING'02



Answer Type Detection

- Answer Type Taxonomy
  - Xin Li, Dan Roth. 2002. Learning Question Classifiers. COLING'02
- Answer types in Jeopardy
  - **2500 types** --> 50% covered by 200 most frequent
  - Ferrucci et al. 2010. Building Watson: An Overview of the DeepQA Project. AI Magazine. Fall 2010. 59-79.

Answer Type Detection

- **Regular expression**-based rules can get some cases:
  - In what country was X born?
  - X (PERSON) was born in Y(PLACE)

- Use **question headword**:
  - Headword of the **first noun phrase after the wh-word**
    - Which city in China has the largest number of foreign financial companies?
    - What is the state flower of California?

Answer Type Detection

- Treat the problem as **machine learning classification**
  - Taxonomy of question types: Target variable
  - Annotated training data
  - Train classifiers for each question type using a rich set of **features**.

    features include those hand-written rules!

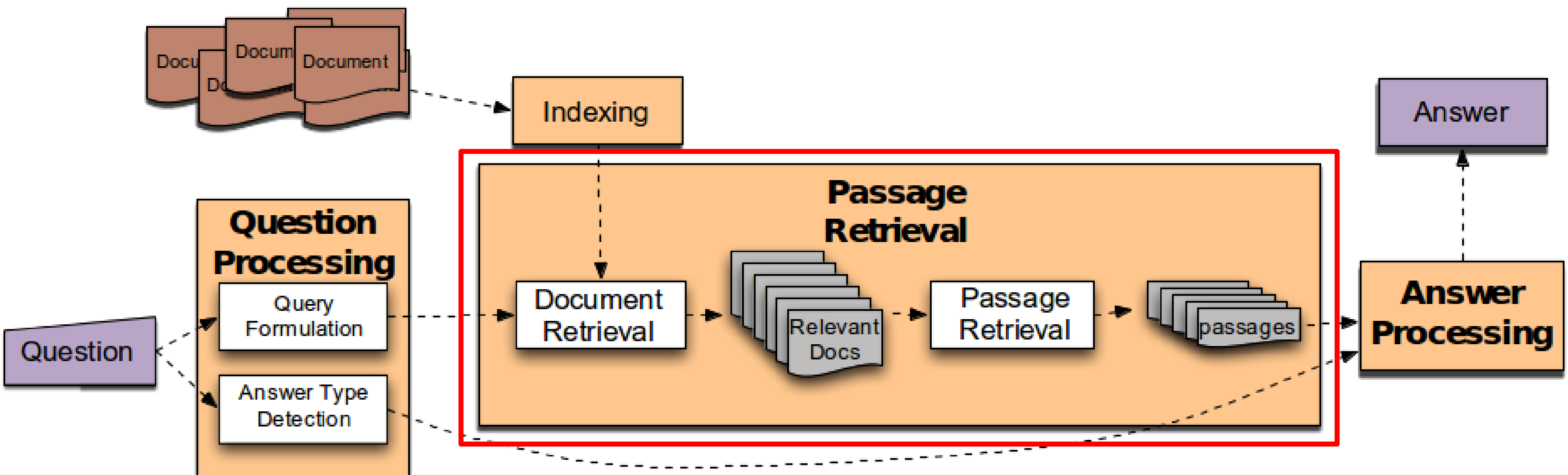
    - Question words and phrases
    - Part-of-speech tags
    - Parse features (headwords)
    - Named Entities
    - Semantically related words

Query Formulation

- **Keyword Selection**
  - Dan Moldovan, Sanda Harabagiu, Marius Paca, Rada Mihalcea, Richard Goodrum, Roxana Girju and Vasile Rus. 1999. Proceedings of TREC-8.

1. Select all non-stop words in quotations
2. Select all NNP words in recognized named entities
3. Select all complex nominals with their adjectival modifiers
4. Select all other complex nominals
5. Select all nouns with their adjectival modifiers
6. Select all other nouns
7. Select all verbs
8. Select all adverbs
9. Select the QFW word (skipped in all previous steps)
10. Select all other words

IR-based QA



Question Answering [Dan Jurafsky, Stanford]

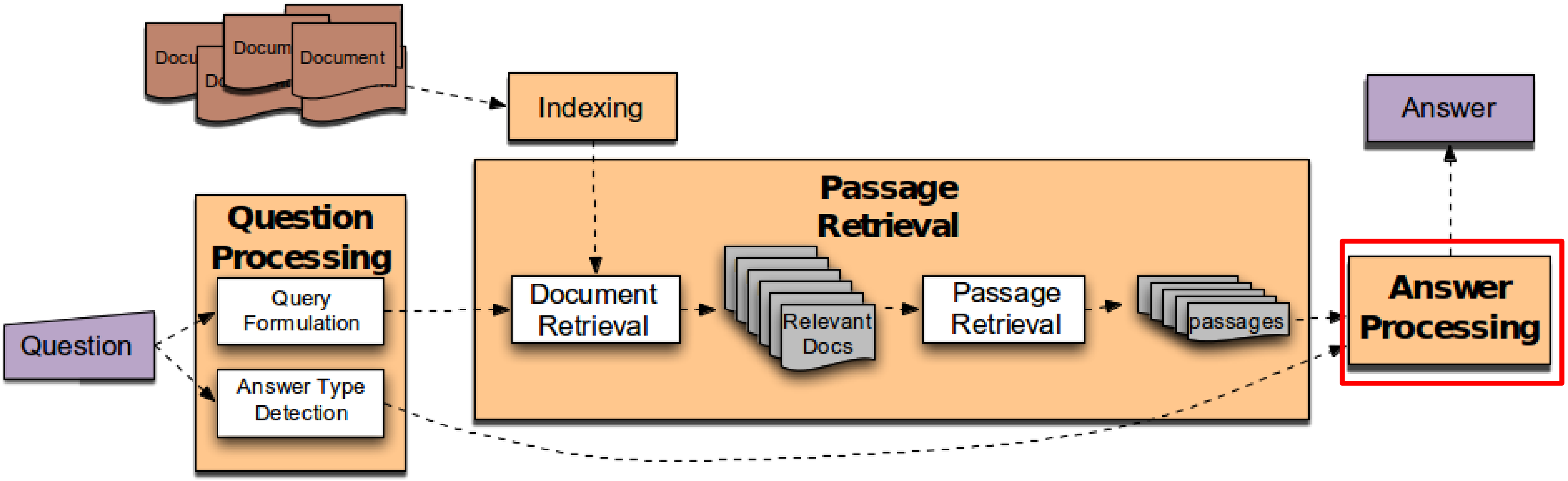
Passage Retrieval

- **Step 1:** IR engine **retrieves documents** using query terms
  - The query formulated at the previous step
- **Step 2: Segment the documents** into shorter units
  - Paragraph-based Tokenization
- **Step 3: Passage ranking**
  - Use answer type to help re-rank passages

