Finding News Citations for Wikipedia

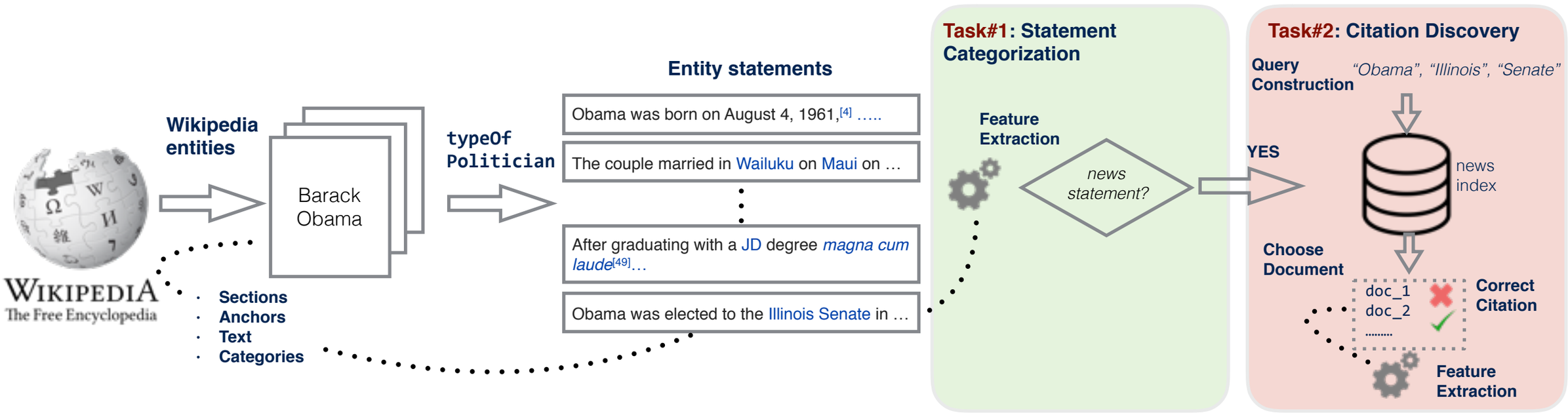
Besnik Fetahu, Katja Markert, Wolfgang Nejdl, and Avishek Anand

@FetahuBesnik
fetahu@l3s.uni-hannover.de
http://www.l3s.de/~fetahu/

Motivation

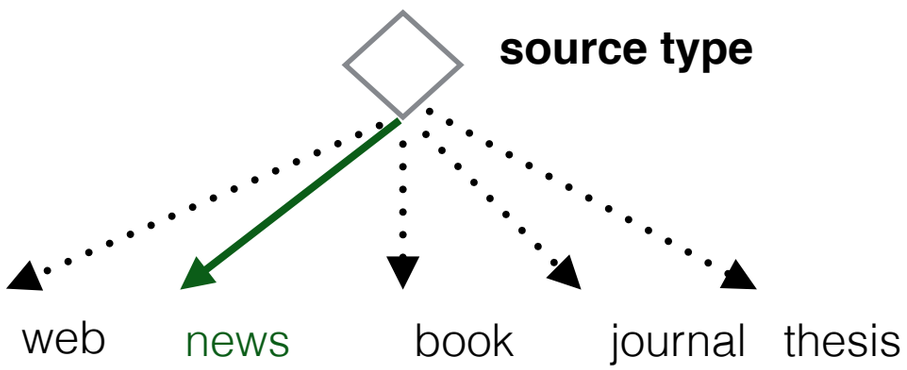
- "*Verifiability"* one of the core principles in Wikipedia
- Citations to external references as *evidence* for statements in Wikipedia articles
- *Reliability* and *authority* of sources for citations of Wikipedia statements
- Citations that point to *outdated* or *dead* URLs
- *Citation needed* for long—tail entities and newly added entities
- An *automated* process of enriching Wikipedia and helping editors in the process of discovering citations

Finding News Citation for Wikipedia



Task#1: Statement Categorization

Obama was elected to the Illinois Senate in …[2]



Task#2: Citation Discovery

news collection

[1] Illinois Sen. Barack Obama's Announcement Speech - Washington Post, 2007.

[2] Jackson, David; Ray Long (April 3, 2007). "Obama Knows His Way Around a Ballot". Chicago Tribune.

[10] In Illinois, Obama Proved Pragmatic and Shrewd - The New York Times, 2007.

Task#1 Statement Categorization

Why Statement Categorization?

Section: "*2008 Presidential Campaign*"

Obama emphasized issues of rapidly ending the Iraq War, increasing energy independence, and reforming the health care system,[1] in a campaign that projected themes of hope and change.[2]

Section: "*Legislative Career*"

On June 3, 2008, Senator Obama—along with Senators Tom Carper, Tom Coburn, and John McCain—introduced follow-up legislation: Strengthening Transparency and Accountability in Federal Spending Act of 2008.[1]

Section: "*Early Life and Career*"

In mid—1988, he traveled for the first time in Europe for three weeks and then for five weeks in Kenya, where he met many of his paternal relatives for the first time.[1,2]

news

2008 Presidential Campaign

report

Legislative Career

book

Early Life and Career

[1] "Barack Obama on the Issues: What Would Be Your Top Three Overall Priorities If Elected?". The Washington Post.

[2] "The Obama promise of hope and change". The Independent. London. November 1, 2008.

"S. 3077: Strengthening Transparency and Accountability in Federal Spending Act of 2008: 2007–2008 (110th Congress)". Govtrack.us. June 3, 2008.

Obama, Auma (2012). And then life happens: a memoir. New York: St. Martin's Press. pp. 189–208, 212–216. ISBN 978-1-250-01005-6.

Why Statement Categorization?

Section: "*2008 Presidential Campaign*"

Obama emphasized issues of rapidly ending the Iraq War, increasing energy independence, and reforming the health care system,[1] in a campaign that projected themes of hope and change.[2]

Section: "*Legislative Career*"

On June 3, 2008, Senator Obama—along with Senators Tom Carper, Tom Coburn, and John McCain—introduced follow-up legislation: Strengthening Transparency and Accountability in Federal Spending Act of 2008.[1]

Section: "*Early Life and Career*"

In mid—1988, he traveled for the first time in Europe for three weeks and then for five weeks in Kenya, where he met many of his paternal relatives for the first time.[1,2]

news

2008 Presidential Campaign

report

lative Career

ngthening Transparency and
in Federal Spending Act of
08 (110th Congress)".
une 3, 2008.

book

Early Life and Career

Obama, Auma (2012). And then life happens: a memoir. New York: St. Martin's Press. pp. 189–208, 212–216. ISBN 978-1-250-01005-6.

[1] "Barack Obama on the Issues: What Would Be Your Top Three Overall Priorities If Elected?". The Washington Post.

[2] "The Obama promise of hope and change". The Independent. London. November 1, 2008.

Task#1: Statement Categorization

- Statement Categorization (**SC**) as a supervised model.

$$SC : f(s, e) \rightarrow c, \text{where } c \in \{\texttt{web, news,...}\}$$

- Model a statement based on language and entity structure features

feature	description
#verbs_attr	the number of verbs of attribution
#POS	the frequency of POS tags in s
$\lambda(s)$	temporal proximity of s to time point
discourse	discourse annotations of s
#quotations	the frequency of quotations in s
$\theta(s, N_t)$	LM score of s
$LDA(s, N_t)$	similarity of s to a topic model
$p(s = news\|\psi)$	section and type news-priors, with
$p(s = news\|t)$	*min*, *max* and *avg* scores of $p(s = news\|\psi)$ and $p(s = news\|t)$ for e
$p(s = news\|t', t)$	type co-occurrence probability between $t \in T(e)$ and $t' \in T(s)$
$p(s = news\|t, \psi)$	type-section joint probability scores

Table 1: Feature list for statement categorization.

