

# Zero-shot Entity Extraction from Web Pages



ACL

June 23, 2014

Panupong Pasupat and Percy Liang

# 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

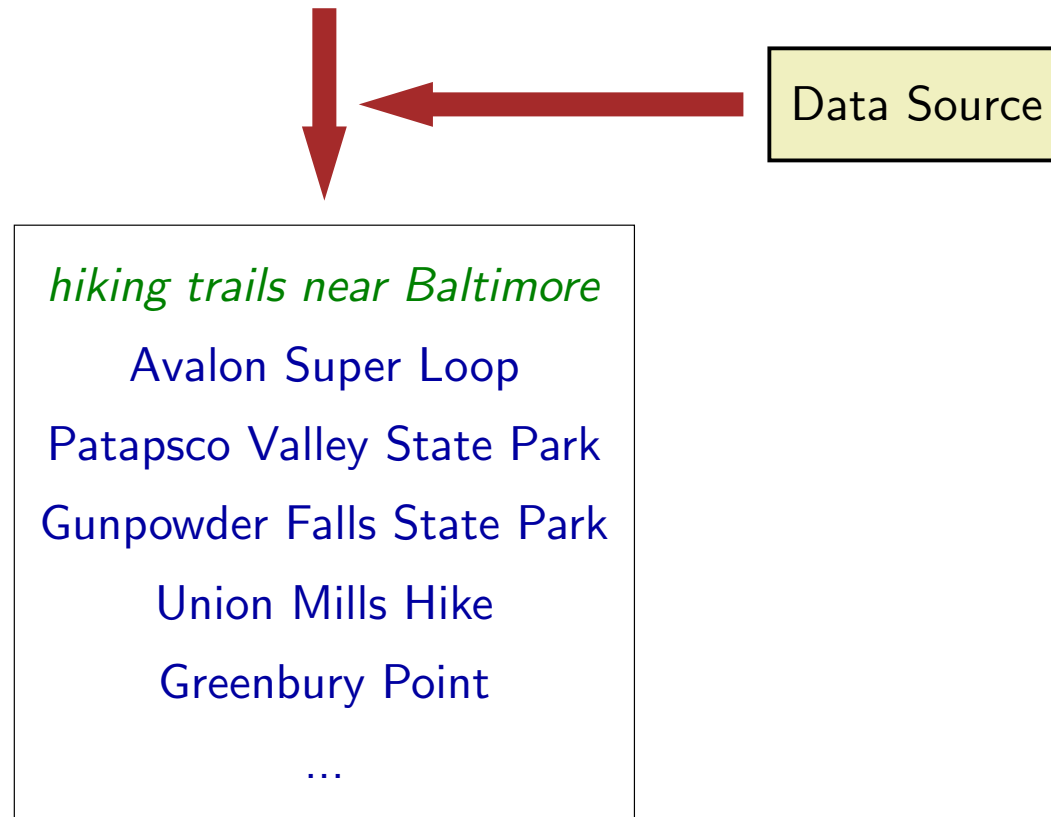
Union Mills Hike

Greenbury Point

...

# Focus: Entity Extraction

*What are the longest* **hiking trails near Baltimore?**



**Applications:** question answering / semantic parsing / taxonomy construction / ontology expansion / knowledge base population / ...

# Semi-Structured Data on the Web

## ACCEPTED LONG PAPERS

### ACL 2014

- A Bayesian Mixed Effects Model of Literary Character  
*David Bamman, Ted Underwood and Noah A. Smith*
- A dense connected measure of inter-annotator agreement for syntax  
*Arne Szegemot*
- A Decision Theoretic Approach to Natural Language Generation  
*Nathan McKinley and Soumya Ray*
- A Decomposable Graph-Based Parser for the Abstract Meaning Representation  
*Jeffrey Flanagan, Sam Thompson, Jesse Carbone, Chris Dyer and Mark*
- A Generalized Language Model  
*Rene Rickford, Thomas Goto*
- A Joint Graph Model for Phrasal  
*Zhangyi Ji and Hai Zhao*
- A Language Relexation Algorithm  
*Alexander M. Rush, Michael Collins*
- A Linear-Time Bottom-Up Discourse Parser  
*Vanessa Wei Feng and Graham*
- A practical but linguistically-motivated  
*Denis Paperno, The Ngia Piao*
- A Provably Correct Learning Algorithm  
*Shay B. Cohen and Michael Collins*

## Regular Faculty

54 people

### Name

Alex Aiken

Berfin Batozoglou

Gil Bejerano

Michael Bernstein

Dan Bonah

David Chertou

Steve Cooper

Bill Dalry

David Dill

Ron Dror

Dawson Engler

Ron Fedkiw

## Most Popular Action Feature Films

1.		<b>Godzilla (2014)</b> ★★★★★ 7.2/10 The world's most famous monster is pitted against more violent creatures who threaten our very existence. Dir: Gareth Edwards Writ: Aaron Ryder, Johnson, Elizabeth Clady, Bryan Cranston Action   Sci-Fi   Thriller
2.		<b>X-Men: Days of Future Past</b> ★★★★★ The X-Men and Wolverine travel back in time to stop the mutant genocide. Dir: Bryan Singer Writ: James Mangold Action   Adventure   Sci-Fi
3.		<b>The Amazing Spider-Man</b> ★★★★★ Peter Parker runs the gauntlet against his inner demons. Dir: Marc Webb Writ: Matt S. Johnson Action   Adventure   Sci-Fi
4.		<b>Transformers: Age of Extinction</b> An automobile mechanic becomes a part of the war between Autobots and Decepticons. Dir: Michael Bay Writ: John Gatins Action   Adventure   Sci-Fi

No.	President	Took office
1	<b>George Washington</b> (1732-1799) (1789-1797)	April 30, 1789 (3-1)
2	<b>John Adams</b> (1735-1826) (1797-1801)	March 4, 1797
3	<b>Thomas Jefferson</b> (1743-1826) (1801-1809)	March 4, 1801
4	<b>James Madison</b> (1751-1836) (1809-1817)	March 4, 1809

Day	Group	Country	Time	Country
Thursday 12 June	GROUP A	BRAZIL	17:00	CROATIA
Friday 13 June	GROUP B	MEXICO	13:00	CAMEROON
	GROUP B	SPAIN	16:00	NETHERLANDS
	GROUP B	CHILE	18:00	AUSTRALIA
Saturday 14 June	GROUP C	COLOMBIA	13:00	GREECE
	GROUP B	URUGUAY	16:00	COSTA RICA

# 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

*person location organization*

*airport battleship acid pitcher*

*settlement headgear metaphor haircut*

*poker hand biome enzyme superstition*

*tutorials at ACL 2014*

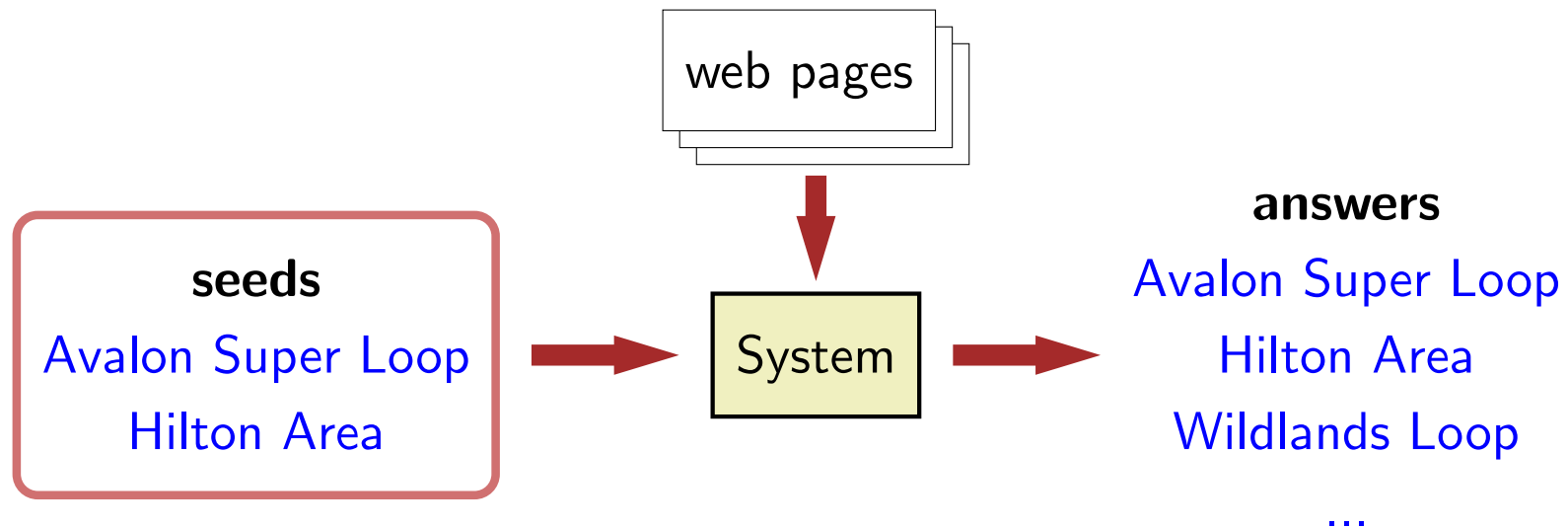
*dishes at Pu Pu Hot Pot*

*Stanford computer science professors*

We want to generalize to **unseen categories**

# Relevant Approaches

Bootstrapping from Seed Examples:

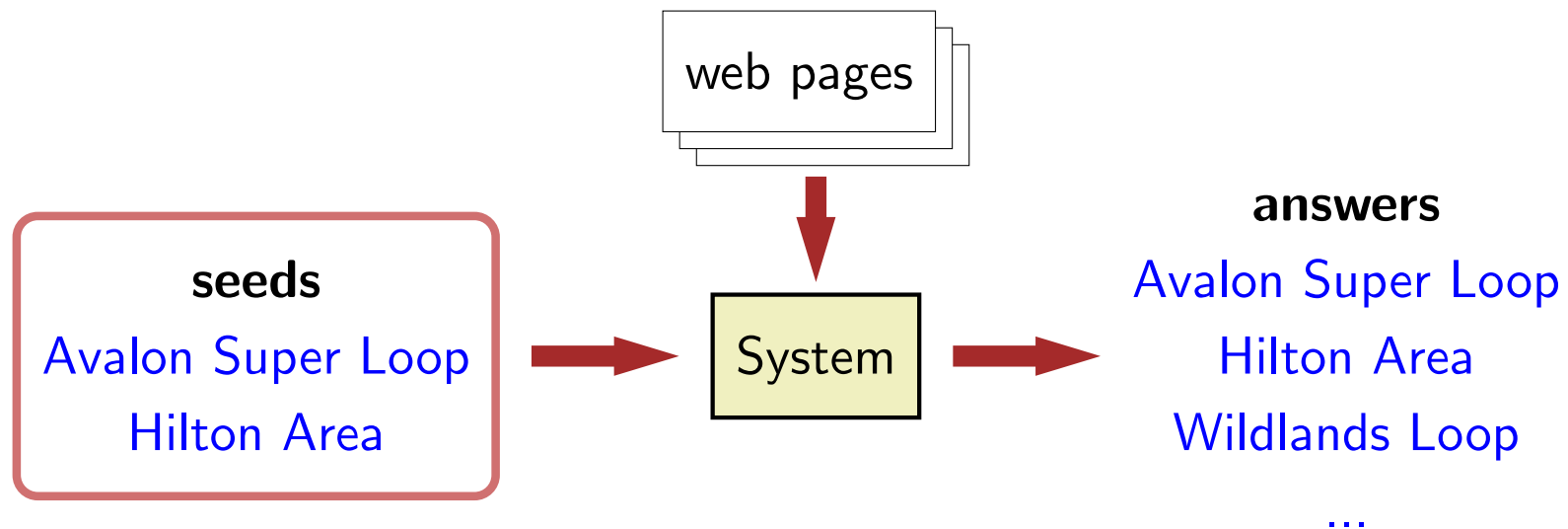


Use **seed examples** to specify the entity category



# Relevant Approaches

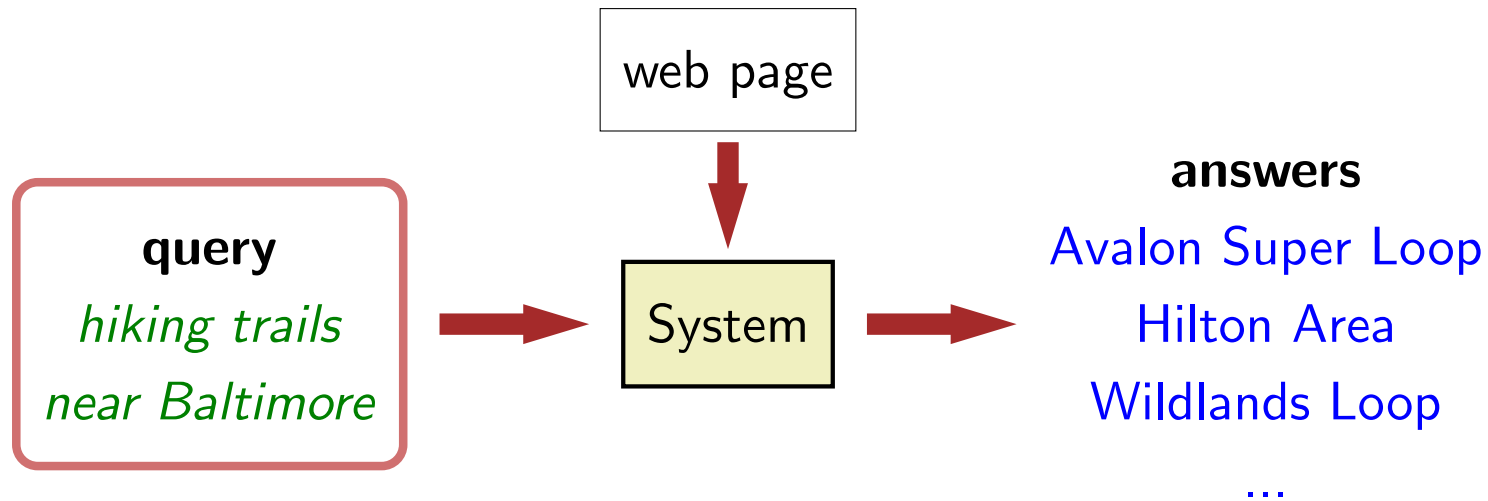
Bootstrapping from Seed Examples:



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

# 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

The screenshot displays the EveryTrail website interface. At the top, there is a navigation bar with links for HOME, EXPLORE, MOBILE APPS, CREATE TRIP, and MY EVERYTRAIL. A search bar is located on the right side of the navigation bar. Below the navigation bar, there is a header section for "Hiking near Baltimore, Maryland" with a "Like" button (49 people like this) and a "Tweet" button (1 tweet). A paragraph of text explains that the list shows the most popular hiking trails based on user reviews, votes, and mobile downloads. Below this text, there is a "Sort" dropdown menu set to "Rating" and a checked checkbox for "show community trips".