Passage Retrieval

- **Features for Passage Ranking**
  - Number of Named Entities of the right type in passage
  - Number of query words in passage
  - Number of question N-grams also in passage
  - Proximity of query keywords to each other in passage
  - Longest sequence of question words
  - Rank of the document containing passage

IR-based QA



Document
Document
Document
Indexing
Answer
Passage
Retrieval
Question
Processing
Query
Formulation
Answer Type
Detection
Question
Document
Retrieval
Relevant
Docs
Passage
Retrieval
passages
Answer
Processing

Answer Extraction

- Run an **answer-type named-entity tagger** on the passages
  - Each answer type requires a named-entity tagger that detects it
  - If answer type is CITY, tagger has to tag CITY
    - Can be full NER, simple regular expressions, or hybrid

- Return **the string with the right type**:
  - **Who is the prime minister of India (PERSON)**

Manmohan Singh, Prime Minister of India, had told left leaders that the deal would not be renegotiated.

  - **How tall is Mt. Everest? (LENGTH)**

The official height of Mount Everest is **29035 feet**

Answer Extraction

- **Multiple candidate answers**

Q: Who was Queen Victoria's second son?

Answer Type: **Person**

The Marie biscuit is named after **Marie Alexandrovna**, the daughter of C**zar Alexander II of Russia** and wife of **Alfred**, the second son of **Queen Victoria** and **Prince Albert**

Features for ranking candidate answers

- **Answer type match:** Candidate contains a phrase with the correct answer type.
- **Pattern match:** Regular expression pattern matches the candidate.
- **Question keywords:** # of question keywords in the candidate.
- **Keyword distance:** Distance in words between the candidate and query keywords
- **Novelty factor:** A word in the candidate is not in the query.
- **Apposition features:** The candidate is an appositive to question terms
- **Punctuation location:** The candidate is immediately followed by a comma, period, quotation marks, semicolon, or exclamation mark.
- **Sequences of question terms:** The length of the longest sequence of question terms that occurs in the candidate answer.

Answer Sentence Selection

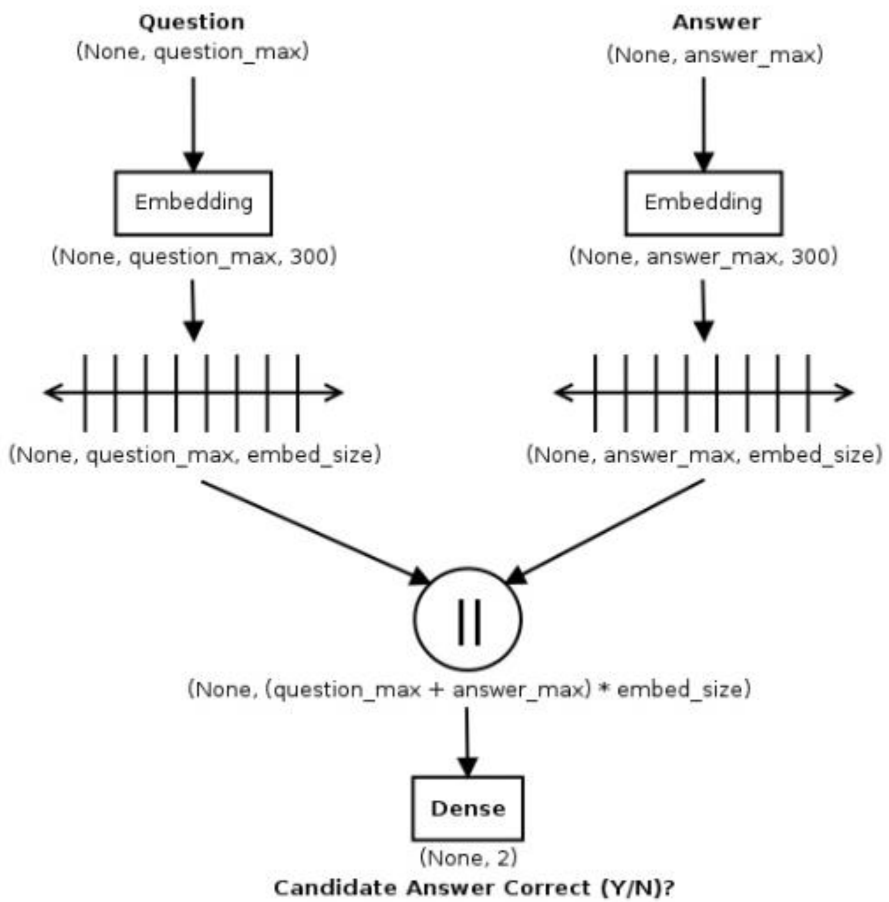
Algorithm	Reference	MAP	MRR
Punyakanok (2004)	Wang et al. (2007)	0.419	0.494
Cui (2005)	Wang et al. (2007)	0.427	0.526
Wang (2007)	Wang et al. (2007)	0.603	0.685
H&S (2010)	Heilman and Smith (2010)	0.609	0.692
W&M (2010)	Wang and Manning (2010)	0.595	0.695
Yao (2013)	Yao et al. (2013)	0.631	0.748
S&M (2013)	Severyn and Moschitti (2013)	0.678	0.736
Shnarch (2013) - Backward	Shnarch (2013)	0.686	0.754
Yih (2013) - LCLR	Yih et al. (2013)	0.709	0.770
Yu (2014) - TRAIN-ALL bigram+count	Yu et al. (2014)	0.711	0.785
W&N (2015) - Three-Layer BLSTM+BM25	Wang and Nyberg (2015)	0.713	0.791
Feng (2015) - Architecture-II	Tan et al. (2015)	0.711	0.800
S&M (2015)	Severyn and Moschitti (2015)	0.746	0.808
W&I (2015)	Wang and Ittycheriah (2015)	0.746	0.820
Tan (2015) - QA-LSTM/CNN+attention	Tan et al. (2015)	0.728	0.832
dos Santos (2016) - Attentive Pooling CNN	dos Santos et al. (2016)	0.753	0.851
Wang et al. (2016) - Lexical Decomposition and Composition	Wang et al. (2016)	0.771	0.845

