Zero-shot Entity Extraction from Web Pages

ACL
June 23, 2014
Panupong Pasupat and Percy Liang
What are the longest hiking trails near Baltimore?

hiking trails near Baltimore
   Avalon Super Loop
   Patapsco Valley State Park
   Gunpowder Falls State Park
   Union Mills Hike
   Greenbury Point
   ...

Data Source
Focus: Entity Extraction

What are the longest hiking trails near Baltimore?

Data Source

hiking trails near Baltimore
Avalon Super Loop
Patapsco Valley State Park
Gunpowder Falls State Park
Union Mills Hike
Greenbury Point
...

Applications: question answering / semantic parsing / taxonomy construction / ontology expansion / knowledge base population / ...
Semi-Structured Data on the Web
Challenge: Long Tail of Categories

*person*    *location*    *organization*
Challenge: Long Tail of Categories

person  location  organization

airport  battleship  acid  pitcher
settlement  headgear  metaphor  haircut
poker hand  biome  enzyme  superstition
Challenge: Long Tail of Categories

We want to generalize to unseen categories.
Relevant Approaches

Bootstrapping from Seed Examples:

Use seed examples to specify the entity category

[Wang and Cohen, 2009; Google Sets; Sarmento et al. 2007; …]
Relevant Approaches

Bootstrapping from Seed Examples:

Use seed examples to specify the entity category

... but we might not have seeds (e.g. in question answering)
Our Work

Use a natural language query to specify the entity category

query

hiking trails
near Baltimore

web page

System

answers

Avalon Super Loop
Hilton Area
Wildlands Loop
...

Use a natural language query to specify the entity category
Outline

1. Setup
   - Problem Setup
   - Dataset

2. Approach

3. Results
Problem Setup

Input:

- query $x$
  
  hiking trails near Baltimore

- web page $w$
Problem Setup

Input:

• query x
hiking trails near Baltimore

• web page
EveryTrail

Hiking near Baltimore, Maryland

This list shows the most popular hiking near Baltimore, Maryland based on user reviews, votes, and mobile downloads. Plan your next trip with EveryTrail guides by downloading a guide to your mobile phone with the EveryTrail iPhone or Android app.

Sort: Rating

Filter Trails

Guides

Avalon Super Loop - Patapsco State Park
Patapsco State Park, Maryland, United States (7.5 miles away)

Difficult: 12.7 miles, Full day
lots of ruins, waterfalls, trains, and river views

Do the entire Avalon Patapsco state park in 1 day! This loop covers the majority of the Avalon area, with multiple ruins, waterfalls and other artifacts to find along the way. Starting at the parking lot, you hike up the road a ways to the Ridge trail sign. The next leg is the maintenance loop which has an old old tractor to look at and some...

Patapsco Valley State Park - Hilton Area 8 Miles/Moderate
Catonsville, Maryland, United States (7.7 miles away)

Moderate: 7.8 miles, Half day
8 mile circuit hike including sections in the Avalon, Orange Grove and Glen Artyne areas of PVSP.

OVERVIEW: One of the more scenic routes in the Patapsco Valley State Park in the Hilton Area which includes multiple stream crossings, viewpoints and waterfalls including Cascade waterfalls, two swinging bridge crossings, Ilchester Overlook, and Bloedes Dam. This is a moderate hike and can be hiked in either direction. Counterclockwise is an easier hike...
Problem Setup

Input

- query: hiking trails near Baltimore
- web page: EveryTrail webpage for hiking near Baltimore, Maryland

Guides:
- **Avalon Super Loop** - Patapsco State Park
  - Difficult: 12.7 miles, Full day
  - lots of ruins, waterfalls, trains, and river views

- **Patapsco Valley State Park** - Hilton Area 8 Miles/Moderate
  - Moderate: 7.8 miles, Half day
  - 8 mile circuit hike including sections in the Avalon, Orange Grove and Glen Arm areas of PVSP

Overview: One of the more scenic routes in the Patapsco Valley State Park in the Hilton Area which includes multiple stream crossings, views and waterfalls including Cascade waterfalls, two swinging bridge crossings, Ilchester Overlook, and Bloedes Dam. This is a moderate hike and can be hiked in either direction. Counterclockwise is an easier hike...
Problem Setup

Input:
• query $x$
  
  hiking trails near Baltimore

• web page $w$

Output:
• list of entities $y$
  
  [Avalon Super Loop, Patapsco Valley State Park, ...]
Dataset

We created the OpenWeb dataset with diverse queries and web pages.

- airlines of italy
- natural causes of global warming
- lsu football coaches
- bf3 submachine guns
- badminton tournaments
- foods high in dha
- technical colleges in south carolina
- songs on glee season 5
- singers who use auto tune
- san francisco radio stations
We created the **OpenWeb** dataset with diverse queries and web pages.

