human language technology
center of excellence

An Evaluation of Technologies for Knowledge Base Population

19 May 2010

Paul McNamee (JHU), Hoa Dang (NIST), Heather Simpson (LDC), Patrick Schone (DoD), Stephanie Strassel (LDC)

Talk Outline

- Background
- Language Resources
- Entity Linking Task
- Slot Filling Task
- Lessons Learned
- TAC KBP 2010

human language technology
center of excellence

Motivation

- **IE & QA technologies have been studied in isolation**
  - **Not focused on discovery of information for inclusion in an existing and evolving knowledge base**
  - **No consideration of novelty, contradiction**
- **Issues when filling in a KB**
  - **Global resolution of entities**
  - **Accurate extraction of facts**
  - **Maintaining provenance of asserted information**
  - **Avoiding contradiction / detecting change in information**
  - **Temporal qualification of assertions**
  - **Leveraging existing knowledge to assist with extraction**
  - **Scalability**
- **Pilot track at NIST Text Analysis Conference 2009**

human language technology
center of excellence

Comparison to ACE & TREC-QA

- **Corpus vs. document focus**
  - **ACE: component tasks (NER, relation extraction) for a set of individual documents**
  - **KBP: learn facts from a corpus. Repetition not valued. Asserting wrong information is bad.**
- **Context**
  - **In KBP, there is a reference knowledge base, so avoiding redundancy and detecting contradiction are important**
  - **In KBP slots are fixed and targets change. In TREC QA, the targets dictated which questions were asked.**
- **Knowing when you don't know**
  - **TREC QA had a small percentage of NIL questions (4-10%). But 80% of slots in KBP-09 had no learnable value.**

human language technology
center of excellence

Language Resources

- LDC provided 1.3M document English collection
  - 99% Newswire from 2007-2008
- Reference KB populated with semi-structured facts
  - Derived from English Wikipedia (Oct 2008)
  - 818k entries: 200k people, 200k GPEs, 60k orgs, 300+k misc/non-entities
- Test queries & judgments
  - ~4000 queries (Entity Linking)
  - 53 targets entities (Slot Filling)
- Systems reports (available at NIST website)
- Forthcoming release of evaluation data by LDC

human language technology
center of excellence

Sample KB Entry

```
<entity wiki_title="Michael_Phelps”
        type="PER”
        id="E0318992”
        name="Michael Phelps">
<facts class="Infobox Swimmer">
<fact name="swimmername">Michael Phelps</fact>
<fact name="fullname">Michael Fred Phelps</fact>
<fact name="nicknames">The Baltimore Bullet</fact>
<fact name="nationality”>United States</fact>
<fact name="strokes”>Butterfly, Individual Medley, Freestyle, Backst
<fact name="club">Club Wolverine, University of Michigan</fact>
<fact name="birthdate">June 30, 1985 (1985-06-30) (age 23)</fact>
<fact name="birthplace”>Baltimore, Maryland, United States</fact>
<fact name="height">6 ft 4 in (1.93 m)</fact>
<fact name="weight">200 pounds (91 kg)</fact>
</facts>
<wiki_text><![CDATA[Michael Phelps
Michael Fred Phelps (born June 30, 1985) is an American swimmer. H
Olympic gold medals, the most by any Olympian. As of August 2008,
world records in swimming. Phelps holds the record for the most gol
single Olympics with the eight golds he won at the 2008 Olympic Gan
```

Michael Phelps

Michael Phelps at the 2008 Beijing Olympics

Personal information	
Full name:	Michael Fred Phelps
Nickname(s):	The Baltimore Bullet[1]
Nationality:	United States
Stroke(s):	Butterfly, Individual Medley, Freestyle, Backstroke
Club:	Club Wolverine, University of Michigan
Date of birth:	June 30, 1985 (age 23)
Place of birth:	Baltimore, Maryland, United States
Height:	6 ft 4 in (1.93 m)
Weight:	200 pounds (91 kg)
Medal record	[show]

human language technology
center of excellence

Entity Linking Task

Michael Phelps

Debbie Phelps, the mother of swimming star **Michael Phelps**, who won a record eight gold medals in Beijing, is the author of a new memoir, ...