Bag of words, Word alignment, Dependency Tree Matching

Deep Neural Networks, LSTM

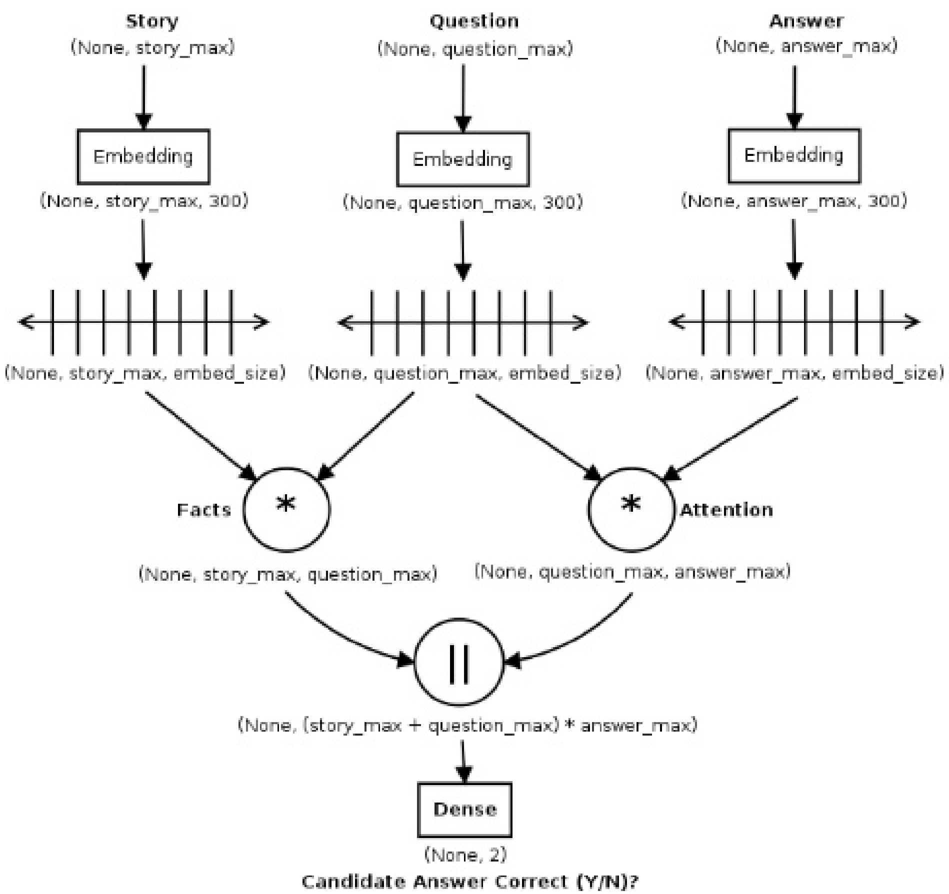
http://aclweb.org/aclwiki/index.php?title=Question_Answering_(State_of_the_art)

QA-LSTM Model



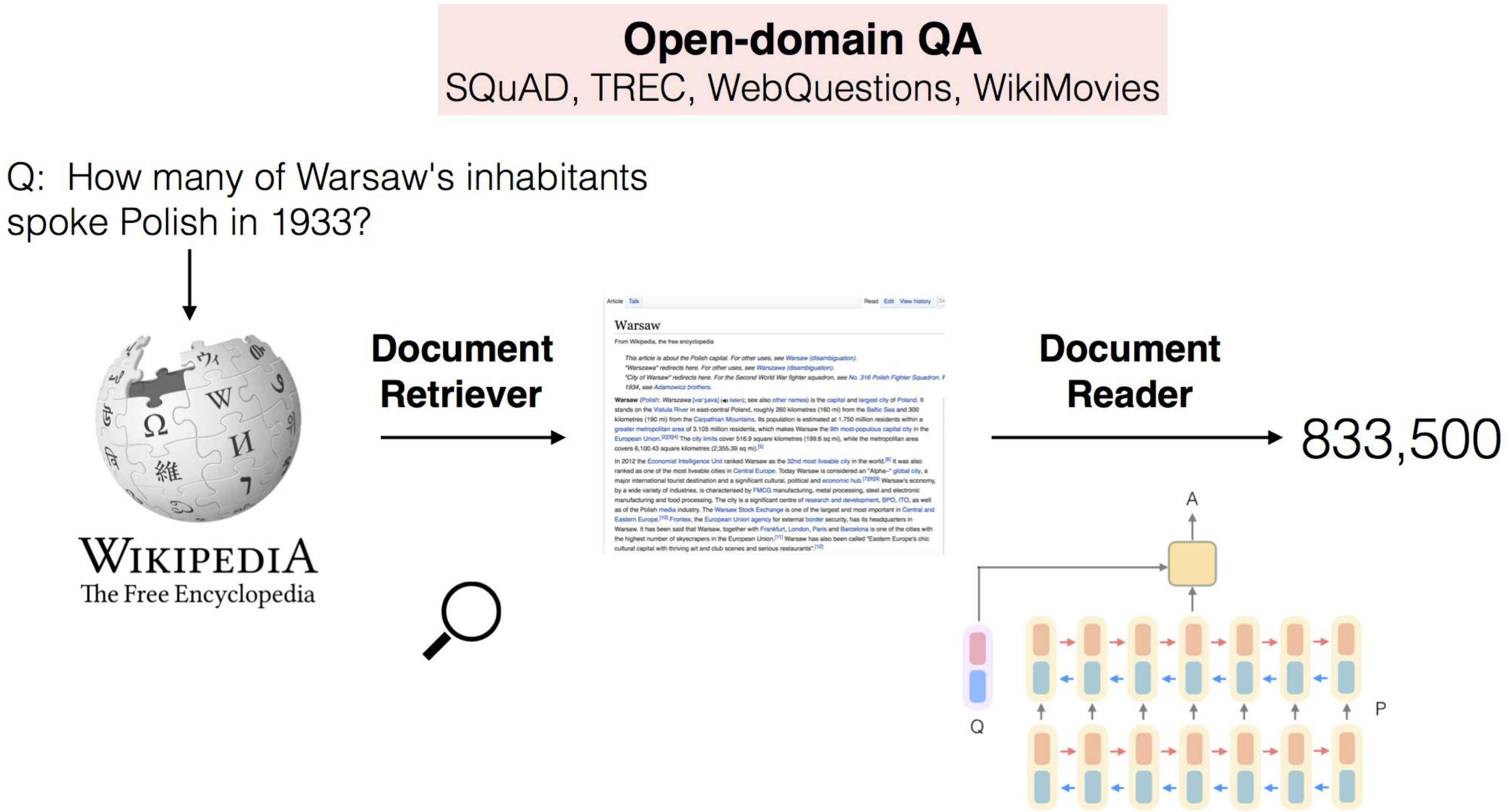
https://www.slideshare.net/sujitpal/deep-learning-models-for-question-answering

