The MSR System for Entity Linking at TAC 2013

Silviu Cucerzan
Microsoft Research
Machine Learning and Intelligence / MS_MLI

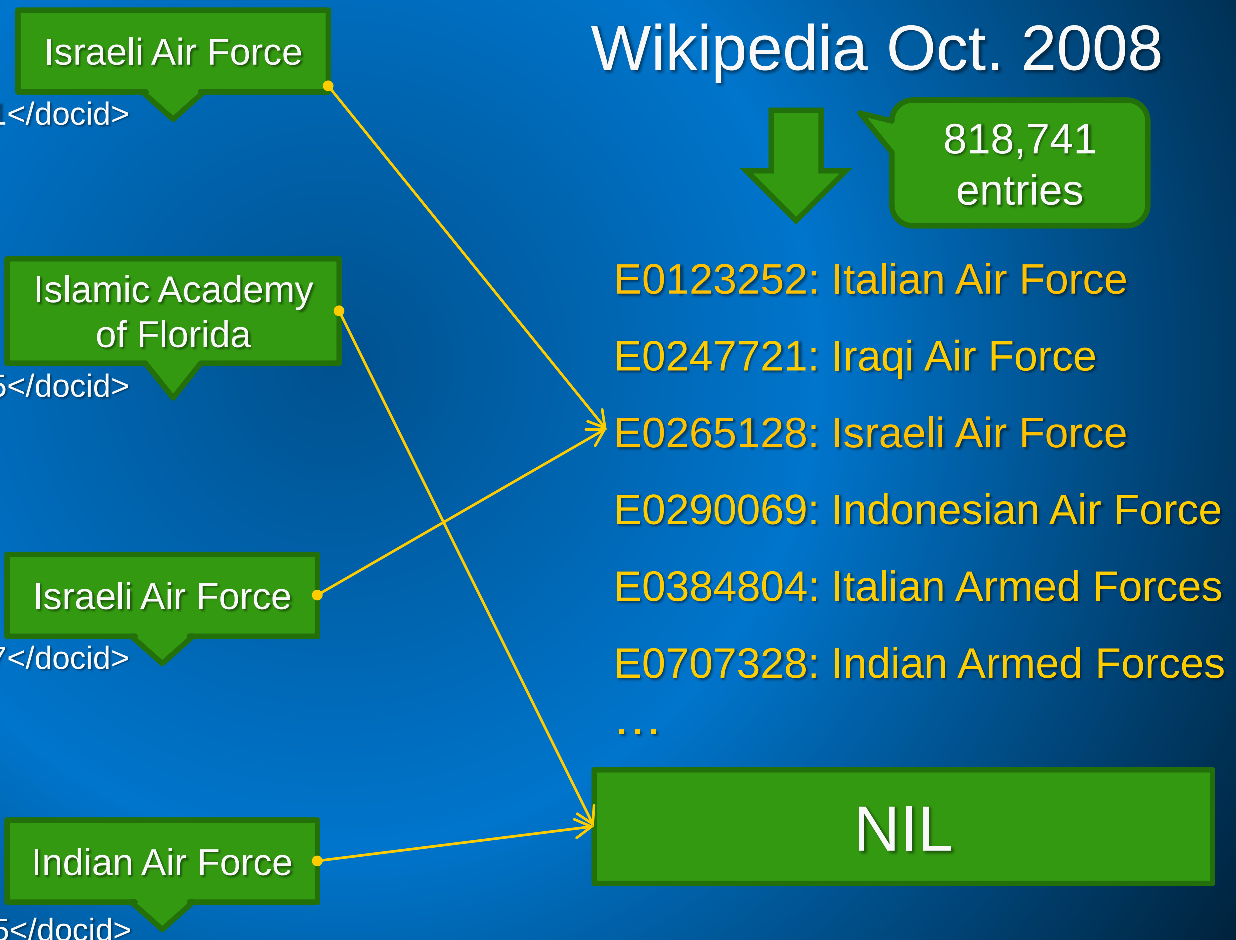
Gaithersburg, MD

11/19/2013

Task Description

For a string at a given offset in a document, determine which entity from the provided knowledge base (if any) is being referred to by the string. Cluster all entities in the test set.

```
<query id="EL005833">
   <name>IAF</name>
   <docid>eng-WL-11-174596-12954631</docid>
   <offset>…</offset>
</query>
<query id="EL005836">
   <name>IAF</name>
   <docid>eng-NG-31-142148-10021195</docid>
   <offset>…</offset>
</query>
<query id="EL05838">
   <name>IAF</name>
   <docid>eng-WL-11-174596-12954257</docid>
   <offset>…</offset>
</query>
<query id="EL05847">
   <name>IAF</name>
   <docid>eng-NG-31-147166-10475895</docid>
   <offset>…</offset>
</query>
```