Michael Phelps

Michael Phelps is the scientist most often identified as the inventor of PET, a technique that permits the imaging of biological processes in the organ systems of living individuals. **Phelps** has ...

818k+ entries

Michael Phelps	swimmer	1985-
Michael E Phelps	biophysicist	1939-
Mike Phelps	basketball player	1961-
Edmund Phelps	economist	1933-
...		

Identify matching entry, or determine that entity is missing from KB.
Non-trivial due to name ambiguity, name variation, & KB absence.

human language technology
center of excellence

Related Work

- **Cluster Documents Mentioning Entities**
  - **Mann & Yarowsky (CoNLL 2003), Gooi & Allan (HLT 2004), WePS (Web People Search) workshops (2007, 2009)**
- **Cross-Document Entity Coreference**
  - **Group together mentions of the same named entity across documents in a large corpus**
  - **Studied at ACE 2008 (English and Arabic)**
- **Link entities to matching Wikipedia article**
  - **Bunescu & Pasca (EACL 2006), Cucerzan (EMNLP 2007)**
    - **Cucerzan reported 92% accuracy in news (NEs in 20 docs)**
  - **Differences with KBP 2009**
    - **Ignore absent entities**
    - **KBP only worked with PER/ORG/GPEs; intentionally selected challenging target names**
      - **Simpson et al. (LREC 2010): Wikipedia and the Web of Confusable Entities: Experience from Entity Linking Query Creation for TAC 2009 Knowledge Base Population**

Entity Linking Queries, Performance

	All	in-KB	Absent
All	0.8217 (3904)	0.7654 (1675)	0.8641 (2229)
PER	0.8309 (627)	0.8039 (255)	0.8495 (372)
ORG	0.8151 (2710)	0.7305 (1013)	0.8696 (1697)
GPE	0.8480 (567)	0.8280 (407)	0.8812 (160)

Accuracy for top-scoring run: Siel_093

- **3904 Queries (560 distinct entities)**
  - **15% People**
  - **70% Organizations**
  - **15% Geo-Political Entities**

- **43% Present in KB / 57% Absent**

Top 5 Systems

Team	All	in KB	Absent
Siel_093	0.8217	0.7654	0.8641
QUANTA1	0.8033	0.7725	0.8264
hltcoe1	0.7984	0.7063	0.8677
Stanford_UBC2	0.7884	0.7588	0.8107
NLPR_KBP1	0.7672	0.6925	0.8232
'NIL' Baseline	0.5710	0.0000	1.0000

Micro-averaged accuracy

human language technology
center of excellence

Typical Approach

Query = "CDC"
1. California Dept. of Corrections
2. US Center for Disease Control
3. Cedar City Regional Airport (IATA code)
4. Communicable Disease Centre (Singapore)
5. Congress for Democratic Change (Liberian political party)
6. Cult of the Dead Cow (Hacker organization)
7. Control Data Corporation
8. NIL (Absence from KB)
9. Consumers for Dental Choice (non-profit)
10. Cheerdance Competition (Philippine organization)

- **Two phased**
  - **1. Candidate identification based on target name**
  - **2. Candidate selection (or ranking) exploiting document features using supervised machine learning**
  - **3. Choosing absence (NIL)**

"According to the CDC the prevalence of H1N1 influenza in California prisons has..."

"William C. Norris, 95, founder of the mainframe computer firm CDC., died Aug. 21 in a nursing home ... "

human language technology
center of excellence

Most Difficult Queries

- **Subsidiary organization**
  - 3871 – Xinhua Finance Ltd vs. Xinhua Finance Media Ltd
- **Typographical mistake / ambiguous acronym**
  - 1213 – DCR for Democratic Republic of Congo
  - 3141 – MND (Taiwan Ministry of National Defense) referred to as NDM in text
- **Foreign names, acronyms**
  - 2885 – PCC – Spanish acronym for Colombian Communist Party (Partido Comunista Colombiano)
- **Metaphorical 'names'**
  - 1717/1718 Iron Lady (several strong female politicians)
- **Unclear referent, metonymy**
  - 2599 – New Caledonia (country or soccer team)
- **Mistakes in human assessments**
  - 3333,3334 – NYC Dept of Health, not US Dept of Health
  - 3335 – NY State Dept of of Health, not US Dept of Health

human language technology
center of excellence

Slot Filling Task

Target: EPA
(plus 1 document)