- **airlines of Italy**
- **natural causes of global warming**
- **LSU football coaches**
Query Generation

Breadth-first search on Google Suggest

list of _____

Google Suggest

list of Indian movies ...

[Berant et al., 2013]
Query Generation

Breadth-first search on Google Suggest

list of ____

Google Suggest

list of Indian movies
...

list of ____ movies
list of movies ____
list of Indian ____
...

Template Extraction

[Berant et al., 2013]
Query Generation

Breadth-first search on Google Suggest

- list of _____
- Google Suggest
- list of Indian movies
- list of Indian movies _____
- list of Indian _____
- ...
Dataset Annotation

Annotate the first, second, and last entities matching the query using Amazon Mechanical Turk.
Annotate the first, second, and last entities matching the query using Amazon Mechanical Turk.

Dataset Annotation

Annotate the first, second, and last entities matching the query using Amazon Mechanical Turk.

airlines of italy

Annotation

First: Air Dolomiti
Second: Air Europe
Last: Wind Jet
Dataset Statistics

2773 examples

2269 unique queries

894 unique headwords ← long tail!

1483 unique web domains ← long tail!

(≠ wrapper induction)
Outline

1. Setup

2. Approach
   • Extraction Predicate
   • Framework
   • Modeling
   • Features

3. Results
Extraction Predicate

How can we choose what to extract from a web page \( w \)?

number of possible entity lists \( \approx 2^{\text{number of nodes}} \)
Idea: Entities usually share the same tag and tree level

$z = /html[1]/body[1]/table[2]/tr/td[1]$
Extraction Predicate

Idea: Entities usually share the same tag and tree level

$z = \texttt{/html[1]/body[1]/table[2]/tr/td[1]}$

Captures structures such as table columns, list entries, headers of the same level, ...

Each web page has $\approx 8500$ extraction predicates $z$
Framework

hiking trails near Baltimore

\[
x \quad w
\]

html

head

body

... ...

15
Framework

hiking trails near Baltimore

(|Z| ≈ 8500)
hiking trails near Baltimore

(|Z| \approx 8500)

Framework

x

Generation

Z

Model

w

/html[1]/body[1]/table[2]/tr/td[1]

/z

/html

head

body

... ...
hiking trails near Baltimore

(|Z| \approx 8500)

[Avalon Super Loop, Patapsco Valley State Park, ...]
A graphical model with latent extraction predicate $z$
Modeling

Let \( x \) be a query and \( w \) be a web page.

Define a log-linear distribution over the extraction predicates \( z \in \mathcal{Z} \):

\[
p_{\theta}(z | x, w) \propto \exp\{\theta^\top \phi(x, w, z)\}
\]

- \( \theta \) is a parameter vector
- \( \phi(x, w, z) \) is a feature vector
Modeling

Let $x$ be a query and $w$ be a web page.

Define a log-linear distribution over the extraction predicates $z \in \mathcal{Z}$:

$$p_\theta(z \mid x, w) \propto \exp\{\theta^\top \phi(x, w, z)\}$$

- $\theta$ is a parameter vector
- $\phi(x, w, z)$ is a feature vector
- Find $\theta$ that maximizes the log-likelihood of the training data using AdaGrad [Duchi et al., 2010]
Features

\[ p_\theta(z \mid x, w) \propto \exp\{\theta^\top \phi(x, w, z)\} \]
Features

\[ p_\theta(z \mid x, w) \propto \exp\{\theta^\top \phi(x, w, z)\} \]

Structural Features: context

<table>
<thead>
<tr>
<th>№</th>
<th>President</th>
<th>Took office</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>George Washington</td>
<td>April 30, 1789</td>
</tr>
<tr>
<td></td>
<td>(1732-1799)</td>
<td>[11][12][13]</td>
</tr>
<tr>
<td>2</td>
<td>John Adams</td>
<td>March 4, 1797</td>
</tr>
<tr>
<td></td>
<td>(1735-1826)</td>
<td>[15][16][17]</td>
</tr>
<tr>
<td>3</td>
<td>Thomas Jefferson</td>
<td>March 4, 1801</td>
</tr>
<tr>
<td></td>
<td>(1743-1809)</td>
<td>[18][19][20]</td>
</tr>
</tbody>
</table>

The listing below is complete for the current government of the USA. For this country, however, there were prior governments (including that under the Articles of Confederation). Prior to George Washington as first president under the current constitution, there were twelve people in leadership over the government of the United States of America who held the title of "President". Also during the Civil War, there was the position of "President of the Confederate States of America", held by one person.
Features

\[ p_\theta(z \mid x, w) \propto \exp\{\theta^\top \phi(x, w, z)\} \]

Denotation Features: content

hiking trails near Baltimore
Avalon Super Loop
Patapsco Valley State Park
Gunpowder Falls State Park
Rachel Carson Conservation Park
Union Mills Hike
...

hiking trails near Baltimore
Home
About Baltimore Tour
Pricing
Contact
Online Support
...