Israeli Air Force

Islamic Academy of Florida

Israeli Air Force

Indian Air Force

Wikipedia Oct. 2008

818,741 entries

E0123252: Italian Air Force

E0247721: Iraqi Air Force

E0265128: Israeli Air Force

E0290069: Indonesian Air Force

E0384804: Italian Armed Forces

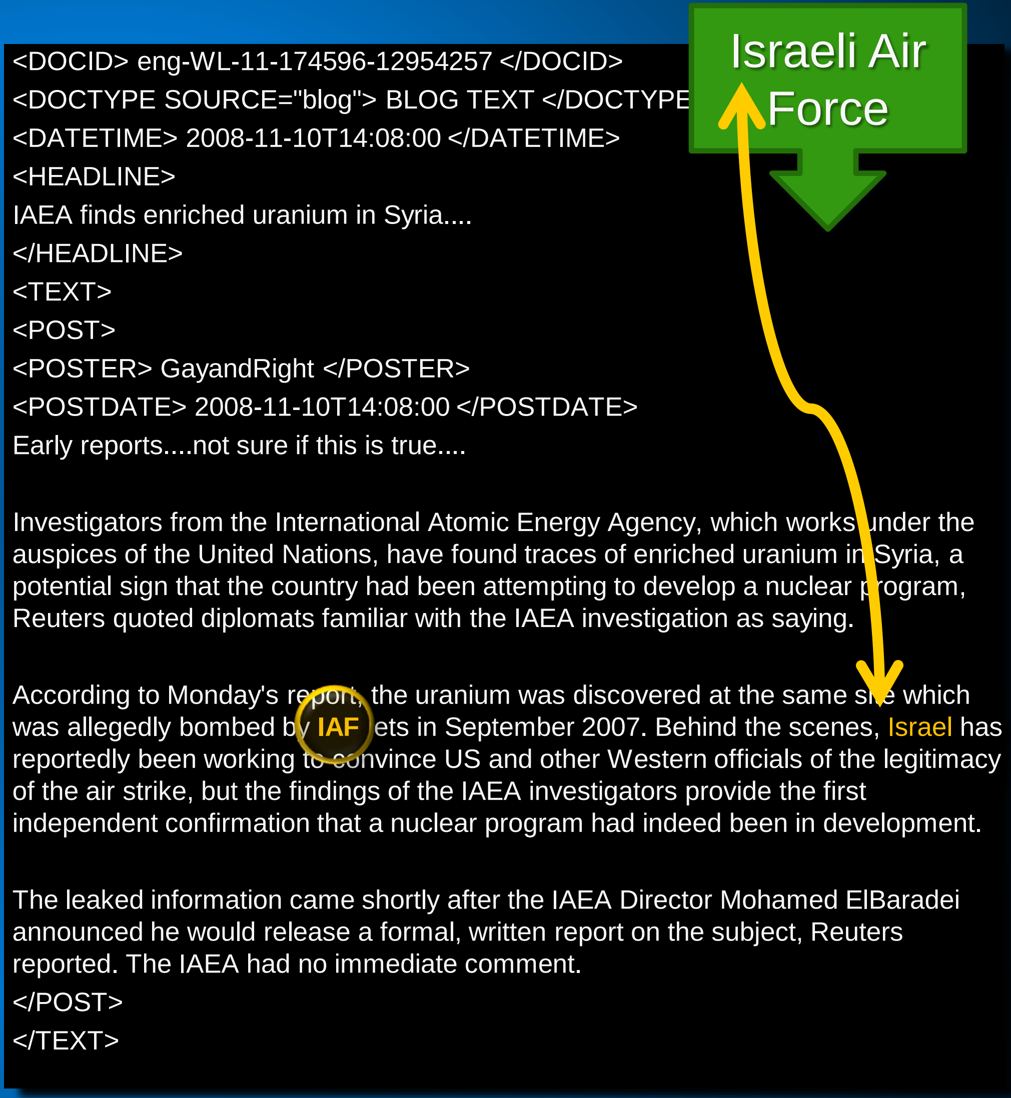
E0707328: Indian Armed Forces

…

NIL

Example: which "IAF"? *Is answering just a text lookup?*

```
<query id="EL005833">
   <name>IAF</name>
   <docid>eng-WL-11-174596-12954631</docid>
   <offset>…</offset>
</query>
<query id="EL005836">
   <name>IAF</name>
   <docid>eng-NG-31-142148-10021195</docid>
   <offset>…</offset>
</query>
<query id="EL05838">
   <name>IAF</name>
   <docid>eng-WL-11-174596-12954257</docid>
   <offset>…</offset>
</query>
<query id="EL05847">
   <name>IAF</name>
   <docid>eng-NG-31-147166-10475895</docid>
   <offset>…</offset>
</query>
```



```
<DOCID> eng-WL-11-174596-12954257 </DOCID>
<DOCTYPE SOURCE="blog"> BLOG TEXT </DOCTYPE
<DATETIME> 2008-11-10T14:08:00 </DATETIME>
<HEADLINE>
IAEA finds enriched uranium in Syria....
</HEADLINE>
<TEXT>
<POST>
<POSTER> GayandRight </POSTER>
<POSTDATE> 2008-11-10T14:08:00 </POSTDATE>
Early reports....not sure if this is true....

Investigators from the International Atomic Energy Agency, which works under the
auspices of the United Nations, have found traces of enriched uranium in Syria, a
potential sign that the country had been attempting to develop a nuclear program,
Reuters quoted diplomats familiar with the IAEA investigation as saying.

According to Monday's report, the uranium was discovered at the same site which
was allegedly bombed by IAF jets in September 2007. Behind the scenes, Israel has
reportedly been working to convince US and other Western officials of the legitimacy
of the air strike, but the findings of the IAEA investigators provide the first
independent confirmation that a nuclear program had indeed been in development.

The leaked information came shortly after the IAEA Director Mohamed ElBaradei
announced he would release a formal, written report on the subject, Reuters
reported. The IAEA had no immediate comment.
</POST>
</TEXT>
```

Example: which “IAF”? *Is answering just a text lookup?*

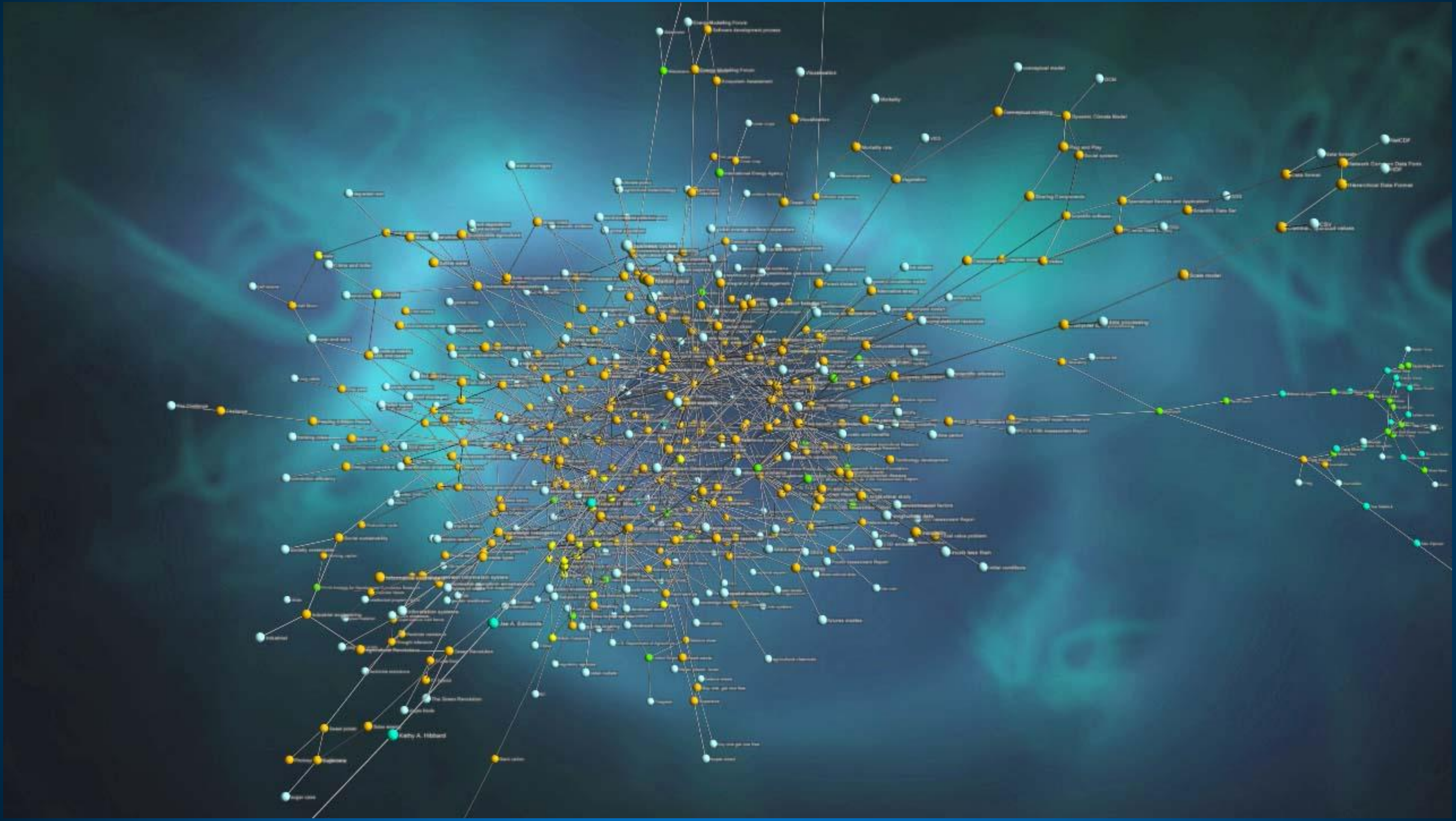
```
<query id="EL005833">
   <name>IAF</name>
   <docid>eng-WL-11-174596-12954631</docid>
   <offset>…</offset>
</query>
<query id="EL005836">
   <name>IAF</name>
   <docid>eng-NG-31-142148-10021195</docid>
   <offset>…</offset>
</query>
<query id="EL05838">
   <name>IAF</name>
   <docid>eng-WL-11-174596-12954257</docid>
   <offset>…</offset>
</query>
<query id="EL05847">
   <name>IAF</name>
   <docid>eng-NG-31-147166-10475895</docid>
   <offset>…</offset>
</query>
```

