“Information retrieval”

The name information retrieval is standard, but as traditionally practiced, it’s not really right

All you get is document retrieval, and beyond that the job is up to you

Getting information

The common person’s view? [From a novel]

“I like the Internet. Really, I do. Any time I need a piece of shareware or I want to find out the weather in Bogota ... I’m the first guy to get the modem humming. But as a source of information, it sucks. You got a billion pieces of data, struggling to be heard and seen and downloaded, and anything I want to know seems to get trampled underfoot in the crowd.”


Web Search in 2025?

The web, it is a changing.

What will people do in 2025?

- Type key words into a search box?
- Use the Semantic Web?
- Ask questions to their computer in natural language?
- Use social or “human powered” search?

What do we know that’s happening?

- Much of what is going on is in the products of companies, and there isn’t exactly careful research explaining or evaluating it
- So most of this is my own meandering observations giving voice over to slides from others

Google

What’s been happening? 2013–2017

- Many updates a year ... and 3rd party sites try to track them
  - e.g., https://moz.com/google-algorithm-change by & aimed at SEOs
  - I just mention a few changes here
- New search index at Google: “Hummingbird” (2013)
  - Answering long, “natural language” questions better
  - Partly to deal with spoken queries on mobile
- More use of the Google Knowledge Graph (2014)
  - Concepts versus words
- RankBrain (second half of 2015):
  - A neural net helps in document matching for the long tail
What’s been happening? 2013–2017

- "Pigeon" update (July 2014):
  - More use of distance and location in ranking signals
- "Mobilegeddon" (Apr 21, 2015):
  - "Mobile friendliness" as a major ranking signal
- "App Indexing" (Android, iOS support May 2015)
  - Search results can take you to an app
- Mobile-friendly 2 (May 12, 2016):
  - About half of all searches are now from mobile
- "Fred" (1st quarter 2017)
  - Various changes discounting spammy, clickbaity, fake? sites

The role of knowledge bases

- Google Knowledge Graph
- Facebook Graph Search
- Bing’s Satori
- Things like Wolfram Alpha

Common theme: Doing graph search over structured knowledge rather than traditional text search

What’s been happening

- More semi-structured information embedded in web pages
  - schema.org

Mobile

Move to mobile favors a move to speech which favors natural language information search

- Will we move to a time when half of searches are spoken?

Mobile

Mobile proved importance of NLU/QA

[What is the best time for wildflowers in the bay area]

Information quality

- There have always been concerns about information provenance (the source) and information reliability, especially among “information professionals” (reporters, lawyers, spies, ...)
- It wasn’t ignored on the web: ideas like PageRank were meant to find good content, and there has been a decade of work targeting link farms, etc.
- However, a lot of recent events have shown the limited effectiveness of that work, and how “fake” information easily gets upvoted and spreads
Towards intelligent agents

Two goals
- Things not strings
- Inference not search

Two paradigms for question answering

- Text-based approaches
  - TREC QA, IBM Watson, DrQA
- Structured knowledge-based approaches
  - Apple Siri, Wolfram Alpha, Facebook Graph Search

(And, of course, there are hybrids, including some of the above.)

At the moment, structured knowledge is back in fashion, but it may or may not last
upset stomach $\Rightarrow$ adverse effect (on stomach)

tylenol $\Rightarrow$ no adverse effects (on stomach) $\Rightarrow$ no upset stomach

<table>
<thead>
<tr>
<th>From</th>
<th>To</th>
<th>Requires</th>
</tr>
</thead>
<tbody>
<tr>
<td>Term</td>
<td>Concept</td>
<td>Parsing, disambiguation, coreference</td>
</tr>
<tr>
<td>Term identity</td>
<td>Entailment</td>
<td>Concept relations</td>
</tr>
<tr>
<td>Co-occurrence</td>
<td>Syntactic relation</td>
<td>Document structure, parsing</td>
</tr>
<tr>
<td>Term index</td>
<td>Semantic index</td>
<td>Concept disambiguation, inference</td>
</tr>
</tbody>
</table>

