# Coverage of Information Extraction from Sentences and Paragraphs

Simon Razniewski[1], Nitisha Jain[2],
Paramita Mirza[1], Gerhard Weikum[1]

[1]Max Planck Institute for Informatics
[2]Hasso Plattner Institute

max planck institut
informatik

HPI

# The classical IE paradigm

*Obama has two children, Malia and Sasha.*

Information extraction tool

*<Obama, child, Malia>*
*<Obama, child, Sasha>*

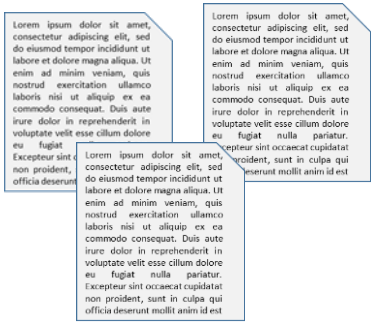
But are these all?

# Coverage differences

- Obama has two children, Malia and Sasha.
- Jolie brought her children Shiloh and Knox to school.

- New York consists of the districts Manhattan, Bronx, Queens, Brooklyn, and Staten Island.
- Important districts of Hong Kong are Wan Chai, Kowloon City, and Yau Tsim Mong.

# Relevance (1/3): IE resource efficiency



IE

Districts(NY) = Manhattan, Bronx, Queens, Brooklyn, Staten Island

Coverage = **High** → Stop further extraction

Districts(Hong Kong) = Wan Chai, Kowloon City, Yau Tsim Mong

Coverage = **Low** → Explore more resources

# Relevance (2/3): Adjust IE thresholds


IE

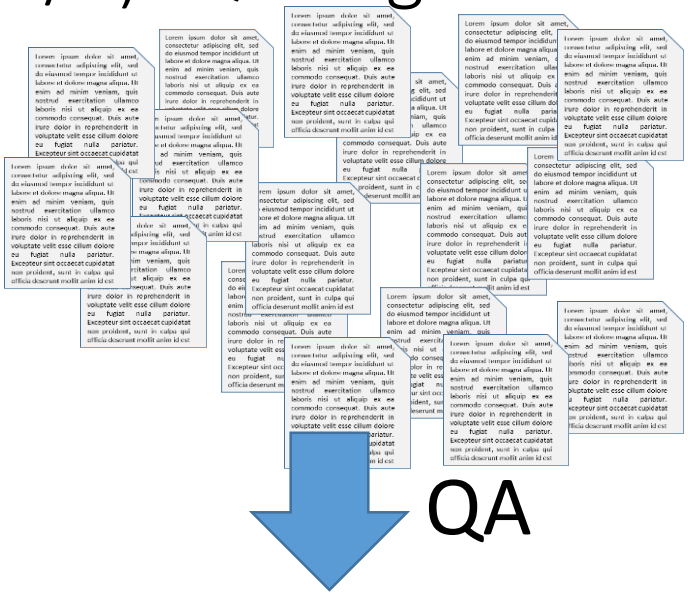
- District(HK, Wan Chai) - confidence 0.93
- District(HK, Kowloon City) - confidence 0.86
- District(HK, Yau Tsim) - confidence 0.74

Accept

- District(HK, Macao) - confidence 0.67
- …

Reject

# Relevance (3/3): QA negation and completeness




- *Which US presidents were married only once?*
- *Which countries participated in no UN mission?*
- *For which cities do we know all districts?*

Without coverage awareness, QA systems cannot answer these

Focus of our research [SIGMOD'15, WSDM'17, ACL'17, ISWC'18, …]

# Coverage estimation how?

Grice's maxims of cooperative communication

[Logic and conversation, 1975]

Maxim of quantity:

- Make your contribution as informative as required

Maxim of relevance:

- Be relevant

*Obama has two children, Malia and Sasha.*

*Jolie brought her children Shiloh and Knox to school.*

Can we automatically determine where full coverage is relevant?

# Formal problem: Full coverage prediction

Subject s

Predicate p

Real-world set of p-objects for s: RW{o | sp}

Ground truth extraction (perfect IE): GTE{o | sp, t}

Given a text segment t:

GTE{o | sp, t} = RW{o | sp}?

If yes →
t has full coverage
of o for s, p

# Setup

GTE{o | sp, t} = RW{o | sp}?

- p ∈ {child, spouse, bandmember, educatedAt, employer}
- s ∈ popular entities from Wikidata
- t: Wikipedia sentences/paragraphs

- GTE{o | sp, t}: OpenIE + predicate dictionary + surface name matching
- RW{o | sp}: Distant supervision
  - Wikidata objects – assumed to be complete for popular s

- Classifier/features:
  - SVM on text n-grams, LSTM on word embeddings

- Baselines:
  - Random
  - Longest text segments are complete
  - Text segments containing most proper names are complete

# Results

| Text unit | Model | child | bandMember | educatedAt |
|---|---|---|---|---|
| Sentence | Random | .06 | .06 | .14 |
| | Length | .05 | .13 | .24 |
| | #pnames | .05 | .17 | .28 |
| | LSTM | **.45** | **.60** | **.64** |
| Paragraph | Random | .12 | | |
| | Length | .17 | | |
| | #pnames | .19 | | |
| | LSTM | **.55** | | |

1. Baselines capture some signal
2. LSTM on text contains much stronger signals
3. Task is not easy

F1-score on predicting text units w/ full coverage.

# Paragraph features

| child | spouse | bandMember | educatedAt |
|---|---|---|---|
| <num> grandsons | married twice | consists of | briefly attended |
| <pname> sons | second marriage | vocals <pname> | left graduating |
| daugthers: <pname> | later married | lineup <pname> | <pname> left |

# Example predictions

| Sentence | LSTM score |
|---|---|
| He was the father of actor Pierre Renoir (1885-1952), filmmaker Jean Renoir (1894-1979) and ceramic artist Claude Renoir (1901-1969). | 0.54 |
| His daughter Julie Gavras and his son Romain Gavras are also filmmakers. | 0.46 |
| Genghis Khan was aware of the friction between his sons (particularly between Chagatai and Jochi) and worried of possible conflict between them if he died. | 0.42 |
| “From this moment I am no longer the king; the king is Victor my son.” | 0.17 |

# Take-home

- IE so far only focused on confidence (precision)
- Coverage (recall) has importance for resource efficiency, thresholding, QA w/ negation
- Linguistic theories give handles towards coverage estimation
- Experiments:
  - Coverage estimation is feasible
  - N-grams provide informative features

max planck institut informatik

HPI