Generic Entity Classes
Person, Organization, GPE

Missing information to mine from corpus:
- Date formed: 12/2/1970
- Website: http://www.epa.gov/
- Headquarters: Washington, DC
- Nicknames: EPA, USEPA
- Type: federal agency
- Address: 1200 Pennsylvania Avenue NW

Entity Attributes

Person	Organization	Geo-Political Entity
alternate names	alternate names	alternate names
age	political/religious affiliation	capital
birth: date, place	top members/employees	subsidiary orgs
death: date, place, cause	number of employees	top employees
national origin	members	political parties
residences	member of	established
spouse	subsidiaries	population
children	parents	currency
parents	founded by	
siblings	founded	
other family	dissolved	
schools attended	headquarters	
job title	shareholders	
employee-of	website	
member-of		
religion		
criminal charges		

Convocation of Anglicans in North America

Slot	Correct Values in Pools
org:alternate_names	CANA
org:founded	2005
org:founded_by	Peter Akinola
org:headquarters	Nigeria
org:member_of	Anglican Church, Nigerian Anglican Church
org:number_of_employees/ members	100,000
org:parents	diocese of the Church of Nigeria, Nigerian Anglican Church
org:political/religious_affiliation	Anglican, Anglican Communion, Episcopal, Episcopal church, Christianity
org:top_members/employees	Peter Akinola, Bishop Martyn Minns, Kelly Oliver
org:website	www.canaconvocation.org

human language technology
center of excellence

Slot Filling Scoring

- **Responses were marked as one of Correct, Inexact, Redundant, or Wrong**
  - **Must be justified from a single supporting document**
  - **Ground truth unknown – evaluated based on pooled system responses**

- **53 target entities (17 PER, 31 ORG, 5 GPE)**
  - **255 single-value slots – 39 (15%) had correct values in the pooled responses**
  - **499 list slots – 129 (26%) had correct values**
  - **Thus predicting NIL (no response) is correct ~80% of the time**
  - **48/53 entities had at least one learnable attribute**

Easy / Hard Slots

Slot	Filled Entities	Correct Responses	Submitted Responses
per:title	16/17	86	409
per:employee_of	10/17	38	429
per:origin	9/16	16	117
per:member_of	9/17	41	424
org:top_members/ employees	24/31	258	1463
org:alternate_names	23/31	87	710
org:headquarters	11/21	17	131

- **No correct values learned for:**
  - **PER: other_family, parents, spouse**
  - **ORG: shareholders**
  - **GPE: capital, political_parties, population**

human language technology
center of excellence

Slot Filling Evaluation Metric

$$Score_{\text{single}} = \frac{NumCorrect}{NumSingleSlots}$$

$$ListSlotValue = \frac{5 \times IP \times IR}{4 \times IR + IP}$$

F_β=2 to weight precision over recall.
IP = Instance precision.
IR = Instance recall.