The "Filter Trails" section is visible, with a "Guides" sub-section. Two hiking trails are listed:

- 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
- 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 Artney areas of PVSP.

The "OVERVIEW" for the second trail states: "One of the more scenic routes in the Patapsco Valley State Park in the Hilton Area which includes multiple stream crossings, viewings 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..."

On the right side of the page, there is a map of Maryland and surrounding areas (Virginia, Delaware, Pennsylvania) with several red location pins. Below the map, there is a section titled "Popular places for Hiking" with links to "Hiking in Maryland", "Hiking in Patapsco Valley State Park", "Hiking in Calvert Cliffs State Park", and "Hiking in Patuxent River State Park".

# Problem Setup

Input

EveryTrail

HOME | EXPLORE | MOBILE APPS | CREATE TRIP | MY EVERYTRAIL

Search GO

(Update Current Location) Login | Signup

## Hiking near Baltimore, Maryland


Like 49 people like this. Tweet 1

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  show community trips

### 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

Patapsco Valley State Park, Maryland, United States (8 miles away)

Moderate: 7.8 miles, Half day  
8 mile circuit hike including sections in the Avalon, Orange Grove and Glen Artney 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, viewings 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...



#### Popular places for Hiking

- Hiking in Maryland
- Hiking in Patapsco Valley State Park
- Hiking in Calvert Cliffs State Park
- Hiking in Patuxent River State Park

# 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*



# Dataset

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

The screenshot shows the Ranker website interface. At the top, there's a search bar with the text "find a list or topic". Below it, there are navigation tabs for "create a list", "recent", "people", "film", "tv", "music", "sports", and "travel". The main content area displays a list titled "All Italian Airlines" with a sub-header "List of Airlines" and statistics "6,395 votes" and "19 items". The list includes the following entries:

Rank	Name	Hubs
1	Air Dolomiti	Munich Airport, Verona Villafranca Airport, Treviso Suseo Airport
2	Air Europe	Malpensa Airport
3	Air Italy	Malpensa Airport, Leonardo da Vinci-Fiumicino Airport, Verona Villafranca Airport
4	Air One	Malpensa Airport
5	Air Vallée	Parma Airport, Acosta Airport, Federico Fellini International Airport
6	Alidaunia	Foggia "Gino Lisa" Airport
7	Alitalia-Linee Aeree Italiane	Malpensa Airport, Leonardo da Vinci-Fiumicino Airport

*airlines of italy*

The screenshot shows a webpage titled "10 Greenhouse Effect". It features a diagram of the greenhouse effect where yellow arrows representing solar radiation hit the Earth's surface, and some are reflected back towards the Earth by the atmosphere. Below the diagram, the text explains: "Greenhouse effect is the process in which the atmosphere of the Earth trap some of the heat coming from the sun, making the Earth warm but due to burning fuels, cutting trees, the concentration of heat on Earth is increased to abnormal levels making greenhouse effect as one of the major causes of global warming. Carbon Dioxide, methane, nitrous oxide are the greenhouse gases which helps to keep the Earth warm. It is a natural phenomenon that takes place with the adequate concentrations of the greenhouse gases. But when the concentration of these gases rises, they disturb the climatic conditions, making the Earth more warm. These gases are not able to escape, which is the cause of world-wide increase in temperature. So the balance of carbon dioxide and other gases should be maintained so that it does not become the major reason of global warming."

Below this, there is a section titled "9. Air Pollution" with an illustration of a person carrying a large green gas canister on their back, with smoke rising from the ground.

*natural causes of global warming*

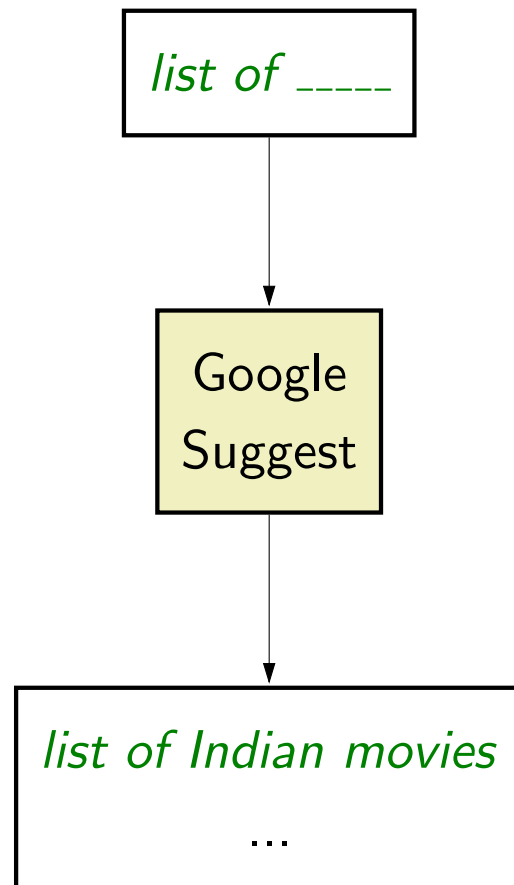
The screenshot shows a website titled "Football" with a navigation menu including "TEAMS", "SCHEDULES", "TICKETS", "FAN ZONE", "VIDEO + AUDIO", "FACILITIES", "DEPARTMENTS", "CONNECT", "VISITORS", and "PR". The main content area is titled "2013 Football Coaches" and displays a grid of 12 coaches with their names and titles:

Name	Title
Les Miles	Head Coach
Cam Cameron	Offensive Coordinator/Quarterbacks Coach
John Chavis	Defensive Coordinator
Frank Wilson	Running Backs Coach-Recruiting Coordinator
Steve Ensminger	Tight Ends Coach
Brick Haley	Defensive Line Coach
Adam Henry	Wide Receivers Coach
Thomas McGaughey	Special Teams Coordinator
Corey Raymond	Defensive Backs Coach
Greg Studrawa	Offensive Line Coach
Steve Kragthorpe	Assistant
Tommy Moffitt	Strength & Conditioning Coordinator
Dr. Sam Nader	Assistant Athletics Director - Football

