Search Engine Support For Software Applications

Jamie Callan
With help from Le Zhao and Paul Ogilvie

Language Technologies Institute
School of Computer Science
Carnegie Mellon University

Motivation for Today's Talk

In recent years I have been part of projects that use a search engine as a 'language database'
- Computer Assisted Language Learning (REAP)
- Question answering (Javelin)
- Read-the-Web

The search engine provides access to text … and information about text

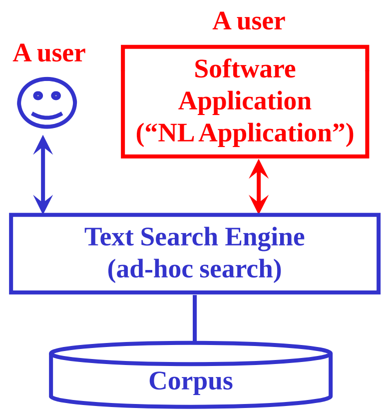
Motivation for Today's Talk

IR typically assumes that the user is a person

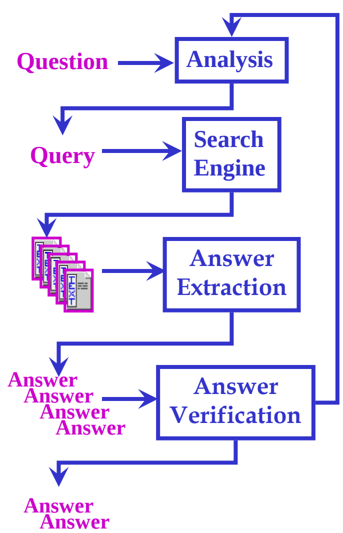
Applications are increasingly built on top of search engines

- Question answering, text mining, tutoring, MT, …

Most applications don't expect much of the search engine



A Common Approach to Building Applications on Top of Search Engines



Question analysis

- Expected answer type
- Answer extraction strategies
- Query creation

Simple keyword search

Answer extraction and verification

- Varying degrees of NL analysis
- Discard the junk

Question Answering Queries

"What year did Wilt Chamberlain score 100 points?"

A bag-of-words query

#date Wilt Chamberlain score 100 points

A query that uses semantic role labels

#combine[target](Score
#combine[./argm-tmp](#any:date)
#combine[./arg0](Wilt Chamberlain)
#combine[./arg1](100 points)))

Problems With This Approach

Queries are usually bag-of-words or simple patterns

– The application's requirements are actually more complex

Search quality is often poor

– Answers may need to satisfy complex constraints (that the search engine does not know about)
– Several queries may be needed to find useful passages

This reinforces the view that text search is inherently limited

Motivation for Today's Talk

Rich language resources are emerging

- WordNet , CIA Factbook, …
- Text annotators (POS, NE, SRL, …)
- Freebase, Dbpedia, TextRunner, Billion Triple, …

We aren't very good at using these effectively

- Special purpose uses: Some progress
- General purpose uses: ???

7

Motivation for Today's Talk

I want the search engine to know as much as possible

- About the application's information need
  - » Probably expressed as a structured query
- About the document contents
  - » Text + text analysis (pick your favorite types)
  - » Probably organized in a structured document
- About what the language might mean

I want general purpose methods of using this information

8

Motivation for Today's Talk

Structured queries & documents are old and well-studied IR topics

– Usage dates back to the earliest Boolean systems

Do we really understand them?

– Basic structure: Yes

– Advanced uses of structure: I'm not so sure

So … let's talk about it

What is a Document?

A document is a container for information

– Any kind of information

A document is a structured object

– Maybe the structure is simple, or maybe not

Some of the information it contains is unstructured

– Maybe all of it is unstructured, or maybe not

A Typical View of a Document

Metadata

– Often <attribute, value> data

– E.g., date, author, source, language, …

Content

– Typically text

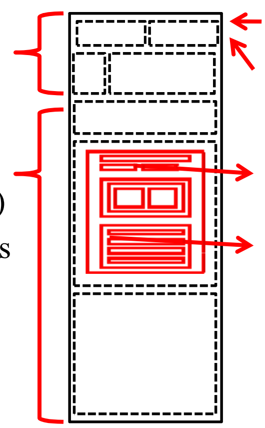
– Maybe organized into fields (elements)

» E.g., title, abstract, body, references

Relations

– E.g., citations, hyperlinks

A Document



Computer Assisted Language Learning: The REAP Project

REAP provides individualized reading practice for (mostly advanced) English language learners

Given

– A detailed model of an individual

– A model of what a fluent speaker should know

Find <u>current</u> and <u>authentic</u> texts that contain vocabulary that she should learn or practice next