Israeli Air Force

```
<DOCID> eng-WL-11-174596-12954631 </DOCID>
<DOCTYPE SOURCE="blog"> BLOG TEXT </DOCTYPE>
<DATETIME> 2008-05-24T12:55:00 </DATETIME>
<HEADLINE> Syria stalls IAEA visit... </HEADLINE>
<TEXT> <POST>
<POSTER> GayandRight </POSTER>
<POSTDATE> 2008-05-24T12:55:00 </POSTDATE>
Gee, I wonder why....

Syria has not yet accepted a request by the International Atomic Energy Agency to
visit the site bombed by the IAF on September 6, which Washington says was a
nuclear reactor, Reuters reported Friday.

The news agency quoted diplomats in Vienna as saying that Damascus was stalling
its approval of the UN delegation visit, demanding more details on the proposed
inspection.

Syrian atomic energy chief Ibrahim Othman came to Vienna earlier this month to
speak with IAEA head Mohamed ElBaradei on the matter, but the two did not agree
on the timing or nature of a visit, diplomats said.

The agency received a letter from Syria several days ago asking for more details on
the trip, one US diplomat said. The IAEA has replied and is now waiting for
Damascus's response, he added.
</POST>
</TEXT>
```

~~Italian Air Force~~, ~~Italian Armed Forces~~, ~~Indonesian Air Force~~, Iraqi Air Force, Israeli Air Force, ~~Indian Armed Forces~~

The Entity Graph



Which "Washington"? *Is the answer unique and absolute?*

386 Wikipedia entities can be referred to as **Washington**

(based on the August 5, 2013 Wikipedia dump).

- Washington, D.C.
- United States
- United States Department of State
- Federal government of the United States

…

… which Syria?

… which Damascus?

The answer depends on:

- The granularity of the knowledge base
- The disambiguation of the other entities in the document

<DOCID> eng-WL-11-174596-12954631 </DOCID>

<DOCTYPE SOURCE="blog"> BLOG TEXT </DOCTYPE>

<DATETIME> 2008-05-24T12:55:00 </DATETIME>

<HEADLINE> Syria stalls IAEA visit... </HEADLINE>

<TEXT> <POST>

<POSTER> GayandRight </POSTER>

<POSTDATE> 2008-05-24T12:55:00 </POSTDATE>

Gee, I wonder why....

Syria has not yet accepted a request by the International Atomic Energy Agency to visit the site bombed by the **IAF** on September 6, which Washington says was a nuclear reactor, Reuters reported Friday.

The news agency quoted diplomats in Vienna as saying that Damascus was stalling its approval of the UN delegation visit, demanding more details on the proposed inspection.

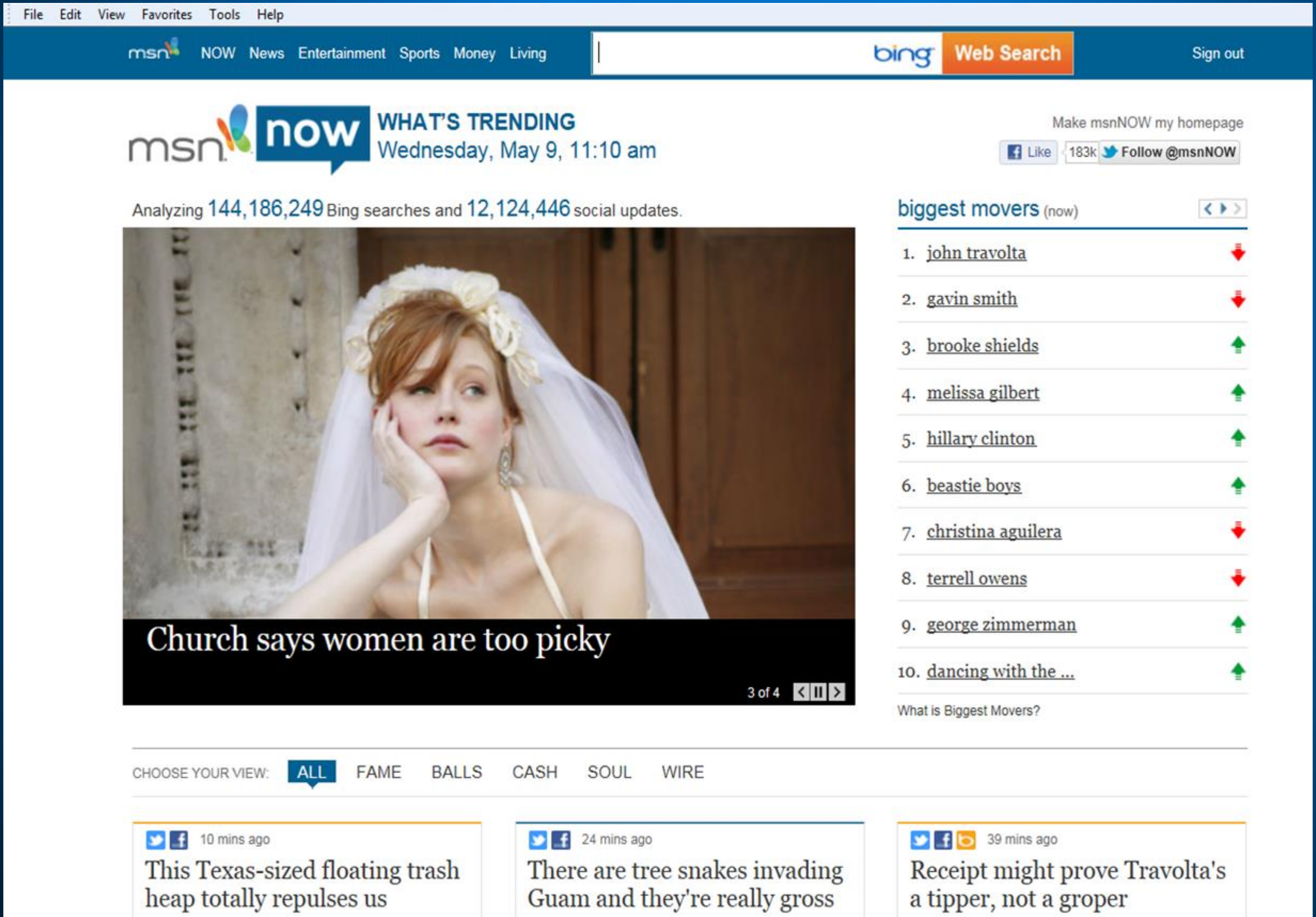
Syrian atomic energy chief Ibrahim Othman came to Vienna earlier this month to speak with IAEA head Mohamed ElBaradei on the matter, but the two did not agree on the timing or nature of a visit, diplomats said.