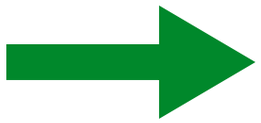
Statement Language Features

Wikipedia Entity Structure

Task#1: Statement Categorization (I)

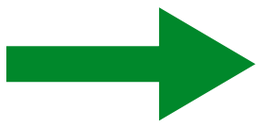
- Language Features:
  - *quotation marks:* an indicator of *paraphrasing*
  - *temporal proximity:* news mostly refer to recent events
  - *discourse:* temporal discourse common in news which report event sequences

He addressed another anti-war rally in March **2003** and told the crowd that ***"it's not too late"*** to stop the war.[73]



news statement:
- close temporal proximity
- quotation marks
- temporal discourse

His next goal was to earn his license and Doctorate in Law, which normally required three years of study. In **1666**, the University of Leipzig turned down Leibniz's doctoral application and refused to grant him a Doctorate in Law, most likely due to his relative youth.



book statement:
- low temporal proximity

Task#1: Statement Categorization (II)

Barack Obama

Contents [hide]

1 Early life and career
- 1.1 Community organizer and Harvard Law School
- 1.2 Chicago Law School and civil rights attorney

2 Legislative career
- 2.1 Illinois State Senator (from 1997)
- 2.2 2004 U.S. Senate campaign
- 2.3 U.S. Senator from Illinois (2005–08)
  - 2.3.1 Legislation
  - 2.3.2 Committees

3 Presidential campaigns
- 3.1 2008 presidential campaign
- 3.2 2012 presidential campaign

4 Presidency (since 2009)
- 4.1 First days
- 4.2 Domestic policy

5 Cultural and political image

6 Family and personal life
- 6.1 Religious views

7 Notes and references
- 7.1 Notes
- 7.2 References
- 7.3 Further reading

8 External links

S1 On February 10, 2007, Obama announced his candidacy for President of the United States in front of the **Old State Capitol** building in **Springfield, Illinois**.[103][104] The choice of the announcement site was viewed as symbolic because it

S2 was also where **Abraham Lincoln** delivered his historic **"House Divided"** speech in 1858.[103][105] Obama emphasized issues of rapidly ending the **Iraq War**, increasing **energy**

S3 **independence**, and **reforming the health care system**,[106] in a campaign that projected themes of hope and change.[107]

Task#1: Statement Categorization (II)

Barack Obama

Contents [hide]

1 Early life and career
- 1.1 Community organizer and Harvard Law School
- 1.2 Chicago Law School and civil rights attorney

2 Legislative career
- 2.1 Illinois State Senator (from 1997)
- 2.2 2004 U.S. Senate campaign
- 2.3 U.S. Senator from Illinois (2005–08)
  - 2.3.1 Legislation
  - 2.3.2 Committees

3 Presidential campaigns
- 3.1 2008 presidential campaign
- 3.2 2012 presidential campaign

4 Presidency (since 2009)
- 4.1 First days
- 4.2 Domestic policy

5 Cultural and political image

6 Family and personal life
- 6.1 Religious views

7 Notes and references
- 7.1 Notes
- 7.2 References
- 7.3 Further reading

8 External links

S_1

On February 10, 2007, Obama announced his candidacy for President of the United States in front of the **Old State Capitol** building in **Springfield, Illinois.**[103][104]

Task#1: Statement Categorization (II)

Barack Obama

Contents [hide]

1 Early life and career
- 1.1 Community organizer and Harvard Law School
- 1.2 Chicago Law School and civil rights attorney

2 Legislative career
- 2.1 Illinois State Senator (from 1997)
- 2.2 2004 U.S. Senate campaign
- 2.3 U.S. Senator from Illinois (2005–08)
  - 2.3.1 Legislation
  - 2.3.2 Committees

3 Presidential campaigns
- 3.1 2008 presidential campaign
- 3.2 2012 presidential campaign

4 Presidency (since 2009)
- 4.1 First days
- 4.2 Domestic policy

5 Cultural and political image

6 Family and personal life
- 6.1 Religious views

7 Notes and references
- 7.1 Notes
- 7.2 References
- 7.3 Further reading

8 External links

s_1

On February 10, 2007, Obama announced his candidacy for President of the United States in front of the **Old State Capitol** building in **Springfield, Illinois.**[103][104]

- **Section-Type Probability**

$$p(s = news|t, \psi) = \frac{\sum_{e \in \mathbf{W} \wedge t \in T(e)} \sum_{s \in S(e,\psi)} \mathbb{1}_{s \text{ typeOf news}}}{\sum_{e \in \mathbf{W} \wedge t \in T(e)} |S(e, \psi)|}$$

"Barack Obama" **isA** Person "*2008 presidential campaign*"

Task#1: Statement Categorization (II)

Barack Obama

Contents [hide]

1 Early life and career
- 1.1 Community organizer and Harvard Law School
- 1.2 Chicago Law School and civil rights attorney

2 Legislative career
- 2.1 Illinois State Senator (from 1997)
- 2.2 2004 U.S. Senate campaign
- 2.3 U.S. Senator from Illinois (2005–08)
  - 2.3.1 Legislation
  - 2.3.2 Committees

3 Presidential campaigns
- 3.1 2008 presidential campaign
- 3.2 2012 presidential campaign

4 Presidency (since 2009)
- 4.1 First days
- 4.2 Domestic policy

5 Cultural and political image

6 Family and personal life
- 6.1 Religious views

7 Notes and references
- 7.1 Notes
- 7.2 References
- 7.3 Further reading

8 External links

s_1

On February 10, 2007, Obama announced his candidacy for President of the United States in front of the **Old State Capitol** building in **Springfield, Illinois.**[103][104]

- **Section-Type Probability**

$$p(s = news|t, \psi) = \frac{\sum_{e \in \mathbf{W} \wedge t \in T(e)} \sum_{s \in S(e,\psi)} \mathbb{1}_{s \text{ typeOf news}}}{\sum_{e \in \mathbf{W} \wedge t \in T(e)} |S(e, \psi)|}$$

"Barack Obama" **isA** Person "*2008 presidential campaign*"

- **Type—Co-occurrence**

$$p(s = news|t', t) = \frac{\sum_{e \in \mathbf{W} \wedge t \in T(e)} \sum_{s \in S(e) \wedge t' \in T(s)} \mathbb{1}_{s \text{ typeOf news}}}{\sum_{e \in \mathbf{W} \wedge t \in T(e)} \sum_{s \in S(e)} \mathbb{1}_{t' \in T(s)}}$$

t' typeOf["Old State Capitol", "Springfield, Illinois"**]**

"Barack Obama" **isA** Person

Task#2: Citation Discovery

Citation Discovery Process

Section: "*2008 Presidential Campaign*"

On February 10, 2007, Obama announced his candidacy for President of the United States in front of the Old State Capitol building in Springfield, Illinois.[103][104]

Citation Discovery Process

Section: "*2008 Presidential Campaign*"

On February 10, 2007, Obama announced his candidacy for President of the United States in front of the Old State Capitol building in Springfield, Illinois.[103][104]

find the appropriate ***news citations***?

query news index

news collection

Citation Discovery Process

Section: "*2008 Presidential Campaign*"

On February 10, 2007, Obama announced his candidacy for President of the United States in front of the Old State Capitol building in Springfield, Illinois.[103][104]

find the appropriate ***news citations***?

query news index

news collection

- What makes a good citation?
  1. the statement is ***entailed*** by the news article
  2. the statement is ***central*** in the news article
  3. the cited news article should be from an ***authoritative*** source

Task#2: Citation Discovery

- Citation Discovery steps for a *news statement* (as categorized in Task#1):

  1. *Query construction* (QC)[1] from the news statement consisting of most important terms

  2. Query top—100 *news article candidates* from a given news index based on the QC step

  3. Compute *similarity features* between the statement and the individual sentences in a news article candidate

  4. *Classify* the news candidates based on how well they fulfill the properties of being a '*good citation*'

[1] Monika Rauch Henzinger, Bay-Wei Chang, Brian Milch, Sergey Brin: Query-free news search. WWW 2003: 1-10

Task#2: Citation Discovery — Entailment

news article citation candidate

Obama: I'm running for president

By **Rick Pearson and Ray Long**
Tribune staff reporters

FEBRUARY 10, 2007, 10:28 AM | SPRINGFIELD

U. S. Sen. Barack Obama formally entered the 2008 race for the presidency today, contending he has the experience to know that "Washington must change" and billing himself as the leader who will bring a new generational attitude to address the nation's challenges.