– Preferably texts on a topic that interests the student

REAP From an IR Perspective: Documents

Crawl 100-200 million documents

Use text categorization to filter out "bad" documents

– A pipeline of filters for different types of "bad"

Use text categorization to create document metadata

– Reading difficulty, topic, …

Index and search with Indri

13

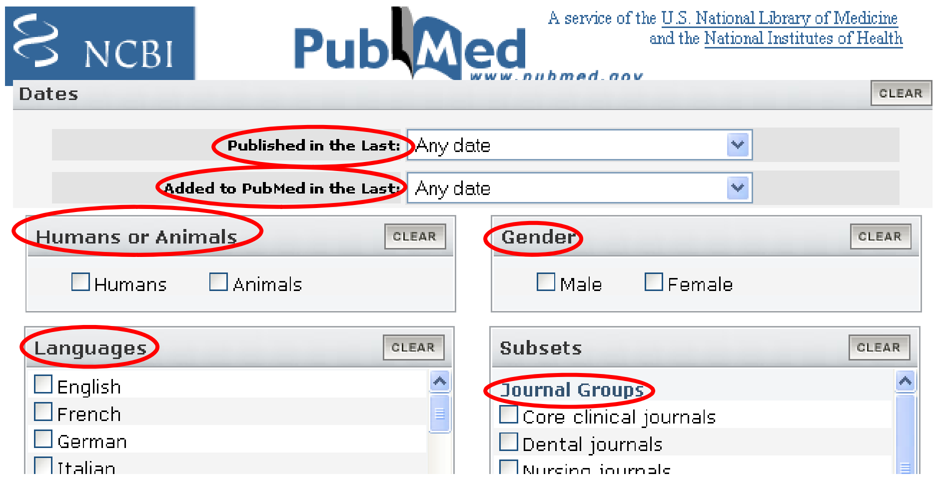
REAP From an IR Perspective: Queries

Retrieve passages that show typical usage of "abate", "smog", and "highlight"

#VB (abate) AND #VB (highlight) — **Focus Words**
AND #NN (smog)
AND #COOCCURS_WITH (abate, highlight, smog) — **Typical Usage**
AND #WEIGHT (0.7 #TOPIC (Technology) 0.3 #TOPIC (Finance)) — **Student Interest**
AND #Length (1, 2000)
AND #Difficulty (7, 9) — **Instructional Constraints**
AND …

14

In The Beginning Were Fields….



In The Beginning Were Fields….

Attributes + text isn't exciting, but it is <u>very</u> common

A typical approach

- <u>Exact-match</u> Boolean retrieval model over attributes
  - » Note assumption that attributes are correct
- Maybe best-match retrieval model over text

Maybe exact-match Boolean is good enough

- » But best-match search on the attributes would be nice
- » E.g., we would accept a slightly easier document

Documents with Text Annotations

Text annotations are becoming more common

- Sentence
- Part-of-speech (POS)
- Named entity (entity)
- Dependency parse
- Semantic role labels (SRL)
- Logical form

…

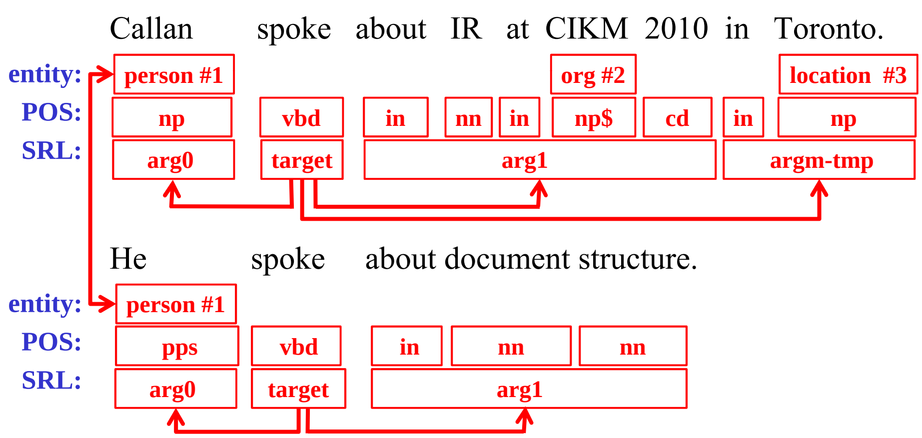
Documents With Text Annotations

	Callan	spoke	about	IR	at	CIKM	2010	in	Toronto.
entity:	person					org			location
POS:	np	vbd	in	nn	in	np$	cd	in	np
SRL:	arg0	target	arg1					argm-tmp	

	He	spoke	about	document	structure.
entity:	person				
POS:	pps	vbd	in	nn	nn
SRL:	arg0	target	arg1		

(Annotations from LingPipe and Assert)

Documents With Relational Text Annotations



19

Supporting Text Annotations

Text annotations can be considered "small fields"

– We think that we understand fields fairly well

– Thus, many existing systems can support annotations

Is this sufficient?

