Fact Harvesting from Natural Language Text in Wikipedia

Matteo Cannaviccio (Roma Tre University)
Denilson Barbosa (University of Alberta)
Paolo Merialdo (Roma Tre University)
Knowledge Graphs

Enabling technology for:

- semantic search
  - in terms of entities-relations (not keywords-pages)
- text analytics
- text understanding/summarization
- recommendation systems
  - to identify personalized entities and relations
Knowledge Graphs: Semantic Search

Google search for "who is the director of apocalypse now?"

About 3,120,000 results (0.74 seconds)

Apocalypse Now / Director

Francis Ford Coppola

Francis Ford Coppola is an American screenwriter, film director, and producer. He was considered to be the central figure of the New Hollywood wave of filmmaking. Wikipedia

Born: April 7, 1939 (age 77), Detroit, MI
Spouse: Eleanor Coppola (m. 1963)
Children: Sofia Coppola, Roman Coppola, Gian-Carlo Coppola
Nephews: Nicolas Cage, Jason Schwartzman, More
Siblings: Talia Shire, August Coppola

Movies

View 40+ more
Knowledge Graphs: Semantic Search

Google search for: who is the wife of the director of apocalypse now?

Results:
- Eleanor Coppola
  - Born: May 4, 1936 (age 80), Los Angeles, CA
  - Spouse: Francis Ford Coppola (m. 1963)
  - Children: Sofia Coppola, Roman Coppola, Gian-Carlo Coppola
  - Grandchildren: Gia Coppola, Romy Mars, Cosima Mars
  - Movies: Hearts of Darkness: A Filmmaker's Apocalypse, On the Set of CQ
  - Siblings: Bill Neil

People also search for:
- Francis
- Sofia
- Roman
- Gian-Carlo
- Gia
- George
Knowledge Graphs: Semantic Search

Google search for "us president when director of apocalypse now borned"

United States of America / President (1939)

Franklin D. Roosevelt

Quotes and overview
Knowledge Graphs: Semantic Search

Google search result for: what is the birthplace of the daughter of the director of apocalypse now

About 5,420,000 results (0.99 seconds)

Sofia Coppola / Place of birth

New York City, NY

Home to the Empire State Building, Times Square, Statue of Liberty and other iconic sites, New York City is a fast-paced, globally influential center of art, culture, fashion and finance. The city's 5 boroughs sit where the Hudson River meets the Atlantic Ocean, with the island borough of Manhattan at the "Big Apple's" core.

Land area: 304.6 mi²
Weather: 80°F (27°C), Wind S at 11 mph (18 km/h), 51% Humidity
Hotels: 3-star averaging $210, 5-star averaging $420. View hotels
Getting there: 5 h 19 min flight, around $405. View flights
Local time: Thursday 6:34 PM
Knowledge Graphs: Recommendation Systems
Knowledge Graphs

- Facebook’s Entity Graph
- Microsoft Probase
- Google’s Knowledge Graph
- Wikidata
- Yago
- Freebase
- Knowledge Vault

Linked Datasets as of August 2014
What is a Knowledge Graph (1)

A graph that aims to describe knowledge about real world Entities, entity types

- An **entity** is an instance (with id) of multiple types
  - It represents a real world object
- **Entity types** are organized in a hierarchy

![Entity Types Diagram]

- **people**
  - **person**
  - **director**
- **film**
- **location**
What is a Knowledge Graph (2)

A graph that aims to describe knowledge about real world

Relations and facts
• A relation is triple: subject type – predicate – object type
  It describes a semantic association between two entity types
What is a Knowledge Graph (3)

A graph that aims to describe knowledge about real world

Relations and facts

• A relation is triple: subject type – predicate – object type
  It describes a semantic association between two entity types
• Facts define instances of relations, represent semantic associations between two entities
What is a Knowledge Graph (4)

A graph that aims to describe knowledge about real world Entities (nodes) and facts (edges)
Knowledge Graphs

- **DBpedia**
  - 4M entities in 250 types
  - 500M facts for 6K relations

- **Google’s Knowledge Graph**
  - 600M entities in 15K types
  - 20B facts

- **Knowledge Vault**
  - 4M entities in 250 types
  - 500M facts for 6K relations

- **Freebase**
  - 40M entities in 1.5K types
  - 650M facts for 4K relations
  - Core of Google Knowledge Graph

