Entity Extraction and Linking with Applications to Web Search

Silviu Cucerzan
Machine Learning and Intelligence
MSR
Game changer: Wikipedia + hardware advancements
   small number of entity classes → large entity collection → Watson wins jeopardy

Current state:
   arguably, current entity extractions systems outperform non-expert humans on entity disambiguation for random documents/domains.

Next big steps:
   • move from one large entity collection to many thousands of domain specific collections, business collections, and personal collections
   • learn to link to Web entities
# NIST/LDC Entity Linking Evaluations

Knowledge base derived from Wikipedia Oct. 2008

## Text corpus:
(for the 2013 evaluation)

- Newswire: 1 mil docs
- Web Text: 1 mil docs
- Discussion Forums: .5 mil docs
- 2.5 million documents

## Annotated Data:

<table>
<thead>
<tr>
<th>Year</th>
<th>Type</th>
<th>Total mentions</th>
<th>PER</th>
<th>ORG</th>
<th>GPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009</td>
<td>Evaluation</td>
<td>3904</td>
<td>627</td>
<td>2710</td>
<td>567</td>
</tr>
<tr>
<td></td>
<td>Training</td>
<td>1500</td>
<td>500</td>
<td>500</td>
<td>500</td>
</tr>
<tr>
<td></td>
<td>Evaluation</td>
<td>1500</td>
<td>500</td>
<td>500</td>
<td>500</td>
</tr>
<tr>
<td></td>
<td></td>
<td>750</td>
<td>250</td>
<td>250</td>
<td>250</td>
</tr>
<tr>
<td>2011</td>
<td>Evaluation</td>
<td>1491</td>
<td>500</td>
<td>491</td>
<td>500</td>
</tr>
<tr>
<td></td>
<td></td>
<td>759</td>
<td>250</td>
<td>259</td>
<td>250</td>
</tr>
<tr>
<td>2012</td>
<td>Evaluation</td>
<td>1471</td>
<td>702</td>
<td>388</td>
<td>381</td>
</tr>
<tr>
<td></td>
<td></td>
<td>755</td>
<td>216</td>
<td>318</td>
<td>221</td>
</tr>
<tr>
<td>2013</td>
<td>Evaluation</td>
<td>1477</td>
<td>686</td>
<td>701</td>
<td>803</td>
</tr>
<tr>
<td></td>
<td></td>
<td>713</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>Evaluation</td>
<td>3318</td>
<td>3162</td>
<td>1007</td>
<td>1065</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1916</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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.

NIST/LDC EL - Task Description

LDC Knowledge Base

- 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
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.
Knowledge Graph
Syria stalls IAEA visit... 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.
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.
- Employ document-based features as well as local features.
Knowledge Base

Wikipedia sources:
- anchor text of interlinks
- processed page titles
- redirects
- infobox fields

distribution over entities

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, frequently co-occurring words

- **Triggers**
  Wikipedia bidirectional linkage

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

- **Geo-coordinates**

- **Time information**
Knowledge Base – Topics Stats (2)
Knowledge Base

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**
  - ...
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

**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
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.

For each entity candidate $e$ for a surface $s$, for each latent representation $h$, compute the similarity between $h(e)$ and $h(D) - h(s)$
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; Washington Redskins 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: “Denmark’s Prime Minister visited Romania”
  - conjunctive constructions: “Denmark and Romania”
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.
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 (3)

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)
11. Treaty of Paris (1810)
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)

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

\[ P(1|\hat{y}) > 1 \sim \frac{P(\hat{y}|x)}{P(y|x)} > 1 \sim \sum_{k=1}^{n} \beta_k (\varphi_k(x, \hat{y}) - \varphi_k(x, y)) \]

- Current system trained on ~90 million data points
Which Knowledge Base? Knowing More Is Not Always an Advantage

- Mapping between entity collections is not trivial
  - Almost 100,000 out of 820,000 Wikipedia 2008 pages in the knowledge base changed their title as of 2014

- More comprehensive may lead to NIL answers
  - e.g.: Birmingham


TAC 2011: EL_00256
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
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
## Results