QA-LSTM Model



https://www.slideshare.net/sujitpal/deep-learning-models-for-question-answering

Facebook's DrQA

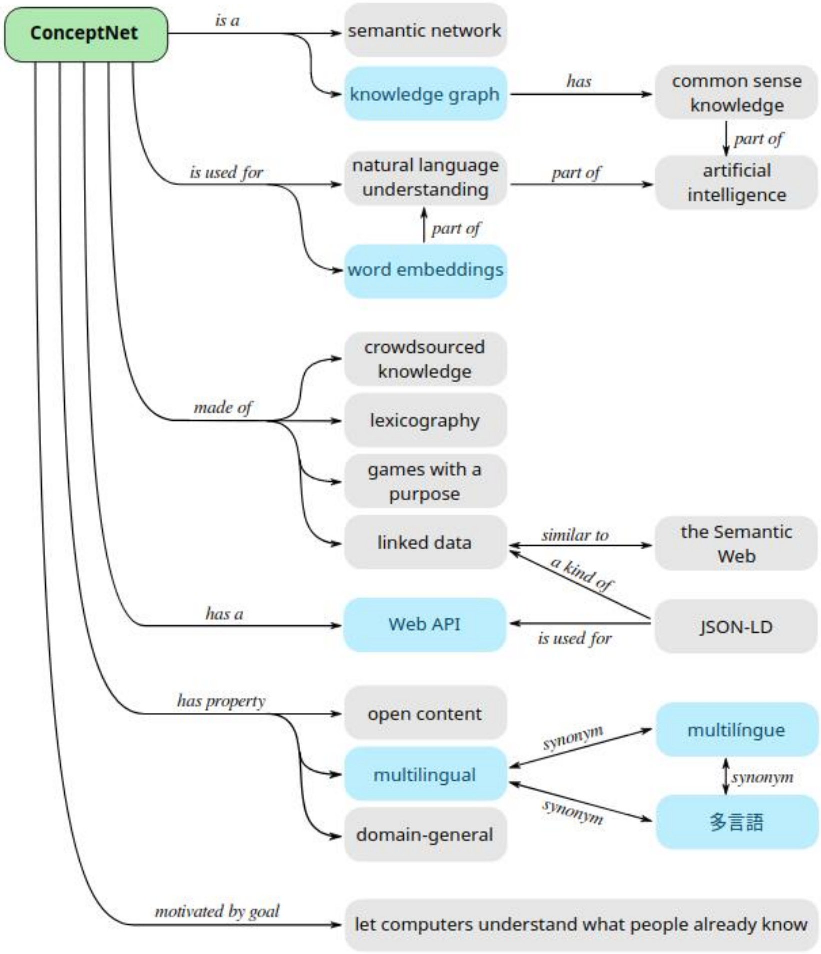


https://research.fb.com/downloads/drqa/

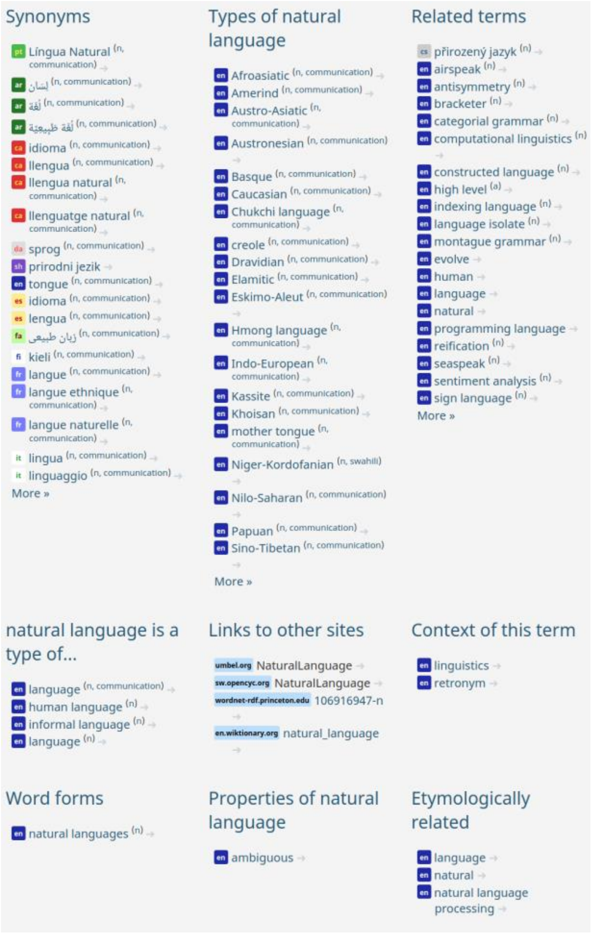
Knowledge-based QA

- Build a **semantic representation of the query**
  - Times, dates, locations, entities, numeric quantities
- **Map from this semantics to query structured database**
  - Geospatial databases (geonames.org, GeoQuery, OpenStreetMap)
  - Ontologies (Wikipedia infoboxes, DBPedia, WordNet, Yago, Freebase, Wikidata)
  - Review sources and reservation services (Yelp, Amazon)
  - Scientific databases

ConceptNet



ConceptNet
is a
semantic network
knowledge graph
has
common sense knowledge
part of
artificial intelligence
is used for
natural language understanding
part of
part of
word embeddings
made of
crowdsourced knowledge
lexicography
games with a purpose
linked data
similar to
the Semantic Web
a kind of
has a
Web API
is used for
JSON-LD
has property
open content
multilingual
synonym
multilíngue
synonym
synonym
多言語
domain-general
motivated by goal
let computers understand what people already know