Defining Features on Lists

George Washington
John Adams
Thomas Jefferson
James Madison
... (39 more) ...
Barack Obama

John Adams
John Adams
John Adams
John Adams
John Adams
... (100 more) ...
John Adams

Blog
Photos and Video
Briefing Room
In the White House
Mobile Apps
Contact Us

good
bad
bad
Defining Features on Lists

- George Washington
- John Adams
- Thomas Jefferson
- James Madison
- ... (39 more) ...
- Barack Obama

- John Adams
- John Adams
- John Adams
- John Adams
- ... (100 more) ...
- John Adams

- Blog
- Photos and Video
- Briefing Room
- In the White House
- Mobile Apps
- Contact Us

- good
- bad
- bad

- identity
- diverse
- identical
- diverse
Defining Features on Lists

- **NNP NNP**
- **NNP NNP**
- **NNP NNP**
- **NNP NNP**
  
  ... (39 more) ...

- **NNP NNP**
- **NNP NNP**
- **NNP NNP**
- **NNP NNPS**

  ... (100 more) ...

- **NN PRP**

  **good**

  **identity**

  **POS**

  **identical**

  **identical**

- **bad**

  **diverse**

- **bad**

  **diverse**
Defining Features on Lists

Avalon Super Loop
Patapsco Valley State Park
Gunpowder Falls State Park
Union Mills Hike
Greenbury Point
1. Abstraction

Map list elements into abstract tokens
Defining Features on Lists

1. Abstraction
Map list elements into abstract tokens

2. Aggregation
Define features using the histogram of the abstract tokens
1. Abstraction

Map list elements into abstract tokens

2. Aggregation

Define features using the histogram of the abstract tokens

Use this method for both structural and denotation features
Outline

1. Setup

2. Approach

3. Results
   - Main Results
   - Error Analysis
   - Feature Analysis
Main Results

Baseline (Most frequent extraction predicates)

Accuracy Accuracy @ 50
10 20
30 40
50 60
Accuracy

10.3
40.5
55.8

(Most frequent extraction predicates)
Error Analysis

Correct: 40.5%

Coverage Errors: 33.4%

Ranking Errors: 26.1%
Examples of Correct Predictions

Query: *disney channel movies*

Examples of Correct Predictions

Query: universities in canada

Our universities

In order to help you choose a university, take a look at the profiles of our member institutions. Here you will find an overview of the university, the number of students enrolled, tuition fees, links to the university websites and more.

Start exploring.

British Columbia
- Emily Carr University of Art + Design
- Kwantlen Polytechnic University
- Royal Roads University
- Simon Fraser University

Alberta
- Athabasca University
- Concordia University College of Alberta
- MacEwan University
- Mount Royal University

/html[1]/body/div/div/div/div/div/div/div/div/div/a/text
Examples of Correct Predictions

Query: *nobel prize winners*

The Nobel Prize in Physics 2013

François Englert, Peter W. Higgs

"for the theoretical discovery of a mechanism that contributes to our understanding of the origin of mass of subatomic particles, and which recently was confirmed through the discovery of the predicted fundamental particle, by the ATLAS and CMS experiments at CERN's Large Hadron Collider"

The Nobel Prize in Physics 2012

Serge Haroche, David J. Wineland

"for ground-breaking experimental methods that enable measuring and manipulation of individual quantum systems"

The Nobel Prize in Physics 2011

Saul Perlmutter, Brian P. Schmidt, Adam G. Riess

"for the discovery of the accelerating expansion of the universe through observations of distant supernovae"
Error Analysis

- Correct: 40.5%
- Coverage: 33.4%
- Ranking: 26.1%
Error Analysis

Correct
40.5%

Coverage Errors
33.4%

Ranking Errors
26.1%

Coverage Errors

No extraction predicate \( z \) produces an entity list \( y \) matching the annotation
Examples of Coverage Errors

Query: *companies named after a person*

- Yves Saint Laurent
- Yves Saint Laurent
- Yuke's
- Yukinori Taniguchi
- Zagato
- Ugo Zagato
- Zakspeed
- Erich ZAKowski
- Zend Technologies
- ZEev Suraski & ANDi Gutmans
- Zust
- Roberto Zust