### NIST/LDC Evaluations

<table>
<thead>
<tr>
<th>Accuracy:</th>
<th>MSR logistic regression</th>
<th>best result in the TAC Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>2013 system</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TAC 2011 test set</td>
<td>89.3 %</td>
<td>86.8% (MSR)</td>
</tr>
<tr>
<td>TAC 2012 test set</td>
<td>80.4 %</td>
<td>76.2% (MSR)</td>
</tr>
<tr>
<td>TAC 2013 test set</td>
<td>83.2 %</td>
<td>83.2% (MSR)</td>
</tr>
<tr>
<td><strong>2014 system</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TAC 2014 train set</td>
<td>87.2 %</td>
<td></td>
</tr>
<tr>
<td>TAC 2014 test set</td>
<td>86.8 %</td>
<td>86.8% (MSR)</td>
</tr>
</tbody>
</table>

### Google-Microsoft-Yahoo Challenge

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERD 2014 Train set</td>
<td>83.7%</td>
<td>72.6%</td>
<td>0.778</td>
</tr>
<tr>
<td>ERD 2014 Test set</td>
<td>83.3%</td>
<td>69.9%</td>
<td>0.760</td>
</tr>
</tbody>
</table>
Entity Extraction and Linking Applications
Applications: Video Annotation
Applications: Trending Topics

PEYTON MANNING

See all social results for this topic

RECENT PUBLIC UPDATES

CloudFantasy: Peyton Manning will be officially cut by the Indianapolis Colts tomorrow. Now everyone is talking about where ... http://www.cloudfantasy.com/201... Facebook · 1 minute ago

WTHi-TV: Peyton Manning may leave Colts | WTHiTV.com
http://www.wthitv.com/dp... Facebook · 1 hour ago

2DaySports.com: Report: Peyton Manning is Leaving the Colts (read)
http://www.2daysports.com/pay... Facebook · 34 minutes ago

Yahoo Search: Say it ain’t so! Peyton Manning won’t be playing for the Colts this season. Where do you think he’ll end up? http://search.yahoo.com/se... Facebook · 53 minutes ago

ESPN 1420: It’s official; the Colts will release Peyton Manning.
http://espn1420.com/pey... Facebook · 49 minutes ago

wrapupp.com: Colts to Cut Peyton Manning | wrapupp
http://www.wrapupp.com/nfl... Facebook · 51 minutes ago

100.9 The Eagle: Peyton Manning, Colts To Split? Star QB Reportedly Done in Indianapolis
http://www.1009theeagle.com/201... Facebook · 50 minutes ago

Fanatic Sports and Cards: Colts to release Peyton Manning | Fanatic Sports and Cards
http://www.fanaticsportsandcards.com/cot... Facebook · 51 minutes ago

WITHIN THIS TOPIC

Keywords: Indianapolis Colts, ESPN, Chris Mortensen, Andrew Luck, ESPNews

See more social results for Peyton Manning
Applications: Facebook Collage

Emre Kiciman, Chun-Kai Wang

http://apps.facebook.com/msrcollage/
Applications: Search Ads Matching

Normalization of keyphrases and queries:

- e.g. big lots sales → big lots, sale
- on sale at big lots → big lots, sale
- big sale of lots → big, lot, sale

Broad matching:

- replace entities in a query with related entities

- e.g. kia sorento mpg → hyundai santa fe mpg
Applications: Query Suggestion

e.g.

John Travolta flics → John Travolta movies
John Travolta filmography
Robert De Niro flics
Robert De Niro movies
John Wayne flics

Bank of America credit card →
Bank of America Visa
Wells Fargo credit card
Bank of Brazil credit card
Bank of America credit valentine
Applications: Spelling Correction

selin dion color of love → celine dion colour of my love

gensis selin → genesis selling england by the pound

sher biliv
cher believe

sher cofetion
usher confessions
Applications: Web Factoid Question Answering

Who discovered America - Answers.com
www.answers.com › ... › Continents › North America
Neither. Lief Ericson discovered North America second while Christopher Columbus discovered it third. A Viking did discover North America and his name was Bjarne ...

Who Discovered America? - LiveScience.com
www.livescience.com/32430-who-discovered-america.html
We all learned that Columbus sailed the ocean blue ... and then American history gets very murky. Though much happened before Christopher Columbus' famous journey ...

Columbus May Not Have Been First to America: Discovery News
news.discovery.com/.../columbus-cabot-new-world-discovery-120503.htm
Oct 14, 2013 · Columbus may not have been the first to America. Learn more about Columbus not being the first to America in this article.