Synonyms
Língua Natural (n, communication)
لِسان (n, communication)
لُغَة (n, communication)
لُغَة طَبِيعِيّة (n, communication)
idioma (n, communication)
llengua (n, communication)
llengua natural (n, communication)
llenguatge natural (n, communication)
sprog (n, communication)
prirodni jezik
tongue (n, communication)
idioma (n, communication)
lengua (n, communication)
زبان طبیعی (n, communication)
kieli (n, communication)
langue (n, communication)
langue ethnique (n, communication)
langue naturelle (n, communication)
lingua (n, communication)
linguaggio (n, communication)
More »
Types of natural language
Afroasiatic (n, communication)
Amerind (n, communication)
Austro-Asiatic (n, communication)
Austronesian (n, communication)
Basque (n, communication)
Caucasian (n, communication)
Chukchi language (n, communication)
creole (n, communication)
Dravidian (n, communication)
Elamitic (n, communication)
Eskimo-Aleut (n, communication)
Hmong language (n, communication)
Indo-European (n, communication)
Kassite (n, communication)
Khoisan (n, communication)
mother tongue (n, communication)
Niger-Kordofanian (n, swahili)
Nilo-Saharan (n, communication)
Papuan (n, communication)
Sino-Tibetan (n, communication)
More »
Related terms
přirozený jazyk (n)
airspeak (n)
antisymmetry (n)
bracketer (n)
categorial grammar (n)
computational linguistics (n)
constructed language (n)
high level (a)
indexing language (n)
language isolate (n)
montague grammar (n)
evolve
human
language
natural
programming language
reification (n)
seaspeak (n)
sentiment analysis (n)
sign language (n)
More »
natural language is a type of...
language (n, communication)
human language (n)
informal language (n)
language (n)
Links to other sites
umbel.org NaturalLanguage
sw.opencyc.org NaturalLanguage
wordnet-rdf.princeton.edu 106916947-n
en.wiktionary.org natural_language
Context of this term
linguistics
retronym
Word forms
natural languages (n)
Properties of natural language
ambiguous
Etymologically related
language
natural
natural language processing

Entity-centric Knowledge Databases

Address: 400 Broad St, Seattle, 98109
Phone: (800) 937-9582
Opened: Apr 21, 1962
Height: 605 feet (184.41 m)
Floors: 6

NFL championships: 2013
Head coach: Pete Carroll
Founded: 1976
Division: NFC West

Location

Home Field

Seattle

Population: 652,405 (2013)
Area: 142.55 sq miles (369.20 km²)
Mayor: Ed Murray

Founded: Mar 30, 1971 · Pike Place Market
Customer service: +1 800-782-7282
CEO: Howard Schultz
Founders: Jerry Baldwin · Zev Siegl · Gordon Bowker

Headquarters

Build the representation

- **Answers:** Databases of Relations
  - Wikipedia infoboxes, DBpedia, FreeBase, etc.

```
born-in("Emma Goldman", "June 27 1869")
author-of("Cao Xue Qin", "Dream of the Red Chamber")
```

- **Questions:** Find these relations in Questions

```
Whose granddaughter starred in E.T.?
(acted-in ?x "E.T.")
(granddaughter-of ?x ?y)
```

Build the representation

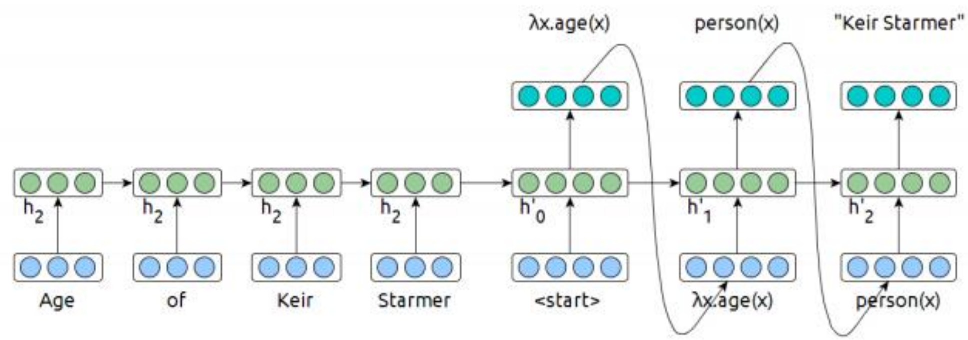
- **Ad-hoc manual rules**
  - Scalability Issues
  - Does not support "open-domain" questions
  - Coverage
- **Semantic Parsing**
  - Learn a translation function
  - Language Mismatch
  - Incomplete Knowledge Bases

Freebase

Relation	Percentage unknown	
	All 3M	*Top 100K*
PROFESSION	68%	24%
PLACE OF BIRTH	71%	13%
NATIONALITY	75%	21%
EDUCATION	91%	63%
SPOUSES	92%	68%
PARENTS	94%	77%
CHILDREN	94%	80%
SIBLINGS	96%	83%
ETHNICITY	99%	86%

Build the representation

- **A Deep Learning Approach**



  - Like in MT, using attention can be helpful

    Dong and Lapata (2016): Language to Logical Form with Neural Attention
  - Exploit the highly rigid structure in the target side to constrain generation

    Ling et al. (2016): Latent predictor networks for code generation
  - Make use of semi-supervised training to counter sparsity

    Kocisky et al. (2016): Semantic Parsing with Semi-Supervised Sequential Autoencoders