- **yago**
  - 45M entities in 1.1K types
  - 271M facts for 4.5K relations

- **Knowledge Graph**
  - 10M entities in 350K types
  - 120M facts for 100 relations

[Dong16, Weikum16]
Knowledge Graphs: incompleteness

#Facts/Entities in Freebase (as of March 2016) [Dong16]

- 40% of entities with no facts
- 56% of entities with <3 facts

<table>
<thead>
<tr>
<th>Relation</th>
<th>Percentage unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All 3M</td>
</tr>
<tr>
<td>PROFESSION</td>
<td>68%</td>
</tr>
<tr>
<td>PLACE OF BIRTH</td>
<td>71%</td>
</tr>
<tr>
<td>NATIONALITY</td>
<td>75%</td>
</tr>
<tr>
<td>EDUCATION</td>
<td>91%</td>
</tr>
<tr>
<td>SPOUSES</td>
<td>92%</td>
</tr>
<tr>
<td>PARENTS</td>
<td>94%</td>
</tr>
<tr>
<td>CHILDREN</td>
<td>94%</td>
</tr>
<tr>
<td>SIBLINGS</td>
<td>96%</td>
</tr>
<tr>
<td>ETHNICITY</td>
<td>99%</td>
</tr>
</tbody>
</table>

[West+14]
Knowledge Graphs: incompleteness

Michelle Obama
From Wikipedia, the free encyclopedia

Michelle LaVaughn Robinson Obama (born January 17, 1964) is an American lawyer and writer. She is married to the 44th and current President of the United States, Barack Obama, and is the first African-American First Lady of the United States. Raised on the South Side of Chicago, Illinois, Obama is a graduate of Princeton University and Harvard Law School, and spent the early part of her legal career working at the law firm Sidley Austin, where she met her husband.

Ben Ripley
From Wikipedia, the free encyclopedia

Ben Ripley is an American screenwriter best known for writing the science-fiction thriller Source Code directed by Duncan Jones. Ripley is a graduate of Stanford University and the University of Southern California’s USC School of Cinema-Television.

Ben Ripley
Born Lawrenceville, New Jersey
Occupation Screenwriter
Known for Source Code
Wikipedia-derived Knowledge Graphs

Our Focus

Articles with no Infobox
- 56% in 2008
- 66% in 2010

Goal:
- Derive a KG from Wikipedia

Source:
- Structured components (category, infoboxes, …)

Process:
- Assign a type to the main entity
- Map attributes to KG relations

DeYoung, Illinois, U.S.
(now Calumet Park)

Political party Democratic
Spouse(s) Barack Obama (m. 1992)
Children Malia
Sasha
Alma mater Princeton University
Harvard University
Religion Protestantism
(Formerly Methodism)
Signature [Signature]
Michelle Obama

yago select knowledge

DBpedia

Lector:
- Text as source of facts
- Encyclopedic nature (many facts)
- Restricted community (homogeneous language)
Lector: Harvesting facts from text

Our purpose: Increase a KG with facts extracted from Wikipedia text.

Experiment: Facts in the domain of people:
- 12 Freebase relations

Result: Lector can extract more than 200K facts:
- Absent in Freebase, DBPedia and YAGO
- Many relations reach an estimated accuracy of 95%

Our method: We rely on the duality between:
- **phrases**: spans of text between two entities
- **relations**: canonical relations from a KG
Duality of Phrases and Relations

Michelle Obama
From Wikipedia, the free encyclopedia

Michelle LaVaughn Robinson Obama (born January 17, 1964) is an American lawyer and writer. She is married to the 44th and current President of the United States, Barack Obama, and is the first African-American First Lady of the United States. Raised on the South Side of Chicago, Illinois, Obama is a graduate of Princeton University and Harvard Law School, and spent the early part of her legal career working at the law firm Sidley Austin, where she met her husband.