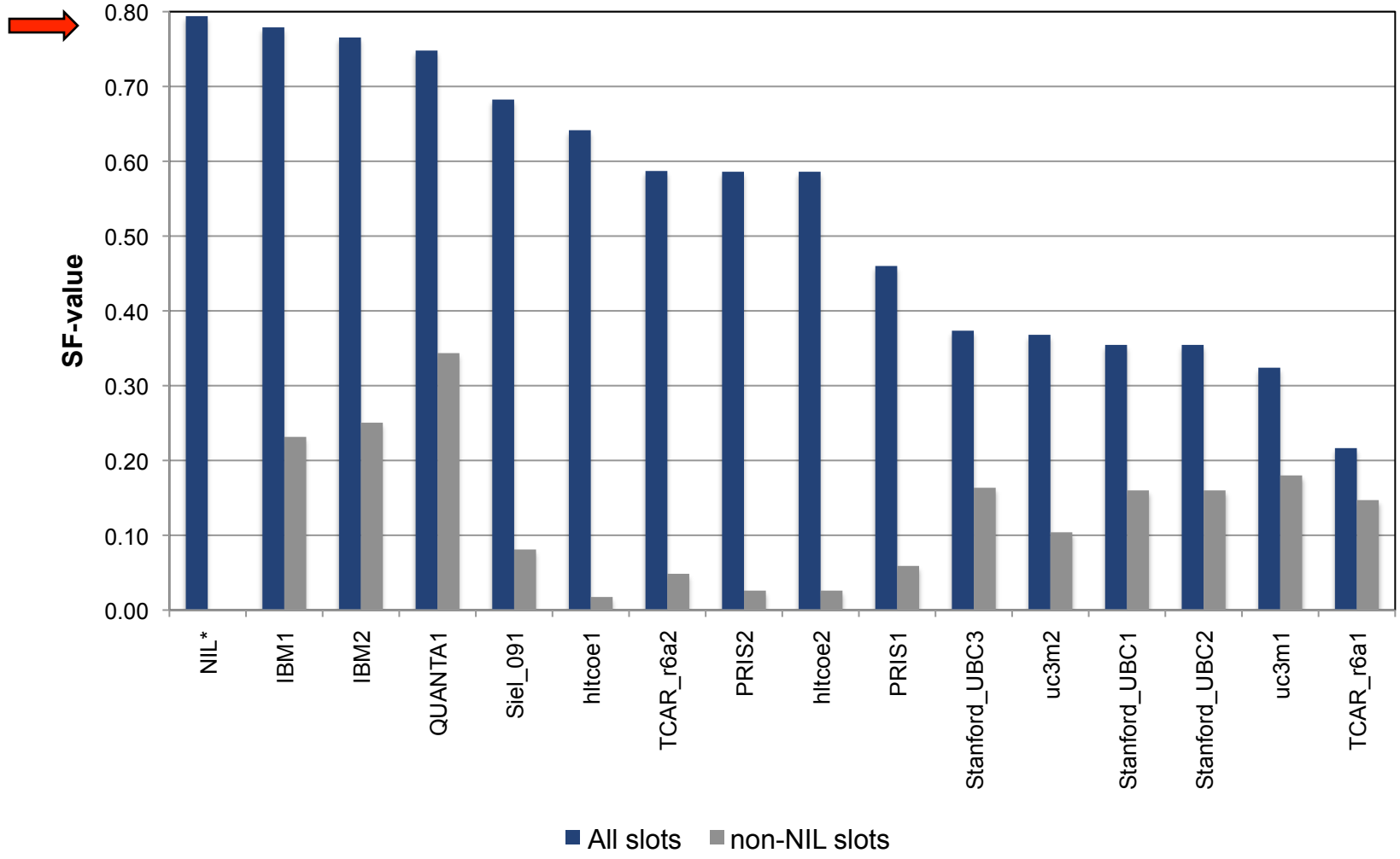
$$Score_{\text{list}} = \frac{\sum ListSlotValue}{NumListSlots}$$

$$SF_{\text{value}} = \frac{1}{2}\left(Score_{\text{single}} + Score_{list}\right)$$

"Correct" included appropriately predicting 'no answer'

human language technology
center of excellence

SF Results



human language technology
center of excellence

Lessons Learned

- GPEs have few learnable attributes in news
  - latitude, longitude, elevation not commonly reported
  - population is, but usually available in KB/Wikipedia
- Difficult to estimate how much information is available (and novel) for a candidate target entity
  - Manual search needed both to facilitate target selection and enrich pools
  - Evaluation measure was unbalanced between slots with discoverable vs. NIL values
- End-to-end assessment of 'KB improvement' is difficult. Component evaluation for KBP is worth considering.
  - Can a passage support a given slot for a given entity? (A text retrieval task)
  - Is a particular slot fill justified from a passage? (An RTE task)
  - Is this slot fill redundant with another value?