Relationship Reasoning

- **Temporal Reasoning**

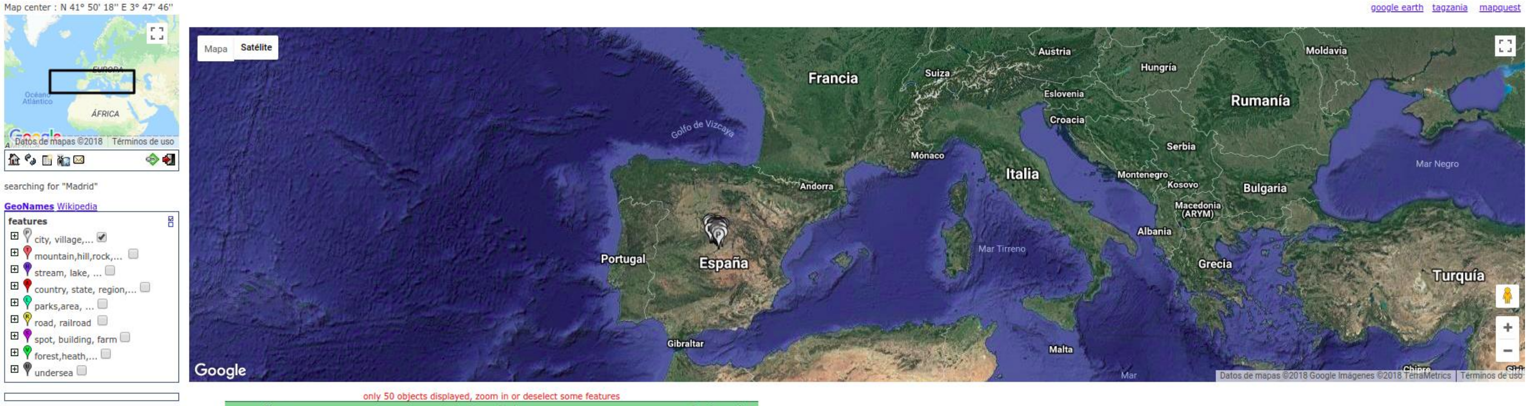
"In **1594** he took a job as a tax collector in Andalusia"

Thoreau is a bad answer (born in 1817)

Cervantes is possible (was alive in 1594)

Relationship Reasoning

- **Geospatial Reasoning**
  - Beijing is a good answer for "Asian city"
  - California is "southwest of Montana"



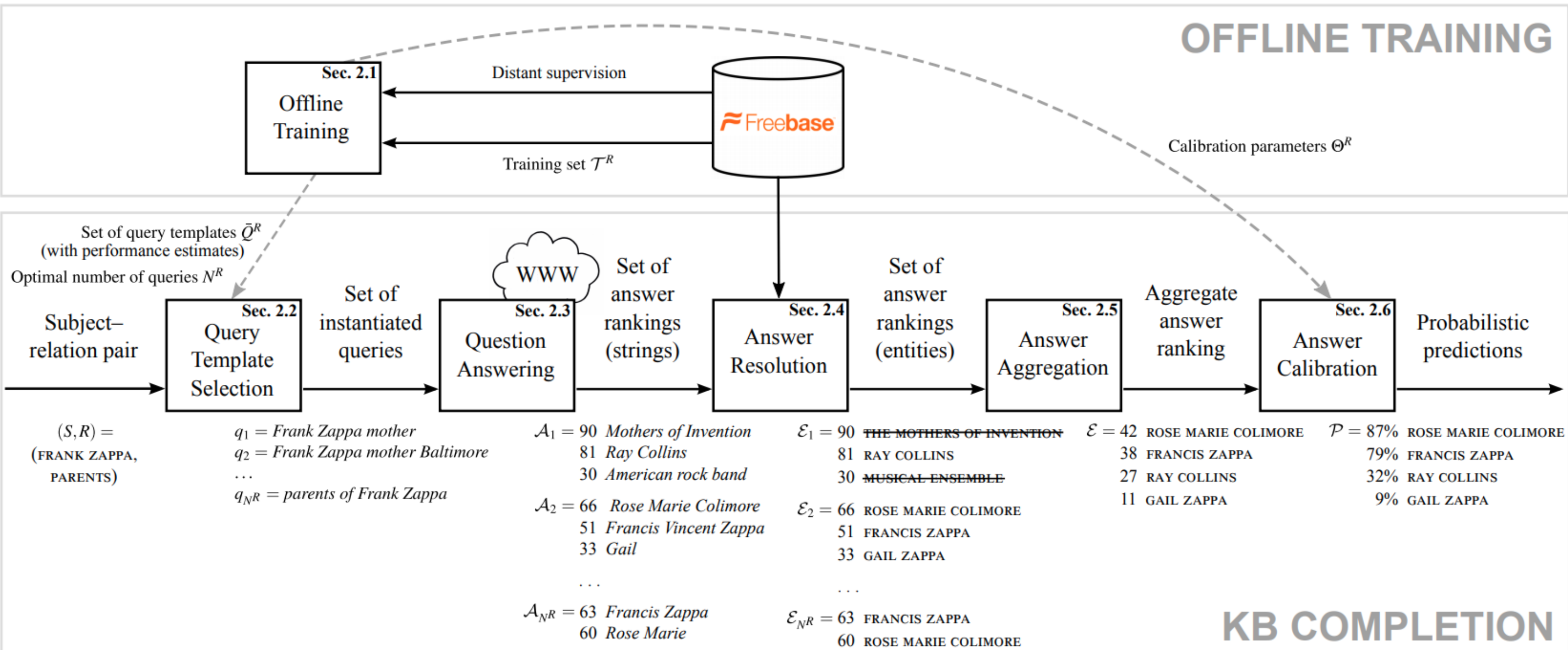
only 50 objects displayed, zoom in or deselect some features

	Name	country	feature	km to center
1	Madrid	Spain	capital of a political entity	647.41 km
2	Comunidad de Madrid	Spain	first-order administrative division	646.16 km
3	Getafe	Spain	seat of a third-order administrative division	653.48 km
4	Fuenlabrada	Spain	seat of a third-order administrative division	659.15 km
5	Leganés	Spain	seat of a third-order administrative division	655.24 km
6	Móstoles	Spain	seat of a third-order administrative division	663.61 km
7	Alcorcón	Spain	seat of a third-order administrative division	659.6 km
8	Alcobendas	Spain	seat of a third-order administrative division	638.47 km
9	City Center	Spain	section of populated place	647.36 km
10	Las Rozas de Madrid	Spain	populated place	658.96 km
11	Hortaleza	Spain	section of populated place	640.6 km