Ben Ripley
From Wikipedia, the free encyclopedia

Ben Ripley is an American screenwriter best known for writing the science-fiction thriller Source Code[2] directed by Duncan Jones.[3][4] Ripley is a graduate of Stanford University and the University of Southern California’s USC School of Cinema-Television.
Duality of Patterns and Relations:

**Facts & Fact Candidates**

- (Michelle, Harward)
- (Hillary, Yale)
- (Michelle, Harward)
- (Hillary, Yale)
- (Alberto, PoliMi)
- (Wesley, UofTexas)

**Patterns**

- X studied at Y
- X graduated from Y
- X earned his degree from Y
- X was a student at Y
- X visited Y

Adapted from an example by Gerhard Weikum
Duality of Patterns and Relations: an Adult Approach…

Dipre (1998)

• seminal work


• build on Dipre


• Open IE: discover new relations (open)
Duality of Patterns and Relations: …with a Teenage Attitude

Facts & Fact Candidates

(Michelle, Harward)
(Hillary, Yale)
(Michelle, Harward)
(Hillary, Yale)
(Alberto, PoliMi)
(Wesley, UofTexas)
(Michelle, Harward)
(Hillary, Yale)
(Alberto, PoliMi)
(Divesh, RomaTre)

Patterns

X studied at Y
X graduated from Y
X earned his degree from Y
X was a student at Y
X visited Y

…
• good for recall
• not for precision: (noisy, drifting)

Adapted from an example by Gerhard Weikum
With a Teenager: better to Introduce a soft Distant Supervision

(Many) Facts from the KG
(Michelle, Harward)
(Hillary, Yale)
...

(Good) Phrases from Articles
X studied at Y
X graduated from Y
X earned his degree from Y
...

New Facts
(Michelle, Harward)
(Hillary, Yale)
(Alberto, PoliMi)
...

• High precision
• (no drifting)

Adapted from an example by Gerhard Weikum
Our approach
Annotate articles with FB entities

We rely on:

- Wikipedia entities (highlighted in the text)
- RDF interlink between Wikipedia and Freebase

Wikipedia original entities:

- Primary entity (subject of the article)
- Secondary entities (entities linked in the article)
<Michelle_Obama> (born January 17, 1964) is an American lawyer. <Michelle_Obama> is married to Barack Obama, the 44th President of the United States. Raised on the south side of Chicago, Illinois, <Michelle_Obama> is a graduate of Princeton University. Subsequently, <Michelle_Obama> worked for the University of Chicago Medical Center. After Barack Obama’s election to the U.S. Senate, the <Michelle_Obama> family continued to live in Chicago …

We match the primary entity using:

- Full name (Michelle Obama)
- Last name (Obama)
- Complete name (Michelle LaVaughn Robinson Obama)
- Personal pronouns (She)
Annotate articles with FB entities

Secondary entities

... but only the first occurrence!

<Michelle_Obama>
<Michelle_Obama> (born January 17, 1964) is an American <Lawyer>.
<Michelle_Obama> is married to <Barack_Obama>, the 44th President of the
<United_States>. Raised on the south side of <Chicago>, Illinois, <Michelle_Obama> is
a graduate of <Princeton_University>. Subsequently, <Michelle_Obama> worked for the
<University_of_Chicago>. After <Barack_Obama>'s election to the U.S. Senate, the
Obama family continued to live in <Chicago> ...

We match secondary entities using:

- Anchor text (University of Chicago Medical Center)
- Wikipedia id (University of Chicago)
Our approach

1. original articles

2. annotated articles

- was born in
- was a native of
- is originally from
- hails from

- attended
- was educated at
- studied at the
- enrolled at

birthPlace

almaMater

Freebase

en1

en2

en3

en4

...
Extracting phrases

For each **sentence** in all the articles (containing en1 and en2):

1. extract the span of text between en1 and en2
2. generalize it (G) and check if it is relational (R)
3. if it is, associate it with all the relations that link en1 to en2 in the KG

### Generalizing phrases (G)

- "was the first”, "was the 41st”  →  “was the **ORD**”
- "is an American”, “is a Canadian”  →  “is a **NAT**”

### Filtering relational phrases (R)

- Conform with POS-level patterns [Mesquita+13]
  - "is married to”  →  [VB], [VB], [TO]  →  relational
  - "together with”  →  [RP], [IN]  →  not relational
Extracting phrases (cont’d)

Considering only witness count is not reliable:

“was born in” \(\rightarrow\) birthPlace

... \(\rightarrow\)

depthPlace

For each relation, we rank the phrases:

- scoring the specificity of a phrase \((p)\) with a relation \((r_i)\):

\[
score(p, r_i) = \log c(p, r_i) \cdot P(r_i|p)
\]

where:

\[
P(r_i|p) = \frac{c(p, r_i)}{\sum_{j\in R} c(p, r_j)}
\]