See also

- List of company name etymologies

/html/body/div[3]/div[3]/div[4]/ul/li/a

Need richer extraction predicates!
Examples of Coverage Errors

Query: *hedge funds in new york*

<table>
<thead>
<tr>
<th>Rank</th>
<th>Firm</th>
<th>Headquarters</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bridgewater Associates</td>
<td>Westport, CT</td>
</tr>
<tr>
<td>2</td>
<td>Man Group</td>
<td>London</td>
</tr>
<tr>
<td>3</td>
<td>J.P. Morgan Asset Management</td>
<td>New York</td>
</tr>
<tr>
<td>4</td>
<td>Brevan Howard Asset Management</td>
<td>London</td>
</tr>
<tr>
<td>5</td>
<td>Och-Ziff Capital Management Group</td>
<td>New York</td>
</tr>
<tr>
<td>6</td>
<td>Paulson &amp; Co.</td>
<td>New York</td>
</tr>
<tr>
<td>7</td>
<td>BlackRock Advisors</td>
<td>New York</td>
</tr>
</tbody>
</table>


Need compositionality!
Error Analysis

Correct
40.5%

Coverage Errors
33.4%

Ranking Errors
26.1%

Coverage Errors
No extraction predicate $z$ produces an entity list $y$ matching the annotation
Error Analysis

<table>
<thead>
<tr>
<th>Correct</th>
<th>40.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coverage Errors</td>
<td>33.4%</td>
</tr>
<tr>
<td>Ranking Errors</td>
<td>26.1%</td>
</tr>
</tbody>
</table>

Coverage Errors

No extraction predicate $z$ produces an entity list $y$ matching the annotation.

Ranking Errors

The system finds a list $y$ matching the annotation, but it does not have the highest model score.
### Examples of Ranking Errors

**Query:** *doctors at emory*

<table>
<thead>
<tr>
<th>Name</th>
<th>Specialization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aaron, Maria MD</td>
<td>Ophthalmology</td>
</tr>
<tr>
<td>Abboushi, Nour MD</td>
<td>Plastic Surgery</td>
</tr>
<tr>
<td>Abdou, Mahmoud MD</td>
<td>Cardiovascular Disease</td>
</tr>
<tr>
<td>Abramowsky, Carlos MD</td>
<td>Pathology</td>
</tr>
<tr>
<td>Abruzzo, Todd MD</td>
<td>Radiology</td>
</tr>
</tbody>
</table>

Augmenting Denotation Features

**Observation:** Entities of different categories have different linguistic properties.

*mayors of Chicago*  
Rahm Emanuel  
Richard M. Daley  
Eugene Sawyer  
...

*universities in Chicago*  
Aurora University  
DePaul University  
Illinois Institute of Technology  
...

...
Augmenting Denotation Features

Observation: Entities of different categories have different linguistic properties.

mayors of Chicago universities in Chicago
Rahm Emanuel Aurora University
Richard M. Daley DePaul University
Eugene Sawyer Illinois Institute of Technology
...
...

Experiment: Augment denotation features with the query category.

\[
\text{POS majority} = \text{NNP NNP} \quad (\text{POS majority} = \text{NNP NNP}, \quad \text{query category} = \text{people})
\]
Augmenting Denotation Features

Accuracy (dev)

- Denotation: 19.8
- Augmented Denotation: 25

32
Augmenting Denotation Features

![Chart showing accuracy comparison between structural and augmented denotation features.](chart.png)

- Structural + Denotation (default) accuracy: 41.1
- Structural + Augmented Denotation accuracy: 41.7
Hypothesis: Structural features have high influence when the web page comes from Web search result.
Augmenting Denotation Features

Hypothesis: Structural features have high influence when the web page comes from Web search result.
Augmenting Denotation Features

Hypothesis: Structural features have high influence when the web page comes from Web search result.

Verify the hypothesis: Concatenate a random web page

hiking trails near Baltimore

Verify the hypothesis: Concatenate a random web page
Augmenting Denotation Features

**Hypothesis:** Structural features have high influence when the web page comes from Web search result.

**hiking trails near Baltimore**

Verify the hypothesis: Concatenate a random web page

- Creates noise: entity lists with high structural feature scores might not be the correct list
Augmenting Denotation Features

hiking trails near Baltimore

![Bar chart showing accuracy improvements with augmented denotation features](chart.png)
A framework for extracting entities from a natural language query and a single web page
Summary

Focus on the long tail of entity categories

tutorials at ACL
Summary

tutorials at ACL

Focus on the long tail of entity categories

Consider both structural and denotation features
Summary

Focus on the long tail of entity categories

Consider both structural and denotation features

Handle lists of different sizes with abstraction and aggregation
Future Work

• Model relationship between entities and category strings
• Compositionality in natural language
Download code and dataset:


Thank you!