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• Citations that point to *outdated* or *dead* URLs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
Task#1: Statement Categorization

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

Task#2: Citation Discovery

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”

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]


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”**

**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]


Task#1: Statement Categorization

- Statement Categorization (SC) as a supervised model.

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

- Model a statement based on language and entity structure features

<table>
<thead>
<tr>
<th>feature</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>#verbs_attr</td>
<td>the number of verbs of attribution</td>
</tr>
<tr>
<td>#POS</td>
<td>the frequency of POS tags in ( s )</td>
</tr>
<tr>
<td>( \lambda(s) )</td>
<td>temporal proximity of ( s ) to time point</td>
</tr>
<tr>
<td>discourse</td>
<td>discourse annotations of ( s )</td>
</tr>
<tr>
<td>#quotations</td>
<td>the frequency of quotations in ( s )</td>
</tr>
<tr>
<td>( \theta(s, N_t) )</td>
<td>LM score of ( s )</td>
</tr>
<tr>
<td>LDA(( s, N_t ))</td>
<td>similarity of ( s ) to a topic model</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>feature</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p(s = \text{news}</td>
<td>\psi) )</td>
</tr>
<tr>
<td>( p(s = \text{news}</td>
<td>t) )</td>
</tr>
<tr>
<td>( p(s = \text{news}</td>
<td>\psi) )</td>
</tr>
<tr>
<td>( p(s = \text{news}</td>
<td>t', t) )</td>
</tr>
<tr>
<td>( p(s = \text{news}</td>
<td>t, \psi) )</td>
</tr>
</tbody>
</table>

Table 1: Feature list for statement categorization.
He addressed another anti-war rally in March 2003 and told the crowd that "it's not too late" to stop the war.[73]

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

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

---

**Section-Type Probability**

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

“Barack Obama” isA Person  
“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]

**S1**

- **Section-Type Probability**
  \[
  p(s = news|t, \psi) = \frac{\sum_{e \in W \land t \in T(e)} \sum_{s \in S(e, \psi)} 1_s \text{ typeOf } \text{news}}{\sum_{e \in W \land t \in T(e)} |S(e, \psi)|}
  \]

  "Barack Obama" isA Person  "2008 presidential campaign"

- **Type—Co-occurrence**
  \[
  p(s = news|t', t) = \frac{\sum_{e \in W \land t \in T(e)} \sum_{s \in S(e) \land t' \in T(s)} 1_s \text{ typeOf } \text{news}}{\sum_{e \in W \land t \in T(e)} \sum_{s \in S(e)} 1_t' \in T(s)}
  \]

  t' typeOf[“Old State Capitol”, “Springfield, Illinois”]  "Barack Obama" isA Person
Task#2: Citation Discovery
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]
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.\textsuperscript{[103][104]}

Citation Discovery Process

Section: “2008 Presidential Campaign”

find the appropriate news citations?

query news index

news collection
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]

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*’

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.\textsuperscript{[103]}\textsuperscript{[104]}
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.\cite{103}\cite{104}
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]
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]
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 — Centrality

news article citation candidate

**σ₁**

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.

**σ₂**

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.

**σ₃**

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.

**σ₄**

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.

Entailment features to the most central sentence[1]

\[
\Gamma(\sigma_i) = (1 - d) + \frac{d \times \sum_{\sigma_j \in \text{In}(\sigma_i)} \frac{J(\sigma_i, \sigma_j)}{\sum_{\sigma_k \in \text{Out}(\sigma_j)} J(\sigma_j, \sigma_k)} \Gamma(\sigma_j)}{1}
\]

Entity Salience in the candidate news article[2]

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

Task#2: Citation Discovery — Authority

**Barack Obama**

- **Barack Obama’s Sections**
  - 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
    - 4.3 First days
    - 4.4 Domestic policy
    - 4.5 Foreign policy
  - 5 Cultural and political image

**Stephen Curry**

- **Stephen Curry’s Sections**
  - 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)
    - 3.2 National team career
    - 3.3 Player profile
    - 3.4 Personal life
    - 3.5 NBA career statistics
    - 3.6 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]

---

[^1]: [http://gdeltproject.org](http://gdeltproject.org)

---

**GDelt domain stats**

<table>
<thead>
<tr>
<th>news domain</th>
<th>news articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>yahoo.com</td>
<td>1244781</td>
</tr>
<tr>
<td>allafrica.com</td>
<td>1035646</td>
</tr>
<tr>
<td>reuters.com</td>
<td>828133</td>
</tr>
<tr>
<td>dailymail.co.uk</td>
<td>815372</td>
</tr>
<tr>
<td>indiatimes.com</td>
<td>743991</td>
</tr>
<tr>
<td>wn.com</td>
<td>587607</td>
</tr>
</tbody>
</table>
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)

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].

