Question-Answering from Web Resources

Tat-Seng Chua
School of Computing
National University of Singapore
17 March 2010, CAMP’10, Shah Alam, Malaysia
Information-Rich World

- Web has opened up vast resources of knowledge to us
  - Can find almost any contents on the Web …
  - Access Web via search engines, forum sites, social networks
- Info on the Web also becoming more multimedia & multi-faceted
  - Wikipedia has over 1 million geo-tagged articles
  - Flickr receives over 3 million geo-tagged photos per month, & YouTube
From Document Retrieval to Question-Answering

- We are used to imprecise results from search engines
  - We learn to adapt!
  - But often need to browse thru many pages and modify query iteratively to get what we want eventually

- One direction is to provide more precise or exact info
  - Passage retrieval
  - Question-Answering, for example, POWERSET
A QA User Scenario

Adult, a native speaker of English, and an “average” reader of newspapers.

User / Assessor

Information Need

 mozart

MOZART

“Who is Mozart”

List his operas

Where was he born?

Others?
How, why or what do you think?

QA System

Answers
Types of QA

• Research on QA focuses on finding:
  o Factoid: “Where is he born?”
  o Definition: “Tell me something about him?”
  o How-to: “How to learn to play like him?”
  o Opinion-oriented: “What is your opinion of his opera?”
  o Analysis answers etc: “Why is his music so popular?”

• In increasing order of difficulties…
  o Needs advanced NLP techniques and reliable sources to deliver quality answers
Traditional QA Architecture

- Given a “free-text” question
  - Question Analysis: “type, key element, context”
  - Document Retrieval: “retrieve top N docs for detailed analysis”
  - Passage Retrieval: “narrow down to relevant passages”
  - Answer Extraction: “depending on task..”
  - Answer validation: “validate/proof that the answer is correct??

- We have been working on QA research since 2002..
  - Our past research on factoid and Definition QA has achieved top positions (in 2003-05) in large-scale international TREC QA Evaluations
QA on the Web

- Because of the difficulty, the retrieval accuracy is limited
- Given the availability of large public info resources...
  - Emerging research focuses on finding ready answers from reliable public sources
Factoid QA on Web Resource - 1

- Focus on QA on Wikipedia & large on-line QA sites

Infobox derived from template

```json
{{Infobox Person
  name = William Henry Gates III
  image = Bill Gates at an IT forum in Danmark, 2004
  caption =
  birth_date =
  birth_place =
  death_date =
  death_place =
  body_discovered =
  death_cause =
  resting_place =
  resting_place_coords =
  residence =
  nationality =
  ethnicity =
  citizenship =
  other_names =
  known_for =
  education =
  alma_mater =
  employer =
  occupation =
  home_town =
  title =
  salary =
  networth =
  height =
  weight =
  term =
  predecessor =
  successor =
  party =
  boards =
  religion =
  spouse =
  partner =
  children =
  parents =
  relations =
  nationality =
  signature =
  website =
  footnotes =
}}
```

Full version (44 properties)

Use domain knowledge and structure to extract information
Factoid QA on Web Resource - 2

Exact answers will be there (for over 2 million available topics)

Who accompanied her when she died?

When did she die?

Carolyn Bessette-Kennedy

Born

January 7, 1966

White Plains, Westchester County, New York

 Died

July 16, 1999 (aged 33)

Atlantic Ocean

Arguments/constraints

Deep parsing results of sentences about the focus

On July 16, 1999, Bessette-Kennedy left Essex County Airport in Fairfield Township, New Jersey, for Hyannis, Massachusetts, to be with her husband.

Their ashes were scattered at sea on July 22, 1999.
Who are Lee Kuan Yew’s sons?