*lsu football coaches*

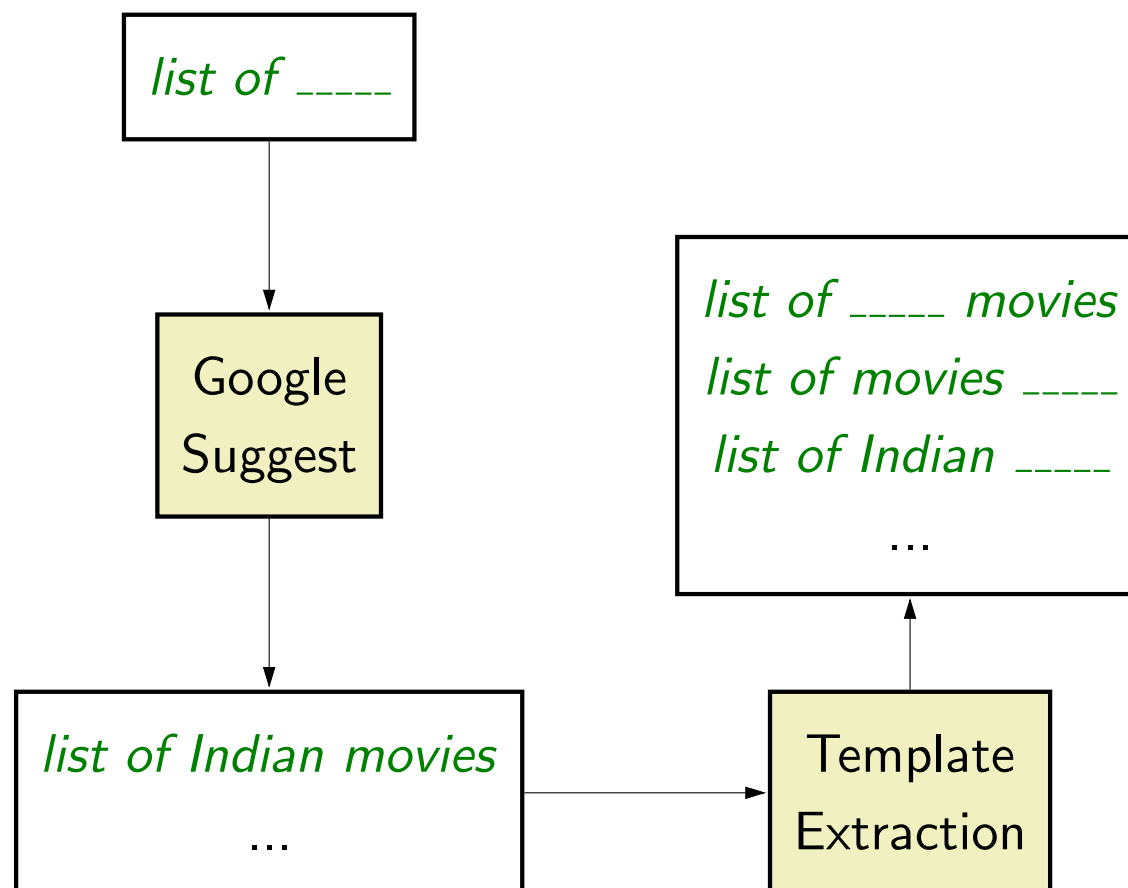
# Query Generation

Breadth-first search on Google Suggest



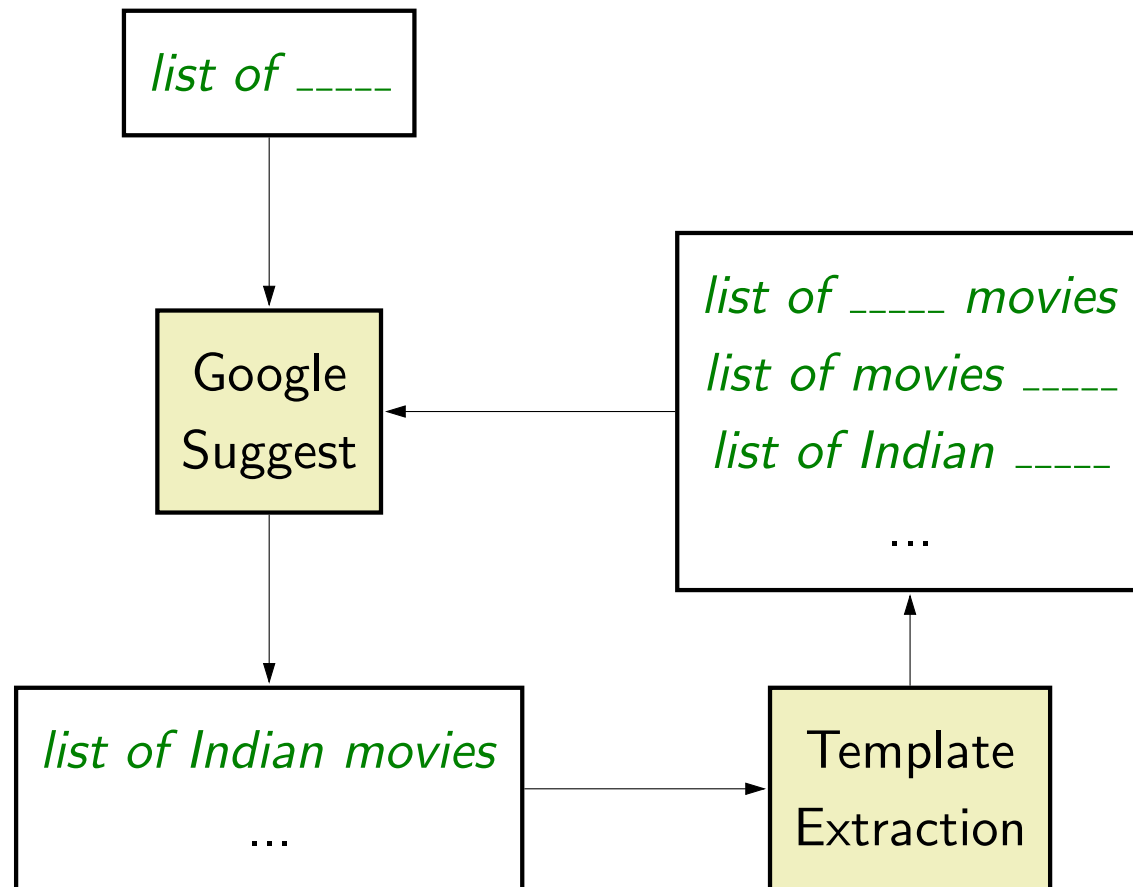
# Query Generation

Breadth-first search on Google Suggest



# Query Generation

Breadth-first search on Google Suggest

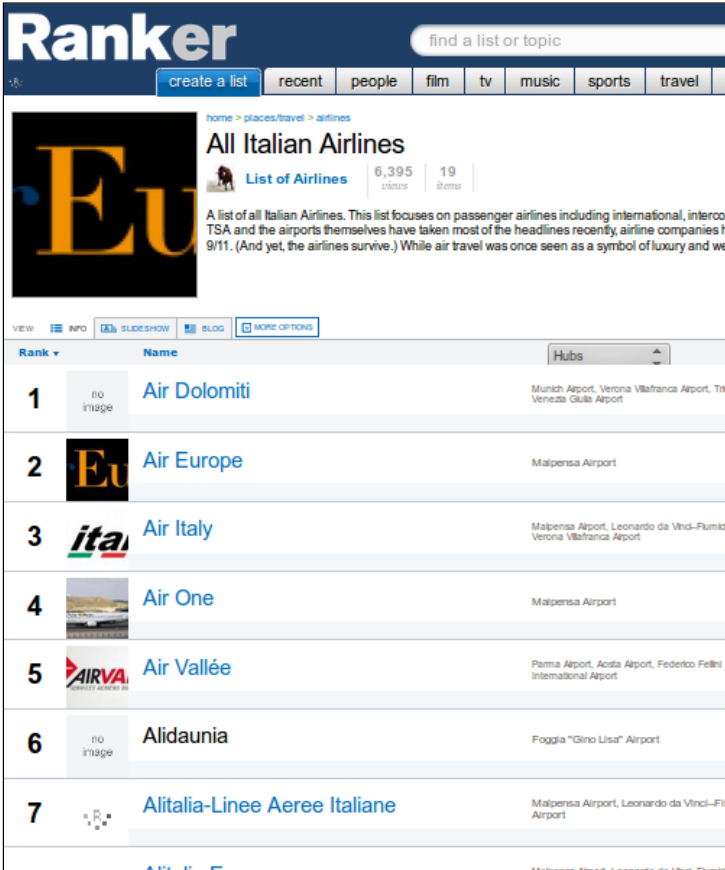


# Dataset Annotation








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.



The screenshot shows the Ranker website interface. At the top, there is a search bar with the text "find a list or topic" and a navigation menu with categories: "create a list", "recent", "people", "film", "tv", "music", "sports", and "travel". Below the navigation, there is a breadcrumb trail: "home > places/travel > airlines". The main heading is "All Italian Airlines" with a sub-heading "List of Airlines" and statistics "6,395 views" and "19 items". A large image of "Eu" is visible on the left. Below the heading, there is a description: "A list of all Italian Airlines. This list focuses on passenger airlines including international, interco... TSA and the airports themselves have taken most of the headlines recently, airline companies h... 9/11. (And yet, the airlines survive.) While air travel was once seen as a symbol of luxury and we...". Below the description, there are options for "VIEW: INFO", "SLIDESHOW", "BLOG", and "MORE OPTIONS". The main content is a table with columns "Rank", "Name", and "Hubs".

Rank	Name	Hubs
1	 Air Dolomiti	Munich Airport, Verona Villafranca Airport, Tre Venezie G. Kufner Airport
2	 Air Europe	Malpensa Airport
3	 Air Italy	Malpensa Airport, Leonardo da Vinci-Fiumicino Airport, Verona Villafranca Airport
4	 Air One	Malpensa Airport
5	 Air Vallée	Parma Airport, Acosta Airport, Federico Fellini International Airport
6	 Alidaunia	Foggia "Gino Lisa" Airport
7	 Alitalia-Linee Aeree Italiane	Malpensa Airport, Leonardo da Vinci-Fiumicino Airport

*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

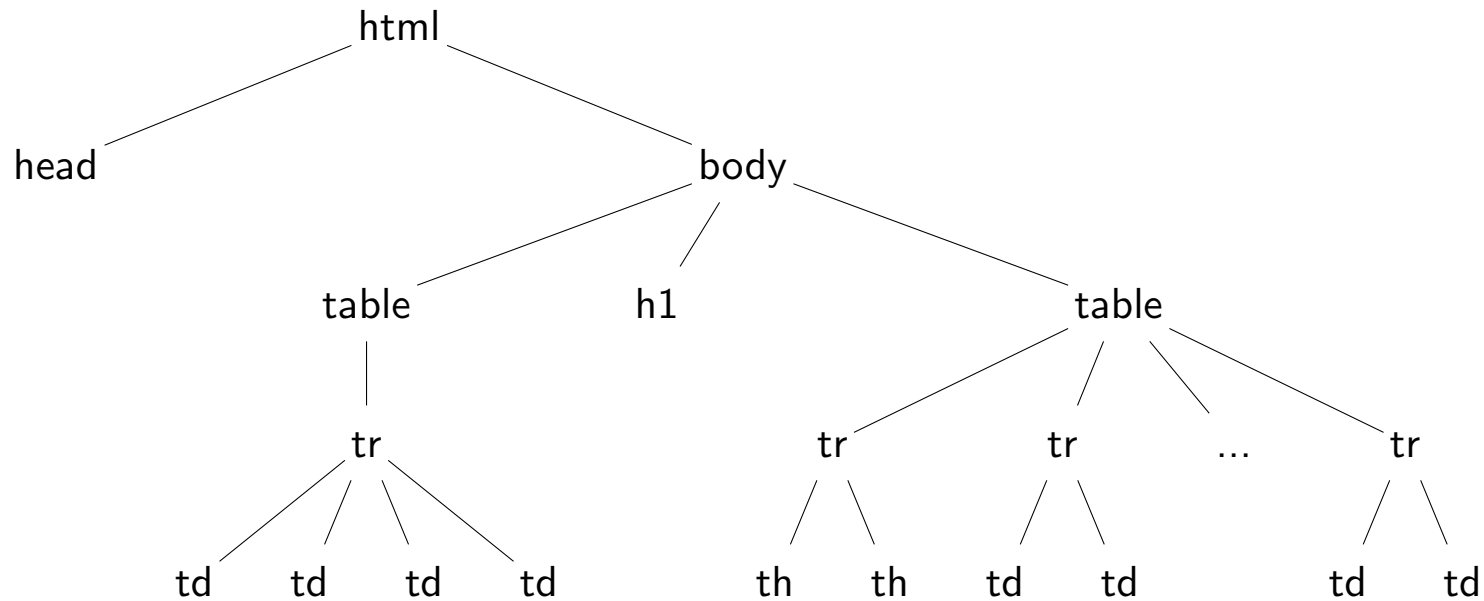
- Extraction Predicate
- Framework
- Modeling
- Features

3. Results



# Extraction Predicate

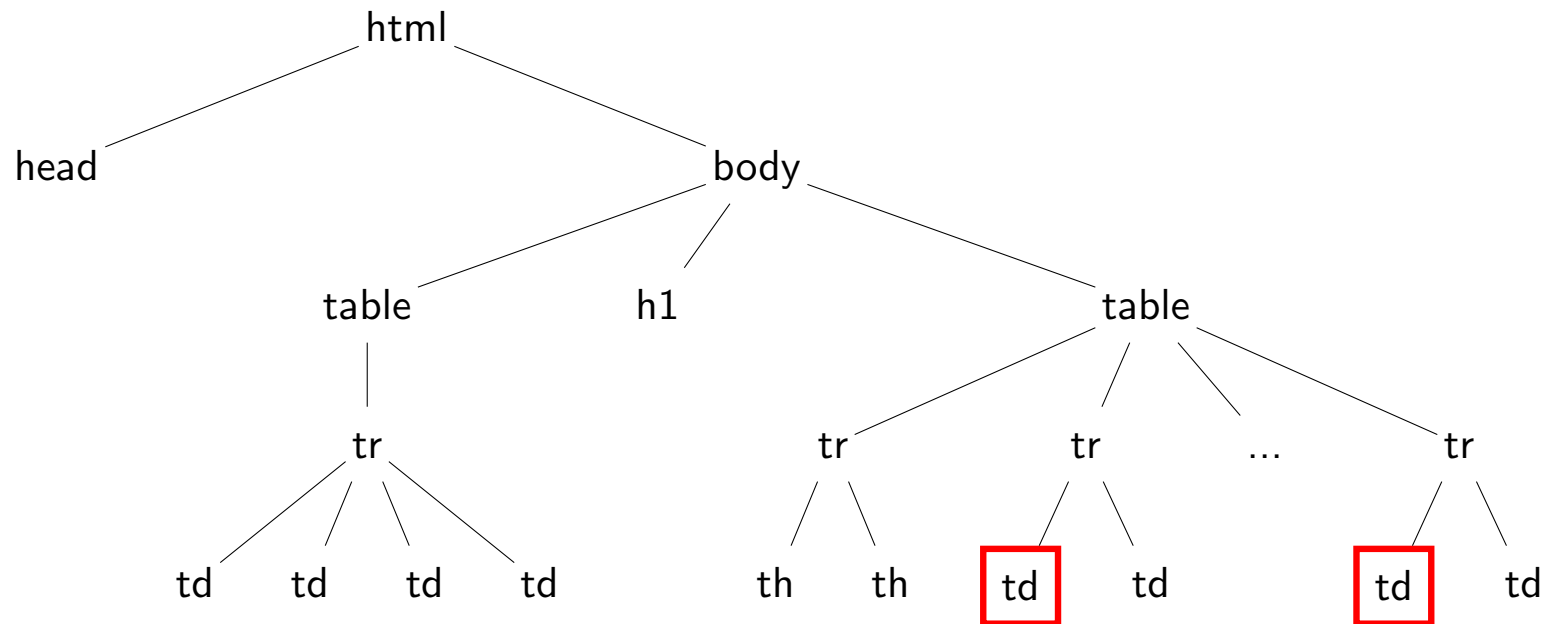
How can we choose what to extract from a web page  $w$ ?



number of possible entity lists  $\approx 2^{\text{number of nodes}}$

# Extraction Predicate

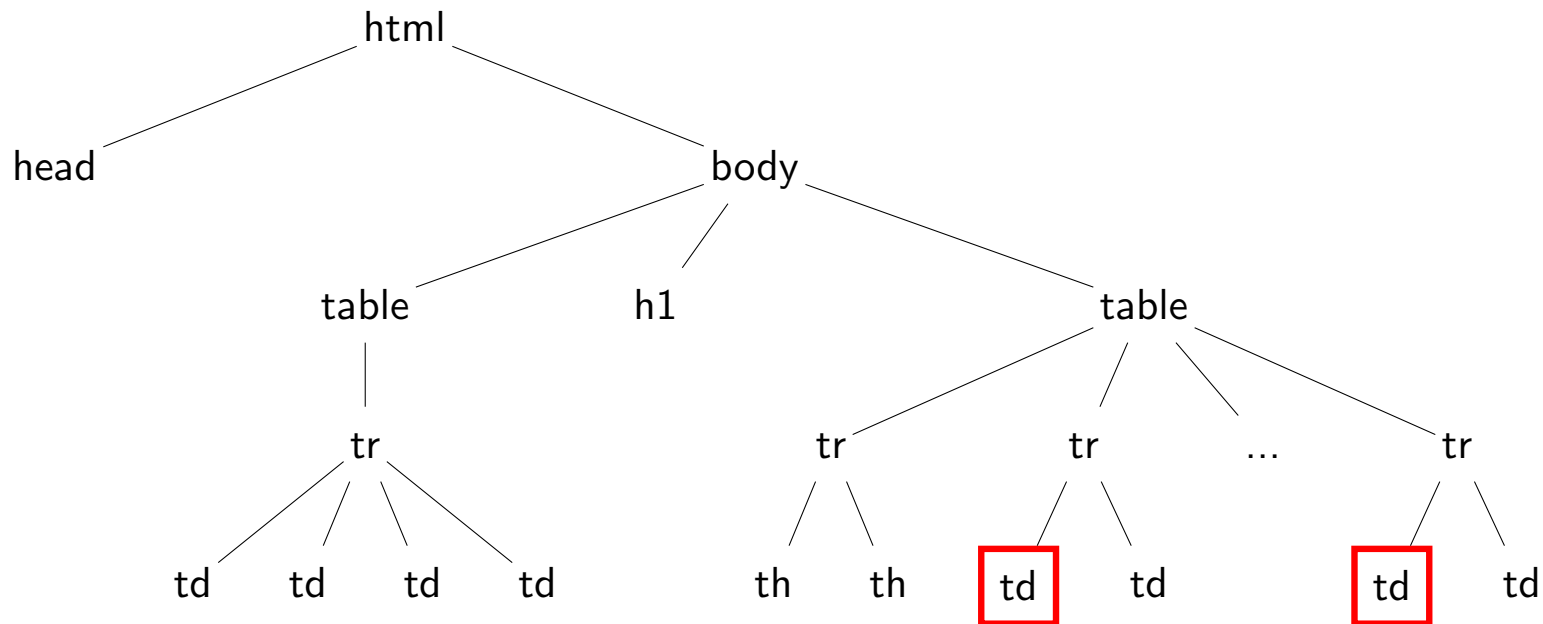
**Idea:** Entities usually share the same tag and tree level