Who discovered America ?? - Yahoo Answers
https://answers.yahoo.com/question/index?qid=20080818174220AA4VyrZ
Resolved · 22 posts · 21 total answers · Published Aug 18, 2008
Aug 18, 2008 · Columbus in 1492 discovered some of the Caribbean Islands such as Hispaniola. There is John Cabot who discovered the east coast of America. Vikings ...

Who Discovered America? - Buzzle
www.buzzle.com/articles/who-discovered-america.html
America, also known as the New World is claimed to have been discovered by many people at many times. However, one can't say for sure who discovered America first.

Who Discovered America Before Columbus? - Yukozimo
discovery.yukozimo.com/who-discovered-america-before-columbus
The map above shows the Viking's journey as they went to discover North America about 500 years before Columbus found it. In 985 AD, Erik the Red reached Greenland...
Applications: Context-aware Search

(CBS/AP) Anyone who's ever tried to fold a well-used map can probably relate to the challenge Discovery astronauts face in the next phase of their mission.

After a successful first spacewalk, the crew aboard the space shuttle was set to spend much of Wednesday preparing to rewire the international space station, one of their most important tasks.

One of the station's solar panels needs to be retracted to make room for another panel that will begin rotating for the first time after the rewiring takes place. The panels follow the sun's movement to maximize the amount of power they can generate for the orbiting space lab.

But retracting a solar array isn't easy. It can lose its original fitness, just like a once-crisply strung t-shirt. NASA managers are wary about putting too much pressure on the array while trying to get it back in its box.

"A map never goes back to the way you bought it. It just doesn't," said John Curry, NASA's lead flight director for the space station.

Though NASA has developed contingency plans for various scenarios, managers aren't sure how the process will play out.

The rewiring is a vital step in the growth of the space station. Reconfiguring the power system will give the station the capability to provide electricity to laboratories that will be added to the structure in upcoming years.

Meanwhile, NASA says the heat shield appears to be fine, reports CBS News correspondent Peter King, and while there are still more pictures and more data to look at, NASA says it won't ask the crew to do a more focused inspection, because there's no need. However, there are still plans for one more detailed inspection after the shuttle leaves the space station.

NASA managers said dings on its left wing and belly were not of concern. The
Communicating Concepts as Context

How Important Is the Browsing Context?

Case study: news browsing

URL of news site browsed page

Query submitted to Search Engine (Bing, Google, Yahoo)

- 53.9% relevant
- 33.0% irrelevant
- 5.8% cannot say navigational for news
- 7.3%

M. Rahurkar and S. Cucerzan. Predicting when Browsing Context Is Relevant to Search. SIGIR 2008
Applications: Topic Pages

- William Alan Shatner (born on March 22, 1931) is a Canadian actor who gained fame for playing Captain James Tiberius Kirk, captain of the starship USS Enterprise in the television show Star Trek from 1966 to 1969 and in seven of the subsequent movies.

- "Shatner is the epitome of the post-ironic, 21st-century television and pop-culture icon," says Robert Thompson, a Syracuse University professor.

- Decades after his much-maligned album, "The Transformed Man," boldly took "Lucy in the Sky With Diamonds" into strange new worlds, Shatner performed songs from Has Been before 4,000 cheering fans.

- Actor William Shatner has sold his kidney stone for $25,000, with the money going to a housing charity, it was announced Tuesday.

- "Up Till Now" by William Shatner allegedly is riddled with discrepancies about the fateful night of August 9, 1999 when Shatner found Nerine in their pool at their studio city home.

- At age 73, Shatner reinvented himself yet again with a recurring role as a nutty attorney on the last season of The Practice, which snagged him his first Emmy in 2004 and another in 2005 for playing the same part on the spin-off Boston Legal.

- Shatner’s Star Trek sidekick Leonard Nimoy will be the only original cast member on board for the prequel—he’ll reprise his role as Mr. Spock in portions of the film.

- In 1999, Shatner suffered public personal tragedy when his third wife, Nerine, accidentally drowned in their swimming pool.

- In the latest Priceline ads, Shatner bursts forth as the Priceline Negotiator, a mashup of James Bond and Ron Popeil who will do anything to help people broker better deals.

- William Shatner has signed up to host a celebrity-interview show on the Biography Channel titled "Shatner’s Raw Nerve" which will premiere sometime next year, reports Daily Variety.