- \(P(r_i|p) > 0.5\) minimum probability threshold
Our approach

original articles

was born in
was born near
is a native of
is originally from
hails from

birthPlace

was a graduate of

almaMater

attended
was educated at
is a graduate of
studied at the
enrolled at

Freebase

new facts

3

en1

almaMater

en4

... birthPlace

en3

...
Experiments

- 12 Freebase relations in the domain of people:
  - people/person/place_of_birth
  - people/person/place_of_death
  - people/person/nationality
  - sports/pro_athlete/teams
  - people/person/education
  - people/person/spouse
  - people/person/parents
  - people/person/children
  - people/person/ethnicity
  - people/person/religion
  - award/award_winner/awards_won
  - government/politician/party

- \( K = 20 \) maximum number of phrases for each relation
- 977K entities person (interlinked in multiple KGs)

Aim of the experiment

- **Quantify** the number of facts extracted by Lector (not in Freebase)
- **Accuracy** of the facts:
  - manually evaluation of a random sample (1800 extracted facts)
  - estimating precision (we use Wilson score interval for C.L. = 95%)
### Lector new facts

<table>
<thead>
<tr>
<th>Freebase relations</th>
<th>already in Freebase</th>
<th>extracted by Lector (not yet in FB)</th>
<th>evaluated facts</th>
<th>estimated accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>people/person/place_of_birth</td>
<td>662,192</td>
<td>57,140</td>
<td>347</td>
<td>88.9</td>
</tr>
<tr>
<td>people/person/place_of_death</td>
<td>178,849</td>
<td>18,458</td>
<td>104</td>
<td>80.6</td>
</tr>
<tr>
<td>people/person/nationality</td>
<td>584,792</td>
<td>50,234</td>
<td>290</td>
<td>95.6</td>
</tr>
<tr>
<td>sports/pro_athlete/teams</td>
<td>145,080</td>
<td>49,809</td>
<td>286</td>
<td>96.5</td>
</tr>
<tr>
<td>people/person/education</td>
<td>378,043</td>
<td>46,342</td>
<td>286</td>
<td>98.3</td>
</tr>
<tr>
<td>people/person/spouse</td>
<td>130,425</td>
<td>14,939</td>
<td>97</td>
<td>31.6</td>
</tr>
<tr>
<td>people/person/parents</td>
<td>123,747</td>
<td>5,648</td>
<td>50</td>
<td>77.8</td>
</tr>
<tr>
<td>people/person/children</td>
<td>141,860</td>
<td>3,149</td>
<td>50</td>
<td>38.8</td>
</tr>
<tr>
<td>people/person/ethnicity</td>
<td>39,869</td>
<td>2,989</td>
<td>50</td>
<td>92.7</td>
</tr>
<tr>
<td>people/person/religion</td>
<td>47,016</td>
<td>1,437</td>
<td>50</td>
<td>94.5</td>
</tr>
<tr>
<td>award/award_winner/awards_won</td>
<td>98,625</td>
<td>1,934</td>
<td>50</td>
<td>96.4</td>
</tr>
<tr>
<td>government/politician/party</td>
<td>65,300</td>
<td>3,684</td>
<td>50</td>
<td>94.5</td>
</tr>
</tbody>
</table>

All the numbers are calculated over the 977K person from RDF interlinks (owl:sameAs).
Limitations

Ambiguous phrases:

- ./spouse: “met”
- ./children: “was succeeded by”
- ./place_of_birth: “grew up in”

Impact of K (number of phrases for relation)

- We try different values $K \in \{1, 5, 10, 15, 20\}$
- Groundtruth: 1800 manually evaluated facts

(+) accuracy: 97.24% ±1.49%
(-) extracted facts: 57K to 50K (-8%)
… and in other KGs?