Statement Categorization — Evaluation

YAGO type hierarchy

Separate models for entity types

Statement Categorization results based on RF

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

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

<table>
<thead>
<tr>
<th>type</th>
<th>B1</th>
<th>B2</th>
<th>E1</th>
<th>E1 + FP</th>
<th>E2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F1</td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>player</td>
<td>0.37</td>
<td>0.36</td>
<td>0.37</td>
<td>0.31</td>
<td>0.28</td>
</tr>
<tr>
<td>entertainer</td>
<td>0.32</td>
<td>0.31</td>
<td>0.31</td>
<td>0.16</td>
<td>0.18</td>
</tr>
<tr>
<td>causal agent</td>
<td>0.26</td>
<td>0.26</td>
<td>0.26</td>
<td>0.17</td>
<td>0.21</td>
</tr>
<tr>
<td>location</td>
<td>0.21</td>
<td>0.19</td>
<td>0.20</td>
<td>0.21</td>
<td>0.23</td>
</tr>
<tr>
<td>artist</td>
<td>0.31</td>
<td>0.31</td>
<td>0.31</td>
<td>0.24</td>
<td>0.27</td>
</tr>
<tr>
<td>football player</td>
<td>0.31</td>
<td>0.31</td>
<td>0.31</td>
<td>0.29</td>
<td>0.38</td>
</tr>
<tr>
<td>wcat Living people</td>
<td>0.27</td>
<td>0.26</td>
<td>0.26</td>
<td>0.21</td>
<td>0.18</td>
</tr>
<tr>
<td>creator</td>
<td>0.34</td>
<td>0.32</td>
<td>0.33</td>
<td>0.25</td>
<td>0.24</td>
</tr>
<tr>
<td>organism</td>
<td>0.29</td>
<td>0.28</td>
<td>0.28</td>
<td>0.22</td>
<td>0.19</td>
</tr>
<tr>
<td>person</td>
<td>0.26</td>
<td>0.24</td>
<td>0.25</td>
<td>0.21</td>
<td>0.23</td>
</tr>
<tr>
<td><strong>micro-average</strong></td>
<td>0.25</td>
<td>0.21</td>
<td>0.67</td>
<td>0.71 ▲ (5.6%)</td>
<td></td>
</tr>
</tbody>
</table>

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

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. 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. Obama emphasized issues of rapidly ending the Iraq War, increasing energy independence, and reforming the health care system in a campaign that projected themes of hope and change.
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. 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. Obama emphasized issues of rapidly ending the Iraq War, increasing energy independence, and reforming the health care system in a campaign that projected themes of hope and change.
Why other and crowdsourced evaluation strategies?

- 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
Why other and crowdsourced evaluation strategies?

• In the English Wikipedia it is desirable for statements to refer to English sources
• Ground-truth ($E_1$) serves as a lower—bound in terms of accuracy for citation discovery
• Multiple relevant citations for a given statement
Introduction

WHERE CITATIONS COME FROM:

CITOGENESIS STEP #1:
THROUGH A CONVOLUTED PROCESS, A USER'S BRAIN GENERATES FACTS. THESE ARE TYPED INTO WIKIPEDIA.

THE "SCROLL LOCK" KEY WAS DESIGNED BY FUTURE ENERGY SECRETARY STEVEN CHU IN A COLLEGE PROJECT.

SURPRISED READERS CHECK WIKIPEDIA, SEE THE CLAIM, AND FLAG IT FOR REVIEW. A PASSING EDITOR FINDS THE PIECE AND ADDS IT AS A CITATION.

GOOGLE IS YOUR FRIEND, PEOPLE.

<REF>{{CITE WEB|URL=

STEP #3

STEP #2

A RUSHED WRITER CHECKS WIKIPEDIA FOR A SUMMARY OF THEIR SUBJECT.

US ENERGY SECRETARY STEVEN CHU, (NOBEL PRIZEWINNER AND CREATOR OF THE UBQUITOUS "SCROLL LOCK" KEY) TESTIFIED BEFORE CONGRESS TODAY....

STEP #4

NOW THAT OTHER WRITERS HAVE A REAL SOURCE, THEY REPEAT THE FACT.

REFERENCES PROLIFERATE, COMPLETING THE CITOGENESIS PROCESS.
“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.
“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]
“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\cite{1,2}

what are the appropriate citations for this definition?

\begin{itemize}
  \item [1] https://xkcd.com/978/
\end{itemize}