Thanks to 13 Participating Teams

BUAP_1	B. Autonomous University of Puebla
CSLU.OHSU	Oregon Health and Science University
DAMSEL	Macquarie University
HLTCOE	JHU Human Language Technology Center of Excellence
IBM	TJ Watson IBM Research
Janya	Janya Inc.
NLPR_KBP	National Laboratory of Pattern Recognition, China
PRIS	Beijing University of Posts and Telecommunications
QUANTA	Tsinghua University
Siel_09	International Institute of Information Technology
Stanford_UBC	Stanford University
TCAR_r6a	National Security Agency
UC3M	Universidad Carlos III de Madrid

human language technology
center of excellence

Summary

- Pilot evaluation for adding information to an existing knowledge base
- 2 initial tasks
  - Linking name mentions to KB entries
  - Augmenting profiles for target entities
- KBP 2010
  - Heng Ji and Ralph Grishman are coordinating KBP 2010
    - Including non-news (Web data)
    - 'Surprise' slot filling task (novel entity types)
  - Website:
    - http://www.nist.gov/tac/
    - http://nlp.cs.qc.cuny.edu/kbp/2010/
  - Registration deadline: this week
  - Results due: ~ by end of July

Questions?

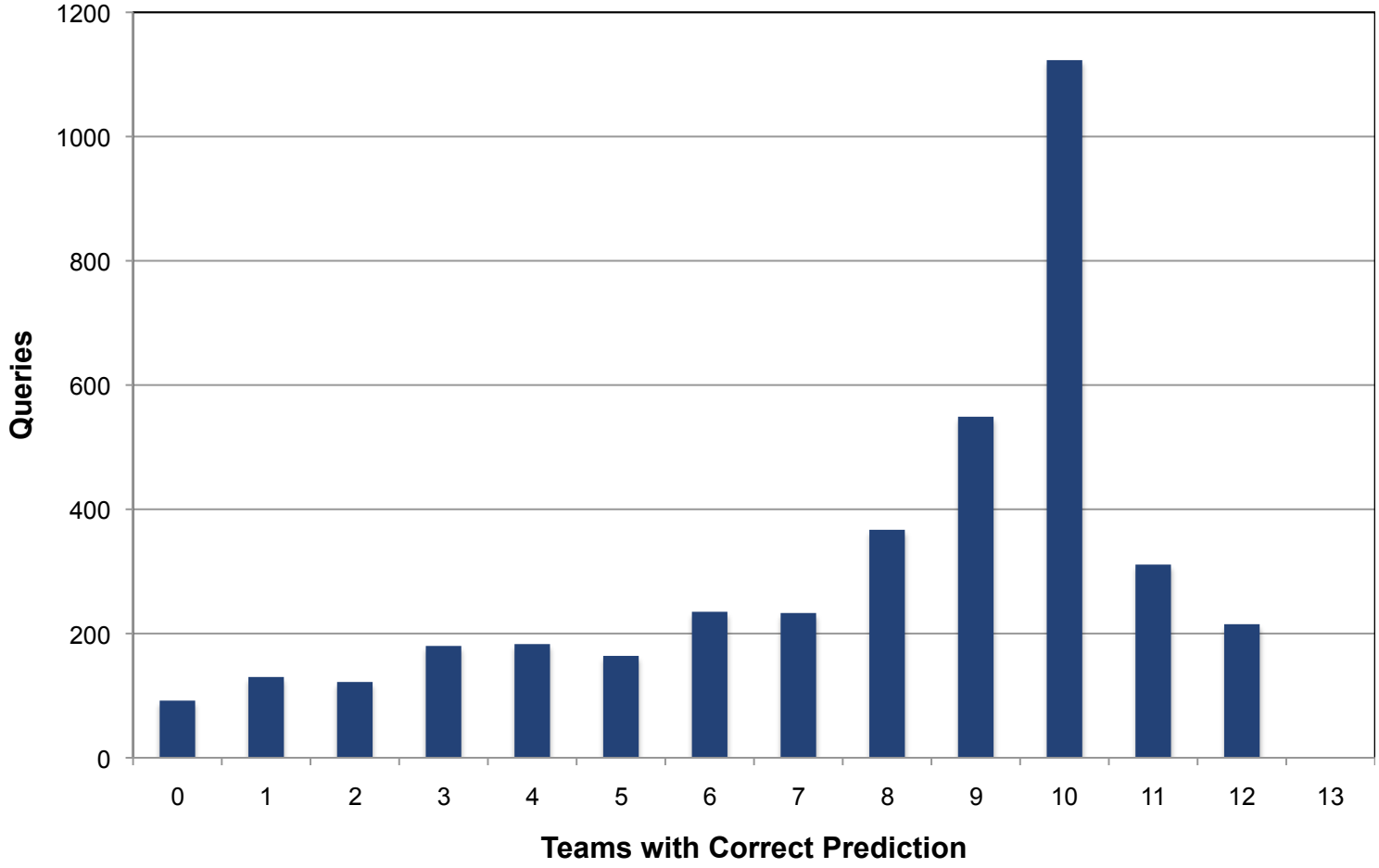
human language technology
center of excellence

