DeFacto
Temporal and Multilingual Deep Fact Validation

Daniel Gerber, Diego Esteves, Jens Lehmann, Lorenz Bühmann, Ricardo Usbeck, Axel-Cyrille Ngonga Ngomo, René Speck

Institute for Applied Informatics
(InfAI)

October 19, 2016

ISWC 2016

October 19, 2016
Lily Aldridge is married to Kings of Leon Arthur Shrewsbury's death place is the English Cricket Team Michelle Obama is married to the presidency of Barack Obama People can be born in Lacrosse, the sport...
Lily Aldridge is married to Kings of Leon.
Lily Aldridge is married to Kings of Leon.

Arthur Shrewsbury’s death place is the English Cricket Team.
Introduction

Lily Aldridge is married to Kings of Leon.

Arthur Shrewsbury’s death place is the English Cricket Team.

- Michelle Obama is married to the presidency of Barack Obama.
- People can be born in Lacrosse, the sport...
- ...

Gerber et al. (InfAI) DeFacto October 19, 2016 2 / 36
Introduction

Problem

- 130+ billion facts
- Automatically generated facts
- Manual checking too tedious
### Problem
- 130+ billion facts
- Automatically generated facts
- Manual checking too tedious

### Solution
- Deep Fact Validation
- Use verbalization and ML to find evidence for facts being true
- Support temporal scoping
Albert Einstein
award
Nobel Prize
BOA Pattern
Library

"Nobel Prize" "was awarded to" "Albert Einstein"

Search Engine

Trust-worthiness

Proof Scoring

Index

1921

RDF-Provenance

TRUE

FALSE

Temporal/Fact Confirmation

Gerber et al. (InfAI)
Overview

Albert Einstein

Nobel Prize

BOA Pattern

Library

"Nobel Prize" "was awarded to" "Albert Einstein"

Search Engine

Trustworthiness

Proof Scoring

Temporal/Fact Confirmation

Index

1921

RDF-Provenance

TRUE

FALSE

Gerber et al. (InfAI)
Billions of triples available
Apply distance learning to extract RDF
Idea

- Billions of triples available
- Apply distance learning to extract RDF
Idea

- Billions of triples available
- Apply distance learning to extract RDF

Gerber et al. (InfAI)
Barack Obama was born in Honolulu, Hawaii.

Barack Hussein Obama is a politician of the Democratic Party.

Barack Obama was born in Honolulu, Hawaii.

Barack Hussein Obama is a politician of the Democratic Party.

Barack Obama was born in Honolulu, Hawaii.

Barack Hussein Obama is a politician of the Democratic Party.

was born in Berlin on 13 December 1924.

is a politician of the German Green Party.

married Jackie Bouvier Kennedy Onassis who married John F. Kennedy was tied to the Auchinclosses via her sister’s marriage into the Auchincloss family.
D with his wife R

↓

Pacheco arrived with his wife Leyla Rodriguez Stahl and several...

↓

Pacheco\_PER arrived\_O with\_O his\_O wife\_O Leyla\_PER Rodriguez\_PER Stahl\_PER and\_O several...

```
\begin{itemize}
  \item \texttt{dbpedia:Abel_Pacheco}
  \item \texttt{dbpedia-owl:spouse}
  \item \texttt{boa:Leyla_Rodriguez_Stahl}
  \item \texttt{rdfs:label}
  \item \texttt{rdf:type}
  \item \texttt{rdfs:label}

  \texttt{“Abel Pacheco”@en}
  \texttt{“Leyla Rodriguez Stahl”@en}
\end{itemize}
```
<table>
<thead>
<tr>
<th>English</th>
<th>German</th>
<th>French</th>
</tr>
</thead>
<tbody>
<tr>
<td>publication</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R ’s novel “D”</td>
<td>R in seinem Roman “D”</td>
<td>D est un roman R</td>
</tr>
<tr>
<td>R ’s book “D”</td>
<td>R in seinem Buch “D”</td>
<td>R dans son roman D</td>
</tr>
<tr>
<td>R, author of “D”</td>
<td>R in seinem Werk “D”</td>
<td>R intitulé D</td>
</tr>
<tr>
<td>marriage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R married D</td>
<td>D seiner Frau R</td>
<td>R épouse D</td>
</tr>
<tr>
<td>R, his wife D</td>
<td>D seiner Ehefrau R</td>
<td>R , veuve D</td>
</tr>
<tr>
<td>D ’s marriage to R</td>
<td>R und seiner Gattin D</td>
<td>D, la femme de R</td>
</tr>
</tbody>
</table>
O’Hara’s first major film was Alfred Hitchcock directed “Jamaica Inn” which was released in 1939, she had previously . . .
"Nobel Prize" "was awarded to" "Albert Einstein"
DeFacto
Proof Scoring

“Nobel Prize” “was awarded to” “Albert Einstein”