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

# Extraction Predicate

**Idea:** Entities usually share the same tag and tree level



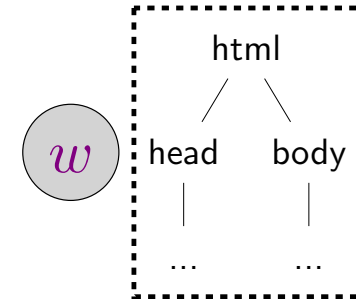
$$z = /html[1]/body[1]/table[2]/tr/td[1]$$

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

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

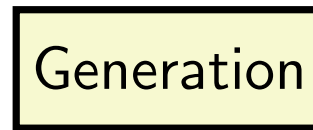
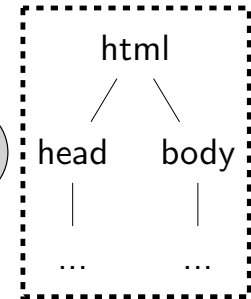
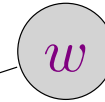
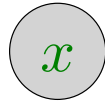
# Framework

*hiking trails  
near Baltimore*  $x$



# Framework

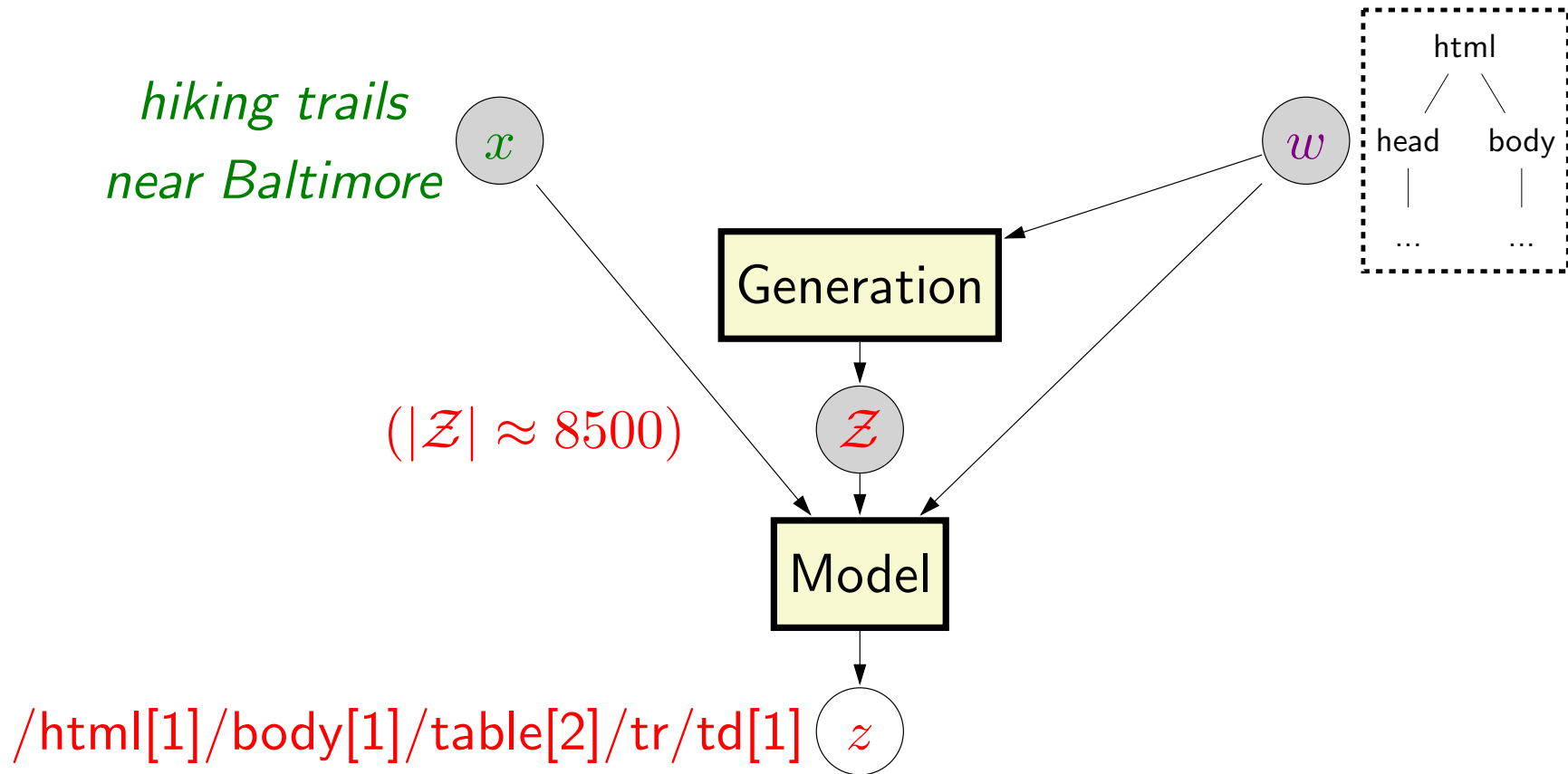
*hiking trails  
near Baltimore*



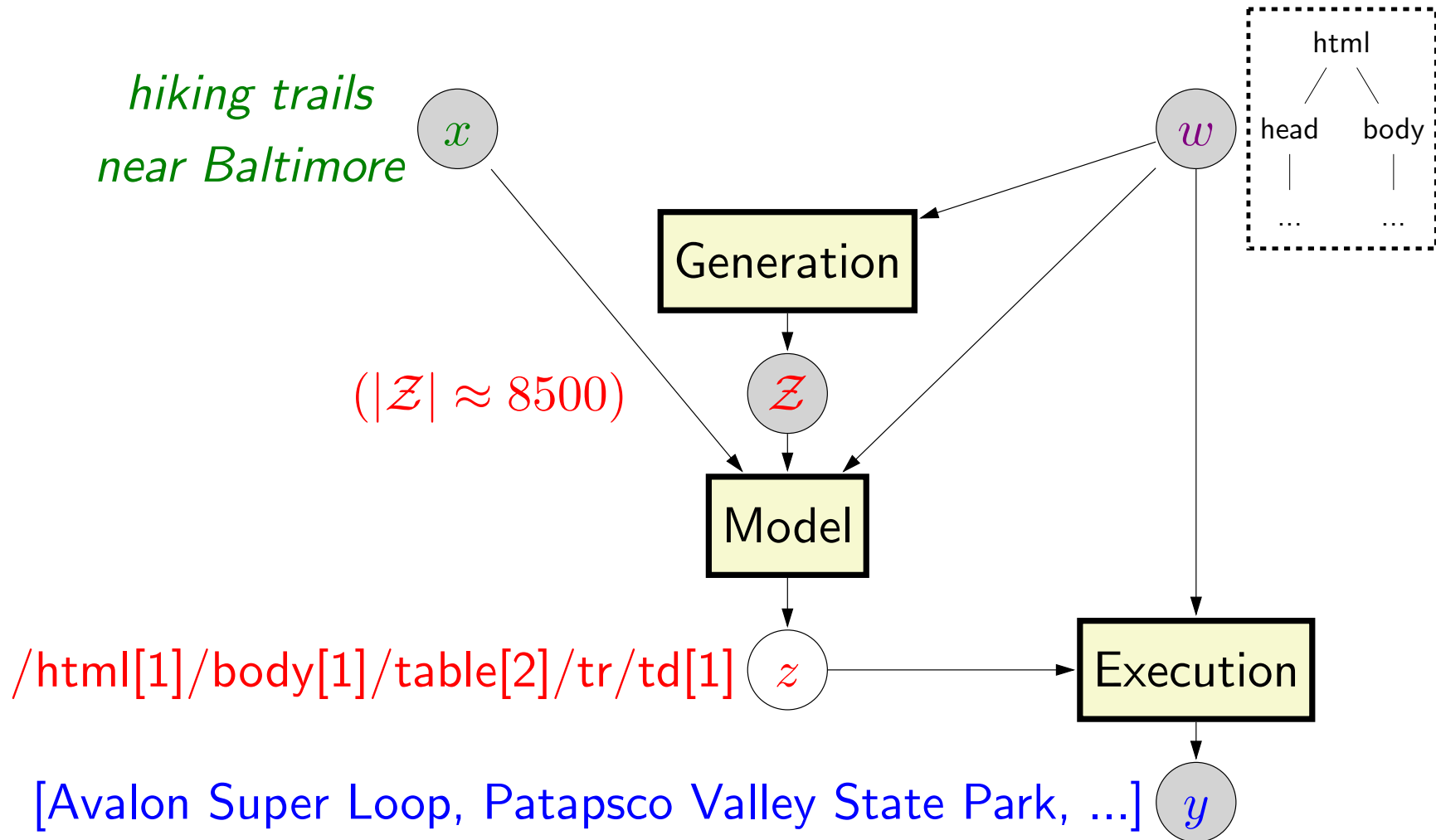
$(|\mathcal{Z}| \approx 8500)$



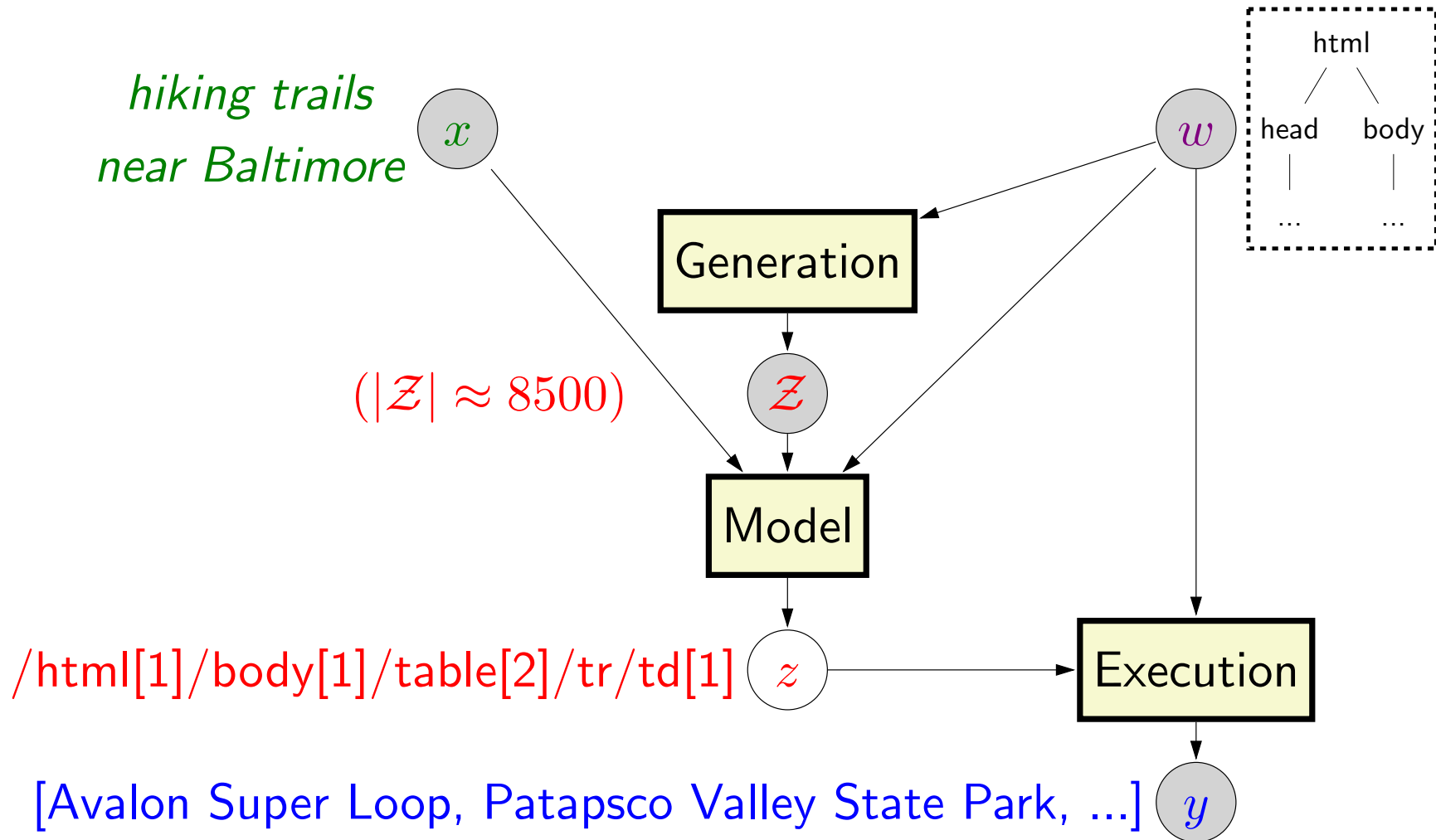
# Framework



# Framework



# Framework



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  $z \in \mathcal{Z}$ :

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

- $\theta$  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

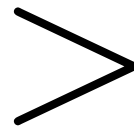
$$p_{\theta}(z \mid x, w) \propto \exp\{\theta^{\top} \phi(x, w, z)\}$$

# Features

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

## Structural Features: context

<b>N<sub>e</sub></b>	<b>President</b>	<b>Took office</b>
1	 <b>George Washington</b> <small>(1732-1799) [11][12][13]</small>	April 30, 1789 <small>[n 2]</small>
2	 <b>John Adams</b> <small>(1735-1826) [15][16][17]</small>	March 4, 1797
3	 <b>Thomas Jefferson</b> <small>(1743-1826) [18][19][20]</small>	March 4, 1801



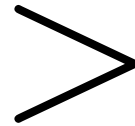
The listing below is complete for the current government of the USA. For this country, however, there were prior governments (including that under the **Articles of Confederation**). Prior to George Washington as first president under the current constitution, there were **twelve people** in leadership 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"** in an entity separate from the USA, and this position was held by **one person**.