Other Evaluation Issues

- **Imperfect KB**
  - **Wikipedia focuses on presentation, not representation**
    - **irrelevant slots (colors, image sizes), values are not normalized (e.g., dates)**
  - **Many non-entities**
- **Use of external resources (e.g., Internet search)**
- **Generic entities (i.e., PER vs. scientist, athlete, writer)**
  - **Slot names were inconsistent (birthdate, date-of-birth)**
- **Response granularity**
  - **USA, Hawaii, Honolulu – which should be considered correct birthplaces for President Obama?**
- **Dealing with time**
  - **Key USA leadership: G. Washington or B. Obama**
- **Query Difficulty (and high NIL percentage)**
- **Assessing KB Growth**
  - **Difficult to directly measure benefit from adding to KB**

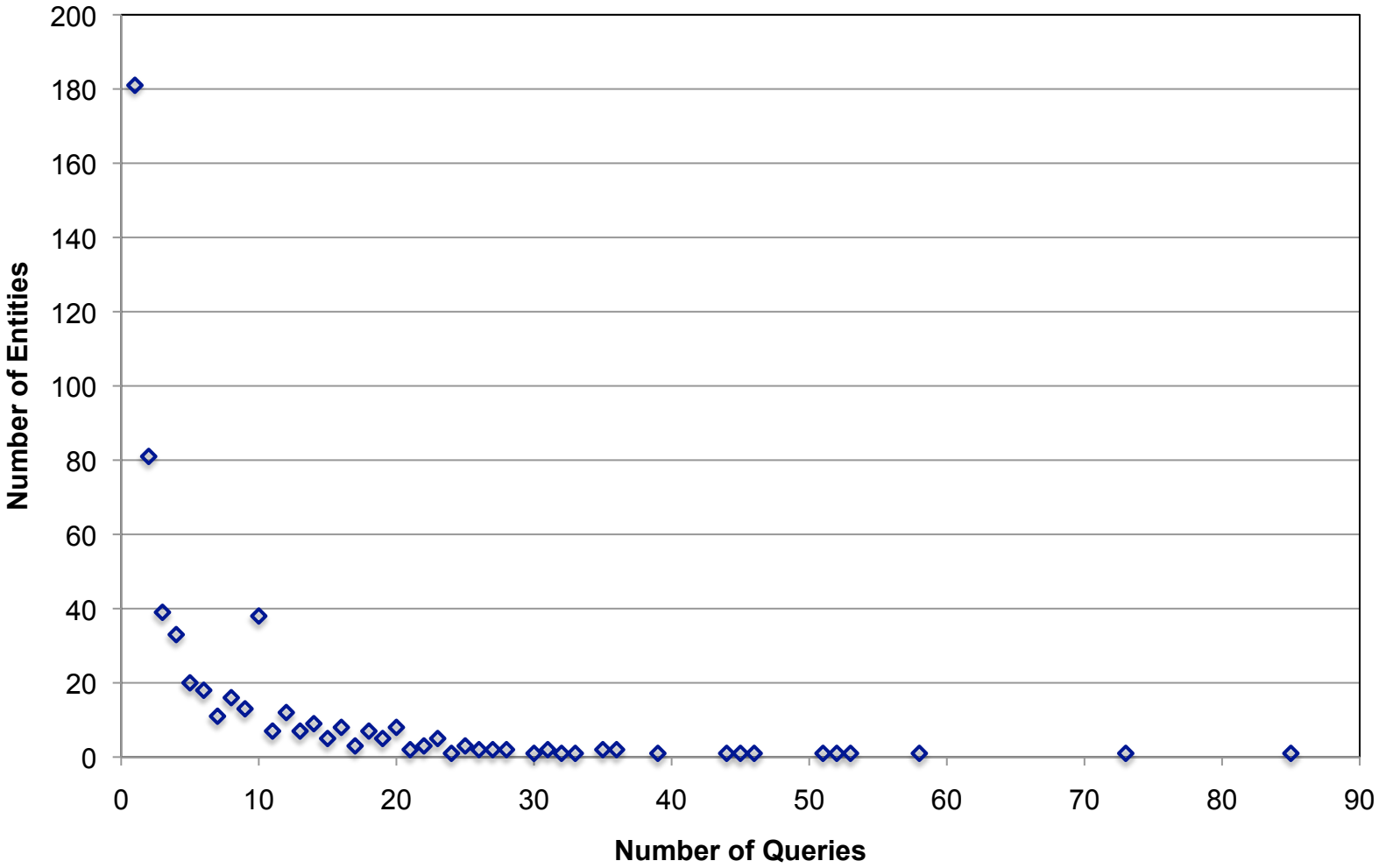
human language technology
center of excellence