Lee Hsien Yang
Lee’s younger son, Lee Hsien Yang, is a former Brigadier-General and also a former President and Chief Executive Officer of SingTel, a pan-Asian telecommunications giant and Singapore’s largest company by market capitalisation (listed on the Singapore Exchange, SGX).
http://en.wikipedia.org/wiki/Lee_Hsien_Yang

Lee Hsien Loong
On 12 August 2004 Goh Chok Tong stepped down in favour of Lee’s eldest son, Lee Hsien Loong.
http://en.wikipedia.org/wiki/Lee_Hsien_Loong

Lee
Temasek Holdings in turn is run by Executive Director and CEO Ho Ching, the wife of Lee’s elder son, the Prime Minister.
http://en.wikipedia.org/wiki/Lee

his son Lee Hsien Loong
He currently holds the self-created post of Minister Mentor under his son Lee Hsien Loong, who became the nation’s third prime minister on August 12, 2004.
http://en.wikipedia.org/wiki/Lee_Hsien_Loong

his grandfather, Lee Hoon Leong
As a child Lee was strongly influenced by British culture, due in part to his grandfather, Lee Hoon Leong, who had given his sons an English education.
http://en.wikipedia.org/wiki/Lee_Hoon_Leong

Get answers directly
Do not need to go through source text again !!!
Factoid QA on Web Resource - 3

- Traditional QA: doc retrieval + passage retrieval + answer extraction (key-term based)
- Our approach: utilize topic + domain + constraints to search the exact answers in the huge semantic ontology assemblies that are constructed based on Wikipedia and the Web
  - Apply to Wikipedia and Baidu-pedia
What is Definitional QA?

- Definitional QA was introduced in TREC 2003
  - What is X? Who is X? What defines X?
- Quality of answers is defined in terms of answer nuggets
  - Sentence fragments that convey pieces of relevant information about the target topic
- Example: The answer nuggets for “George Foreman” is:
  - Was graduate of Job Corps.
  - Has lent name to line of food preparation products.
  - Waved American flag after winning 1968 Olympics championship.
  - Returned to boxing after 10 yr hiatus.
  - Became oldest world champion in boxing history.

- A query-oriented summarization
Summarization using DCL - 1

• Document concept lattice (DCL) model
  o Aim to find list of sentences that result in minimal info lost
  o Extract key concept terms, and build concept lattice
  o Perform sentence extraction at global level to cover max # of concept terms
Perform optimization at lattice level to extract sentences
Summarization using DCL - 3

• **Document concept lattice (DCL) model**
  - Aim to find list of sentences that result in minimal info lost
  - Extract key concept terms, and build concept lattice
  - Perform sentence extraction at global level to cover max # of concept terms
  - Participated in DUC, came in 1st (2005) and 2nd (2006)

• **Extend to Definition QA (or Query-focus summarization)**
  - Given a topic, extract min. passages to define topic → query-oriented summarization
  - Employ pattern-based model to extract definition-like sentences (SIGIR05)
  - Incorporate human interest model (HIM) (SIGIR07)
  - Employ combinations of pattern, HIM & Lattice models for summarization
Start with an Article on Mobile Operators in Singapore

DCL Online Summarizer

Query (Optional)

price

Go!

Home

Singaporean mobile operators MobileOne (M1) and StarHub will start selling Apple's iPhone on Dec. 9, but users won't see big drops in price for monthly plans.

Until now, Singapore Telecom (SingTel) was the only operator with rights to sell the iPhone in Singapore. That changed in October, when M1 announced a deal with Apple to sell the iPhone. StarHub followed suit in November with an identical announcement.

Hopes that increased availability of the iPhone might lead to lower prices appear unfounded. However, M1 and StarHub are being more generous with the amount of data that subscribers can download before paying excess data charges.

The lowest subscription prices on offer come from M1, which will offer four plans, ranging from S$36 to S$198 (US$23 to $142) per month. With the cheapest monthly plan and a two-year contract, subscribers pay S$818 for a 16GB iPhone 3G and S$658 for a 32GB iPhone 3GS. The handsets are free for users who sign up for the most expensive plan.

StarHub's four plans are slightly more expensive than M1, but are just below those offered by SingTel.

Subscribers of SingTel's two cheapest monthly plans get just 500MB and 1GB of data downloads, with the two most expensive monthly plans offering 2GB and 3GB, respectively.