Hybrid-based QA

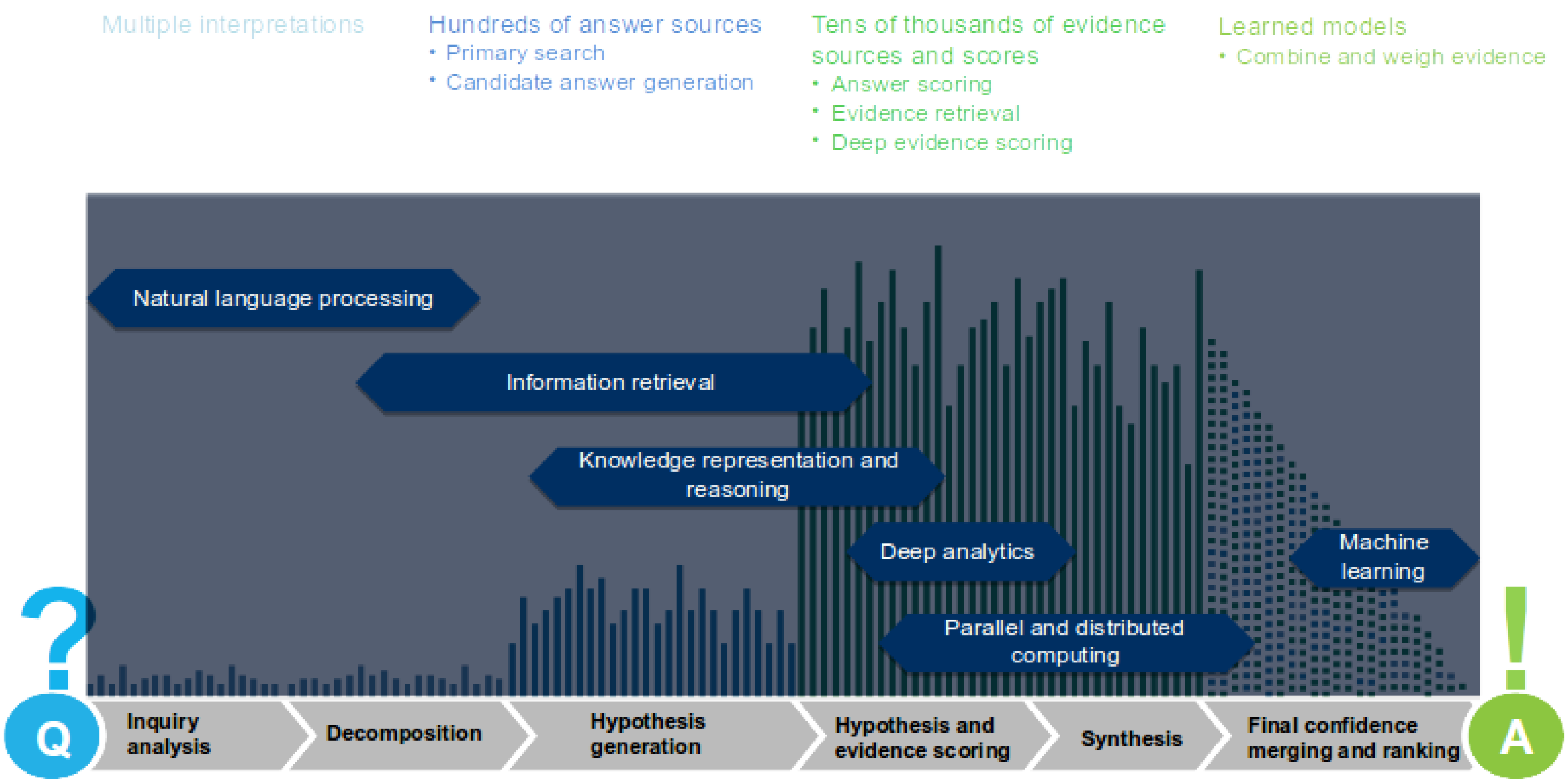
- **Build a shallow semantic representation of the query**
- Generate **answer candidates using IR methods**
  - Augmented with ontologies and semi-structured data
- **Score each candidate using richer knowledge sources**
  - Geospatial databases
  - Temporal reasoning
  - Taxonomical classification