Speaking in a single-digit, morning chill and sunshine to thousands of supporters outside the Old State Capitol, the first-term Democratic senator delivered an address that built upon his biography as a community organizer in Chicago, state legislator and U.S. senator to call for quick action on issues ranging from bringing a close to the Iraq war to the need for universal health care and an end to foreign-oil dependence.

The historic announcement by the state's 45-year-old junior Democratic senator--launching a three-day wave of campaign events in Iowa and New Hampshire with a Chicago fundraiser in between--is heavily tinged in symbolism for the first black candidate with a realistic chance of obtaining the broad-based support necessary for securing a major party presidential nomination.

Using the home of Lincoln's 1858 "House Divided" speech as a backdrop, Obama frequently paid homage to the 16th president for using his will and words to bring a country divided by war together as one through the goal of freedom.

Wikipedia Statement

On February 10, 2007, Obama announced his candidacy for President of the United States in front of the Old State Capitol building in Springfield, Illinois.**[103][104]**

Task#2: Citation Discovery — Entailment

news article citation candidate

Obama: I'm running for president

By **Rick Pearson and Ray Long**
Tribune staff reporters

FEBRUARY 10, 2007, 10:28 AM | SPRINGFIELD

U. S. Sen. Barack Obama formally entered the 2008 race for the presidency today, contending he has the experience to know that "Washington must change" and billing himself as the leader who will bring a new generational attitude to address the nation's challenges.

Speaking in a single-digit, morning chill and sunshine to thousands of supporters outside the Old State Capitol, the first-term Democratic senator delivered an address that built upon his biography as a community organizer in Chicago, state legislator and U.S. senator to call for quick action on issues ranging from bringing a close to the Iraq war to the need for universal health care and an end to foreign-oil dependence.

The historic announcement by the state's 45-year-old junior Democratic senator--launching a three-day wave of campaign events in Iowa and New Hampshire with a Chicago fundraiser in between--is heavily tinged in symbolism for the first black candidate with a realistic chance of obtaining the broad-based support necessary for securing a major party presidential nomination.

Using the home of Lincoln's 1858 "House Divided" speech as a backdrop, Obama frequently paid homage to the 16th president for using his will and words to bring a country divided by war together as one through the goal of freedom.

chunk into sentences

Wikipedia Statement

On February 10, 2007, Obama announced his candidacy for President of the United States in front of the Old State Capitol building in Springfield, Illinois.[103][104]

Task#2: Citation Discovery — Entailment

news article citation candidate

U. S. Sen. Barack Obama formally entered the 2008 race for the presidency today, contending he has the experience to know that "Washington must change" and billing himself as the leader who will bring a new generational attitude to address the nation's challenges.

Speaking in a single-digit, morning chill and sunshine to thousands of supporters outside the Old State Capitol, the first-term Democratic senator delivered an address that built upon his biography as a community organizer in Chicago, state legislator and U.S. senator to call for quick action on issues ranging from bringing a close to the Iraq war to the need for universal health care and an end to foreign-oil dependence.

The historic announcement by the state's 45-year-old junior Democratic senator--launching a three-day wave of campaign events in Iowa and New Hampshire with a Chicago fundraiser in between--is heavily tinged in symbolism for the first black candidate with a realistic chance of obtaining the broad-based support necessary for securing a major party presidential nomination.

Using the home of Lincoln's 1858 "House Divided" speech as a backdrop, Obama frequently paid homage to the 16th president for using his will and words to bring a country divided by war together as one through the goal of freedom.

Wikipedia Statement

On February 10, 2007, Obama announced his candidacy for President of the United States in front of the Old State Capitol building in Springfield, Illinois.[103][104]

Task#2: Citation Discovery — Entailment

news article citation candidate

U. S. Sen. Barack Obama formally entered the 2008 race for the presidency today, contending he has the experience to know that "Washington must change" and billing himself as the leader who will bring a new generational attitude to address the nation's challenges.

Speaking in a single-digit, morning chill and sunshine to thousands of supporters outside the Old State Capitol, the first-term Democratic senator delivered an address that built upon his biography as a community organizer in Chicago, state legislator and U.S. senator to call for quick action on issues ranging from bringing a close to the Iraq war to the need for universal health care and an end to foreign-oil dependence.

The historic announcement by the state's 45-year-old junior Democratic senator--launching a three-day wave of campaign events in Iowa and New Hampshire with a Chicago fundraiser in between--is heavily tinged in symbolism for the first black candidate with a realistic chance of obtaining the broad-based support necessary for securing a major party presidential nomination.

Using the home of Lincoln's 1858 "House Divided" speech as a backdrop, Obama frequently paid homage to the 16th president for using his will and words to bring a country divided by war together as one through the goal of freedom.

Wikipedia Statement

On February 10, 2007, Obama announced his candidacy for President of the United States in front of the Old State Capitol building in Springfield, Illinois.[103][104]

Entailment features

- *Jaccard similarity* between sentences and the statement
- *NNP phrase* overlap between sentences and the statement
- *Tree Kernel* similarity between the statement and the sentences
- Score statement against the *language model* from the news article
- *Language model* similarity between previously referred news articles and the current candidate
- *Query—similarity* score and *rank* of news article for statement

Task#2: Citation Discovery — Entailment

news article citation candidate

U. S. Sen. Barack Obama formally entered the 2008 race for the presidency today, contending he has the experience to know that "Washington must change" and billing himself as the leader who will bring a new generational attitude to address the nation's challenges.

Speaking in a single-digit, morning chill and sunshine to thousands of supporters outside the Old State Capitol, the first-term Democratic senator delivered an address that built upon his biography as a community organizer in Chicago, state legislator and U.S. senator to call for quick action on issues ranging from bringing a close to the Iraq war to the need for universal health care and an end to foreign-oil dependence.

The historic announcement by the state's 45-year-old junior Democratic senator--launching a three-day wave of campaign events in Iowa and New Hampshire with a Chicago fundraiser in between--is heavily tinged in symbolism for the first black candidate with a realistic chance of obtaining the broad-based support necessary for securing a major party presidential nomination.

Using the home of Lincoln's 1858 "House Divided" speech as a backdrop, Obama frequently paid homage to the 16th president for using his will and words to bring a country divided by war together as one through the goal of freedom.

Wikipedia Statement

On February 10, 2007, Obama announced his candidacy for President of the United States in front of the Old State Capitol building in Springfield, Illinois.[103][104]

Entailment features

- *Jaccard similarity* between sentences and the statement
- *NNP phrase* overlap between sentences and the statement
- *Tree Kernel* similarity between the statement and the sentences
- Score statement against the *language model* from the news article
- *Language model* similarity between previously referred news articles and the current candidate
- *Query—similarity* score and *rank* of news article for statement

normalize the computed features for a news article candidate

- Number of sentences varies across articles
- Consider the *average, min, max, weighted average* for each feature group

Task#2: Citation Discovery — Centrality

news article citation candidate

σ_1 U. S. Sen. Barack Obama formally entered the 2008 race for the presidency today, contending he has the experience to know that "Washington must change" and billing himself as the leader who will bring a new generational attitude to address the nation's challenges.