SingTel's four plans range from S$39 to S$205, with users paying S$548 and S$678 for the 16GB and 32GB versions of the iPhone 3GS, respectively, under the cheapest monthly plan.

Define focus of summary: “Price of iPhone in Singapore”
Now users' focus turns to “StarHub” for relevant information
Or other focuses like “Singtel” etc.
Definition QA on Wikipedia (ACL 2010)

- Utilize Wikipedia structures: Infobox & Outline
- Utilize lattice model to find list of sentences with min info lost
- Leverage structure of topics, domain knowledge + context
- => effective query-focused summarizer for Wikipedia
### Profile with 9,777 words

Images, from top, left to right: Merlion by the CBD, Singapore Zoo entrance, Esplanade - Theatres on the Bay, Gateway of Sentosa, Statue of Thomas Stamford Raffles, Downtown Core of Singapore, Raffles Hotel.

Singapore, officially the Republic of Singapore, is an island city-state off the southern tip of the Malay Peninsula. 137 kilometres (85mi) north of the equator, south of the Malaysian state of Johor and north of Indonesia's Riau Islands.

At 710.2km² (274.2sqmi), [9] Singapore is a microstate and the smallest nation in Southeast Asia.

It is substantially larger than Monaco and Vatican City, the only other surviving sovereign city-states.

Before European settlement, the island now known as Singapore was the site of a Malay fishing village at the mouth of the Singapore River.

Several hundred indigenous Orang Laut people also lived along the nearby coast, rivers and on smaller islands.

In 1819, the British East India Company, led by Sir Stamford Raffles, established a trading post on the island, which was used as a port along the spice route.

[9] Singapore became one of the most important commercial and military centres of the British Empire, and the hub of British power in Southeast Asia.

During the Second World War, the colony was occupied by the Japanese after the Battle of Singapore, which Winston Churchill called "Britain's greatest defeat.

[10] Singapore reverted to British rule in 1945, immediately after the war.

Eighteen years later, in 1963, the city, having achieved independence from Britain, merged with Malaya, Sabah, and Sarawak to form Malaysia.

However, Singapore's merger proved unsuccessful, and, less than two years later, it seceded from the federation and became an independent republic within the Commonwealth of Nations on 9 August 1965.

Singapore was admitted to the United Nations on 21 September of that year. Since independence, Singapore's standard of living has risen dramatically.

### Definition

English is the main language of Singapore and has been heavily promoted as such since the country's independence.

The Singapore Armed Forces is one of the three services of the Singapore Armed Forces.

Singapore is a parliamentary republic, and the Constitution of Singapore establishes representative democracy as the nation's political system.

Singapore is also considered to be one of the top centres of trade in the world.

The use of English became widespread in Singapore after it was implemented as a first language medium in the education system, and English is the most common language in Singaporean literature.

Singapore supports the concept of Southeast Asian regionalism and plays an active role in the Association of Southeast Asian Nations (ASEAN), of which Singapore is a founding member.

The Singapore Armed Forces, the military forces of Singapore, takes charge of the overall defence of the country.

The final branch, the Republic of Singapore Navy (RSN), is the navy of the Singapore Armed Forces, responsible for the defence of Singapore against airborne threats and protection of its sea lines of communications.
Answering “How-To” Questions -1

• Very difficulty – problem in understanding the question and generating the answers

• Community QA systems are good resource to offer this service
  ○ Yahoo!Answer has tremendous QA resources in archives: 60M QA pairs in YA (the latest statistic obtained on 12-12-08 manually!)

• Simplify problem as similar question search problem

• But it is still a difficult problem
  1. Long verbose natural language query
  2. Noisy and Ungrammatical language
  3. Multiple forms of questions
  4. Uneven quality of answer

• Examples of multiple forms of questions:
  ○ How can I lose weight in a few month?
  ○ What should I do to lose pound in a short period?
Answering “How-To” Questions -2

- Explore Syntactic Tree Matching (SIGIR’09)
  - To tackle problems 2 & 3
  - Tree decomposition
  - Tree fragment matching at production rule level
  - Matching towards leaf level – robust against grammatical errors

- Parsing Tree for the sentence: “Good workout plan for losing a little bit of weight?”
- Parsing Tree for another variant of sentence: “Good workout plan to losing a little bit of weight?”