The agency received a letter from Syria several days ago asking for more details on the trip, one US diplomat said. The IAEA has replied and is now waiting for Damascus's response, he added.

</POST>

</TEXT>

The MSR System:
TAC-independent Foundation


File Edit View Favorites Tools Help
msn NOW News Entertainment Sports Money Living
bing Web Search
Sign out
msn now
Make msnNOW my homepage
Like 183k Follow @msnNOW
3 of 4
CHOOSE YOUR VIEW: ALL FAME BALLS CASH SOUL WIRE

WHAT'S TRENDING
Wednesday, May 9, 11:10 am

Analyzing 144,186,249 Bing searches and 12,124,446 social updates.

Church says women are too picky

biggest movers (now)

1. john travolta ↓
2. gavin smith ↓
3. brooke shields ↑
4. melissa gilbert ↑
5. hillary clinton ↑
6. beastie boys ↑
7. christina aguilera ↓
8. terrell owens ↓
9. george zimmerman ↑
10. dancing with the ... ↑

What is Biggest Movers?

10 mins ago
This Texas-sized floating trash heap totally repulses us

24 mins ago
There are tree snakes invading Guam and they're really gross

39 mins ago
Receipt might prove Travolta's a tipper, not a groper

The MSR System – Framework (1)

- The best evidence for entity disambiguation is the set of co-occurring entities (rather than the plain text)

↓

- Extract and disambiguate all entities in a target document

↓

- Match the target string against the surface forms extracted from the document

The MSR System – Framework (2)

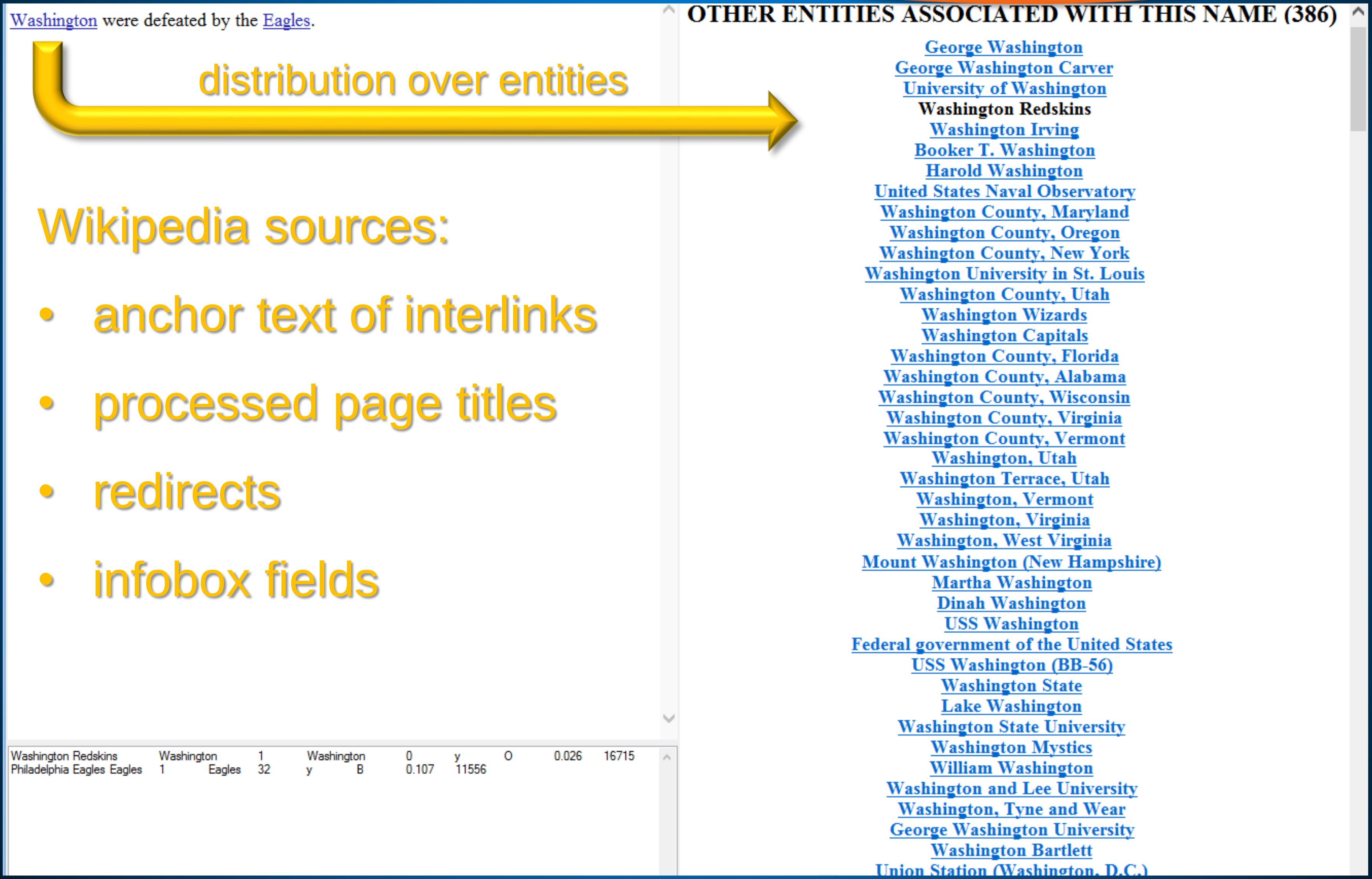
- Employ the most-recent Wikipedia collection:
  Wikipedia collection dump from August 5, 2013
- Build the knowledge base for the system

TAC-independent

- For each TAC query, process the target document, output the entity that corresponds to the target string
- Map the entity to the TAC 2008 entity collection; do not do anything else for clustering

Knowledge Base

Surface Forms



Washington were defeated by the Eagles.

distribution over entities

Wikipedia sources:

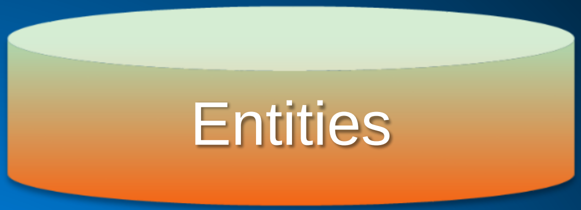
- anchor text of interlinks
- processed page titles
- redirects
- infobox fields

```
Washington Redskins     Washington      1       Washington      0       y       O       0.026   16715
Philadelphia Eagles Eagles      1       Eagles  32      y       B       0.107   11556
```

OTHER ENTITIES ASSOCIATED WITH THIS NAME (386)

George Washington
George Washington Carver
University of Washington
Washington Redskins
Washington Irving
Booker T. Washington
Harold Washington
United States Naval Observatory
Washington County, Maryland
Washington County, Oregon
Washington County, New York
Washington University in St. Louis
Washington County, Utah
Washington Wizards
Washington Capitals
Washington County, Florida
Washington County, Alabama
Washington County, Wisconsin
Washington County, Virginia
Washington County, Vermont
Washington, Utah
Washington Terrace, Utah
Washington, Vermont
Washington, Virginia
Washington, West Virginia
Mount Washington (New Hampshire)
Martha Washington
Dinah Washington
USS Washington
Federal government of the United States
USS Washington (BB-56)
Washington State
Lake Washington
Washington State University
Washington Mystics
William Washington
Washington and Lee University
Washington, Tyne and Wear
George Washington University
Washington Bartlett
Union Station (Washington, D.C.)