20

Retrieval of Annotated Text

Query: #sentence (#person (obama))

- **S1:** President Obama to Appear on Mythbusters.
- **S2:** President Barack Obama checks out some …
- **S3:** What myth will Obama be debunking …
- **S4:** President Obama challenged Jamie and Adam …

person name annotations

Often the field retrieval model is unranked exact match

– That would work here

…but isn't a general solution for annotations

Retrieval of Annotated Text

Query: #sentence(#**target** (take #./**arg1**(measures)))

- **S1:** It must **take measures**.
- **S2:** U.S. investments worldwide could be in jeopardy if other countries **take** up **similar measures**.
- **S3:** **Chanting** the slogan "**take measures** before they **take our measurements**," the Greenpeace activists **set** up a coffin outside the ministry to **draw** attention to the deadly combination of atmospheric pollution and **rising** temperatures in Athens, which are **expected** to **reach** 42 degrees centigrade at the weekend.
- **S4:** The Singapore government will **take measures** to **discourage** speculation in the private residential property market and **tighten** credit, particularly for foreigners, Deputy Prime Minister Lee Hsien Loong **announced** here today.

(Zhao and Callan, 2008)

Retrieval of Annotated Text

Term weighting in short fields is difficult

- Current normalization models don't handle this range well
- Scores have high variance

Weighting must address

- Variation in length
- Variation in reliability

S1: President Obama to Appear on Mythbusters.
S2: President Barack Obama checks out some …
S3: What myth will Obama be debunking …
S4: President Obama challenged Jamie and Adam …

S1: It must take measures.
S2: U.S. investments worldwide could be in jeopardy if other countries take up similar measures.
S3: Chanting the slogan "take measures before they take our measurements," the Greenpeace activists set up a coffin outside the ministry to draw attention to the deadly combination of atmospheric pollution and rising temperatures in Athens, which are expected to reach 42 degrees centigrade at the weekend.
S4: The Singapore government will take measures to discourage speculation in the private residential property market and tighten credit, particularly for foreigners, Deputy Prime Minister Lee Hsien Loong announced here today.

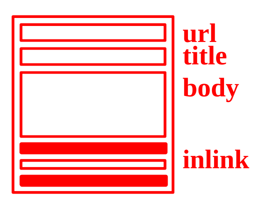
Callan spoke about IR at CIKM 2010 in Toronto.

Retrieval of Annotated Text: Multiple Matches

Query: #document (#inlink (fairmont royal york hotel))

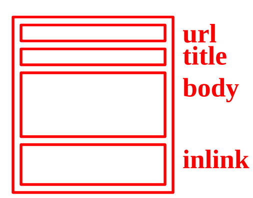
This document has several inlink fields

- What if two (or more) match?
- How is the evidence combined?



One common solution

- Only allow one field per datatype
- Fine for some cases
- Not a general solution



Retrieval of Annotated Text: Multiple Matches

Query: #sentence(#**target** (take #./**arg1**(measures)))

S3: **Chanting** the slogan “**take measures** before they **take our measurements**," the Greenpeace …

This sentence has several target annotations (fields)

- Two match
- How is the evidence combined?

Retrieval of Annotated Text: Multiple Matches

We know how to think about “ordinary” fields

- #and (#title (…) #abstract (…) #author (…))

Does this make sense for text annotations?

- A sentence might have several target fields that match
- #sentence (#combine (#**target** (take #./**arg1**(measures))))
- What form should #combine take?
  - » Probabilistic AND? Probabilistic OR? Average?
  - » Would we prefer something more like tf?

(Zhao and Callan, 2008)

Retrieval of Annotated Text

Query: #sentence (#person (obama) #person (jamie))

- **S3:** What myth will Obama be debunking …
- **S4:** President Obama challenged Jamie and Adam …

Both S3 and S4 contain a matching #person annotation

– But, S4 is the better match

This is a major problem in using text annotations

– If the field (annotation) isn't present, nothing matches

– Exact-match on structure

Retrieval of Annotated Text

Annotations are less reliable than traditional structure

– Not created by the document author or publisher

– Created by software that makes mistakes

– Maybe identifying properties that people don't agree on

Treating them like fields overlooks these differences

– Annotations are noisy structure

Matching Noisy Annotations: Current Practice

A more robust query

– #sentence (#weight (0.3 #person (obama) 0.7 obama))

Smoothing

$$P_{smooth}(q_i|e) = \lambda_1 P_{MLE}(q_i|e) + \lambda_2 P_{MLE}(q_i|s) + \lambda_3 P_{MLE}(q_i|c)$$

See Elsas, et al., this conference for a related approach

Still very much an open problem

Common Types of Annotation Errors

Missing annotation

– E.g., named entities, …

President Obama challenged Jamie and Adam …

Bad annotation boundary

– E.g., semantic role labels, …

Callan spoke about IR at CIKM 2010 in Toronto.