- Structures of sentence at production rule level remain the same!!
Answering “How-To” Questions -3

- Semantic-smoothed Tree Matching
  - Perform semantic similarity matching between leaf words

- Perform Answer Matching
  - Questions with similar answers tend to be semantically similar too

- Overall Retrieval Model

![Diagram showing the overall retrieval model with steps](image)
**Result Analysis**

**Query question:** what is the reason that my gums bleeding when I brush my teeth?

<table>
<thead>
<tr>
<th>Rank</th>
<th>Matched Questions</th>
<th>Results returned by BoW method</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>qid=20080330042256AAYGcpJ bleeding gums whenever brush teeth?</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>qid=20080102233315AArC6GR What are the reasons my gums could be bleeding?</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>qid=20080505223640AAPMuYc what is the reason for bleeding in gums?</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>qid=20080221211719AA5PMD0 why do my gums bleed when i brush my teeth??</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>qid=2008050503109AAKs4Ba why are my gums bleeding when I brush my teeth?</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>qid=20080113135256AAAX72kx My gums bleed when i brush my teeth, why?</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>qid=20080518173957AAZlYkN whenever i brush my teeth, my gums start to bleed, is that a problem, and how do i stop it?</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>qid=20080107034333AAC7hFl I was brushing my teeth and my gums started bleeding</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>qid=20080317152130AAQLpW4 My gums bleed when i brush my teeth.. not all the time but quite often..?</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>qid=20080506115638AAHgc62 When I brush my teeth, my gums bleed... at first it was a little but now its more...?</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rank</th>
<th>Matched Questions</th>
<th>Results returned by STM method</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>qid=20080113135256AAAX72kx My gums bleed when i brush my teeth, why?</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>qid=20080221211719AA5PMD0 why do my gums bleed when i brush my teeth??</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>qid=2008050503109AAKs4Ba why are my gums bleeding when I brush my teeth?</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>qid=20080113135256AAAX72kx My gums bleed when i brush my teeth, why?</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>qid=20080505223640AAPMuYc what is the reason for bleeding in gums?</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>qid=20080326043133AAx74HO Do your gums bleed when you wash your teeth?</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>qid=20080330042256AAYGcpJ bleeding gums whenever brush teeth?</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>qid=20080518173957AAZlYkN whenever i brush my teeth, my gums start to bleed, is that a problem, and how do i stop it?</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>qid=20080317152130AAQLpW4 My gums bleed when i brush my teeth.. not all the time but quite often..?</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>qid=20080105084121AAS7kYG my teeth bleed when i brush them?</td>
<td></td>
</tr>
</tbody>
</table>
Result Analysis – robustness

Original Query: what is the reason that my gums bleeding when I brush my teeth?
Query with grammatical errors: wat the reazon my gum bleding when I brush my teeth?

<table>
<thead>
<tr>
<th>Original result list</th>
<th>New result list (with errors)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Rank</strong></td>
<td><strong>Matched Questions</strong></td>
</tr>
<tr>
<td>1</td>
<td>qid=20080113135256AAAX72kx My gums bleed when i brush my teeth, why?</td>
</tr>
<tr>
<td>2</td>
<td>qid=20080221211719AA5PMD0 why do my gums bleed when i brush my teeth??</td>
</tr>
<tr>
<td>3</td>
<td>qid=2008050503109AAKs4Ba why are my gums bleeding when I brush my teeth?</td>
</tr>
<tr>
<td>4</td>
<td>qid=20080102233315AArC6GR What are the reasons my gums could be bleeding?</td>
</tr>
<tr>
<td>5</td>
<td>qid=2008050523640AAPMuYc what is the reason for bleeding in gums?</td>
</tr>
<tr>
<td>6</td>
<td>qid=20080326043133AAx74HO Do your gums bleed when you wash your teeth?</td>
</tr>
<tr>
<td>7</td>
<td>qid=2008030042256AYAYGcpJ bleeding gums whenever brush teeth?</td>
</tr>
<tr>
<td>8</td>
<td>qid=20080518173957AAZiYkN whenever i brush my teeth, my gums that a problem, and how to fix it</td>
</tr>
<tr>
<td>9</td>
<td>qid=2008031752130AAQLpW4 My gums bleed when i brush my teeth.. not all the time but quite often..?</td>
</tr>
<tr>
<td>10</td>
<td>qid=20080105084121AAS7kYG my teeth bleed when i brush them?</td>
</tr>
</tbody>
</table>

