“Tell Me More”
Using Ladders in Wikipedia

Siarhei Bykau (Bloomberg L.P.)
Jihwan Lee (Purdue University)
Divesh Srivastava (AT&T Labs-Research)
Yannis Velegrakis (University of Trento)
What is “Tell Me More”?

♦ We often ask entity queries such as “Who is Barack Obama?”

♦ Given the query answer, we would typically like to know more.
What is “Tell Me More”?

♦ Consider a role query “Who is the mayor of Chicago?”

♦ In this case, what should “tell me more” return?
  – More about the answer entity?

Born: November 29, 1959 (age 57), Chicago, IL
Height: 5' 7"
Nationality: American
Spouse: Amy Rule (m. 1994)
What is “Tell Me More”?

- Consider a role query “Who is the mayor of Chicago?”

- In this case, what should “tell me more” return?
  - More about the answer entity? More about the role?

Born: November 29, 1959 (age 57), Chicago, IL

Height: 5’ 7”

Nationality: American

Spouse: Amy Rule (m. 1994)

On February 23, 2011, former Illinois Congressman and White House Chief of Staff, Rahm Emanuel, won the mayoral election, beating five rivals with 55 percent of the vote, and was sworn in as Mayor on May 16, 2011.
Focus of this Paper

♦ “Tell me more” **about the role** of role queries.
  – Role query answerable using Infobox attribute in Wikipedia page.
  – E.g., “Mayor” in “Chicago” page.

♦ Look at text on Wikipedia page.
  – Find **attribute** in text.
  – E.g., find mayor Rahm Emmanuel.
  – Return surrounding text.

♦ Raises interesting challenges.
Role Linkage in Wikipedia Page

♦ Approach 0: find Infobox attribute value in Wikipedia page text.
  – Works for “Rahm Emanuel is the mayor of Chicago”.
  – Value is typically a Wikilink.

On February 23, 2011, former Illinois Congressman and White House Chief of Staff, Rahm Emanuel, won the mayoral election, beating five rivals with 55 percent of the vote,[82] and was sworn in as Mayor on May 16, 2011.
Role Linkage in Wikipedia Page

- Approach 0: find Infobox attribute value in Wikipedia page text.
  - Does it work for “New York City is the largest city of US”?

largest immigrant population.[26] The capital is Washington, D.C., and the largest city is New York City; nine other major metropolitan areas—each with at least 4.5 million inhabitants and the largest having more than 13 million people—are Los Angeles, Chicago, Dallas, Houston, Philadelphia, Miami, Atlanta, Boston, and San Francisco.
Role Linkage in Wikipedia Page

♦ Approach 0: find Infobox attribute value in Wikipedia page text.
  – Does it work for “New York City is the largest city of US”?

On September 11, 2001, Al-Qaeda terrorists struck the World Trade Center in New York City and the Pentagon near Washington, D.C., killing nearly 3,000 people. In response, Ellis Island in New York City was a major gateway for European immigration.

<table>
<thead>
<tr>
<th>Capital</th>
<th>Washington, D.C.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>38°53'N 77°01'W</td>
</tr>
<tr>
<td>Largest city</td>
<td>New York City</td>
</tr>
<tr>
<td></td>
<td>40°43'N 74°00'W</td>
</tr>
<tr>
<td>Official languages</td>
<td>None at federal level[fn 1]</td>
</tr>
</tbody>
</table>
| National language| English[fn 2] }
Role Linkage in Wikipedia Page

♦ Approach 1: Approach 0 + synonym of attribute name in vicinity.
  – Does it work for “New York City is the largest city of US”?