# Features

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

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

good

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

bad

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

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  
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 NNP  
NNP NNP  
NNP NNP  
NNP NNP  
... (100 more) ...  
NNP NNP

NN  
NNS CC NNP  
NN NN  
IN DT NNP NNP  
NNP NNPS  
NN PRP

good

bad

bad

identity

diverse

identical

diverse

POS

identical

identical

diverse



# Defining Features on Lists

Avalon Super Loop

Patapsco Valley State Park

Gunpowder Falls State Park

Union Mills Hike

Greenbury Point

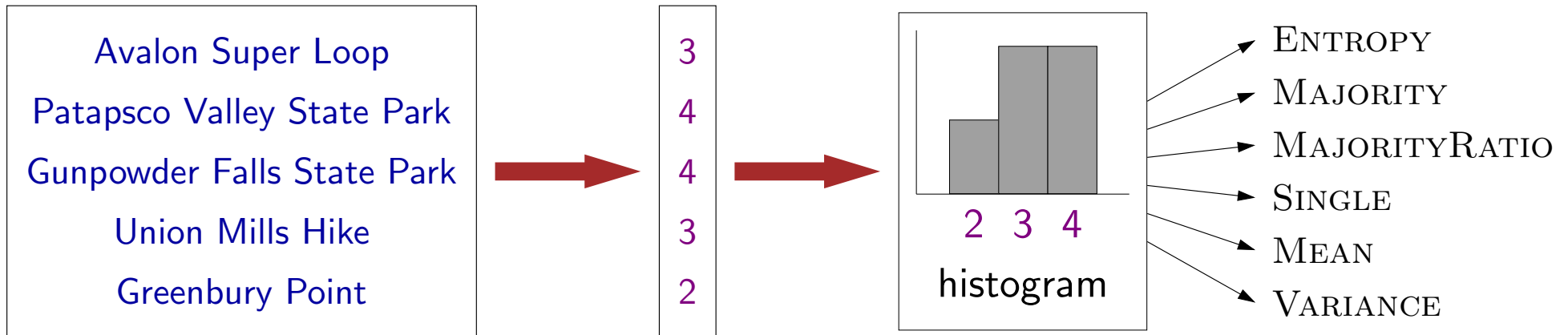
# Defining Features on Lists



## 1. Abstraction

Map list elements into abstract tokens

# Defining Features on Lists



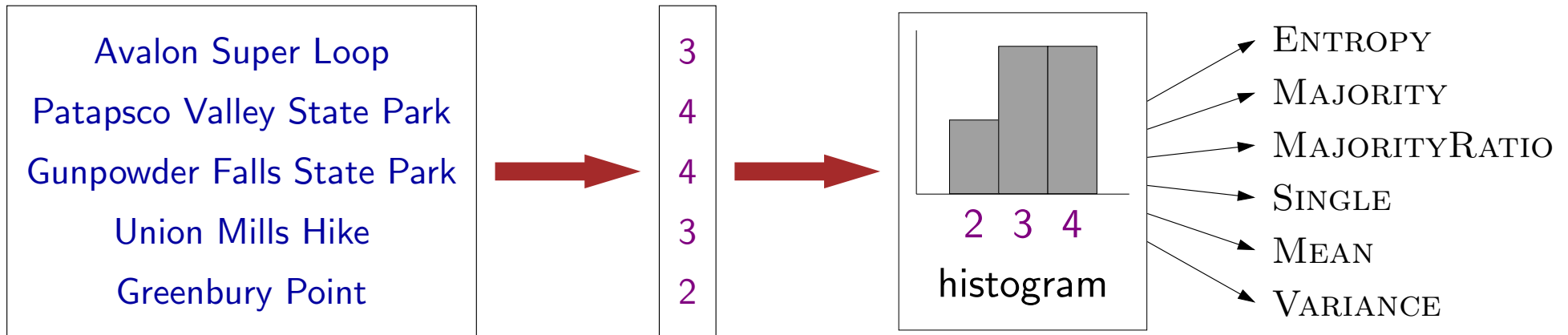
## 1. Abstraction

Map list elements into abstract tokens

## 2. Aggregation

Define features using the histogram of the abstract tokens

# Defining Features on Lists



## 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

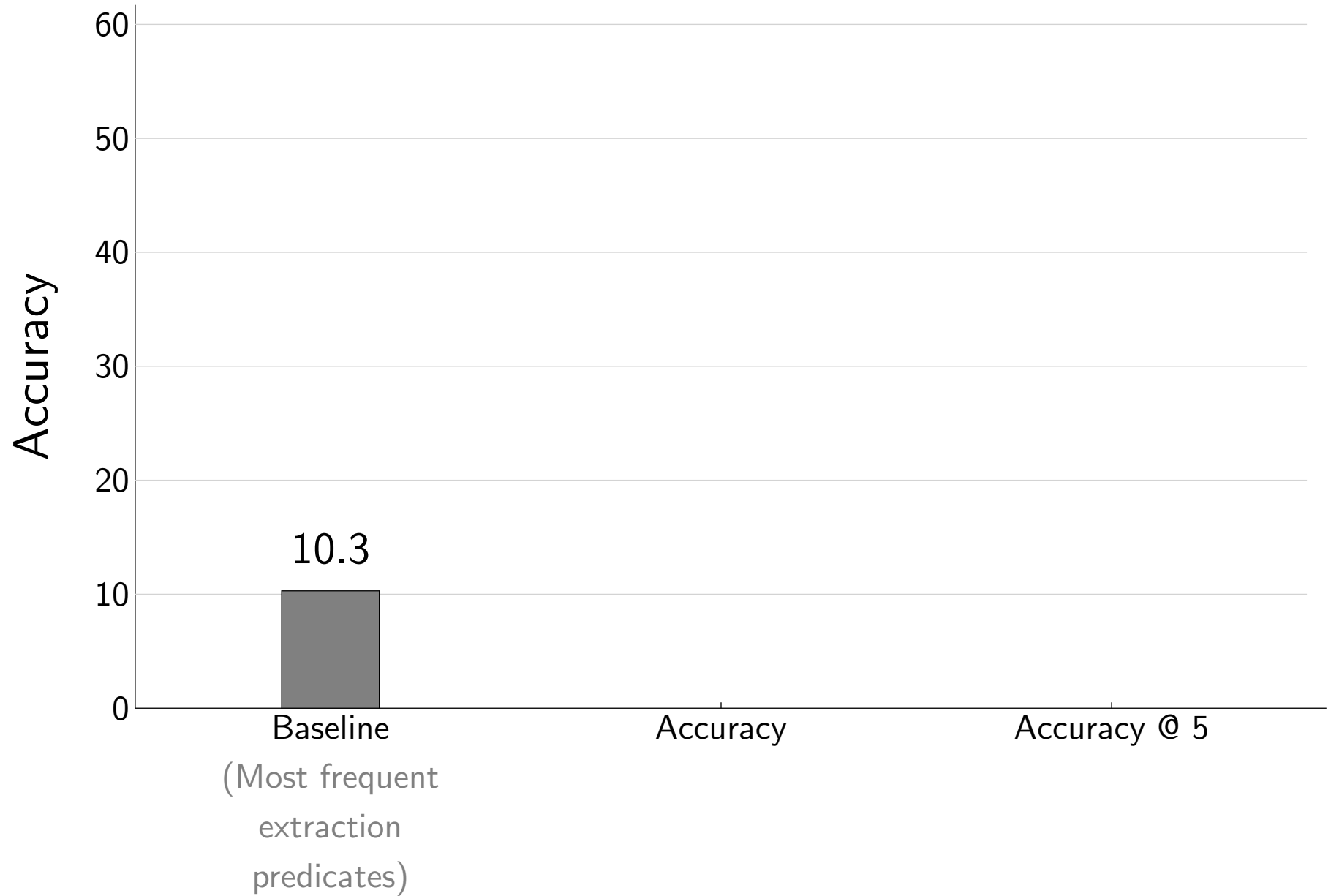
1. Setup

2. Approach

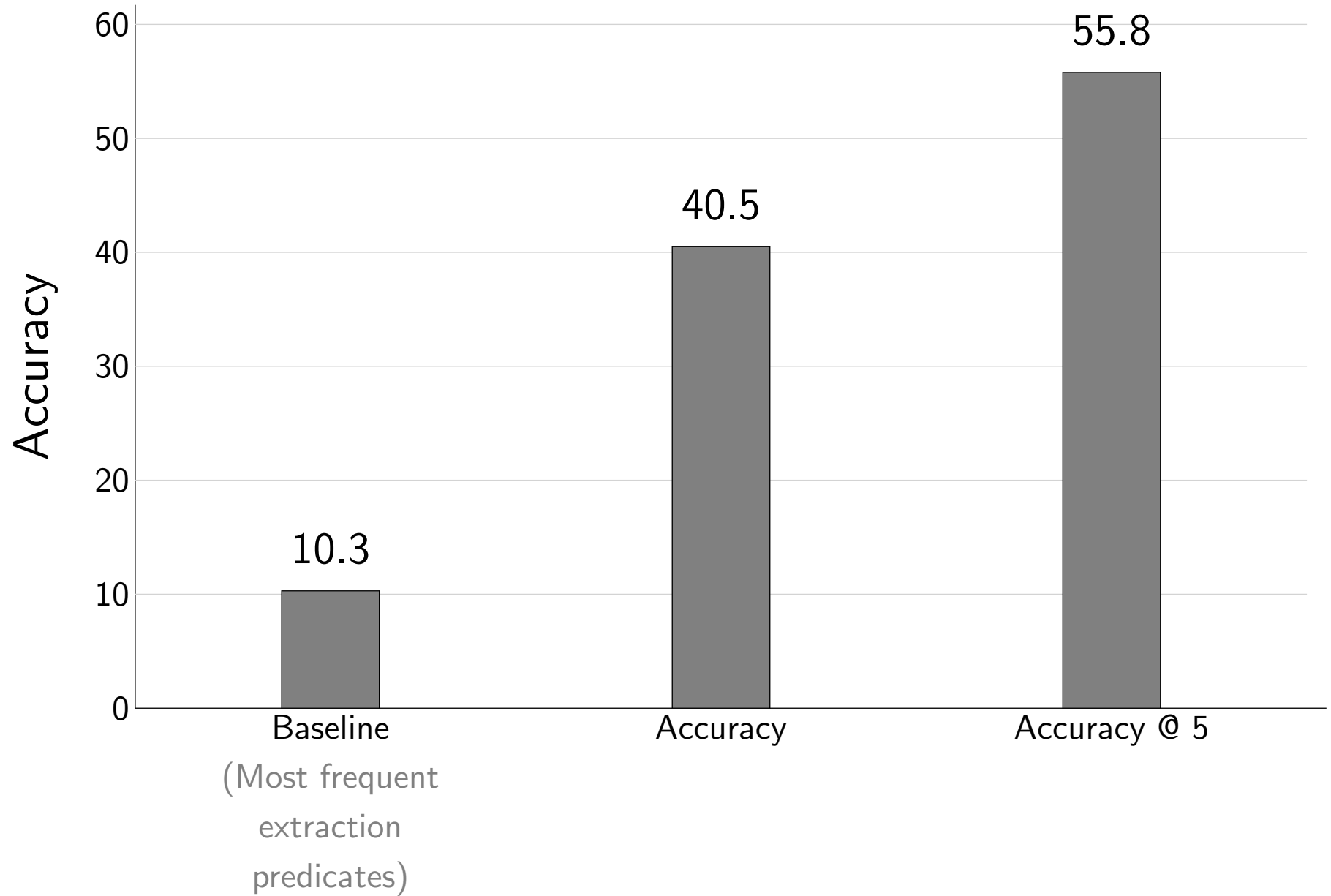
3. Results

- Main Results
- Error Analysis
- Feature Analysis

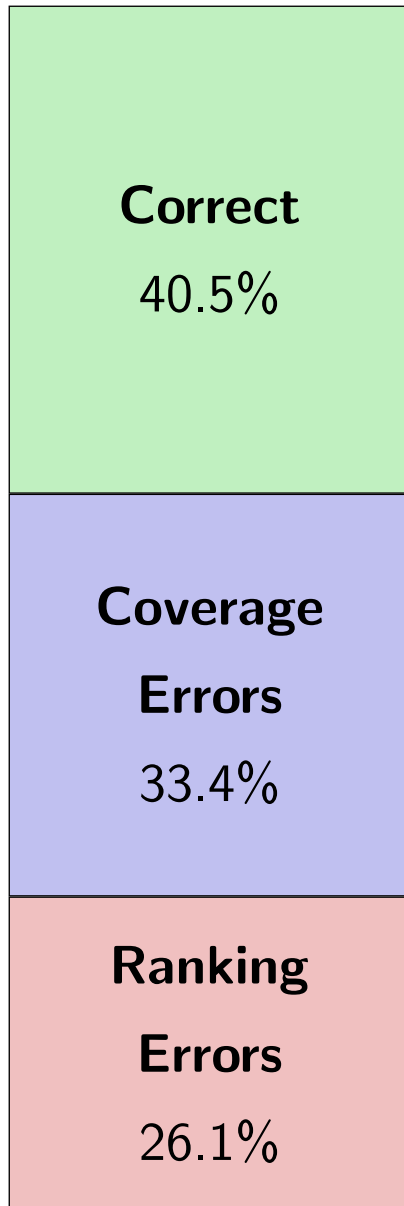
# Main Results



# Main Results



# Error Analysis





# Examples of Correct Predictions

Query: *disney channel movies*

The screenshot shows a web interface for a movie list. The title is "All Disney Channel Original Movies" by hannersbananers, created on 23 Apr 2011 and last updated on 24 Apr 2011. The list is sorted by date. The first movie is "Northern Lights" (1997 TV Movie), a 111-minute film directed by Linda Yellen, starring Diane Keaton, Maury Chaykin, Joseph Cross, and Kathleen York. The second movie is "Under Wraps" (1997 TV Movie), a 95-minute film directed by Greg Beeman, starring Adam Wylie, Mario Yedidia, Clara Bryant, and Ken Hudson Campbell. Both movie titles are highlighted with blue boxes.