BOA Pattern Library

Search Engine

Index

Trustworthiness

Proof Scoring

Temporal/Fact Confirmation

RDF-Provenance

TRUE

FALSE

1921
Idea

- Search for surface forms $\lambda(s)$ and $\lambda(o)$ for triple $(s, p, o)$
- if $dist(e_1, e_2) < \vartheta$ then extract proof features
- Example: Albert Einstein was awarded the Nobel Prize in 1921
DeFacto
Proof Scoring

Idea

- Search for surface forms $\lambda(s)$ and $\lambda(o)$ for triple $(s, p, o)$
- if $\text{dist}(e_1, e_2) < \vartheta$ then extract proof features
- Example: Albert Einstein was awarded the Nobel Prize in 1921

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BOA Pattern</td>
<td>0, 1</td>
</tr>
<tr>
<td>BOA Score</td>
<td>[0,1]</td>
</tr>
<tr>
<td>Entity$_{\text{dist}}$</td>
<td>[0, $\vartheta$]</td>
</tr>
<tr>
<td>Wordnet$_{\text{sim}}$</td>
<td>[0,1]</td>
</tr>
<tr>
<td>Frequency</td>
<td>[1,n]</td>
</tr>
<tr>
<td>Title$_{\text{sim}}$</td>
<td>[0,1]</td>
</tr>
<tr>
<td>Punctuation</td>
<td>0, 1</td>
</tr>
<tr>
<td>Text</td>
<td>Vector</td>
</tr>
<tr>
<td>Predicate</td>
<td>Word</td>
</tr>
</tbody>
</table>
DeFacto
Trustworthiness

"Nobel Prize" "was awarded to" "Albert Einstein"

Albert Einstein
Nobel Prize
award
1921

Search Engine
Index
BOA Pattern Library

Proof Scoring

Trustworthiness

TRUE

Temporal/Fact Confirmation

FALSE

RDF-Provenance

Gerber et al. (InfAI)
DeFacto
October 19, 2016 21 / 36
DeFacto
Trustworthiness

Idea

- Approximate reliability of document (Nakamura et al., 2007)
- Rely on distributional features
- Example: Albert Einstein was awarded the Nobel Prize in 1921
### Idea

- Approximate reliability of document (Nakamura et al., 2007)
- Rely on distributional features
- Example: **Albert Einstein** was awarded the **Nobel Prize** in 1921

<table>
<thead>
<tr>
<th>Feature</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic majority in search results</td>
<td>Number of pages in search results with similar topics</td>
</tr>
<tr>
<td>Topic majority in WWW</td>
<td>Number of websites with similar topics</td>
</tr>
<tr>
<td>Pagerank</td>
<td>Google Page Rank of website</td>
</tr>
</tbody>
</table>
### Idea
- **Approximate reliability of document** (Nakamura et al., 2007)
- **Rely on distributional features**
- **Example:** Albert Einstein **was awarded** the Nobel Prize in 1921

### Feature | Explanation
--- | ---
Topic majority in search results | Number of pages in search results with similar topics
Topic majority in WWW | Number of websites with similar topics
Pagerank | Google Page Rank of website
$|\{p \mid p \in P\}|$ | Number of proofs found in website
Number of search results | Total number of all queries (approx.)
*rdfs:domain, rdfs:range* | Binary validation of domain and range
Background knowledge | Co-occurrence classes and predicates
Facts are not always true...

:Tom_Cruise :spouse :Katie_Holmes>

:Franz_Beckenbauer :plays_for_club :New_York_Cosmos
Facts are not always true ...

:Tom_Cruise :spouse :Katie_Holmes>
:Franz_Beckenbauer :plays_for_club :New_York_Cosmos

Extension to temporal checking

- Use distribution of years across the Web
- Extraction of years using patterns
- Search year literals between within context window
- Learn patterns for valid interval
- Use year distribution to normalize score