Knowledge Base



Information associated with each entity:

- Topics

  Wikipedia categories, list pages, interlinked entity mentions in enumerations, interlinked entity mentions in tables, …

- Contexts

  parentheticals in titles, infobox information, …

- Triggers

  Wikipedia bidirectional linkage

- Entry/Entity Types

  14 types: Disambiguation, Common, Person, Geo-political entity, Location, Organization, Event, Vehicle, Work of art, …, Other

- Geo-coordinates

Knowledge Base

Linguistic Resources

Derived from the Wikipedia collection:

- name normalization

e.g.:

Ben	Benjamin, Benjy, Benedict
Bernard	Bernie, Barney, Bernardus, Bernhard
Betty	Elizabeth, Beth, Betsy
Bill	William, Billy, Bil

- entity-type contexts

e.g.:

result in →	B 216	C 456	D 4	L 8	M 4	O 2	P 14	T 6
← was founded.	B 18	G 2	L 42	M 2	O 198	P 8		

- word-capitalization statistics

…

System Architecture

The analysis of an input text is done in three stages, with the following main roles:

- Stage 1
  - text normalization
  - sentence breaking
- Stage 2:
  - surface form boundary detection
- Stage 3: disambiguation
  - latent document model construction
  - feature computation
  - entity candidate ranking

Surface Form Detection

- Capitalization, lexical resources, known surface forms
- Soft boundaries

TAC 2011 test: query: EL_00279
string: Bordeaux
doc: AFP_ENG_20081006.0534.LDC2009T13.sgm

e.g.:

Bordeaux-based wine merchant, Jeffrey Davies, said that while the crisis triggered by the terror attacks on New York and Washington had hit US wine sales, the economic meltdown had global implications. […]

"The big spenders that were ordering the top wines in top restaurants have been taken out," Davies said.

After the attacks, sales of Bordeaux wine to the United States fell by 29 percent in volume during the final quarter of 2001 -- the key Thanksgiving, Christmas and New Year period, which accounts for half of annual sales.

Composite(*Bordeaux*, *Bordeaux wine*)

Composite(*US*, *US wine*)

Employ composite surface forms and let the disambiguation process determine the best entity in the context → resolve boundaries afterwards