/html [1] /body/div [2] /div/div/div [3] /div [1] /div/div/div/div/b

# 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	Alberta
<a href="#">Emily Carr University of Art + Design</a>	<a href="#">Athabasca University</a>
<a href="#">Kwantlen Polytechnic University</a>	<a href="#">Concordia University College of Alberta</a>
<a href="#">Royal Roads University</a>	<a href="#">MacEwan University</a>
<a href="#">Simon Fraser University</a>	<a href="#">Mount Royal University</a>

/html [1] /body/div/div/div/div/div/div/div/a/text

# Examples of Correct Predictions

Query: *nobel prize winners*

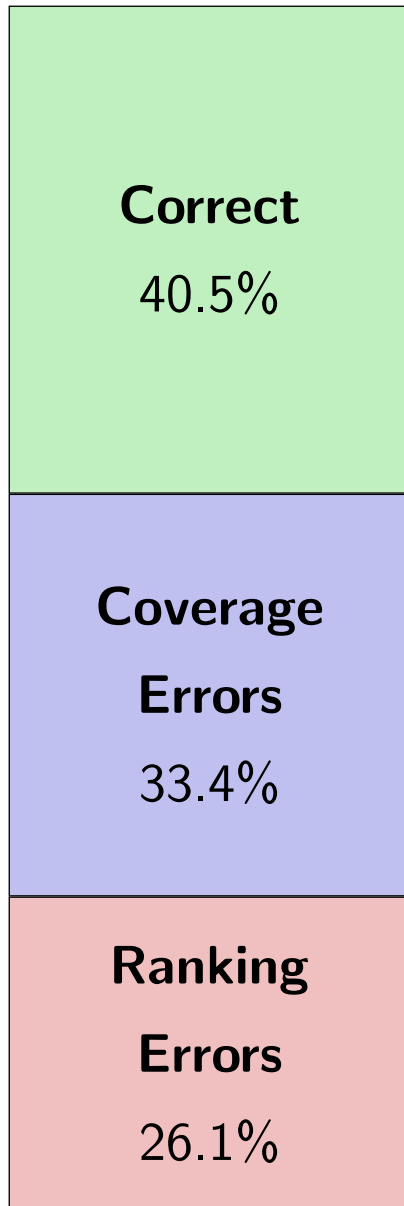
**The Nobel Prize in Physics 2013**  
**François Englert** and **Peter W. Higgs**  
"for the theoretical discovery of a mechanism that contributes to our understanding of the origin of mass of subatomic particles, and which recently was confirmed through the discovery of the predicted fundamental particle, by the ATLAS and CMS experiments at CERN's Large Hadron Collider"

**The Nobel Prize in Physics 2012**  
**Serge Haroche** and **David J. Wineland**  
"for ground-breaking experimental methods that enable measuring and manipulation of individual quantum systems"

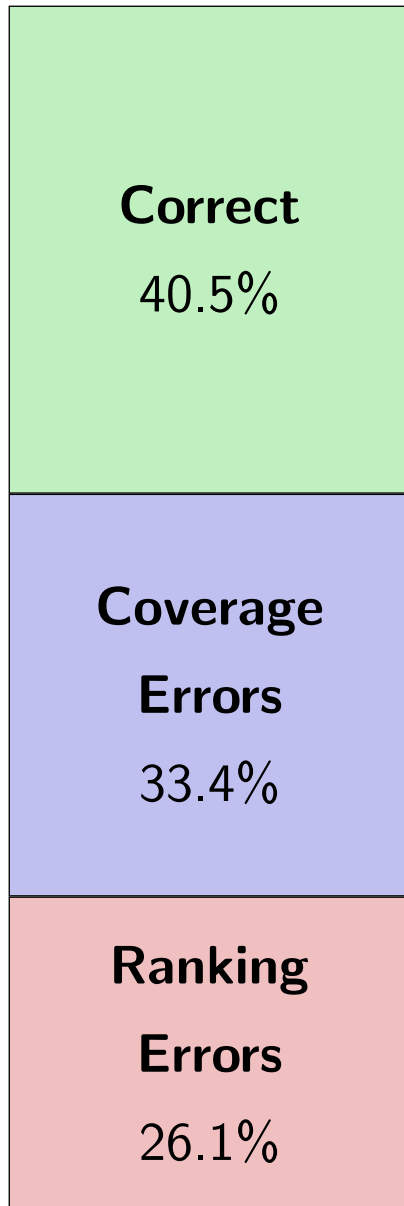
**The Nobel Prize in Physics 2011**  
**Saul Perlmutter, Brian P. Schmidt** and **Adam G. Riess**  
"for the discovery of the accelerating expansion of the universe through observations of distant supernovae"

/html[1]/body/div/div[2]/div/div/div/h6/a/text

# Error Analysis



# Error Analysis



## Coverage Errors

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

# Examples of Coverage Errors

Query: *companies named after a person*

The screenshot shows a search result for 'Z' on a Wikipedia page. The results are as follows:

- Yves Saint Laurent - Yves Saint Laurent (highlighted with a blue box)
- Yuke's - Yukinori Taniguchi (highlighted with a blue box)
- Z** [edit]
- Zagato - Ugo Zagato (highlighted with a blue box)
- Zakspeed - Erich **ZAK**owski (highlighted with a blue box)
- Zend Technologies - **Z**Eev Suraski & **AND**i Gutmans (highlighted with a blue box)
- Zust - Roberto **Z**ust (highlighted with a blue box)
- See also [edit]
- List of company name etymologies (highlighted with a red box)

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

Need richer extraction predicates!

# Examples of Coverage Errors

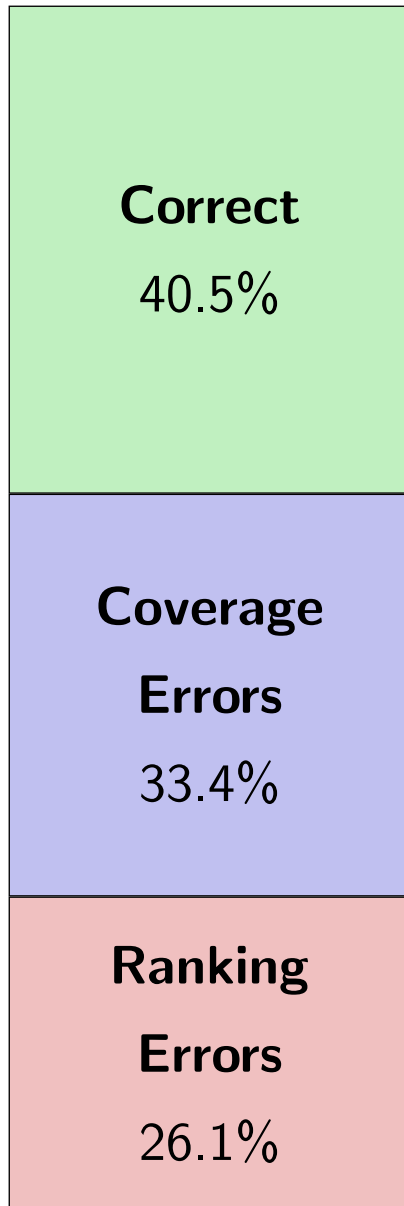
Query: *hedge funds in new york*

Rank ↕	Firm	Headquarters ↕
1	Bridgewater Associates	 Westport, CT
2	Man Group	 London
3	J.P. Morgan Asset Management	 New York
4	Brevan Howard Asset Management	 London
5	Och-Ziff Capital Management Group	 New York
6	Paulson & Co.	 New York
7	BlackRock Advisors	 New York

/html/body/div[3]/div[3]/div[4]/.../table/tbody/tr/td[2]/a

Need compositionality!

# Error Analysis

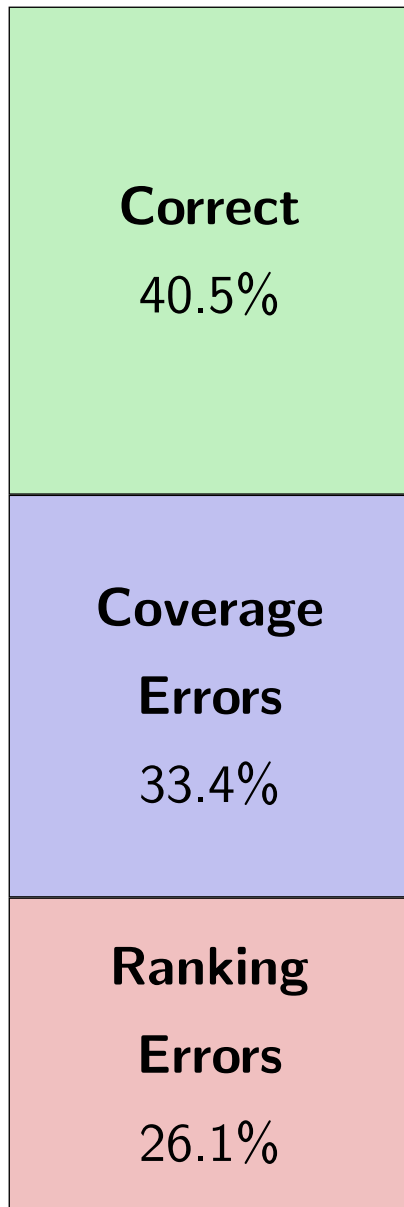


## Coverage Errors

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



# Error Analysis



## 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*

<b>Aaron, Maria</b> MD	Ophthalmology
<b>Abboushi, Nour</b> MD	Plastic Surgery
<b>Abdou, Mahmoud</b> MD	Cardiovascular Disease
<b>Abramowsky, Carlos</b> MD	Pathology
<b>Abruzzo, Todd</b> MD	Radiology

`/html/body/div[3]/div[4]/table/tbody/tr/td[2]`

# 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*

Rahm Emanuel

Richard M. Daley

Eugene Sawyer

...

*universities in Chicago*


Aurora University

DePaul University

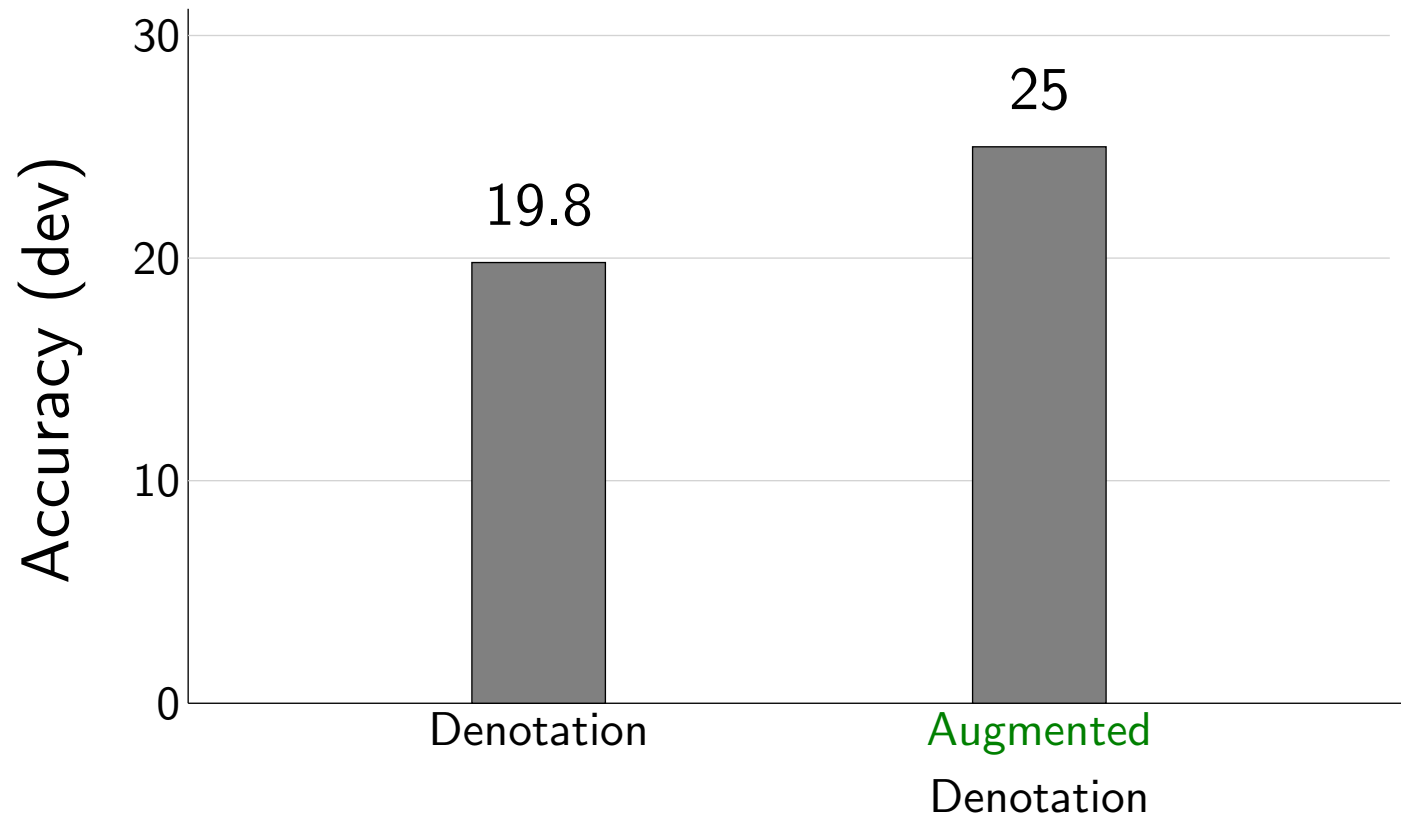
Illinois Institute of Technology

...