σ_2 Speaking in a single-digit, morning chill and sunshine to thousands of supporters outside the Old State Capitol, the first-term Democratic senator delivered an address that built upon his biography as a community organizer in Chicago, state legislator and U.S. senator to call for quick action on issues ranging from bringing a close to the Iraq war to the need for universal health care and an end to foreign-oil dependence.

σ_4 The historic announcement by the state's 45-year-old junior Democratic senator--launching a three-day wave of campaign events in Iowa and New Hampshire with a Chicago fundraiser in between--is heavily tinged in symbolism for the first black candidate with a realistic chance of obtaining the broad-based support necessary for securing a major party presidential nomination.

σ_3 Using the home of Lincoln's 1858 "House Divided" speech as a backdrop, Obama frequently paid homage to the 16th president for using his will and words to bring a country divided by war together as one through the goal of freedom.

Entailment features to the most central sentence[1]

$$\Gamma(\sigma_i) = (1-d) \; + \; d * \sum_{\sigma_j \in In(\sigma_i)} \frac{\mathbf{J}(\sigma_i, \sigma_j)}{\sum_{\sigma_k \in Out(\sigma_j)} \mathbf{J}(\sigma_j, \sigma_k)} \Gamma(\sigma_j)$$

Entity Salience in the candidate news article[2]

$$\phi(e,n) = \frac{|\rho(e,n)|}{|\rho(n)|} \sum_{\rho \in \rho(n)} \left(\frac{tf(e,\rho)}{\sum_{e' \neq e} tf(e',\rho)} \right)^{\frac{1}{\rho}}$$

[1] Rada Mihalcea, Paul Tarau: TextRank: Bringing Order into Text. EMNLP 2004: 404-411
[2] Besnik Fetahu, Katja Markert, Avishek Anand: Automated News Suggestions for Populating Wikipedia Entity Pages. CIKM 2015: 323-332

Task#2: Citation Discovery — Authority

Barack Obama

44th President of the United States

Incumbent

Assumed office
January 20, 2009

Vice President Joe Biden

Preceded by George W. Bush

United States Senator from Illinois

Barack Obama's Sections

Contents [hide]

1 Early life and career
- 1.1 Community organizer and Harvard Law School
- 1.2 Chicago Law School and civil rights attorney

2 Legislative career
- 2.1 Illinois State Senator (from 1997)
- 2.2 2004 U.S. Senate campaign
- 2.3 U.S. Senator from Illinois (2005–08)

3 Presidential campaigns
- 3.1 2008 presidential campaign
- 3.2 2012 presidential campaign

4 Presidency (since 2009)
- 4.1 First days
- 4.2 Domestic policy
- 4.1 First days
- 4.2 Domestic policy
- 4.3 Foreign policy

5 Cultural and political image

Stephen Curry

Curry with the Warriors in 2016

No. 30 – Golden State Warriors

Position Point guard

League NBA

Stephen Curry's Sections

Contents [hide]

1 Early life
2 College career
- 2.1 Freshman season
- 2.2 Sophomore season
- 2.3 Junior season
- 2.4 College statistics

3 Professional career
- 3.1 Golden State Warriors (2009–present)
  - 3.1.1 Early seasons (2009–11)
  - 3.1.2 Injury-riddled year (2011–12)
  - 3.1.3 Getting back on track (2012–14)
  - 3.1.4 NBA championship and MVP (2014–15)
  - 3.1.5 Unanimous MVP (2015–16)

4 National team career
5 Player profile
6 Personal life
7 NBA career statistics
8 Awards and honors

isA Politician

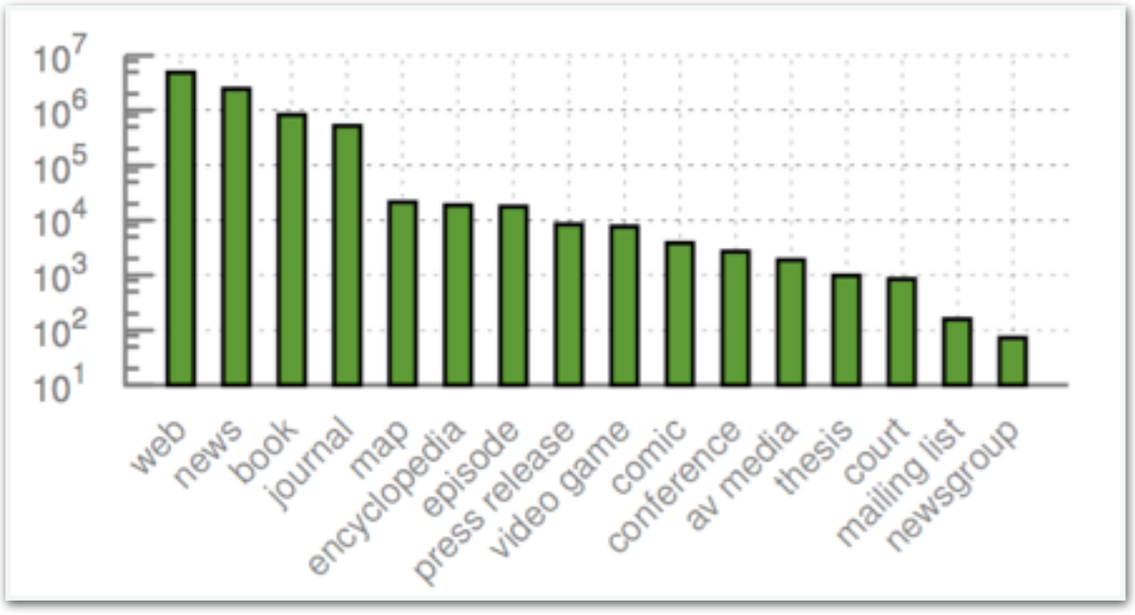
isA Athlete

Experimental Setup — Datasets

- **Statement Categorization Data**
  - 6.9 million Wikipedia statements
  - 8.8 million citations
  - 1.6 million entities
  - 668k sections

- **Citation Discovery Data**
  - 1.88 million news articles cited from Wikipedia news statements
  - 27k news articles within the range of 2013—2015
  - 20 million news articles from a real world news collection[1]

Statement distribution per Category



GDelt domain stats

news domain	news articles
yahoo.com	1244781
allafrica.com	1035646
reuters.com	828133
dailymail.co.uk	815372
indiatimes.com	743991
wn.com	587607

[1] http://gdeltproject.org

Experimental Setup — Ground-truth

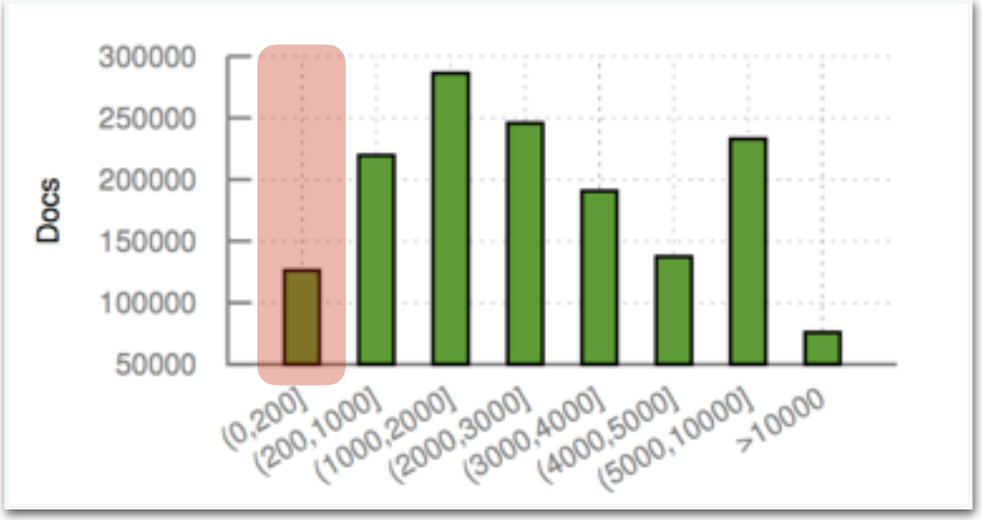
- **Statement Categorization Ground—truth**
  - Wikipedia editors often mix the citation category
  - Simple heuristics to fix such violations
    1. Majority Voting
    2. URL Patterns