Aspect Models Derived from Query Logs

Topic: Michael Phelps
Aspects: sport, girlfriend, ...

Topic: George Clooney
Aspects: movies, pet, pictures, ...

Topic: Oprah Winfrey
Aspects: show, tv, website, ...

Distribution of topic only queries and topic + aspect queries in a query log sample.
Sentence Selection by Exploiting Web Redundancy

Sentence Selection Goals

| Grammatical | Classify sentences using lexical and language modeling features |
| Relevant    | Aspect specific context vectors.                                 |
| Diverse     | Explicitly maintain covered aspects. Iteratively select sentences that cover new aspects. |

Sentence Selection Methods

- **Full Aspect (Full-context):**
  \[ f(s) = \cos \text{ine}(A, S) \]

- **Novelty (NS-Full Context):**
  \[
  \text{until } |I| < n \text{ do } \\
  (1) n(s) = \lambda \cdot f(s) + (1 - \lambda) \cdot \min_{s_i \in S} (1 - \cos \text{ine}(s, s_i)) \\
  (2) I = I \cup \arg \max_{s \in S} n(s)
  \]

- **Typical:**
  \[
  \text{for each } a \in A \text{ do } \\
  (1) V_a = \sum_{s \in T_a} S \\
  (2) C_a = \lambda \cdot A + (1 - \lambda) \cdot V_a \\
  (3) \forall s \in S, t(s) = \cos \text{ine}(s, C_a) \\
  (4) I = I \cup \arg \max_{s \in S} t(s)
  \]

- **Diversity:**
  \[
  \text{until } |I| < n \text{ do } \\
  (1) I = I \cup \arg \max_{s \in S} f(s) \\
  (2) A = A - \delta \cdot A_s
  \]

- **Diversity + Typical (DS-Typical):**
  Apply **Diversity** to **Typical** ranking.

What is a *typical* sentence for a given aspect?

1. *After The Pianist Adrien Brody* has appeared in four very different movies.

2. *Adrien Brody* received widespread recognition when he was cast as the lead in Roman Polanski’s *The Pianist* (2002).

3. *Adrien Brody* is a New York actor who is known to international audiences as the star of Roman Polanski’s 2002 film, *The Pianist*.

Rachel Weisz (*The Constant Gardener*), *Adrien Brody* (*The Pianist*), Mark Buffalo (*Zodiac*), and Rinko Kikuchi (*Babel*) star in THE BROTHERS BLOOM.[...]

[Image of the page with text and table]
Results

Bette Midler, Harvey Keitel, Mario Cuomo, Newt Gingrich, Billy Bragg, Holly Hunter, Mario Lemieux, Reese Witherspoon, Bob Brady, Joe Theismann, Marion Jones, Roberto Benigni, Carmen Electra, Julie Walters, Matthew Santos, Saxby Chambliss, Elton Brand, Lindsey Graham, Monica Lewinsky, Sean Young

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Grammar</th>
<th>Non-Redund.</th>
<th>Novelty</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Web Search</td>
<td>0.37 (0.12)</td>
<td>0.61 (0.22)</td>
<td>0.57 (0.23)</td>
<td>0.07 (0.09)</td>
<td>0.30 (0.16)</td>
</tr>
<tr>
<td>LexRank</td>
<td>0.22 (0.11)</td>
<td>0.61 (0.18)</td>
<td>0.71 (0.26)</td>
<td>0.19 (0.20)</td>
<td>0.34 (0.21)</td>
</tr>
<tr>
<td>DS-Typical</td>
<td>0.50 (0.17)</td>
<td>0.83 (0.13)</td>
<td>0.93 (0.15)</td>
<td>0.29 (0.17)</td>
<td>0.53 (0.19)</td>
</tr>
</tbody>
</table>