Less affected! Overall improvement in results is 12%
Multi-Sentence Segmentation -1

- cQA questions are often long and contain multiple sub-questions
- Example of user query: (do you have health insurance?)\(^1\) (is it provided by your employer?)\(^2\) (do you supply your own?)\(^3\) (if so how much do you have to pay)\(^4\) and (is it worth it for services you get?)\(^5\)?

- Retrieval results without segmentation:

<table>
<thead>
<tr>
<th>Rank</th>
<th>Matched Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>qid=20071203201922AAamL5K How do you pay your health insurance deductibles?</td>
</tr>
<tr>
<td>2</td>
<td>qid=20080528102757AA5qfMK How can I get help paying for health insurance?</td>
</tr>
<tr>
<td>3</td>
<td>qid=20080103120157AAjhfrw Health Insurance Through Employers?</td>
</tr>
<tr>
<td>4</td>
<td>qid=20080208204910AAC1wWb health insurance during a switch of employment?</td>
</tr>
<tr>
<td>5</td>
<td>qid=20080425051938AA8be20 How much would I pay for a tonsillectomy if I did not have health insurance?</td>
</tr>
<tr>
<td>6</td>
<td>qid=20080416173725AAPDCPI how to get health insurance?</td>
</tr>
<tr>
<td>7</td>
<td>qid=20080520055704AAJXnft How to find best health insurance provider?</td>
</tr>
</tbody>
</table>
Multi-Sentence Segmentation -2

- Back to example of user query:
  (do you have health insurance?)$^1$ (is it provided by your employer?)$^2$ (do you supply your own?)$^3$ (if so how much do you have to pay?)$^4$ and (is it worth it for services you get?)$^5$?
  - Several sub-questions are asked in the user’s query!

- With question segmentation
  - Different sub-questions are well separated
  - Can find better matches for each sub-questions
  - More exact answers for different sub-questions

- Our approach:
  - Build a question detector (QD) model
  - Build graph-based model for question threading – group sub-question + context
Result Analysis (Example 2)

- **User Query:**
  - What are the requirements of being a dentist?

- The following question thread contains a good similar question that could be matched with the user’s query:

  I heard somewhere that in order to become a dentist, you need certain hours of volunteering or shadowing. Is that true? If it is, how many hours? I have only a few hours of such activity… and can you write down other requirements that one would need to become a dentist? I know there are a lot of things but if you can write down as much as you can, that'd be a lot of help. Thanks

- **Without question segmentation**, the above answer cannot be matched (as the underlined question is not explicitly separated).
- **Question segmentation** has result in over 10% improvement in cQA matches.
Towards New Search Paradigm on Web Scale Info Collections - 1

- Consolidate QA and opinion mining research to develop a system to answer a wide range of questions
  - Factoid & Definition questions from Wikipedia & Web
  - How-to questions from Community QA archives
  - Opinion questions from forum sites
  - Organize info around a topic into, say FAQ

- Extend to multilingual & multimedia QA
  - Find Chinese answers from say Baidu-I-Know
  - Find media answers from Flickr, Youtube

For example:
  - How to transfer photos from camera to computer?
  - How to check engine oil?
Towards New Search Paradigm on Web Scale Info Collections -2

• Overall Goal, WebQA = wQA + cQA + vQA + iQA
  - Wikipedia (w), Yahoo!Answers (c), YouTube (v), Flickr (image); and equivalent in Chinese

• Target: Multilingual & Multimedia web-QA
Question-Answer?

Contact: Chua Tat-Seng (chuats@comp.nus.edu.sg)