On September 11, 2001, Al Qaeda terrorists struck the World Trade Center in New York City and the Pentagon near Washington, D.C., killing nearly 3,000 people.[165] In response,

<table>
<thead>
<tr>
<th>The United States and its territories</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Capital</strong></td>
</tr>
<tr>
<td>Washington, D.C. 38°53'N 77°01'W</td>
</tr>
<tr>
<td><strong>Largest city</strong></td>
</tr>
<tr>
<td>New York City 40°43'N 74°00'W</td>
</tr>
<tr>
<td><strong>Official languages</strong></td>
</tr>
<tr>
<td>None at federal level[fn 1]</td>
</tr>
<tr>
<td><strong>National language</strong></td>
</tr>
<tr>
<td>English[fn 2]</td>
</tr>
</tbody>
</table>
Role Linkage in Wikipedia Page

♦ Approach 1: Approach 0 + synonym of attribute name in vicinity.
  – Does it work for “New York City is the largest city of US”?

largest immigrant population.[26] The capital is Washington, D.C., and the largest city is New York City; nine other major metropolitan areas—each with at least 4.5 million inhabitants and the largest having more than 13 million people—are Los Angeles, Chicago, Dallas, Houston, Philadelphia, Miami, Atlanta, Boston, and San Francisco.

<table>
<thead>
<tr>
<th>Capital</th>
<th>Washington, D.C.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>38°53′N 77°01′W</td>
</tr>
<tr>
<td>Largest city</td>
<td>New York City</td>
</tr>
<tr>
<td></td>
<td>40°43′N 74°00′W</td>
</tr>
<tr>
<td>Official languages</td>
<td>None at federal level[fn 1]</td>
</tr>
<tr>
<td>National language</td>
<td>English[fn 2]</td>
</tr>
</tbody>
</table>
Role Linkage in Wikipedia Page

♦ Approach 1: Approach 0 + synonym of attribute name in vicinity.
  – Does it work for “Chicago was incorporated on August 12, 1833”?

On **August 12, 1833**, the Town of Chicago was organized with a population of about 200.[28] Within seven years it grew to more than 4,000 people. On June 15, 1835, the first public land sales began with Edmund Dick Taylor as U.S. Receiver of Public Monies. The City of Chicago was incorporated on Saturday, March 4, 1837 at

<table>
<thead>
<tr>
<th>Country</th>
<th>United States</th>
</tr>
</thead>
<tbody>
<tr>
<td>State</td>
<td>Illinois</td>
</tr>
<tr>
<td>Counties</td>
<td>Cook and DuPage</td>
</tr>
<tr>
<td>Settled</td>
<td>1780s</td>
</tr>
<tr>
<td>Incorporated</td>
<td>August 12, 1833</td>
</tr>
<tr>
<td>(town)</td>
<td>March 4, 1837</td>
</tr>
<tr>
<td>Incorporated</td>
<td>Jean Baptiste</td>
</tr>
<tr>
<td>(city)</td>
<td>Point du Sable</td>
</tr>
<tr>
<td>Founded by</td>
<td>Miami-Illinois: shikaakwa</td>
</tr>
<tr>
<td></td>
<td>(wild onion or wild garlic)</td>
</tr>
</tbody>
</table>
Role Linkage in Wikipedia Page

♦ Approach 2: Use ladders to capture coordinated attribute value changes between IB and text across revisions of Wikipedia page.

♦ What is a ladder?
  – Rung: spans between an Infobox attribute and text.
  – Rail: spans between adjacent temporal revisions.

♦ Role linkage has three components:
  – Spatial linkage: use Approach 0 or 1 to find candidate rungs.
  – Infobox temporal linkage: same attribute name across revisions.
  – Text temporal linkage: identify corresponding text across revisions.
Building Ladders in Wikipedia: Append Values
To date, there have been four Mario Kart games released for home consoles, three portable games, and three Namco co-developed arcade games, for a total of ten. Additionally, Mario Kart 8, for the Wii U console, is currently in development. The most recently released game of the Mario Kart series is Mario Kart 7.

To date, there have been four Mario Kart games released for home consoles, three portable games, and three Namco co-developed arcade games, for a total of ten. Additionally, Mario Kart 8, for the Wii U console, is currently in development. The most recently released game of the Mario Kart series is Mario Kart GP DX.