Applications: Fact Checking

Q : \{t_1, \ldots, t_n\}

Web Search

S : w_1 \ldots w_m

Fact Verification Engine

Supporting evidence

Rewrite System

Re-ranker / Aggregator

Web Results
In late 1988, Obama entered Harvard Law School. He was selected as an editor of the Harvard Law Review at the end of his first year, and president of the journal in his second year. During his summers, he returned to Chicago, where he worked as an associate at the law firms of Sidley Austin in 1989 and Hopkins & Sutter in 1990. After graduating with a J.D. magna cum laude from Harvard in 1991, he returned to Chicago. Obama’s election as the first black president of the Harvard Law Review gained national media attention and led to a publishing contract and advance for a book about race relations, which evolved into a personal memoir. The manuscript was published in mid-1995 as Dreams from My Father.
<subject> Martin Wendt </subject>
<statement> As of 2009, his total live tournament winnings exceed $900,000. </statement>
<br>
<ref> http://pokerdb.thehendonmob.com/player.php?n=29302 </ref>
<br>
<subject> Gamma-Aminobutyric acid </subject>
<statement> In general, GABA does not cross the blood-brain barrier, although certain areas of the brain which have no effective blood brain barrier, such as the periventricular nucleus, can be reached by drugs such as systematically injected GABA. </statement>
<br>
<ref> http://physrev.physiology.org/content/79/2/511.full </ref>
<br>
<subject> Jason Beghe </subject>
<statement> Kennedy and Beghe often hung out together outside the Metropolitan Museum of Art and in Central Park, and were monitored by Kennedy’s Secret Service detail. </statement>
<br>
<ref> http://www.cnn.com/US/9907/22/jfk.growing.up.in.nyc/ </ref>
<br>
<subject> Los Angeles </subject>
<statement> Los Angeles enjoys a Mediterranean climate, with an average of 35 days with measurable precipitation annually. </statement>
<br>
<br>
<subject> Murder </subject>
<statement> More than 500,000 people have been killed by firearms in Brazil between 1979 and 2003. </statement>
<br>
<ref> http://news.bbc.co.uk/2/hi/americas/4628813.stm </ref>
<br>
<subject> Fox hunting </subject>
<statement> An amendment which allowed licensed hunting under stricter conditions, advocated by the then Prime Minister Tony Blair and some members of the government’s independent inquiry on fox hunting (including its chairman Lord Burns), was voted down. </statement>
<br>
<ref> http://www.guardian.co.uk/uk/2000/jan/24/hunting.rural... </ref>
<br>
<subject> Armenian Genocide </subject>
<statement> Despite his previous public recognition and support of Genocide bills, as well as the election campaign promises to formally recognize the Armenian Genocide, the U.S. President, Barack Obama, although repeating that his views on the issue have not changed, has thus far abstained from using the term ‘genocide’. </statement>
<br>
<ref> http://www.politifact.com/truth-o-meter/promises/obameter/... </ref>
Rewriting System - Framework

Wikipedia Training

Overlapping Terms

\[ O_1 = \{ t^1_{1}, \ldots, t^1_{m} \} \]
\[ O_2 = \{ t^2_{1}, \ldots, t^2_{p} \} \]
\[ \ldots \]
\[ O_n = \{ t^n_{1}, \ldots, t^n_{q} \} \]

Search(\(O_i\)) = URL_i ?

Features : \{POS, sense #, tf*idf, wikiFreq, keyphraseness, offset, multi-word\}

S : About 31% of the population is made up of [foreign nationals] living in [Saudi Arabia]

O:

0 1 0 0 1 0 0 0 0 1 0 0 1
Results

Percentage of Wikipedia citations retrieved in top $n$ results

Amazon Turk Experiments
(scale 0 – 2)

<table>
<thead>
<tr>
<th>Reference Type</th>
<th>Quality</th>
<th>Ease of Effort</th>
<th>Trust-worthiness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wikipedia</td>
<td>1.37</td>
<td>1.26</td>
<td>1.48</td>
</tr>
<tr>
<td>One Rewrite</td>
<td>1.15</td>
<td>1.18</td>
<td>1.36</td>
</tr>
<tr>
<td>Drop-k PAM</td>
<td>1.15</td>
<td>1.24</td>
<td>1.45</td>
</tr>
</tbody>
</table>

Verbatim (w/o stopwords)

One Rewrite

Drop-k uniform

Drop-k PAM

Oracle (overlapping terms)

Percentage of Wikipedia citations retrieved in top $n$ results:

- Top 1: 3.61
- Top 10: 7.77
- Top 50: 9.71

Percentage of Wikipedia citations retrieved in top $n$ results:

- Top 1: 9.96
- Top 10: 21.40
- Top 50: 28.96

Top 1

Top 10

Top 50
Where We Go From Here

- Web-scale knowledge collections
- Web-scale annotation and fact extraction
- Instant access to knowledge in any environment
- Context-awareness
- Personalized views of the knowledge
- Fact engines