“Things, not strings”
Introduction to Information Retrieval

Task completion

Price prediction

Aggregate ratings

Task completion

Introduction to Information Retrieval

Direct Answer

Structured Data

Patrick Pantel talk

(Then) Current experience
Desired experience: Towards actions

Actions easily accessible

Recognize entity in query

Actions vs. Intents

User Intents and Goals

plan vacation
get in shape

Query

hilton orlando reviews
sea world location
how to lose weight

Query Intent

Informational
Navigational
Transactional

Finer-grained Intents

Advice
Locate
Download
Obtain
Interact

Actions on Entities

get address(landmark)
add to Netflix queue(film)

Learning actions from web usage logs

- Three months of us-en web logs
- Annotate with Freebase entities
- Keep queries with an entity in set of 21 types
- Filter out navigational queries
- Filter out clicked hosts that weren't clicked at least 100 times

- get reviews
- Orlando hotel reviews
- read biography
- Does Hope Solo have a boyfriend?
- download software
- Free Winzip download
- watch shows online
- watch family guy online

A hodgepodge of related strings
Only actionable through search
Entity disambiguation and linking

- Key requirement is that entities get identified
  - Named entity recognition (e.g., Stanford NER!)
- and disambiguated
  - Entity linking (or sometimes “Wikification”)
    - e.g., Michael Jordan the basketballer or the ML guy

- and linked to a canonical reference
  - Freebase, dbPedia, Yago2, (WordNet)

Mentions, Meanings, Mappings

Sergio talked to Ennio about
Eli wrote in the Ecstasy scene
This sequence on the graveyard was a highlight in
Sergio’s trilogy
of western films.

Mentions (surface names)

Entities (meanings)

KB

Sergio means Sergio_Leone
Sergio means Sergio_Gainsbourg
Ennio means Ennio_Antonelli
Ennio means Ennio_Morricone
Eli means Eli (bible)
Eli means Eli_Wallach
Ecstasy means Ecstasy (drug)
Ecstasy means Ecstasy_of_Gold
Ecstasy means Ecstasy_of_the_Gold_trilogy
trilogy means Star_Wars_Trilogy
triology means Lord_of_the_Rings
trilogy means Dollars_Trilogy

Understanding questions

Chris Watson
The 3rd Prime Minister of Australia

List of Prime Ministers of Australia, the 3rd

http://wiki/Dollars_Trilogy
http://wiki/The_Good_the_Bad_and_the_Ugly
http://wiki/Clint_Eastwood
http://wiki/Honorary_Academ

http://wiki/The_Good_the_Bad_and_the_Ugly
http://wiki/Metallica
http://wiki/Bellagio_casino
http://wiki/Extreme_Lightinfrastructure
http://wiki/Eli_Wallach
http://wiki/Sergio_Leone
http://wiki/For_a_Few_Dollars_More
Introduction to Information Retrieval

3 approaches to question answering:
Knowledge-based approaches (Siri)
- Build a semantic representation of the query
  - Times, dates, locations, entities, numeric quantities
  - Map from this semantics to query structured data or resources
    - Geospatial databases
    - Ontologies (Wikipedia infoboxes, dbPedia, WordNet, Yago)
    - Restaurant review sources and reservation services
    - Scientific databases
    - Wolfram Alpha

Text-based (mainly factoid) QA
- QUESTION PROCESSING
  - Detect question type, answer type, focus, relations
  - Formulate queries to send to a search engine
- PASSAGE RETRIEVAL
  - Retrieve ranked documents
  - Break into suitable passages and rerank
- ANSWER PROCESSING
  - Extract candidate answers (as named entities)
  - Rank candidates
    - using evidence from relations in the text and external sources

Hybrid approaches (IBM Watson)
- 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
Knowledge: Jeremy Zawodny says ...