---

```
[0-9]{4}\s*(\|\-\--)\s*[0-9]{4}

[Flf]rom \[0-9]\{4\} until \[0-9]\{4\}

[bB]etween (the years) \[0-9]\{4\} and \[0-9]\{4\}

\[0-9]\{4\} bis einschließlich \[0-9]\{4\}

[zZ]wischen (den Jahren) \[0-9]\{4\} und \[0-9]\{4\}

[dD]urant la période \[0-9]\{4\} - \[0-9]\{4\}

[eE]ntre les années \[0-9]\{4\} et \[0-9]\{4\}

...```
“Nobel Prize” “was awarded to” “Albert Einstein”

Nobel Prize

BoA Pattern Library

Search Engine

Index

Trustworthiness

Proof Scoring

Temporal/Fact Confirmation

TRUE

FALSE

RDF-Provenance

Gerber et al. (InfAI)  DeFacto  October 19, 2016  25 / 36
Dataset: FactBench

Key Performance Indicators
- Accuracy
- Precision
- Recall
- F-measure

Machine learning approaches
- J48
- Naïve Bayes
- Support Vector Machine (with sequential optimization)
- ...
Factbench

Overview

- Benchmark for fact checking
- 10 common relations
- 1,500 facts including validity period
- 750 in train and 750 in test set

- **Award**
  - persons who received a nobel prize (timepoint, freebase)

- **Birth**
  - birth place and date of a person (timepoint, dbpedia)

- **Death**
  - death place and date of a person (timepoint, dbpedia)

- **Foundation Place**
  - place and date of a company's foundation (timepoint, freebase)

- **Leader**
  - presidents of countries (timespan, dbpedia)

- **NBA Team**
  - team associations of NBA players (timespan, dbpedia)

- **Publication Date**
  - author of a book and it's publication date (timepoint, freebase)

- **Spouse**
  - marriage of two persons (timespan, freebase)

- **Starring**
  - actors who starred in films (timepoint, dbpedia)

- **Subsidiary**
  - companies and their subsidiaries (timepoint, freebase)
Manually curated positive examples

Negative examples generated automatically using random selected Triples $t = (s, p, o)$

- **domain** Replace $s \rightarrow (s', p, o)$
- **range** Replace $o \rightarrow (s, p, o')$
- **domainrange** Replace $s$ and $o \rightarrow (s', p, o')$
- **property** Replace $p \rightarrow (s, p', o)$
- **random** Replace $s$, $p$ and $o \rightarrow (s', p', o')$
- **mix** Random sample of 20% of each of the above
## DeFacto

### Fact Checking Evaluation

<table>
<thead>
<tr>
<th></th>
<th>Domain</th>
<th></th>
<th>Range</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F&lt;sub&gt;1&lt;/sub&gt;</td>
<td>P</td>
</tr>
<tr>
<td><strong>J48</strong></td>
<td>0.898</td>
<td>0.897</td>
<td>0.897</td>
<td>0.909</td>
</tr>
<tr>
<td>SimpleLogistic</td>
<td>0.890</td>
<td>0.890</td>
<td>0.890</td>
<td>0.880</td>
</tr>
<tr>
<td>NaiveBayes</td>
<td>0.837</td>
<td>0.812</td>
<td>0.808</td>
<td>0.852</td>
</tr>
<tr>
<td>SMO</td>
<td>0.861</td>
<td>0.854</td>
<td>0.853</td>
<td>0.852</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>DomainRange</th>
<th></th>
<th>Property</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F&lt;sub&gt;1&lt;/sub&gt;</td>
<td>P</td>
</tr>
<tr>
<td><strong>J48</strong></td>
<td>0.910</td>
<td>0.910</td>
<td>0.910</td>
<td>0.786</td>
</tr>
<tr>
<td>SimpleLogistic</td>
<td>0.889</td>
<td>0.889</td>
<td>0.889</td>
<td>0.653</td>
</tr>
<tr>
<td>NaiveBayes</td>
<td>0.861</td>
<td>0.845</td>
<td>0.843</td>
<td>0.620</td>
</tr>
<tr>
<td>SMO</td>
<td>0.853</td>
<td>0.836</td>
<td>0.834</td>
<td>0.673</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Random</th>
<th></th>
<th>Mix</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F&lt;sub&gt;1&lt;/sub&gt;</td>
<td>P</td>
</tr>
<tr>
<td><strong>J48</strong></td>
<td>0.910</td>
<td>0.909</td>
<td>0.909</td>
<td>0.850</td>
</tr>
<tr>
<td>SimpleLogistic</td>
<td>0.879</td>
<td>0.878</td>
<td>0.878</td>
<td>0.810</td>
</tr>
<tr>
<td>NaiveBayes</td>
<td>0.851</td>
<td>0.841</td>
<td>0.839</td>
<td>0.789</td>
</tr>
<tr>
<td>SMO</td>
<td>0.864</td>
<td>0.843</td>
<td>0.841</td>
<td>0.817</td>
</tr>
</tbody>
</table>
1. Using multiple languages better in most cases
2. Close to 100% accuracy (award, multilingual)
Precision $P = \frac{|\{\text{correct} \cap \{\text{found}\}\}|}{|\{\text{found}\}|}$

Recall $R = \frac{|\{\text{correct} \cap \{\text{found}\}\}|}{|\{\text{correct}\}|}$
DeFacto
Validity Span Check

Precision = 1/2
Recall = 1/3
F-measure = 2/5
1. Using multiple languages significantly better
2. Increase from 86.53% to 89.2% on average
3. +6.5% for time points and +6.9% time spans
Conclusion and Future Work

Summary

- Presented DeFacto, a framework for checking RDF facts
- Performs best with J48 classifier
- Achieves between 80% and 100% accuracy on FactBench relations

Future Work

- Extend FactBench
- Combination with Deep Learning
- Integration of more languages
- Reference corpus for faster search
Summary
- Presented DeFacto, a framework for checking RDF facts
- Performs best with J48 classifier
- Achieves between 80% and 100% accuracy on FactBench relations

Future Work
- Extend FactBench
- Combination with Deep Learning
- Integration of more languages
- Reference corpus for faster search
That’s all Folks!

HOBBIT
Holistic Benchmarking of Big Linked Data

Axel Ngonga
AKSW Research Group
Augustusplatz 10, Room P905
04109 Leipzig, Germany
ngonga@informatik.uni-leipzig.de
@NgongaAxel