- **Citation Discovery Data**
  - 1.8 million news articles cited from Wikipedia news statements
  - 19% point to '*dead links*', '*moved content*', etc.
  - 7% are shorter than 200 characters (usually these articles are the *index* page of a given domain)

Statement category changes after applying the heuristics

	book	journal	news	web
book	0	2,650	1,155	71,801
journal	14,905	0	13542	110,133
news	5,698	2,770	0	**391,634**
web	16,549	25,109	**944,977**	0

News article distribution based on number of characters



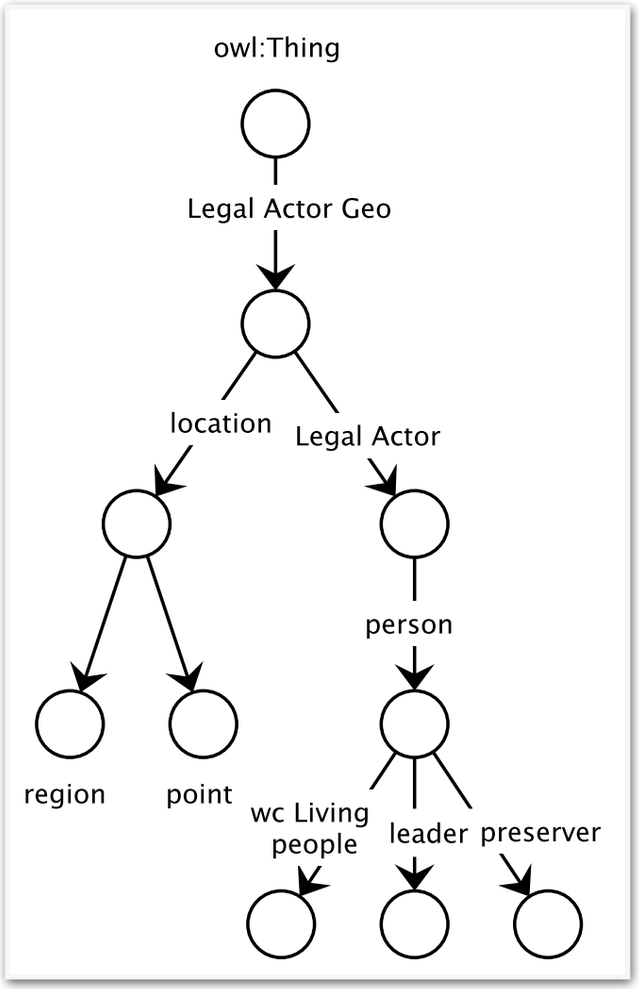
Random example of citation category violation, where instead of *'news'* the news source is categorized as '*web*'

In June 2008, Byrne suggested the "August bank holiday" to be made a weekend of national celebration (the so-called "British Day") in a speech to a New Labour think tank. However, Scotland's August bank holiday is held on a different date from that in Wales and England. He later retracted this - after pressure from the Scottish National Party - saying he was merely trying to "get the debate started" [1].

[1] {{cite type=**web** | url=**http://news.bbc.co.uk/1/hi/uk_politics/7433479.stm**}}

Statement Categorization — Evaluation

YAGO type hierarchy



Separate models for entity types

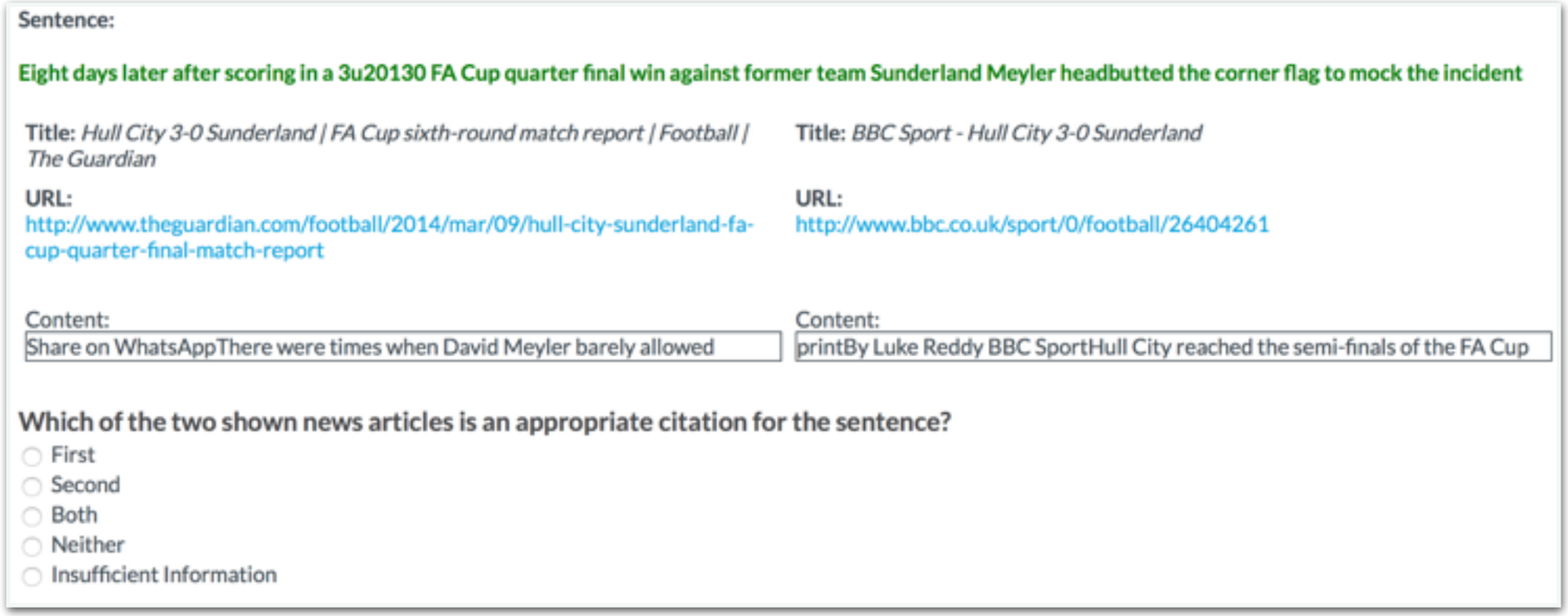
Statement Categorization results based on RF

		yagoLegalActorGeo								
Parent Type	Child Type	$1 \leq \tau \leq 10$			$10 < \tau \leq 50$			$50 < \tau \leq 90$		
		P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
owl:Thing	Legal Actor Geo	0.48	0.36	0.41	0.51	0.43	0.47	0.53	0.47	0.50
Legal Actor Geo	Legal Actor	0.51	0.34	0.41	0.54	0.41	0.47	**0.56**	**0.45**	**0.50**
	location	0.30	0.29	0.29	0.34	0.40	0.37	0.36	0.45	0.40
location	region	0.30	0.28	0.29	0.35	0.40	0.37	0.37	0.44	0.40
	point	0.30	0.10	0.14	0.38	0.22	0.28	0.39	0.26	0.32
Legal Actor	person	0.53	0.36	0.43	0.56	0.43	0.49	**0.58**	**0.46**	**0.51**
person	preserver	0.63	0.31	0.42	0.67	0.46	0.54	**0.67**	**0.49**	**0.57**
	authority	0.53	0.20	0.29	0.62	0.24	0.35	0.65	0.33	0.44
	contestant	0.59	0.43	0.50	0.62	0.52	0.57	0.64	0.56	0.60
	leader	0.53	0.26	0.34	0.59	0.34	0.43	0.61	0.37	0.46
	wc Living people	0.55	0.37	0.44	0.58	0.44	0.50	0.59	0.47	0.52