<table>
<thead>
<tr>
<th>DBpedia relations</th>
<th>not in DBpedia</th>
<th>extracted by Lector (not yet in FB)</th>
<th>not in YAGO</th>
<th>YAGO relations</th>
</tr>
</thead>
<tbody>
<tr>
<td>birthPlace</td>
<td>48,314</td>
<td>57,140</td>
<td>55,577</td>
<td>wasBornIn</td>
</tr>
<tr>
<td>deathPlace</td>
<td>15,818</td>
<td>18,458</td>
<td>18,014</td>
<td>diedIn</td>
</tr>
<tr>
<td>nationality</td>
<td>48,125</td>
<td>50,234</td>
<td>49,977</td>
<td>isCitizenOf</td>
</tr>
<tr>
<td>team</td>
<td>23,640</td>
<td>49,809</td>
<td>35,013</td>
<td>playsFor</td>
</tr>
<tr>
<td>almaMater</td>
<td>45,585</td>
<td>46,342</td>
<td>46,095</td>
<td>graduatedFrom</td>
</tr>
<tr>
<td>spouse</td>
<td>14,662</td>
<td>14,939</td>
<td>14,573</td>
<td>isMarriedTo</td>
</tr>
<tr>
<td>parent</td>
<td>5,631</td>
<td>5,648</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>child</td>
<td>3,140</td>
<td>3,149</td>
<td>2,958</td>
<td>hasChild</td>
</tr>
<tr>
<td>ethnicity</td>
<td>2,890</td>
<td>2,989</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>religion</td>
<td>1,368</td>
<td>1,437</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>award</td>
<td>1,655</td>
<td>1,934</td>
<td>1,370</td>
<td>hasWonPrize</td>
</tr>
<tr>
<td>party</td>
<td>3,594</td>
<td>3,684</td>
<td>3,684</td>
<td>isPoliticianOf</td>
</tr>
</tbody>
</table>
Conclusions

Future works

- Introduce negative counts to filter ambiguous phrases
- Extend and generalize the process to other relations
All the facts produced are available for download at:
http://dx.doi.org/10.7939/DVN/10795

Questions?
Extracting phrases (cont’d)

**deathPlace**
- died in
- was born in
- moved to
- died at
- returned to
- lived in
- settled in
- went to
- retired to
- arrived in

**birthPlace**
- was born in
- was born at
- is a
- returned to
- was a
- grew up in
- is an
- died in
- was an
- is a native of

![Bar charts showing occurrences of different phrases related to death and birth places.](chart.png)
Improve phrases extraction

We “normalize” list of entities using *such as* Hearst pattern

\(<Ronaldo> \text{ played for many teams such as } <\text{FCBarcelona}>, <\text{Real_Madrid}> \text{ and } <\text{InterFC}>\)

\(<Ronaldo> \text{ played for } <\text{FCBarcelona}>\)
\(<Ronaldo> \text{ played for } <\text{Real_Madrid}>\)
\(<Ronaldo> \text{ played for } <\text{InterFC}>\)
Improve phrases extraction

To improve accuracy, check around!

Alice, the sister of Bob, is married with Charlie
Alice is married with Bob’s brother

To improve recall, find subordinate clauses!

Ronaldo played for FCB Barcelona and then moved to Inter FC
### Place of birth ranking

| phrase                        | \(c(p, r_i)\) | \(P(r_i | p)\) | \(\text{score}(r_i, p)\) |
|-------------------------------|----------------|----------------|---------------------------|
| was born in                   | 106,200        | 0.73           | 8.43                      |
| was born at                   | 5,606          | 0.86           | 7.43                      |
| is a native of                | 445            | 0.72           | 4.43                      |
| grew up in                    | 2,399          | 0.54           | 4.22                      |
| was born on                   | 165            | 0.81           | 4.15                      |
| born in                       | 431            | 0.68           | 4.13                      |
| was a native of               | 440            | 0.65           | 4.00                      |
| is originally from            | 163            | 0.70           | 3.60                      |
| hails from                    | 149            | 0.68           | 3.40                      |
|                               |                |                |                           |
| is a                          | 4,890          | 0.07           | 0.61                      |
| returned to                   | 3,459          | 0.21           | 1.72                      |
| died in                       | 2,108          | 0.11           | 0.79                      |
| was raised in                 | 615            | 0.46           | 3.01                      |

*filtered out*
Knowledge Graphs: Semantic Search

Google search for "us president during world war 2"

About 48,400,000 results (1.03 seconds)

Franklin D. Roosevelt

Harry S. Truman in World War II. Harry S. Truman (1884-1972) became the 33rd President of the United States upon the death of Franklin D. Roosevelt in April 1945.

Harry S. Truman in World War II - Shmoop
www.shmoop.com/wwii/harry-s-truman.html
Knowledge Graphs: Semantic Search

Google search query:

what is the place of birth of the son of francis ford coppola

Search results:

Francis Ford Coppola / Sons / Place of birth

Neuilly-sur-Seine, France
Roman Coppola

Los Angeles, CA
Gian-Carlo Coppola