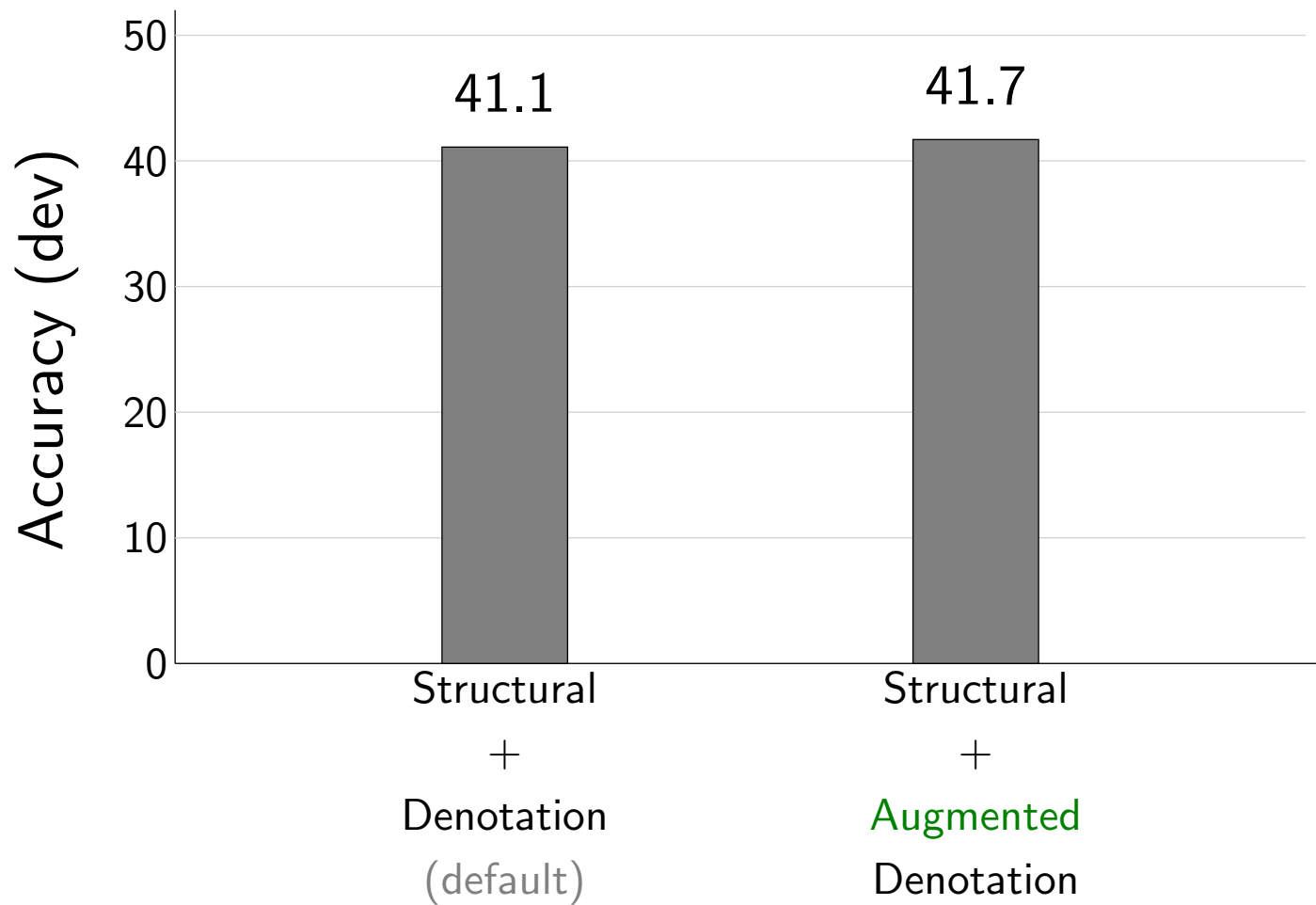
**Experiment:** Augment denotation features with the query category.

POS majority  
= NNP NNP   $\left( \begin{array}{l} \text{POS majority} \\ = \text{NNP NNP} \end{array} , \begin{array}{l} \text{query category} \\ = \text{people} \end{array} \right)$

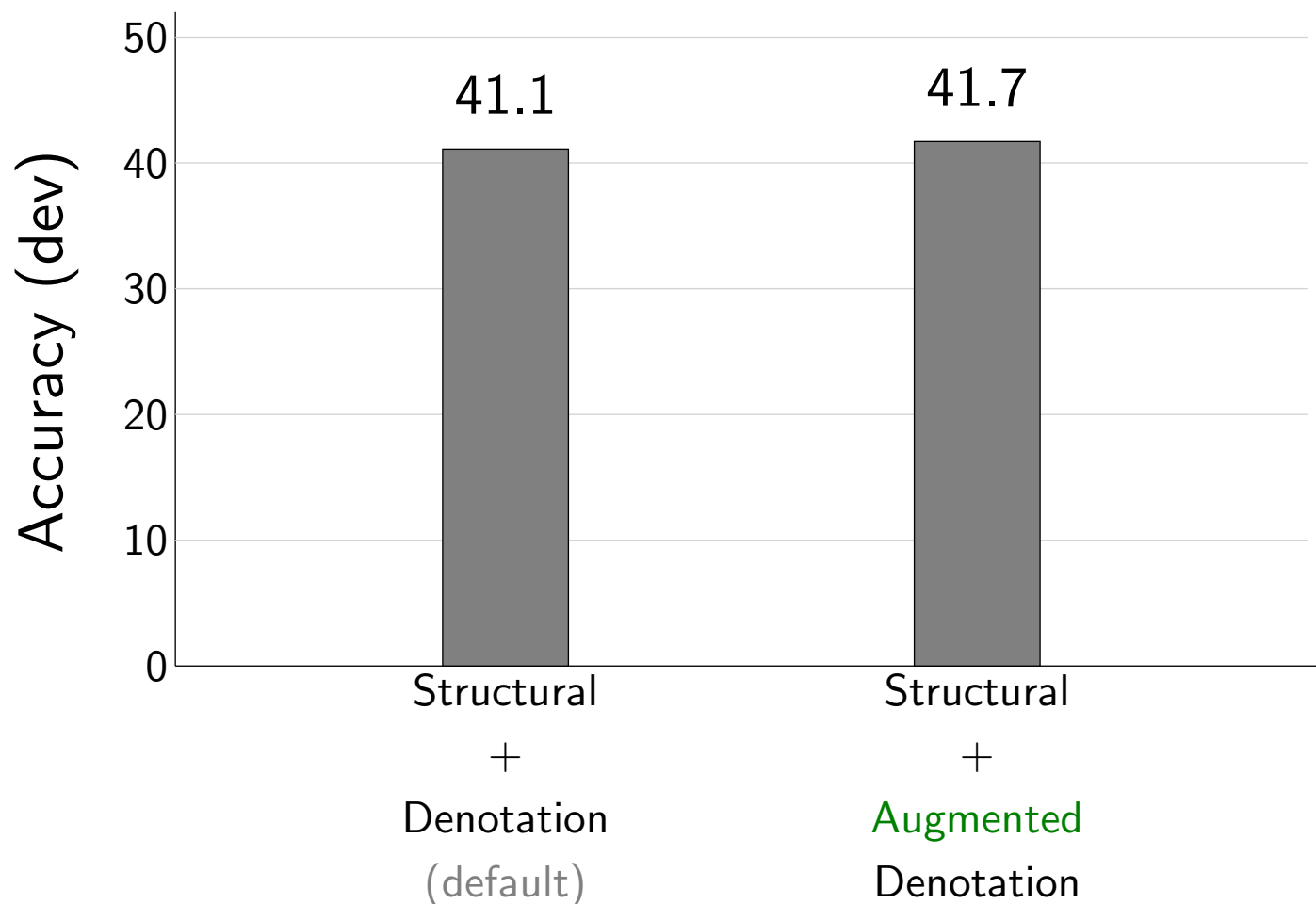
# Augmenting Denotation Features



# Augmenting Denotation Features



# Augmenting Denotation Features



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

# Augmenting Denotation Features

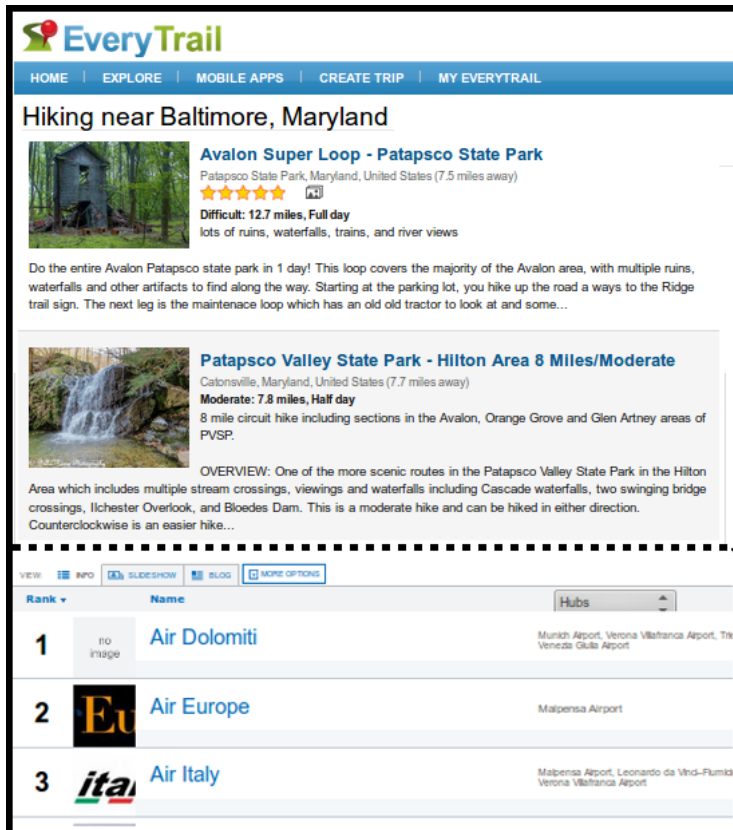
**Hypothesis:** Structural features have high influence when the web page comes from Web search result.



# Augmenting Denotation Features

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

*hiking trails near Baltimore*






The screenshot shows the EveryTrail website interface. At the top, there is a navigation bar with links for HOME, EXPLORE, MOBILE APPS, CREATE TRIP, and MY EVERYTRAIL. Below this, the page title is "Hiking near Baltimore, Maryland".

The first trail listed is "Avalon Super Loop - Patapsco State Park". It includes a small image of a wooden structure in a forest. The text describes it as a 12.7-mile full-day hike with ruins, waterfalls, trains, and river views. It is located in Patapsco State Park, Maryland, United States (7.5 miles away) and has a 5-star rating.

The second trail is "Patapsco Valley State Park - Hilton Area 8 Miles/Moderate". It includes a small image of a waterfall. The text describes it as an 8-mile circuit hike including sections in the Avalon, Orange Grove, and Glen Artney areas of PVSP. It is located in Catonsville, Maryland, United States (7.7 miles away) and is rated as moderate.