For entities of type **Person** the task can be accurately performed

Citation Discovery — Evaluation

- Baselines **B1** and **B2:** top—1 retrieved news article and supervised model based on the retrieval score and rank of news article
- Evaluation Strategy **E1:** a correct citation is considered if the models can recover an already existing news citation for a statement from the news collection (Wiki news and GDelt)
- Evaluation Strategy **E1+FP:** consider as relevant the **FP** news articles (they do not exist in the Wiki News ground-truth) where their similarity is over 0.8 w.r.t the ground truth citation from Wiki news
- Evaluation Strategy **E2:** assess the remaining **FP** news articles through crowdsourcing


Sentence:

Eight days later after scoring in a 3u20130 FA Cup quarter final win against former team Sunderland Meyler headbutted the corner flag to mock the incident

Title: *Hull City 3-0 Sunderland | FA Cup sixth-round match report | Football | The Guardian*

URL:
http://www.theguardian.com/football/2014/mar/09/hull-city-sunderland-fa-cup-quarter-final-match-report

Content:
Share on WhatsAppThere were times when David Meyler barely allowed

Title: *BBC Sport - Hull City 3-0 Sunderland*

URL:
http://www.bbc.co.uk/sport/0/football/26404261

Content:
printBy Luke Reddy BBC SportHull City reached the semi-finals of the FA Cup

Which of the two shown news articles is an appropriate citation for the sentence?

- First
- Second
- Both
- Neither
- Insufficient Information

Citation Discovery — Evaluation (I)

Top—10 best performing entity types. E1+FP and E2 columns show the improvement for P over E1. The last row shows the *micro-average* precision across all models.

	B1			B2			E1			E1 + FP	E2
type	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**P**
player	0.37	0.36	0.37	0.31	0.28	0.29	0.67	0.46	0.55	0.71 ▲ (5.63%)	**0.85** ▲ (21.18%)
entertainer	0.32	0.31	0.31	0.16	0.18	0.17	0.70	0.33	0.45	0.78 ▲ (10.26%)	**0.90** ▲ (22.22%)
causal agent	0.26	0.26	0.26	0.17	0.21	0.19	0.73	0.28	0.41	0.77 ▲ (5.19%)	**0.88** ▲ (17.05%)
location	0.21	0.19	0.20	0.21	0.23	0.22	0.55	0.26	0.35	0.62 ▲ (11.29%)	**0.83** ▲ (33.73%)
artist	0.31	0.31	0.31	0.24	0.27	0.25	0.67	0.21	0.32	0.67	**0.85** ▲ (21.18%)
football player	0.31	0.31	0.31	0.29	0.38	0.33	0.80	0.30	0.43	0.80	**0.90** ▲ (11.11%)
wcat Living people	0.27	0.26	0.26	0.21	0.18	0.2	0.67	0.23	0.34	0.70 ▲ (4.29%)	**0.85** ▲ (21.18%)
creator	0.34	0.32	0.33	0.25	0.24	0.24	0.74	0.25	0.38	0.74	**0.91** ▲ (18.68%)
organism	0.29	0.28	0.28	0.22	0.19	0.2	0.69	0.30	0.41	0.70 ▲ (1.43%)	**0.83** ▲ (16.87%)
person	0.26	0.24	0.25	0.21	0.23	0.22	0.64	0.35	0.46	0.66 ▲ (3.03%)	**0.85** ▲ (24.71%)
micro-average	0.25			0.21			0.67			0.71 ▲ (5.6%)	**0.86** ▲ **(22.00%)**

Citation relevance for **E2** based on the crowdsourcing evaluation.

both	4,506 (38.2%)
ground truth only	3,768 (31.9%)
our suggestion only	2,287 (19.4%)
neither	1,242 (10.5%)
all	11,803 (100%)

Conclusions

- Citations as a core principle for the "*Verifiability*" of statements in Wikipedia
- Entities are added and constantly evolve in Wikipedia, thus, the need for automated citation discovery
- Many existing citations are either outdated or missing
- Citation discovery in two stages: (i) first determine if a statement needs a news citation, and (ii) find an appropriate citation from a real-world news collection
- Challenges on determining *web* from *news* statements
- Important aspects to consider: (i) construct a query from a statement, and (ii) account for noise in the categorization of citations from Wikipedia editors
- Automated citation discovery can be performed with high accuracy for all entity types

Thank you!

Questions?

Task#1: What is a statement?

- What is a *statement* in Wikipedia?

A statement is a piece of text (a single or sequence of sentences) from a Wikipedia page that has or needs a citation and that occur between two consecutive citation markers or a citation marker and paragraph beginning/end.

A statement belongs to an **entity** and a specific **entity section**.

Section: "*2008 Presidential Campaign*"

On February 10, 2007, Obama announced his candidacy for President of the United States in front of the Old State Capitol building in Springfield, Illinois.[103][104] The choice of the announcement site was viewed as symbolic because it was also where Abraham Lincoln delivered his historic "House Divided" speech in 1858.[103][105] Obama emphasized issues of rapidly ending the Iraq War, increasing energy independence, and reforming the health care system,[106] in a campaign that projected themes of hope and change.[107]

Task#1: What is a statement?

- What is a *statement* in Wikipedia?

A statement is a piece of text (a single or sequence of sentences) from a Wikipedia page that has or needs a citation and that occur between two consecutive citation markers or a citation marker and paragraph beginning/end.

A statement belongs to an **entity** and a specific **entity section**.

Section: "*2008 Presidential Campaign*"

S1

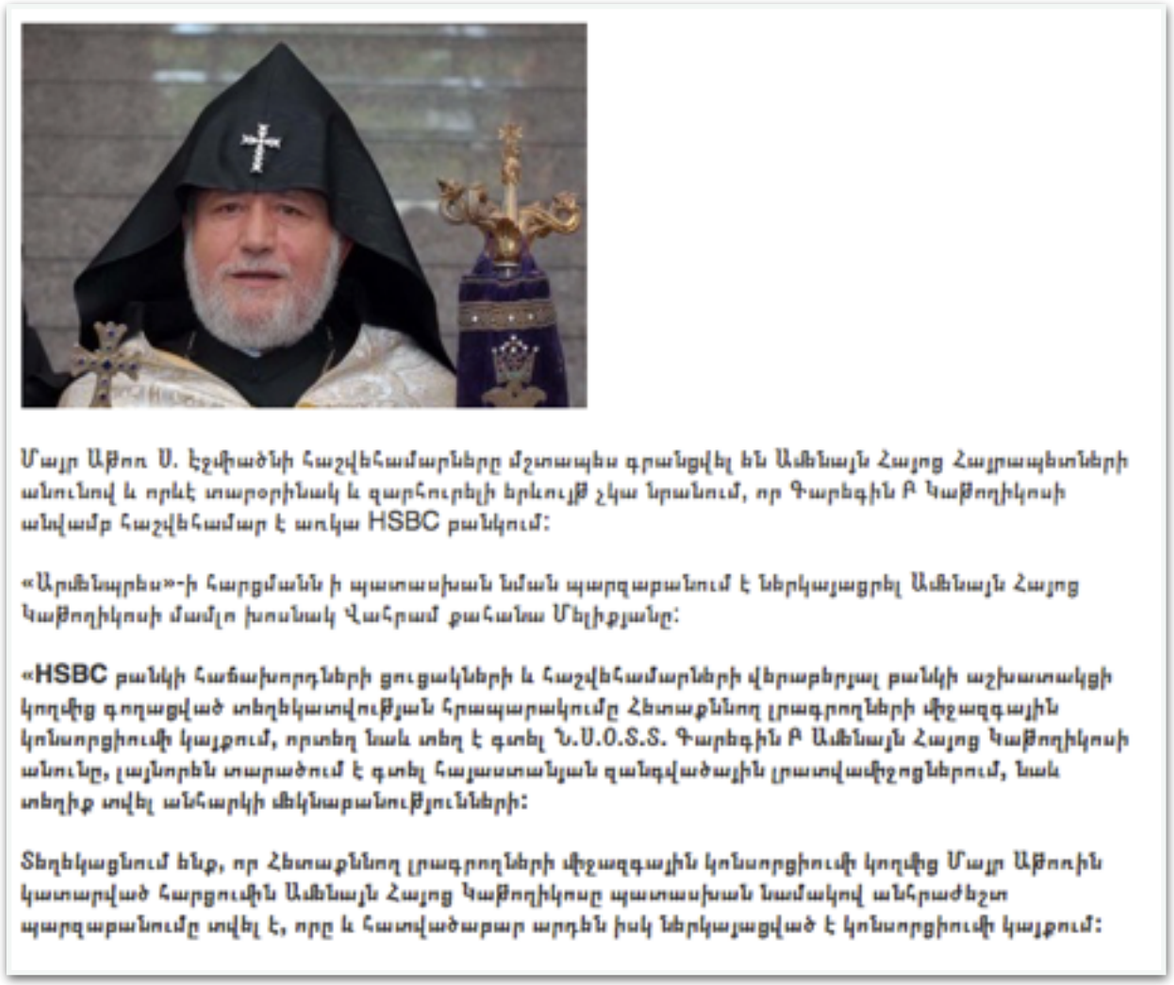
On February 10, 2007, Obama announced his candidacy for President of the United States in front of the Old State Capitol building in Springfield, Illinois.[103][104] The choice of the announcement site was viewed as symbolic because it was also where Abraham Lincoln delivered his historic "House Divided" speech in 1858.[103][105] Obama emphasized issues of rapidly ending the Iraq War, increasing energy independence, and reforming the health care system,[106] in a campaign that projected themes of hope and change.[107]

