Entity-Based Query Interpretation

Bachelor's Defence

Marcel Gohsen

Bauhaus-Universität Weimar

04 July 2018

Problem of Query Interpretation

```
new york times square dance
```

Problem of Query Interpretation

`new york times` `square dance`

"All the News That's Fit to Print"

The New York Times

Late Edition
Today, a shower, yielding to sun, high 64. Tonight, chilly, low 45. Tomorrow, sunny and cool, high 64. Yesterday's high, 70, low, 54. Weather map and details, Page 24.

VOL. CLVIII . No. 54,678 © 2009 The New York Times NEW YORK, SUNDAY, MAY 17, 2009 $5 beyond the greater New York metropolitan area. $4.00

Square Dance

19 U.S. States Have Designated It As Their Official State Dance

From a Theory To a Consensus On Emissions

Permits Gain Political Edge Over Taxation

By JOHN M. BRODER

WASHINGTON — As Congress weighs imposing a mandatory limit on climate-altering gases — an outcome still far from certain — it is likely to turn to a system that sets a government ceiling on total emissions and allows polluting industries to buy and sell permits to meet it.

That approach, known as cap and trade, has been embraced by President Obama, Democratic leaders in Congress, mainstream environmental groups and a growing number of business interests, including energy-consuming industries like autos, steel and aluminum.

But not long ago, many of today's supporters dismissed the idea of tradable emissions permits as an industry-inspired Republican scheme to avoid the real costs of cutting air pollution. The right answer, they said, was strict government regulation, state-of-the-art technology and a federal tax on every ton of harmful emissions.

How did cap and trade, hatched as an academic theory in obscure economic journals half a century ago, become the policy of choice in the debate over how to slow the heating of the planet? And how did it come to eclipse the idea of simply slapping a tax on energy consumption that befouls the public square or leaves the nation hostage to foreign oil

CONSERVATIVES MAP STRATEGIES ON COURT FIGHT

MEMOS OUTLINE ATTACKS

Hoping to Re-Energize G.O.P. by Opposing Obama's Choice

By CHARLIE SAVAGE

WASHINGTON — If President Obama nominates Judge Diane P. Wood to the Supreme Court, conservatives plan to attack her as an "outspoken" supporter of "abortion, including partial-birth abortion."

If he nominates Judge Sonia Sotomayor, they plan to accuse her of being "willing to expand constitutional rights beyond the text of the Constitution."

And if he nominates Kathleen M. Sullivan, a law professor at Stanford, they plan to denounce her as a "prominent supporter of homosexual marriage."

Preparing to oppose the confirmation of Mr. Obama's eventual choice to succeed Justice David H. Souter, who is retiring, conservative groups are working together to stockpile ammunition. Ten memorandums summarizing their research, obtained by The New York Times, provide a window onto how they hope to frame the coming debate.

The memorandums dissect possible nominees' records, noting statements the groups find objectionable on issues like abortion, same-sex marriage, the sep-

Problem of Query Interpretation

new york times square dance

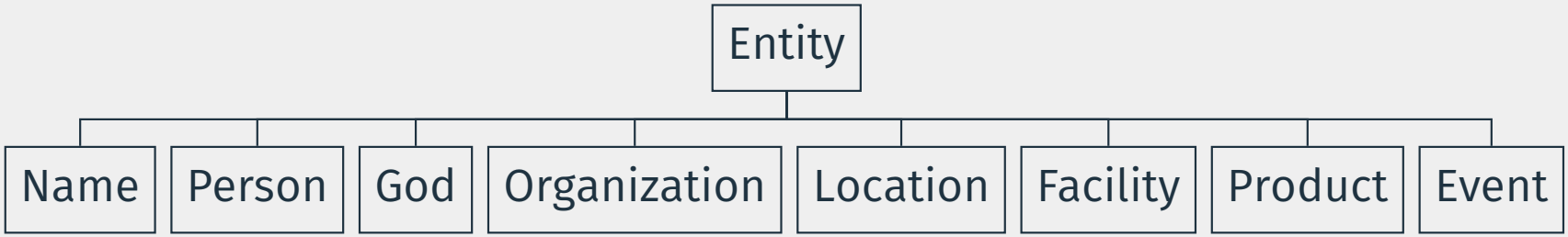
Entities in Queries

- Named Entity
  - object from the real world with a proper name
  - e.g., *person, location, organization*
- Entities in Queries
  - Definitions differ
  - May be limited to proper nouns [1]
  - May include general concepts [2]

[1] [Hasibi et al., 2015]
[2] [Cornolti et al., 2016]

Used Entity Taxonomy

- Based on “Extended Named Entity Hierarchy”
  [Sekine et al., 2002]
- 8 main classes
- 108 specialized subclasses



- for example: removed class *Units* (e.g., `kilogram`)

Traditional Problem Statements

Entity Linking [Hasibi et al., 2015]

Linking an entity in a query to the most likely candidate in some knowledge base.