Below the trail descriptions, there is a section for "Hubs" with a table listing search results:

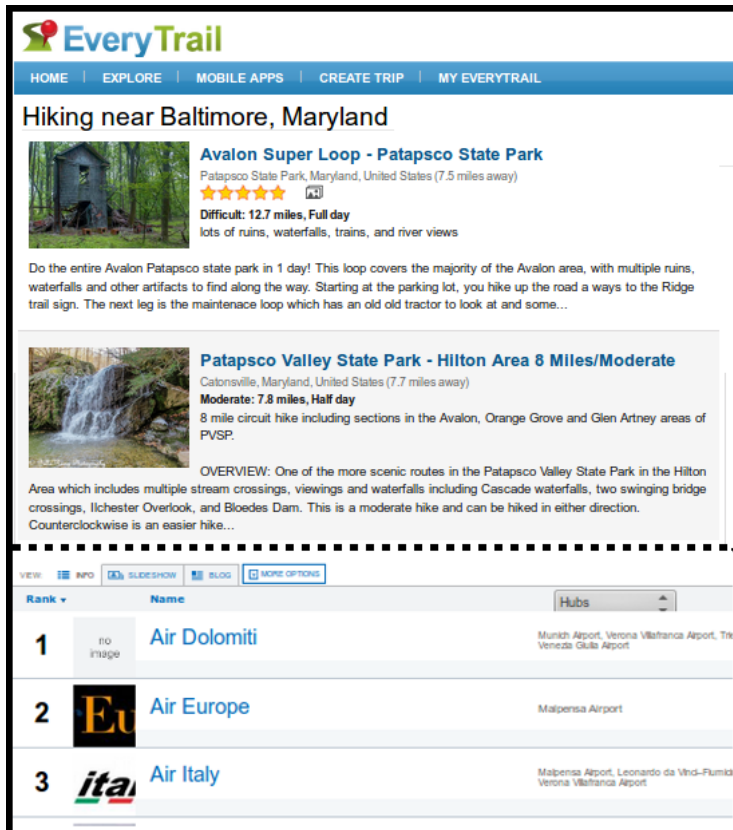
Rank	Name	Hubs
1	 Air Dolomiti	Munich Airport, Verona Villafranca Airport, Tre Venezie Giùla Airport
2	 Air Europe	Malpensa Airport
3	 Air Italy	Malpensa Airport, Leonardo da Vinci-Fiumicino Airport, Verona Villafranca Airport

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*



**EveryTrail**

HOME | EXPLORE | MOBILE APPS | CREATE TRIP | MY EVERYTRAIL

### Hiking near Baltimore, Maryland




**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 Artney 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, viewings 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...

VIEW: INFO | SLIDESHOW | BLOG | MORE OPTIONS

Rank	Name	Hubs
1	 Air Dolomiti	Munich Airport, Verona Villafranca Airport, Tre Venezie Glùbe Airport
2	 Air Europe	Malpensa Airport
3	 Air Italy	Malpensa Airport, Leonardo da Vinci-Fiumicino Airport, Verona Villafranca Airport

**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

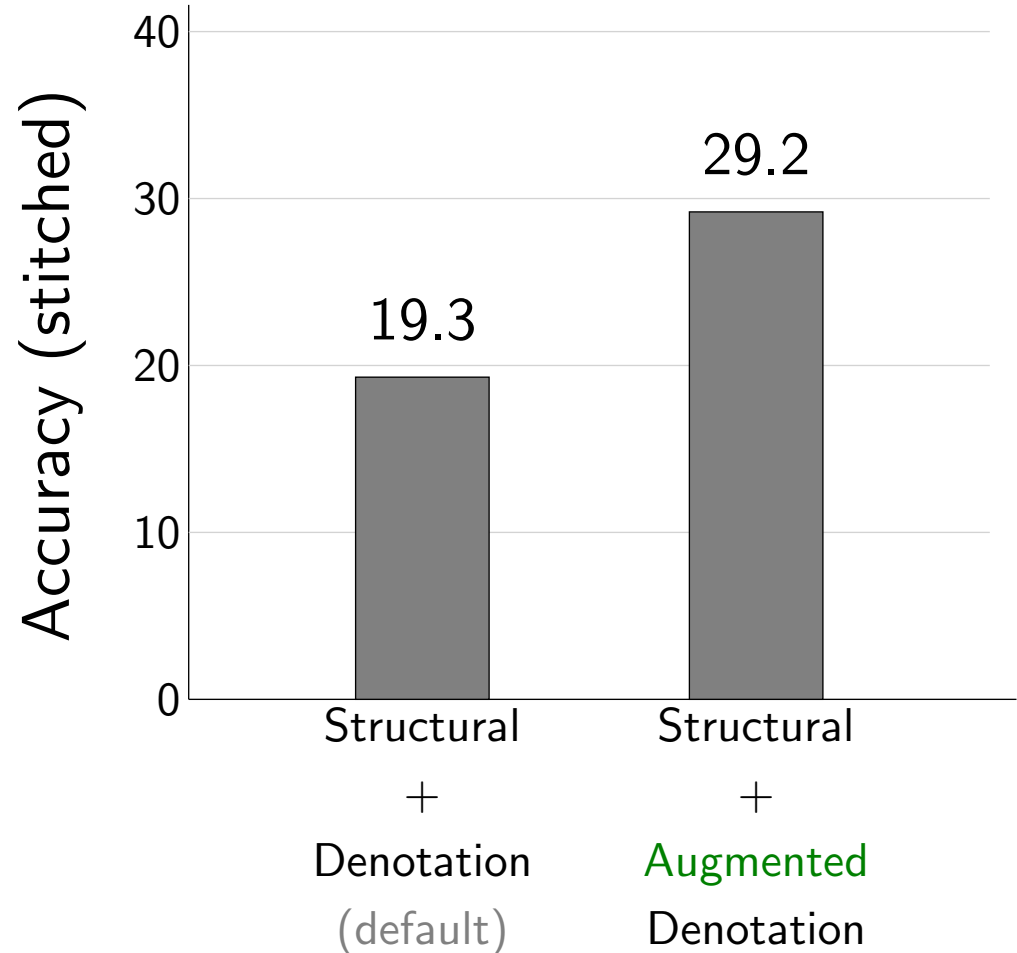
*hiking trails near Baltimore*

The screenshot shows the EveryTrail website interface. At the top, there's a navigation bar with 'HOME', 'EXPLORE', 'MOBILE APPS', 'CREATE TRIP', and 'MY EVERYTRAIL'. The main heading is 'Hiking near Baltimore, Maryland'. Below this, there are two trail listings:

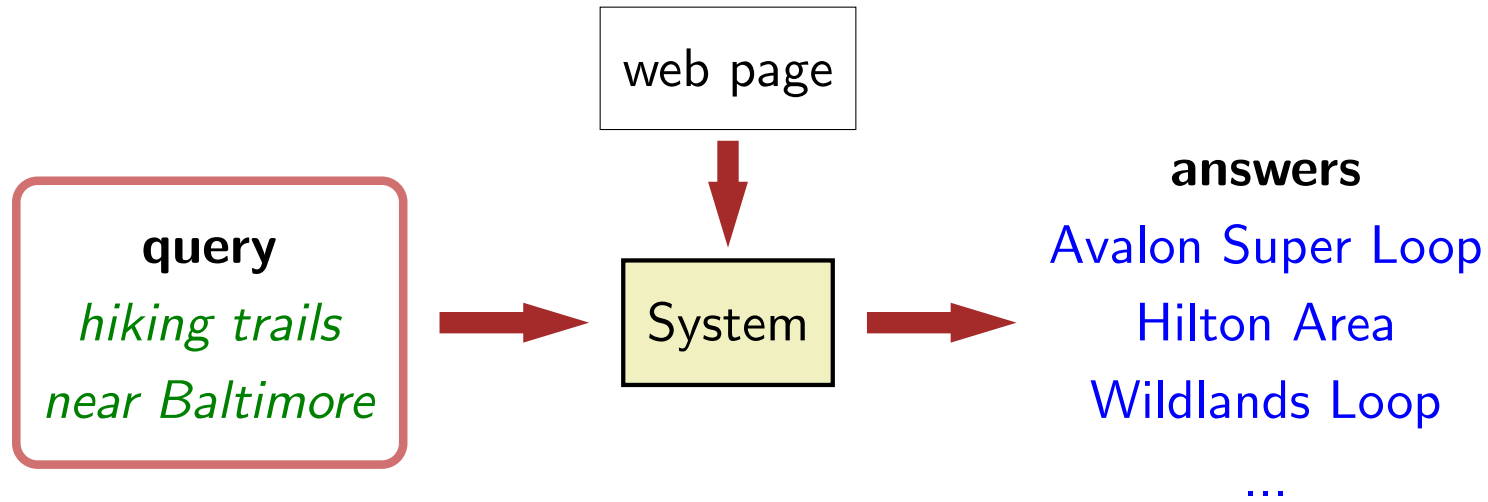
- Avalon Super Loop - Patapsco State Park**: Patapsco State Park, Maryland, United States (7.5 miles away). Difficulty: 12.7 miles, Full day. Description: lots of ruins, waterfalls, trains, and river views.
- Patapsco Valley State Park - Hilton Area 8 Miles/Moderate**: Catonsville, Maryland, United States (7.7 miles away). Difficulty: 7.8 miles, Half day. Description: 8 mile circuit hike including sections in the Avalon, Orange Grove and Glen Artney areas of PVSP.

Below the trail listings, there's a table of results with columns for Rank, Name, and Hubs. The table shows three entries:

Rank	Name	Hubs
1	Air Dolomiti	Munich Airport, Verona Villafranca Airport, Tre Venezie Glisè Airport
2	Air Europe	Malpensa Airport
3	Air Italy	Malpensa Airport, Leonardo da Vinci-Fiumicino Airport, Verona Villafranca Airport

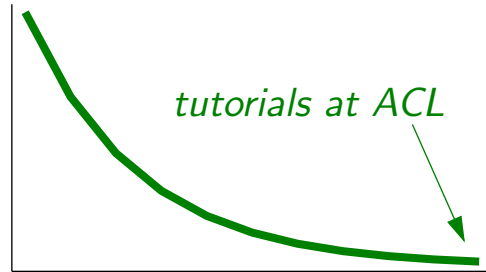


# Summary



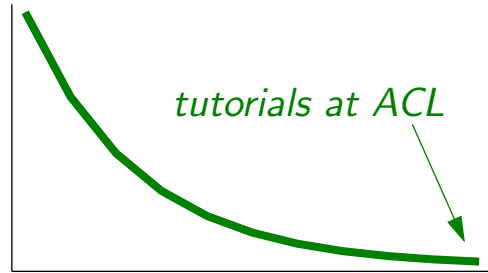
A framework for extracting entities from a **natural language query** and a single web page

# Summary

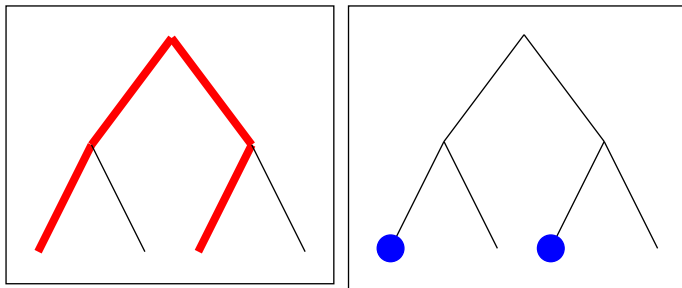


Focus on the **long tail** of entity categories

# Summary

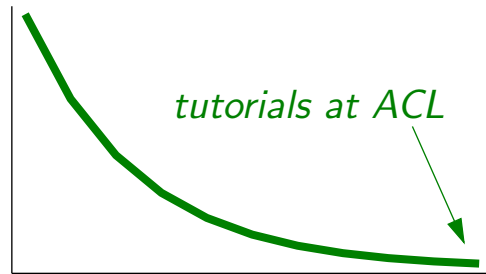


Focus on the **long tail** of entity categories

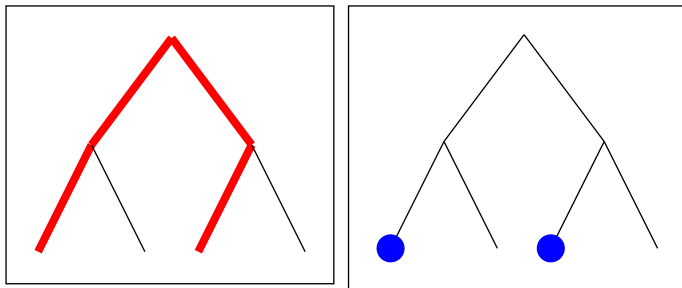


Consider both **structural** and **de-notation** features

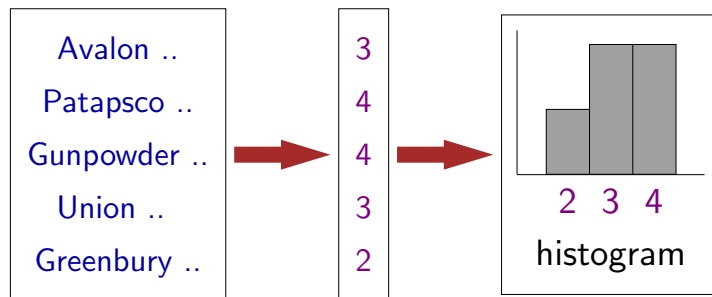
# Summary



Focus on the **long tail** of entity categories



Consider both **structural** and **de-notation** features



Handle lists of different sizes with **abstraction** and **aggregation**

# Future Work

- Model relationship between entities and category strings
- Compositionality in natural language

Rank ↕	Firm	Headquarters ↕
1	Bridgewater Associates	 Westport, CT
2	Man Group	 London
3	J.P. Morgan Asset Management	 New York
4	Brevan Howard Asset Management	 London
5	Och-Ziff Capital Management Group	 New York
6	Paulson & Co.	 New York
7	BlackRock Advisors	 New York



Download code and dataset:

<http://nlp.stanford.edu/software/web-entity-extractor-ACL2014>

**Thank you!**