Hybrid-based QA

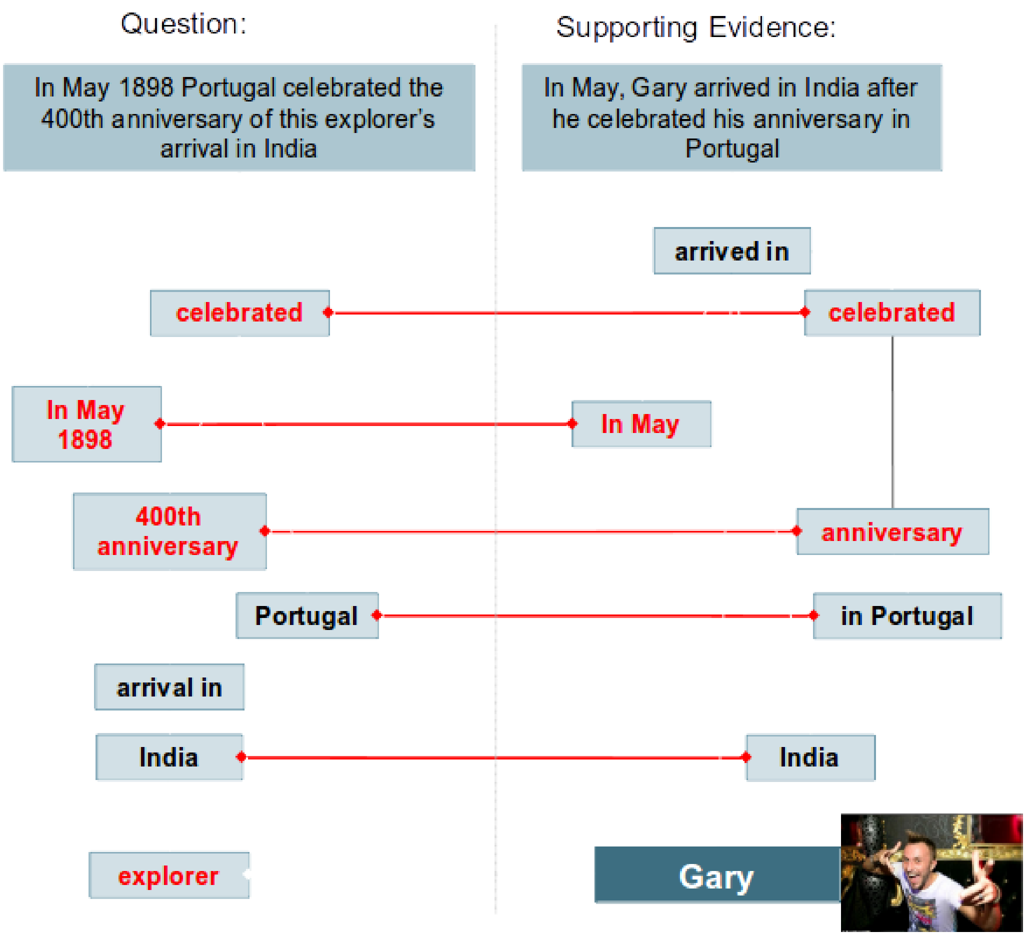


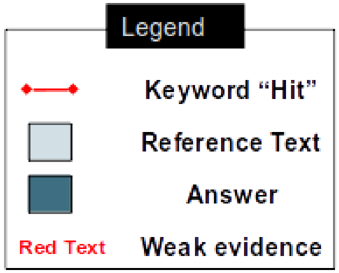
Knowledge Base Completion via Search-Based Question Answering [Robert West, et al., WWW 2014]

IBM Watson Architecture



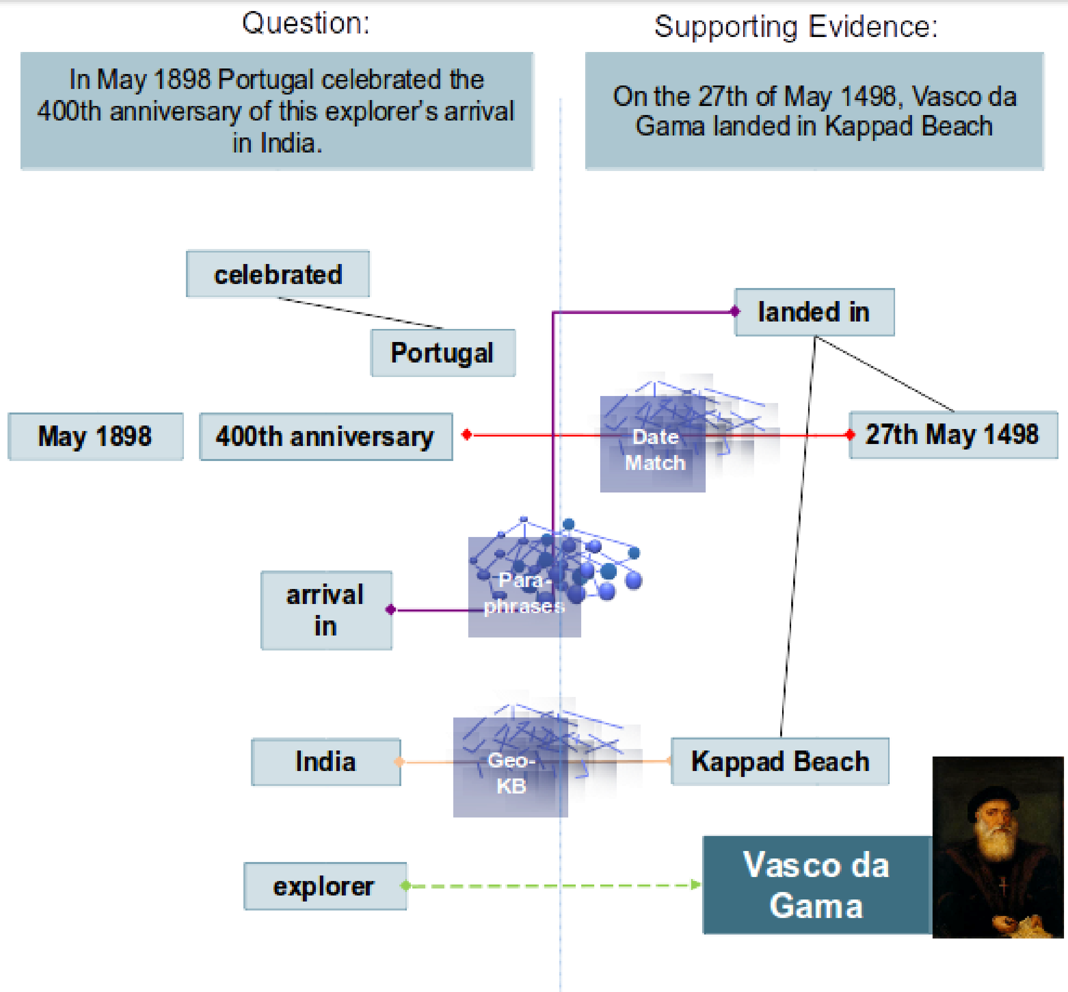
IBM Watson Architecture

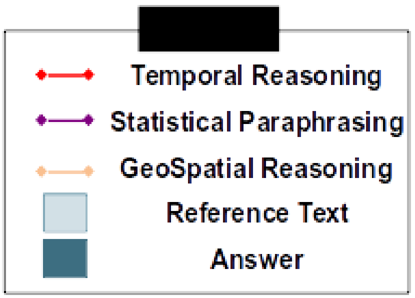




This evidence suggests "Gary" is the answer BUT the system must learn that keyword matching may be weak relative to other types of evidence

IBM Watson Architecture





Stronger evidence can be much harder to find and score…

Search far and wide

Explore many hypotheses

Find judge evidence

Many inference algorithms

Resources

- A survey on question answering systems with classification
  https://www.sciencedirect.com/science/article/pii/S1319157815000890
- How does the [current] best question answering model work?
  https://towardsdatascience.com/how-the-current-best-question-answering-model-works-8bbacf375e2a
- QA State of the Art
  https://aclweb.org/aclwiki/Question_Answering_(State_of_the_art)
- Strategies for Advanced Question Answering
  https://pdfs.semanticscholar.org/e10b/08e1c7c5fd37b19eaf24a9addc503b413c65.pdf
- Facebook DrQA
  https://github.com/facebookresearch/DrQA

Resources

- **Datasets**
  - Nguyen et al. (2016), MS MARCO: A Human Generated Machine Reading Comprehension
  - Dataset Haas and Riezler (2016), A Corpus and Semantic Parser for Multilingual Natural Language Querying of OpenStreetMap
- **Deep Language Modeling for Question Answering using Keras**
  - https://codekansas.github.io/blog/2016/language.html