EL: Agreement with Judged Response



human language technology
center of excellence

EL: Queries per Entity



human language technology
center of excellence

Entity Linking Example

EL1718 – Iron Lady

The furor also brought China's long-running domestic food safety problems to light, just as Beijing prepares to host hundreds of thousands of foreign visitors at the summer Olympics in August.

The seriousness with which the government took the issue was underscored by the appointment of its top problem solver, Vice Premier Wu Yi, to head a Cabinet-level panel overseeing the campaign.

Wu, a stern-looking, 69-year-old woman known as the "Iron Lady," shepherded China's difficult entry into the World Trade Organization, took over as health minister during the SARS epidemic and has been tasked with handling the vociferous U.S. complaints about China's exchange rate policy.

One month into the product safety campaign, Wu herself set out to randomly inspect shops and restaurants in the eastern province of Zhejiang. She had no itinerary and told no one in advance, making the driver stop at her whim.

human language technology
center of excellence

Entity Linking Example

EL3871 – Xinhua Finance

Chinese business news giant Xinhua Finance Media Ltd. is seeking to raise up to 371 million dollars through an initial public offering (IPO) on the Nasdaq stock market, according to a US regulatory filing.

...

"These outlets reach an estimated 210 million potential television viewers, a potential listening audience of 33 million people, and the readers of leading magazines and newspapers," Xinhua Finance Media said.

...

Describing itself as "a leading diversified media company in China," Xinhua Finance said it would use 50 million dollars from its US share listing to repay debts and "an undetermined amount" for future acquisitions.

The firm, which is based in the Cayman Islands, said it would be 36.7 percent owned by parent Xinhua Finance Ltd., 8.0 percent by Patriarch Partners Media Holdings LLC., and 5.8 percent owned by chief executive Fredy Bush, among other shareholders.

Convocation of Anglicans in North America

- **founded_by**
  - **Akinola, AMIA Bishop Chuck Murphy, Bishop Martyn Minns, Episcopal, Helmandollar, Jim Robb, Martyn Minns, Minns, Peter Akinola, Robinson, Stephen**

- **shareholders**
  - **Anglican Church, Bishop Martyn Minns, CANA, Episcopal Church, Martyn Minns, Peter Akinola**

- **headquarters**
  - **America, Nigeria, Quincy, Woodbridge**