Conflated annotations

– E.g., part of speech tags, semantic role labels, …

He/PPS lived/VBD at/IN the/AT white/JJ house/NN.

Matching Noisy Annotations: Possible Practice

Give the system an error model for each annotator

- Different types of annotators make different types of mistakes

Automatic reformulation of query structure based on the probability of different annotation mistakes

An open problem…

(Zhao and Callan, 2009)

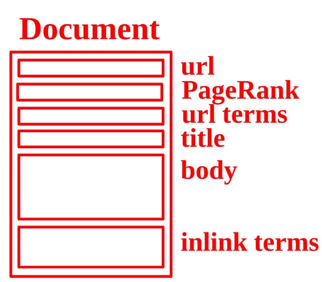
Relations

Relations among documents and elements are common

- Hyperlinks and RDF: Cross-document relations
- XML: Within-document relations

One common approach

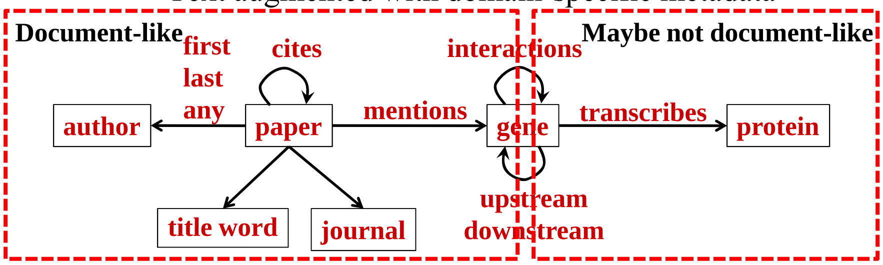
- Materialize the relation
- Works for some special cases
- Not a general solution



Relational Retrieval

Consider a relational model of PubMed abstracts

– Text augmented with domain-specific metadata



Venue recommendation: title, genes, proteins → journal

Expert finding: Title, genes, protein → author

(Lao and Cohen, 2010)

Relational Retrieval

Some of this problem can be cast as typical retrieval

– Title, author, journal, cites, genes

The domain-specific information is harder to integrate

– Gene transcribes protein?

– Gene upstream/downstream of gene?

There have been some successes, e.g., at TREC

– Problem-specific, heuristic, post-processing, …

– No general guidance

Relational Retrieval: How Would We Do It?

The search engine has many types of "documents"

- Author, paper, journal, gene, protein, …
- Documents have typed relations

The query language specifies what and how to retrieve

- Standard retrieval capabilities
- Random walk or other propagation along links

This feels doable

- Is it the right approach? Is it enough?

35

Read the Web's Never Ending Language Learner (NELL)

NELL does open-domain information extraction

- On English ClueWeb09 and Google search results
- Entities and relations
- 440,000 beliefs and growing daily

Knowledge is organized by a loose ontology

36

Inferred Knowledge: What NELL "Knows" About IBM

Member of Category: Organization, Company

Acquired: Cognos, Informix, Filenet, Ascential, …
Acquired By: Lenovo
CEO: Lou Gerstner
Competes With: Google, Oracle, Sun
Economic sector: Information technology, Consulting, …
Offices In: San Jose, Zurich, Austin, Haifa, New York City
…

37

So, What Have I Talked About?

- **Applications built on top of search engines**
- **Exact match field retrieval**
- **Issues with treating text annotations like fields**
  - Weighting
  - Combining evidence from multiple matches
  - Noisy structure, error models
- **Relational retrieval**
- **Integration of (loosely) structured information**

38

Why is this Important?

(Most) IR systems do not decide the meaning of a text before the information need is known

– Ambiguity is retained … and that's a good thing

– Interpret the meaning based upon the information need

– This is very powerful

However, our systems should not be dumb about meaning

– Incorporate state-of-the-art language analysis tools flexibly

– Allow query-time decisions about what and how to use

A Research Agenda for Text Search

Software applications are a new and challenging class of search engine users

– Multiple forms of knowledge and language analysis

– Metadata and structure of varying reliability

More of us should be thinking about how to support them

– Many interesting unsolved core IR problems

– Diverse information resources to exploit

– New retrieval models

– Interesting new applications

Thanks!

41