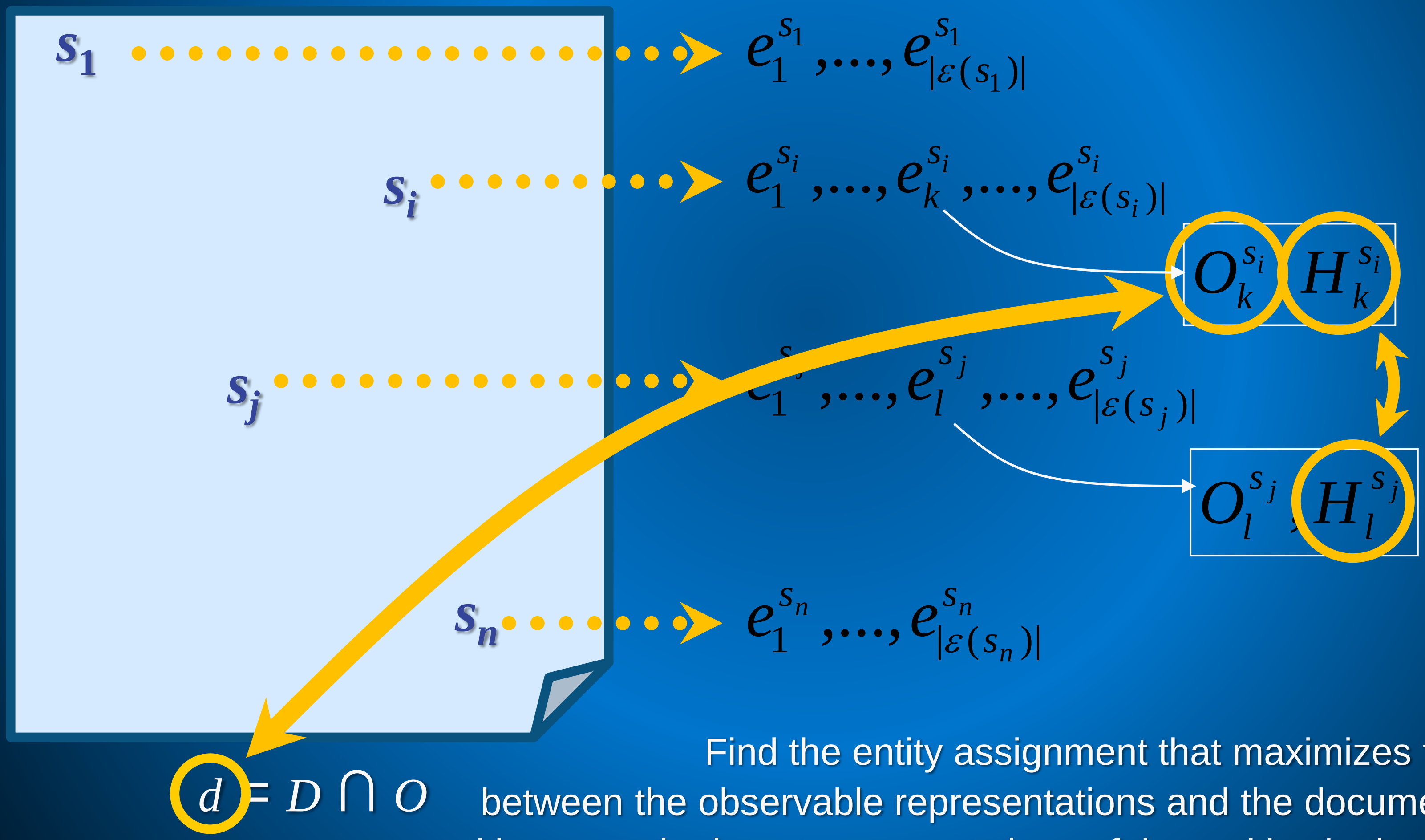
Disambiguation - Intuition

Each entity has multiple vectorial representations

observable

latent / hidden

Text document D



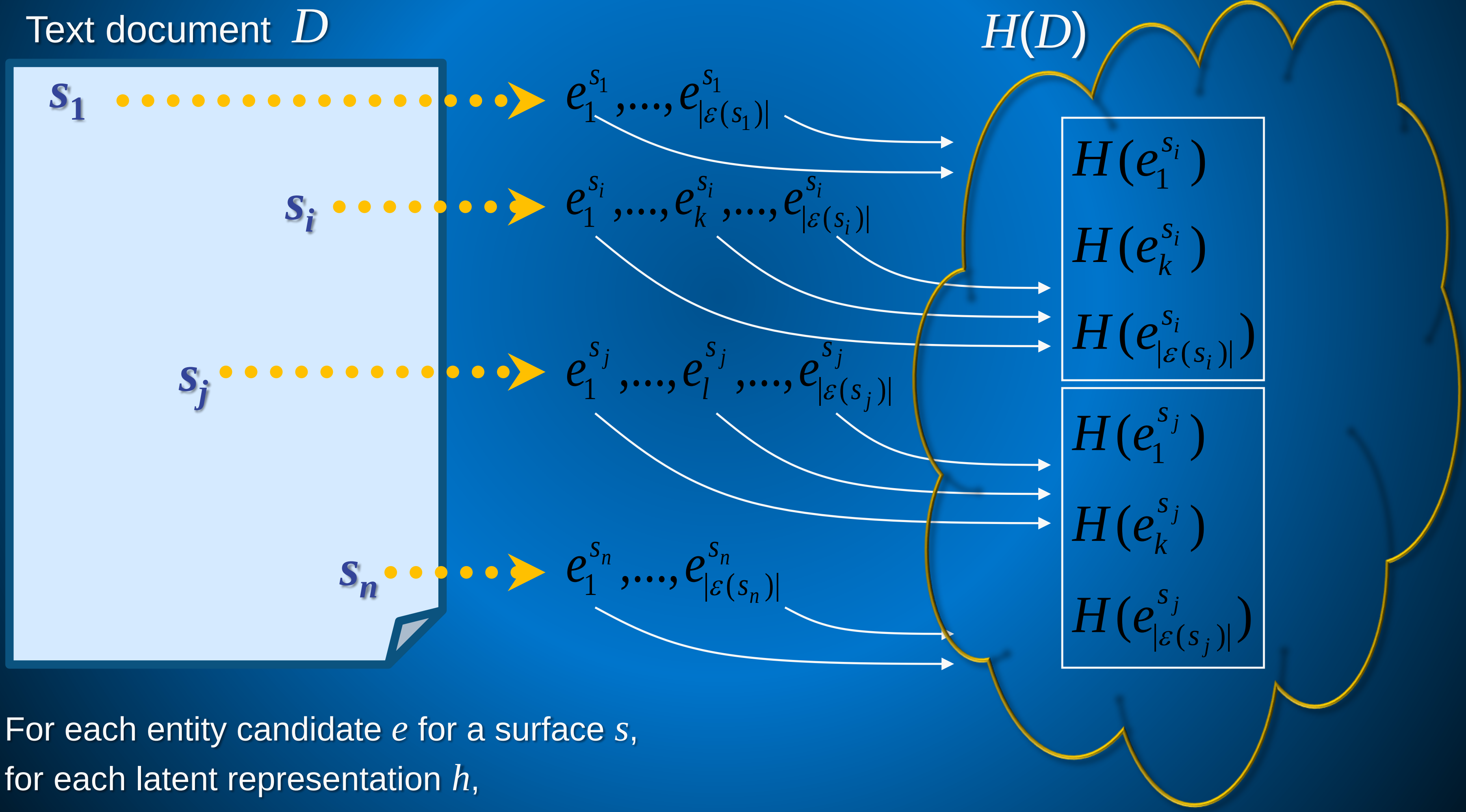
s_1
$e_1^{s_1}, \ldots, e_{|\varepsilon(s_1)|}^{s_1}$
s_i
$e_1^{s_i}, \ldots, e_k^{s_i}, \ldots, e_{|\varepsilon(s_i)|}^{s_i}$
$O_k^{s_i}$ $H_k^{s_i}$
s_j
$e_1^{s_j}, \ldots, e_l^{s_j}, \ldots, e_{|\varepsilon(s_j)|}^{s_j}$
$O_l^{s_j}$, $H_l^{s_j}$
s_n
$e_1^{s_n}, \ldots, e_{|\varepsilon(s_n)|}^{s_n}$
$d = D \cap O$

Find the entity assignment that maximizes the similarity between the observable representations and the document context d and between the latent representations of the entities in the assignment.

S. Cucerzan. Large-scale Entity Disambiguation based on Wikipedia Data. EMNLP 2007

The Latent Document Representation

Build (noisy) latent vectorial representations of the document by aggregating the latent vectors from all entity candidates.