To date, there have been four Mario Kart games released for home consoles, three portable games, and three Namco co-developed arcade games, for a total of ten. Additionally, Mario Kart 8, for the Wii U console, is currently in development. The most recently released game of the Mario Kart series is Mario Kart 8.

There have been five Mario Kart games released for home consoles, three portable games, and three Namco co-developed arcade games, for a total of eleven. The latest title in the series, Mario Kart 8 Deluxe, was released on Nintendo Switch in April 2017. The series has sold over 100 million copies worldwide to date.
Building Ladders in Wikipedia: Vandalism

Mario Kart (Jan 9, 2013)
Infobox ...
*Last release: Mario Kart 7*

... Additionally, two future games are in development; one for the Wii U system, and a third arcade title. The most recently released game of the Mario Kart series is Mario Kart 7.

Mario Kart (Jan 10, 2013)
Infobox ...
*Last release: Mario Kart 7*

... Removed!!!

Mario Kart (Jan 11, 2013)
Infobox ...
*Last release: Mario Kart 7*

... Removed!!!

Mario Kart (Feb 12, 2013)
Infobox ...
*Last release: Mario Kart 7*

... Removed!!!
Aegina (July 16, 2010)

Infobox ...

Administrative region: Attica

The island, along with offshore islets, comprises the Municipality of Aegina in Piraeus Prefecture, a part of the Attica.

The municipality Aegina consists of the island Aegina and a few offshore islets. It is part of the Islands peripheral unit, Attica Periphery.

The municipality Aegina consists of the island Aegina and a few offshore islets. It is part of the Islands regional unit, Attica region.
Giuseppe Pilon was fired, and replaced by Luigi Del Neri.

After the expected departure of several top-quality players, as well as the manager (Luigi Del Neri parted ways with the club, and Giuseppe Iachini replacing Del Neri) and...
Discovering the Text Rail of a Ladder

♦ Goal: locate corresponding text across Wikipedia page revisions.

♦ Challenges: text edited by humans, correspondence is tenuous.
  – Changes to entity in a role in text and IB may not be synchronized.
  – Need to handle vandalism: cannot compare only adjacent revisions.

♦ Opportunities: text edited by humans, for human consumption.
  – Local context changes slowly when an entity in a role persists.
  – Local context doesn’t change much even if entity in a role changes.

♦ Solution: start with a rung, gradually grow rail across revisions.
  – Use decision tree to learn the text rail, based on entity features.
Experimental Setup

- Use real world Wikipedia datasets, crowdsourced ground truth.
  - Extracted 19,511 distinct roles from selected Wikipedia pages.
  - Partitioned based on n_rev, ambiguity, n_trans.

<table>
<thead>
<tr>
<th>dataset</th>
<th>n_rev</th>
<th>ambiguity</th>
<th>n_trans</th>
<th># of pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>dataset1</td>
<td>low</td>
<td>low</td>
<td>low</td>
<td>9,241</td>
</tr>
<tr>
<td>dataset2</td>
<td>high</td>
<td>low</td>
<td>low</td>
<td>2,708</td>
</tr>
<tr>
<td>dataset3</td>
<td>high</td>
<td>low</td>
<td>high</td>
<td>597</td>
</tr>
<tr>
<td>dataset4</td>
<td>high</td>
<td>high</td>
<td>high</td>
<td>218</td>
</tr>
</tbody>
</table>

- Competitive methods:
  - iPopulator: automatically populates Infoboxes using text.
  - Wordnet, link/text clustering.
  - Link/text ladder, link/text ladder + wordnet
Experimental Results: Effectiveness

- Ladder methods: high precision, inapplicable if no value changes.
  - Ladder + wordnet: successfully combines complementary ideas.

- Link-based methods are more effective than text-based methods.
Summary, Future Work

♦ “Tell me more” for role queries is a compelling application.

♦ Role linkage is a novel problem, distinct from entity linkage.
  – Proposed efficient methods to build ladders to solve this problem.
  – Use of ladder-based methods is effective in Wikipedia data.

♦ A lot of interesting work remains to be done.
  – Integrating related Wikipedia pages for better role linkage.