`obama mother`	$\rightarrow$	(“obama”, `Barack Obama`)
`new york pizza manhattan`	$\rightarrow$	(“new york”, `New York City`)
		(“manhattan”, `Manhattan`)

Issues:

- Non-overlapping entities only

Interpretation Finding [Hasibi et al., 2015]

Finding subsets of semantic compatible non-overlapping linked entities

`obama mother`	$\rightarrow$	$\{$`Barack Obama`$\}$
`new york pizza manhattan`	$\rightarrow$	$\{$`New York City`, `Manhattan`$\}$
		$\{$`New York-Style Pizza`, `Manhattan`$\}$

Issues:

- Imprecise interpretations
- Explicit mentioned entities only

Interpretation Finding [Hasibi et al., 2015]

Finding subsets of semantic compatible non-overlapping linked entities

`obama mother`	$\rightarrow$	{`Barack Obama`}	*mother?*
`new york pizza manhattan`	$\rightarrow$	{`New York City`, `Manhattan`}	*pizza?*
		{`New York-Style Pizza`, `Manhattan`}	

Issues:

- Imprecise interpretations
- Explicit mentioned entities only

Redefined Problems

Explicit Entity Recognition

Given:
- Query

Task:
- Identifying explicit mentioned entities in a query
- Segment is an entity's name or surface form

`obama mother` $\rightarrow$	(“obama”, `Barack Obama`)
	(“obama”, `Michelle Obama`)
	(“obama”, `Natsuki Obama`)...
`new york pizza manhattan` $\rightarrow$	(“new york”, `New York City`)
	(“new york”, `New York (state)`)
	(“manhattan”, `Manhattan`
	(“manhattan”, `Manhattan (film)`)...

Implicit Entity Recognition

Given: - Query

Task: - Identifying implicitly referenced entities in a query
- Segment is a description of an entity

`obama mother` $\rightarrow$	(“obama mother”, `Ann Dunham`)
	(“obama mother”, `Marian Shields`)...
`new york pizza manhattan` $\rightarrow$	$\emptyset$
`president of usa` $\rightarrow$	(“president of usa”, `Donald Trump`)
	(“president of usa”, `Barack Obama`)
	(“president of usa”, `George W. Bush`)

Entity-Based Query Interpretation

Given:
- Query
- Explicit entities in query
- Implicit entities in query

Task:
- Semantically segmentation of query
- Replacing explicit and implicit entity-mentions with entities

obama mother → {Barack Obama, Ann Dunham}
{Michelle Obama, Marian Shields}
...

new york pizza manhattan → {New York City, "pizza", Manhattan}
...

Corpora

ERD'14 Challenge Dataset [Carmel et al., 2014]

- Dataset of the ERD'14 Challenge
- 91 queries
  - 45 queries having annotated entities
- Provides query interpretation

$$\texttt{obama family tree} \rightarrow \{\texttt{Barack Obama}\}$$
$$\texttt{east ridge high school} \rightarrow \{\texttt{East Ridge High School (FL)}\}$$
$$\{\texttt{East Ridge High School (MN)}\}$$
$$\{\texttt{East Ridge High School (KY)}\}$$

YSQLE Dataset [Yahoo, 2010]

- "Yahoo Search Query Log to Entities"
- 2635 queries
  - 2583 queries having annotated entities
- No query interpretations

```
france 1998 final →France National Football Team,
                   France, Fifa World Cup 1998 Final
      obama mother →Barack Obama, Ann Dunham
```

DBpedia-Entity v2 Dataset [Hasibi et al., 2017]

- Collection for Entity Search
- 467 queries
- No query interpretations
- Introduced relevance levels
  - 2: highly relevant
  - 1: relevant
  - 0: irrelevant

$$\texttt{john lennon, parents} \rightarrow \{\texttt{Julia Lennon} : 2,$$
$$\texttt{Alfred Lennon} : 1$$
$$\ldots : 0\}$$

Query Interpretation Corpus

- Queries from the three existing corpora
- Manually (re-)annotated:
  - Query difficulty judgments {easy | moderate | hard}
  - Explicit entities with relevance judgments {relevant | plausible}
  - Implicit entities with relevance judgments
  - Entity-based query interpretations with relevance judgments
- 2068 queries
  - 1578 queries with explicit entities
  - 131 queries with implicit entities
  - 1597 queries with query interpretations