Text document D

s_1

s_i

s_j

s_n

$e_1^{s_1}, \ldots, e_{|\varepsilon(s_1)|}^{s_1}$

$e_1^{s_i}, \ldots, e_k^{s_i}, \ldots, e_{|\varepsilon(s_i)|}^{s_i}$

$e_1^{s_j}, \ldots, e_l^{s_j}, \ldots, e_{|\varepsilon(s_j)|}^{s_j}$

$e_1^{s_n}, \ldots, e_{|\varepsilon(s_n)|}^{s_n}$

$H(D)$

$H(e_1^{s_i})$

$H(e_k^{s_i})$

$H(e_{|\varepsilon(s_i)|}^{s_i})$

$H(e_1^{s_j})$

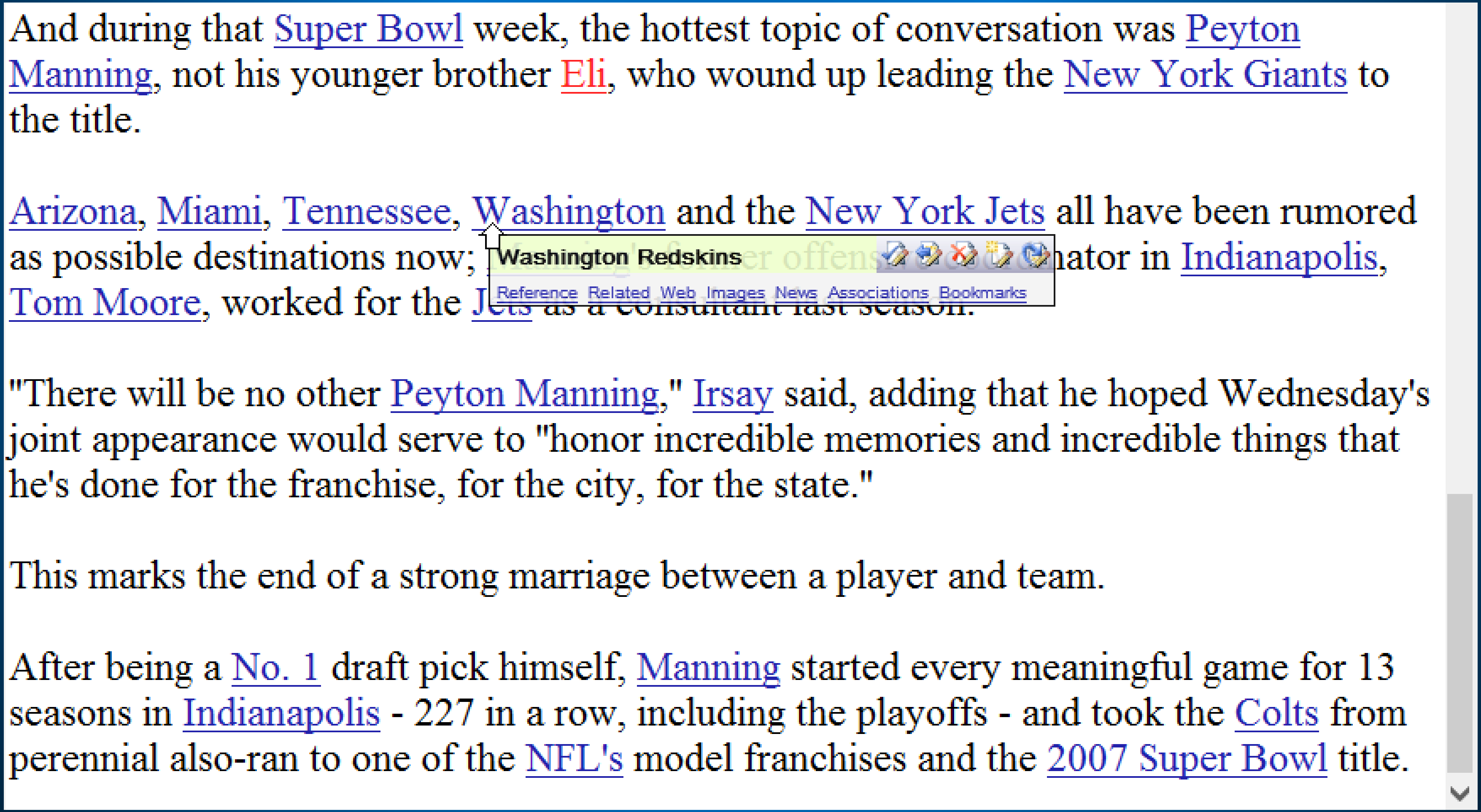
$H(e_k^{s_j})$

$H(e_{|\varepsilon(s_j)|}^{s_j})$

For each entity candidate e for a surface s,
for each latent representation h,
compute the similarity between $h(e)$ and $h(D) – h(s)$

Latent Features Example

demo



And during that Super Bowl week, the hottest topic of conversation was Peyton Manning, not his younger brother Eli, who wound up leading the New York Giants to the title.

Arizona, Miami, Tennessee, Washington and the New York Jets all have been rumored as possible destinations now; Manning's former offensive coordinator in Indianapolis, Tom Moore, worked for the Jets as a consultant last season.

"There will be no other Peyton Manning," Irsay said, adding that he hoped Wednesday's joint appearance would serve to "honor incredible memories and incredible things that he's done for the franchise, for the city, for the state."

This marks the end of a strong marriage between a player and team.

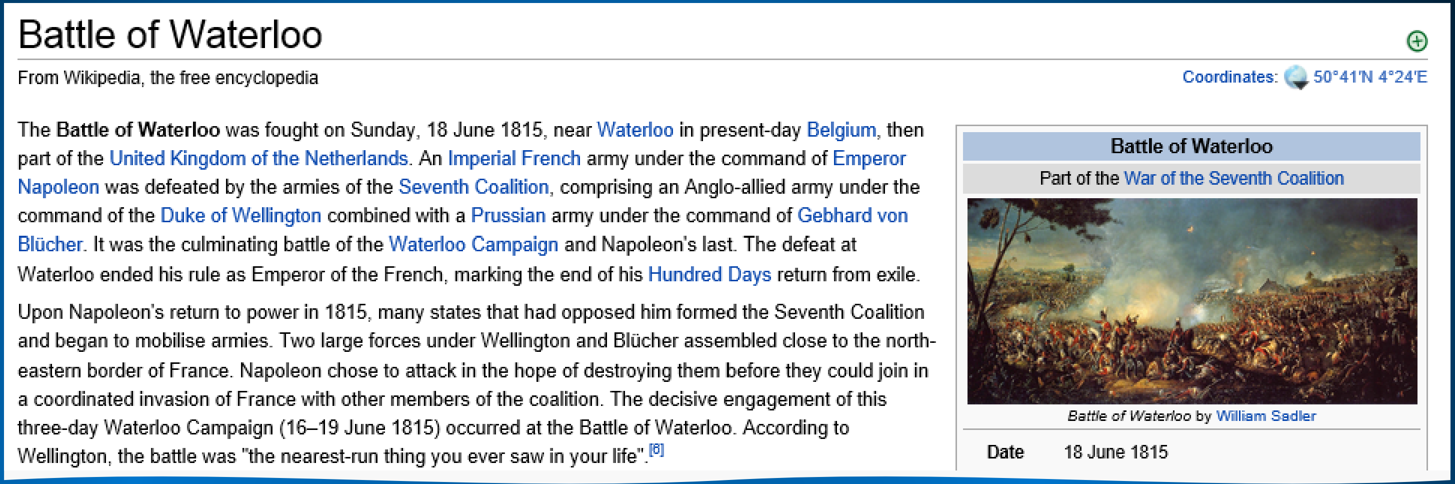
After being a No. 1 draft pick himself, Manning started every meaningful game for 13 seasons in Indianapolis - 227 in a row, including the playoffs - and took the Colts from perennial also-ran to one of the NFL's model franchises and the 2007 Super Bowl title.

Local Features

- Depart from the one-sense-per-discourse paradigm
- Employ a latent paragraph model
  in addition to the latent document model
- Employ lexico-syntactic patterns to weight the latent contributions to the paragraph model

  e.g.: possessive constructions: “*Sweden’s Prime Minister*”
  conjunctive constructions: “*Sweden and Romania*”

Training



Battle of Waterloo

From Wikipedia, the free encyclopedia

Coordinates: 50°41′N 4°24′E

The **Battle of Waterloo** was fought on Sunday, 18 June 1815, near Waterloo in present-day Belgium, then part of the United Kingdom of the Netherlands. An Imperial French army under the command of Emperor Napoleon was defeated by the armies of the Seventh Coalition, comprising an Anglo-allied army under the command of the Duke of Wellington combined with a Prussian army under the command of Gebhard von Blücher. It was the culminating battle of the Waterloo Campaign and Napoleon's last. The defeat at Waterloo ended his rule as Emperor of the French, marking the end of his Hundred Days return from exile.

