Introduction to Information Retrieval

CS276: Information Retrieval and Web Search
Lecture 19: Web Question Answering

Christopher Manning
Pandu Nayak

"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? 2013–2019

- 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? 2013–2019

- “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

Longer snippets in results pages (Nov 2017)
- Mobile-first Index (Mar 2018)
  - Index mobile version of websites in preference to desktop!
  - Revert snippet length in results pages (May 2018)
- “Medic” update (Aug 2018)
  - More emphasis on expertise, authoritativeness, trust
  - Big changes for diet, nutrition, medical products sites
- Core Algorithm Update (Mar 2019)
  - Seems kind of like “Medic 2”
- 2019 seems to have been kinda quiet so far …

Mobile

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

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

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
  - Mobile proved importance of NLU/QA
    - [What is the best time for wildflowers in the bay area]
  - Move to mobile favors a move to speech which favors natural language information search
Introduction to Information Retrieval

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

Example from Fernando Pereira (GOOG)
tylenol ⇒ painkiller

upset stomach ⇒ adverse effect (on stomach)

tylenol ⇒ no adverse effects (on stomach)
 ⇒ no upset stomach

“Things, not strings”

<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>
Patrick Pantel talk
(Then) Current experience

Desired experience: Towards actions

Politician

Actions vs. Intents

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 hedgepostage of related strings
- Only actionable through search

(Broder, 2002)
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)
Introduction to Information Retrieval

Understanding questions

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
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 double.

Source: Jack Lemmon -- Wikipedia

Who won the best actor Oscar in 1973?

- Scott Wen - tau Yih (ACL 2013) paper

---

What is the fastest car in the world?

The Jaguar XJ220 is the dearest, fastest and most sought after car on the planet.

Assume that there is an underlying alignment:

- Describes which words in and can be associated
- See if the (syntactic/semantic) relations support the answer

---

Introduction to Information Retrieval

Full NLP QA: LCC (Harabagiu/Moldovan)
[below is the architecture of LCC’s QA system circa 2003]

Open-domain Question Answering

SQuAD
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)
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

---

DrQA: Open-domain Question Answering
Document Retriever

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Wiki Search</th>
<th>Doc. Search</th>
<th>Retriever + bigrams</th>
</tr>
</thead>
<tbody>
<tr>
<td>SQuAD</td>
<td>62.7</td>
<td>76.1</td>
<td>77.8</td>
</tr>
<tr>
<td>ConLLSTREC</td>
<td>81.0</td>
<td>85.2</td>
<td>86.0</td>
</tr>
<tr>
<td>WebQuestions</td>
<td>73.7</td>
<td>75.5</td>
<td>74.4</td>
</tr>
<tr>
<td>WikiMovies</td>
<td>61.7</td>
<td>54.4</td>
<td>70.3</td>
</tr>
</tbody>
</table>

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

Stanford Attentive Reader

SQuAD 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>BIDAF (UI &amp; Allen Institute)</td>
<td>77.3</td>
</tr>
<tr>
<td>Multi-Perspective Matching (IBM)</td>
<td>78.3</td>
</tr>
<tr>
<td>ReasonNet (MSR Redmond)</td>
<td>79.3</td>
</tr>
<tr>
<td>BiDAF (UW &amp; Allen Institute)</td>
<td>79.4</td>
</tr>
<tr>
<td>Multi-Perspective Matching (IBM)</td>
<td>79.4</td>
</tr>
<tr>
<td>ReasoNet (MSR Redmond)</td>
<td>79.4</td>
</tr>
<tr>
<td>DrQA (Chen et al. 2017)</td>
<td>79.4</td>
</tr>
<tr>
<td>r-net (MSR Asia) (Wang et al., ACL 2017)</td>
<td>79.7</td>
</tr>
<tr>
<td>Human performance</td>
<td>91.2</td>
</tr>
</tbody>
</table>

General questions

Combined with Web search, we can answer 57.5% of trivia questions correctly

Q: The Dodecanese Campaign of WWII that was an attempt by the Allied forces to capture islands in the Aegean Sea was the inspiration for which acclaimed 1961 commando film?
A: The Guns of Navarone

A: Fitness