Algorithmic Approaches

Entity Linking Steps

Typical steps for entity linking frameworks

(i) **Candidate Generation**

(ii) **Scoring**

(iii) **Selecting**

(1) Candidate Generation

- DBpedia Ontology [DBpedia, 2017] used for classification
  - Digital representation of our entity taxonomy
- Index all Wikipedia articles that represent entities
- Retrieve the top 100 articles from the index containing a segment from the query
- Retrieve for each segment of the query

(II) Scoring

$$Jaccard(T_1, T_2) = \frac{|T_1 \cap T_2|}{|T_1 \cup T_2|}$$

$$norm = \frac{|segment|}{|query|}$$

(iii) Selection

- Precision vs. Recall
- Threshold vs. Fixed number of retrieved entities

- Take the top 20 entities by score

Evaluation

Evaluation Results for Explicit Entity Recognition

Algorithm	*rec*	*prec*	F_1	rec^*	F_1^*	*RT*
Nordlys EL	**.55**	**.69**	**.58**	**.50**	**.52**	4400 ms
Explicit Entity Approach	.40	.16	.18	.35	.16	270 ms
Smaph	.38	.45	.37	.32	.31	117000 ms
TagMe	.37	.39	.33	.31	.28	**40 ms**
Nordlys ER	.33	.05	.07	.29	.06	1900 ms
Baseline	.26	.26	.26	.26	.26	-

Conclusion

- Refined problem statements for entity linking
  - Ambiguous explicit and implicit entities
  - More precise and diverse query interpretations
- Query Interpretation Corpus
  - Comparatively large corpus
  - Explicit and implicit entities
  - Query interpretations
- Algorithmic Approaches
  - Efficient explicit entity recognition
  - Implicit entity recognition prototype

Thank you for the attention!

References I

Carmel, D., Chang, M.-W., Gabrilovich, E., Hsu, B.-J. P., and Wang, K. (2014).
ERD'14: Entity recognition and disambiguation challenge.
SIGIR Forum, 48(2):63–77.

Cornolti, M., Ferragina, P., Ciaramita, M., Rüd, S., and Schütze, H. (2016).
A piggyback system for joint entity mention detection and linking in web queries.
In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 567–578, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

DBpedia (2017).
DBpedia Ontology 2016-10.
`https://wiki.dbpedia.org/services-resources/ontology`.

References II

Hasibi, F., Balog, K., and Bratsberg, S. E. (2015).
Entity linking in queries: Tasks and evaluation.
In Allan, J., Croft, W. B., de Vries, A. P., and Zhai, C., editors, *Proceedings of the 2015 International Conference on The Theory of Information Retrieval, ICTIR 2015, Northampton, Massachusetts, USA, September 27-30, 2015*, pages 171–180. ACM.

Hasibi, F., Nikolaev, F., Xiong, C., Balog, K., Bratsberg, S. E., Kotov, A., and Callan, J. (2017).
DBpedia-Entity v2: A test collection for entity search.
In Kando, N., Sakai, T., Joho, H., Li, H., de Vries, A. P., and White, R. W., editors, *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, pages 1265–1268. ACM.

Sekine, S., Sudo, K., and Nobata, C. (2002).
Extended named entity hierarchy.
In *LREC*.

References III

Yahoo (2010).
L24 - Yahoo Search Query Log To Entities v1.0.
https://webscope.sandbox.yahoo.com/.

Evaluation metrics

$$prec = \begin{cases} \frac{|E \cap E'|}{|E|}, & \text{if } |E| > 0 \\ 1, & \text{if } |E| = 0, |E'| = 0 \\ 0, & \text{if } |E| = 0, |E'| > 0 \end{cases} \tag{1}$$

$$rec = \begin{cases} \frac{|E \cap E'|}{|E'|}, & \text{if } |E'| > 0 \\ 1, & \text{if } |E| = 0, |E'| = 0 \\ 0, & \text{if } |E| > 0, |E'| = 0 \end{cases} \tag{2}$$

$$F_1 = \frac{2 \cdot prec \cdot rec}{prec + rec} \tag{3}$$

Evaluation metrics

$$w = \frac{\sum_{e \in E \cap E'} rel(e)}{\sum_{e' \in E'} rel(e')} \tag{4}$$

$$rec^* = w \cdot rec \tag{5}$$

$$F_1^* = \frac{2 \cdot prec \cdot rec^*}{prec + rec^*} \tag{6}$$

Algorithm	*prec*	*rec*	F_1	rec^*	F_1^*
TagMe	.52	.49	.44	.42	.37
Smaph	.58	.48	.47	.40	.39
Explicit Entity Approach	.14	.47	.17	.40	.14
Nordlys EL	.64	.45	.49	.38	.41
Nordlys ER	.04	.43	.07	.37	.07