Is the goal to go from language to knowledge bases?
- For humans, going from the largely unstructured language on the web to actionable information is effortlessly easy
- But for computers, it’s rather difficult!
- This has suggested to many that if we’re going to produce the next generation of intelligent agents, which can make decisions on our behalf
  - Answering our routine email
  - Booking our next trip to Fiji
then we still first need to construct knowledge bases
- To go from languages to information
- But should we rather just have computers work with language?

Knowledge: Not just semantics but pragmatics
Pragmatics = taking account of context in determining meaning
A natural part of language understanding and use
Search engines are great because they inherently take into account pragmatics ("associations and contexts")
- [the national] → The National (a band)
- [the national ohio] → The National - Bloodbuzz Ohio – YouTube
- [the national broadband] → www.broadband.gov

Who won the best actor Oscar in 1973?
Jack Lemmon was awarded the Best Supporting Actor Oscar in 1956 for Mister Roberts (1955) and the Best Actor Oscar for Save the Tiger (1973), becoming the first actor to achieve this rare honor.

Introduction to Information Retrieval
Full NLP QA: LCC (Harabagiu/Moldovan)
[below is the architecture of LCC's QA system circa 2003]
DrQA: Open-domain Question Answering

Open-domain Question Answering

SQaD
Q: How many of Warsaw’s inhabitants spoke Polish in 1933? A: 833,500

TREC
Q: What U.S. state’s motto is “Live free or Die”? A: New Hampshire

WebQuestions (Berant et al, 2013) Freebase
Q: What part of the atom did Chadwick discover? A: Neutron

WikiMovies (Miller et al, 2016)
Q: Who wrote the film Gigli? A: Martin Brest

55

WebQuestions
(Q: What part of the atom did Chadwick discover? A: Neutron)

Q: How many of Warsaw's inhabitants spoke Polish in 1933?

Document Retriever

Dataset Wiki Search Plain +bigrams
SQaD 62.7 76.1 77.8
CuratedTREC 81.0 85.2 86.0
WebQuestions 73.7 75.5 74.4
WikiMovies 61.7 54.4 70.3

70-86% of questions we have that the answer segment appears in the top 5 articles

Document Reader:
Stanford Attentive Reader

Bidirectional LSTMs

Bidirectional LSTMs

Stanford Attentive Reader

Attention
Stanford Attentive Reader++

Results (single model)

<table>
<thead>
<tr>
<th>Model</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic regression</td>
<td>51.0</td>
</tr>
<tr>
<td>Fine-Grained Gating (Carnegie Mellon U)</td>
<td>73.3</td>
</tr>
<tr>
<td>Match-LSTM (Singapore Management U)</td>
<td>73.7</td>
</tr>
<tr>
<td>DCN (Salesforce)</td>
<td>75.9</td>
</tr>
<tr>
<td>Multi-Perspective Matching (IBM)</td>
<td>78.7</td>
</tr>
<tr>
<td>ReaNet (Chen et al. 2017)</td>
<td>79.4</td>
</tr>
<tr>
<td>DCN (Salesforce)</td>
<td>80.8</td>
</tr>
<tr>
<td>Human performance</td>
<td>91.2</td>
</tr>
</tbody>
</table>

Results

- Pre-trained SQuAD
- Fine-tuning
- Multi-task learning

Exact match (top-1 prediction)

Demo

Hello! Please ask a question.

Who invented LSTM?

Hmmm. Tough one. I would say... Sepp Hochreiter

Where is Stanford University located?

I'm quite confident the answer is... California

Demo

What is the highest mountain on Earth?

I'm quite confident the answer is... Mount Everest

What year was the American Declaration of Independence?

I'm quite confident the answer is... 1776

Demo

What is the meaning of life?

I'm quite confident the answer is... What is the meaning in asking?

Who will solve AI?

Hmmm. Tough one. I would say... Computers
Thanks everyone! 😊