Upon Napoleon's return to power in 1815, many states that had opposed him formed the Seventh Coalition and began to mobilise armies. Two large forces under Wellington and Blücher assembled close to the north-eastern border of France. Napoleon chose to attack in the hope of destroying them before they could join in a coordinated invasion of France with other members of the coalition. The decisive engagement of this three-day Waterloo Campaign (16–19 June 1815) occurred at the Battle of Waterloo. According to Wellington, the battle was "the nearest-run thing you ever saw in your life".[6]

Battle of Waterloo	
Part of the War of the Seventh Coalition	
Battle of Waterloo by William Sadler	
Date	18 June 1815

This text is about [[Battle of Waterloo|Waterloo]]. Allegedly, Napoleon tried to escape to North America, but the [[Royal Navy|Royal Navy]] was blockading French ports to forestall such a move. He finally surrendered to [[Captain (Royal Navy)|Captain]] [[Frederick Lewis Maitland (Royal Navy officer)|Frederick Maitland]] of [[Her Majesty's Ship|HMS]] "[[HMS Bellerophon (1786)|Bellerophon]]" on 15 July. There was a campaign against French fortresses that still held out; [[Longwy|Longwy]] capitulated on 13 September 1815, the last to do so. The [[Treaty of Paris (1815)|Treaty of Paris]] was signed on 20 November 1815. [[Louis XVIII of France|Louis XVIII]] was restored to the throne of France, and Napoleon was exiled to [[Saint Helena|Saint Helena]], where he died in 1821.

Training (2)

ORIGINAL TRAINING TEXT:

This text is about [[Battle of Waterloo|Waterloo]]. Allegedly, Napoleon tried to escape to North America, but the [[Royal Navy|Royal Navy]] was blockading French ports to forestall such a move. He finally surrendered to [[Captain (Royal Navy)|Captain]] [[Frederick Lewis Maitland (Royal Navy officer)|Frederick Maitland]] of [[Her Majesty's Ship|HMS]] ''[[HMS Bellerophon (1786)|Bellerophon]]'' on 15 July. There was a campaign against French fortresses that still held out; [[Longwy|Longwy]] capitulated on 13 September 1815, the last to do so. The [[Treaty of Paris (1815)|Treaty of Paris]] was signed on 20 November 1815. [[Louis XVIII of France|Louis XVIII]] was restored to the throne of France, and Napoleon was exiled to [[Saint Helena|Saint Helena]], where he died in 1821.

THE ANALYSIS OF THE TRAINING TEXT:

This text is about Waterloo. Allegedly, Napoleon tried to escape to North America, but the Royal Navy was blockading French ports to forestall such a move. He finally surrendered to Captain Frederick Maitland of HMS "Bellerophon" on 15 July. There was a campaign against French fortresses that still held out; Longwy capitulated on 13 September 1815, the last to do so. The Treaty of Paris was signed on 20 November 1815. Louis XVIII was restored to the throne of France, and Napoleon was exiled to Saint Helena, where he died in 1821.

Treaty of Paris (and 7 other training examples)

Training (3)

[[Treaty of Paris (1815)|Treaty of Paris]]

1. Treaty of Paris
2. Treaty of Paris (1763)
3. Treaty of Paris (1783)
4. Treaty of Paris (1814)
5. Treaty of Paris (1815)
6. Treaty of Paris (1898)
7. Treaty of Paris (1920)
8. Treaty of Paris (1951)
9. Treaty of Paris (1856)
10. Paris Peace Treaties, 1947
11. Treaty of Paris (1810)
12. Paris Peace Conference, 1919
13. Treaty of Peace with Italy, 1947
14. Paris Peace Accords
15. Treaty of Paris (1259)
16. Pacte de Famille
17. Treaty of Paris (1229)
18. Treaty of Paris (1303)
19. Treaty of Paris (1323)
20. Treaty of Paris (1355)
21. Treaty of Paris (1623)
22. Treaty of Paris (1657)
23. Treaty of Paris (1900)
24. Bonn–Paris conventions
25. Treaty of Paris (band)
26. Treaty of Paris (1796)
27. Peace of Paris (1783)
28. Treaty of Paris (1626)
29. Treaty of Paris (1857)
30. Treaty of Paris (1802)

← Features values →

y_1	y_2	y_3	y_4	…	$y_{n\text{-}1}$	y_n
x_1	x_2	x_3	x_4	…	$x_{n\text{-}1}$	x_n

R1. LINEAR CLASSIFIER (LOGISTIC REGRESSION)

- Trained on the pairs (truth, other_candidate):

$$\frac{P(1|X)}{P(1|Y)} > 1 \qquad \rightarrow \qquad \sum_{k=1}^{n} \beta_k (xk - yk)$$

R2. BOOSTED-TREES RANKER

- Trained for NDCG@3 on the pairs (truth, other_candidate)

Training (4)

R1. LINEAR CLASSIFIER (LOGISTIC REGRESSION)

2 million data points

R2. BOOSTED-TREES RANKER

4 million data points

Average ambiguity: 41

TAC data not used for training

~~- mapping between 2008 and 2013 collections~~

~~- handling the NIL labels~~

NIL Clustering …by Knowing More

- Target strings mapped to a much larger knowledge base

	2013	2008
e.g.: Appleton	→ Appleton, Wisconsin	→ E0790618
	→ Appleton, New York	→ NILxxx1
Mandeville	→ Mandeville, Louisiana	→ NILxxx2
	→ Mandeville, Jamaica	→ NILxxx3

- Inside-document coreference

e.g.: Harpootlian	→ Dick Harpootlian	→ NILxxx4
ADF	→ Alliance Defense Fund	→ NILxxx5
	→ Australian Defence Force	→ NILxxx6

Is Knowing More Always an Advantage?

- Mapping back to 2008 is not trivial

more than 95,000 out of 820,000 Wikipedia 2008 pages in the knowledge base changed their title as of 2013

- More comprehensive may lead to NIL answers

e.g.: Birmingham

TAC 2011: EL_00256

NILxxxx ← Birmingham campaign

The army moved to Albany, Ga., in 1961. Some observers say Albany was a failure for Dr. King, but others say it played an important part in preparing the movement for Birmingham.

TAC 2011: EL_00258

Birmingham, Alabama → E0609361

A map of hate groups from the Southern Poverty Law Center in Birmingham, Ala., shows there are 33 active white supremacist groups that have formed in Pennsylvania.

Gold standard for both queries is E0609361: Birmingham, Alabama

TAC Evaluation

Accuracy	Systems corresponding to the submitted MSR runs		Best Result TAC Evaluation
	Run 1	Run 2	
TAC 2011 test set	89.3 %	89.9 %	86.8% (MSR)
TAC 2012 test set	80.4 %	79.3 %	76.2% (MSR1)

TAC 2013 test set	**Run 1**	**Run 2**
B^3+ F1 (Overall -- 2190 queries)	0.720	**0.721**
B^3+ F1 (in KB -- 1090 queries)	0.718	**0.724**
B^3+ F1 (not in KB -- 1100 queries)	**0.720**	0.716
B^3+ F1 (NW docs -- 1134 queries)	0.795	**0.801**
B^3+ F1 (WB docs -- 343 queries)	**0.673**	0.666
B^3+ F1 (DF docs -- 713 queries)	0.623	0.618
B^3+ F1 (PER -- 686 queries)	**0.758**	**0.758**
B^3+ F1 (ORG -- 701 queries)	**0.737**	0.716
B^3+ F1 (GPE -- 803 queries)	0.672	0.693

* numbers corresponding to the best performance in the TAC 2013 evaluation are in bold

81% accuracy