S2

S3

Citation Discovery — Evaluation (II)

Why other and crowdsourced evaluation strategies?



Մայր Աթոռ Ս. Էջմիածնի հաշվեհամարները մշտապես գրանցվել են Ամենայն Հայոց Հայրապետների անունով և որևէ տարօրինակ և զարհուրելի երևույթ չկա նրանում, որ Գարեգին Բ Կաթողիկոսի անվամբ հաշվեհամար է առկա HSBC բանկում:

«Արմենպրես»-ի հարցմանն ի պատասխան նման պարզաբանում է ներկայացրել Ամենայն Հայոց Կաթողիկոսի մամլո խոսնակ Վահրամ քահանա Մելիքյանը:

«**HSBC** բանկի հաճախորդների ցուցակների և հաշվեհամարների վերաբերյալ բանկի աշխատակցի կողմից գողացված տեղեկատվության հրապարակումը Հետաքննող լրագրողների միջազգային կոնսորցիումի կայքում, որտեղ նաև տեղ է գտել Ն.Ս.Օ.Տ.Տ. Գարեգին Բ Ամենայն Հայոց Կաթողիկոսի անունը, լայնորեն տարածում է գտել հայաստանյան զանգվածային լրատվամիջոցներում, նաև տեղիք տվել անհարկի մեկնաբանությունների:

Տեղեկացնում ենք, որ Հետաքննող լրագրողների միջազգային կոնսորցիումի կողմից Մայր Աթոռին կատարված հարցումին Ամենայն Հայոց Կաթողիկոսը պատասխան նամակով անհրաժեշտ պարզաբանումը տվել է, որը և համապատասխան արդեն իսկ ներկայացված է կոնսորցիումի կայքում:

Secret Account? Head of Armenian Church Shows Up in Leaked HSBC Files

13:32, February 9, 2015

The head of the Armenian Apostolic Church, Catholicos Garegin II, has been profiled in a project published by The IJIC (International Consortium of Investigative Journalists) looking into 60,000 leaked files providing details on over 100,000 HSBC clients and their bank accounts.

The HSBC files were obtained through an international collaboration of news outlets, including the Guardian, Le Monde, BBC Panorama and the International Consortium of Investigative Journalists reveal that HSBC's Swiss private bank:

The Guardian writes:

HSBC's Swiss banking arm helped wealthy customers dodge taxes and conceal millions of dollars of assets, doling out bundles of untraceable cash and advising clients on how to circumvent domestic tax authorities, according to a huge cache of leaked secret bank account files.

- In the English Wikipedia it is desirable for statements to refer to English sources
- Ground-truth (**E1**) serves as a *lower—bound* in terms of accuracy for **citation discovery**
- Multiple relevant citations for a given statement

Citation Discovery — Evaluation (II)

Why other and crowdsourced evaluation strategies?

Catholicos Karekin II, the spokeswoman explained in HSBC bank account in the name of the

15:08 • 09.02.15

Like 61 Tweet G+1 1

Fr. All accounts are registered in the name of Echmiadzin, the Armenian Pontiff and there is a strange and terrible thing is that the name of Catholicos Karekin II of the HSBC bank account. "Armenpress" introduced similar clarification in response to the Catholicos of All Armenians spokesman Vahram Melikyan. « **HSBC Bank customer lists and account information stolen by a bank employee publication on the website of the international consortium of Investigative journalists, which also included His Holiness Karekin II, Catholicos of All Armenians, widely spread in the mass media, it also leads to inappropriate comments. Please be informed that the International Consortium of Investigative Journalists question the Catholicos response letter need clarification from the Holy See has given, which in parts is already represented the consortium website. it has been and Vazgen I and Karekin I, Catholicos Karekin II and the case. After the account is transferred to the successor Catholicos Karekin II, asleep in the Lord, and made to account renamed. The same is also true of this account. This account is opened in the Pontificate of His Holiness Karekin I, together with the indicated amount of time an account is transferred to the Holy See Karekin II, Catholicos of All Armenians** , "the explanation reads.

Secret Account? Head of Armenian Church Shows Up in Leaked HSBC Files

13:32, February 9, 2015

The head of the Armenian Apostolic Church, Catholicos Garegin II, has been profiled in a project published by The IJIC (International Consortium of Investigative Journalists) looking into 60,000 leaked files providing details on over 100,000 HSBC clients and their bank accounts.

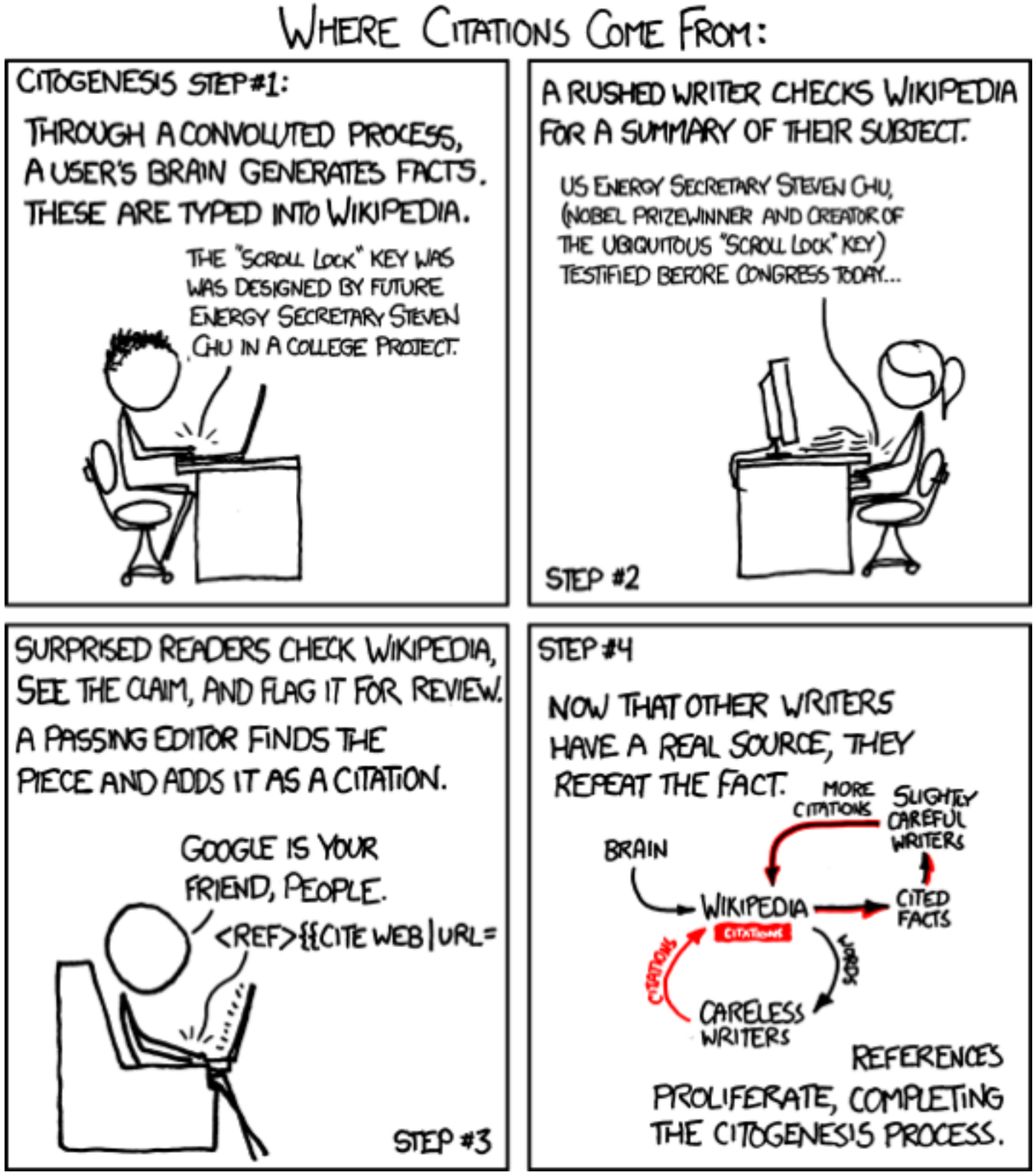
The HSBC files were obtained through an international collaboration of news outlets, including the Guardian, Le Monde, BBC Panorama and the International Consortium of Investigative Journalists reveal that HSBC's Swiss private bank:

The Guardian writes:

HSBC's Swiss banking arm helped wealthy customers dodge taxes and conceal millions of dollars of assets, doling out bundles of untraceable cash and advising clients on how to circumvent domestic tax authorities, according to a huge cache of leaked secret bank account files.

- In the English Wikipedia it is desirable for statements to refer to English sources
- Ground-truth (**E1**) serves as a *lower—bound* in terms of accuracy for **citation discovery**
- Multiple relevant citations for a given statement

Introduction

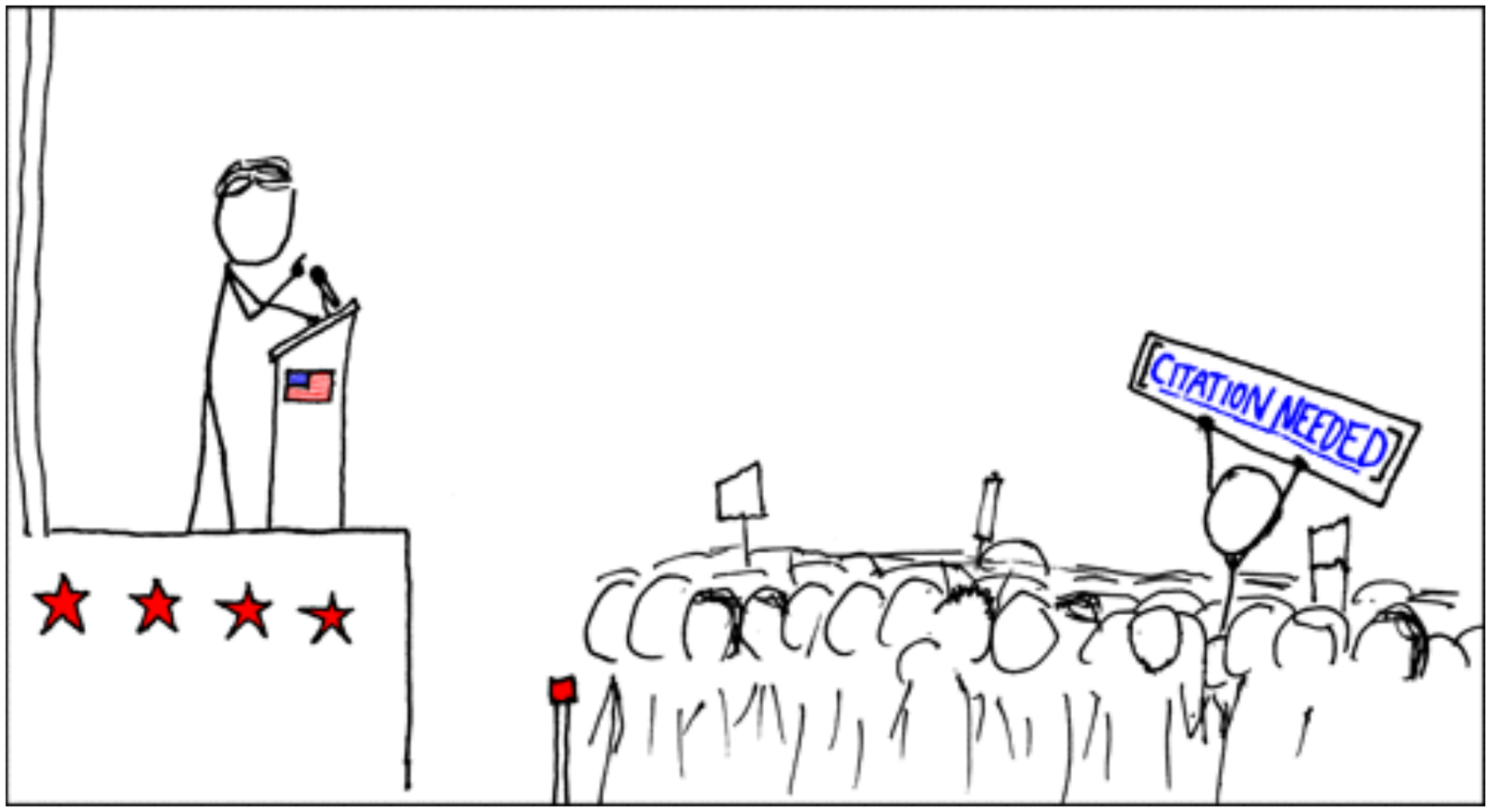


Introduction

“Citogenesis”: Citogenesis, on the other hand is a portmanteau of 'Citation' and 'Genesis'. A Citation is a reference to a source, used to back up a specific claim. Genesis means the origin of something. By extension, citogenesis is the creation of text in a reliable source that can be cited to back-up a claim

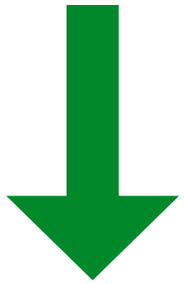
Introduction

“Citogenesis”: Citogenesis, on the other hand is a portmanteau of 'Citation' and 'Genesis'. A Citation is a reference to a source, used to back up a specific claim. Genesis means the origin of something. By extension, citogenesis is the creation of text in a reliable source that can be cited to back-up a claim[citation needed]



Introduction

“Citogenesis”: Citogenesis, on the other hand is a portmanteau of 'Citation' and 'Genesis'. A Citation is a reference to a source, used to back up a specific claim. Genesis means the origin of something. By extension, citogenesis is the creation of text in a reliable source that can be cited to back-up a claim[1,2]



what are the appropriate citations for this definition?

[1] https://xkcd.com/978/

[2] https://www.explainxkcd.com/wiki/index.php/978:_Citogenesis