# Canonicalizing Open Knowledge Bases

Luis Galarraga, Kevin Murphy, Geremy heitz, Fabian Suchanek

CIKM-2014, Shanghai, China

**Presenter**

**Rudra Pratap Deb Nath**

**9th January 2015**

# PRESENTATION OUTLINE

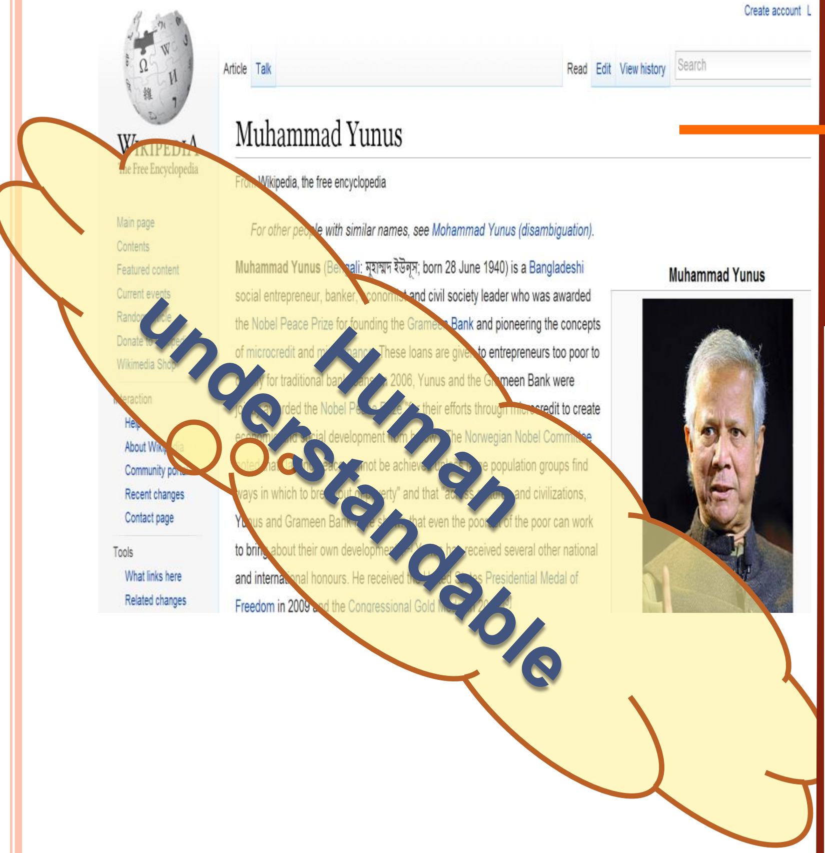
- Motivation
  - Information Extraction
  - Problems in Open Knowledge Bases
  - Contribution
- Canonicalization of Noun Phrases
  - Mention
  - Clustering
  - Similarity Functions
- Canonicalization of Verbal Phrases
  - Procedure
  - Rule Mining
  - Phrase Clustering
  - Canonicalization
- Experiment
  - Evaluation Metrics
  - Results of Entity Clustering
  - Results of Relation Clustering
- Conclusion

# Presentation Outline

- **Motivation**
  - **Information Extraction**
  - Problems in Open Knowledge Bases
  - Contribution
- Canonicalization of Noun Phrases
  - Mention
  - Clustering
  - Similarity Functions
- Canonicalization of Verbal Phrases
  - Procedure
  - Rule Mining
  - Phrase Clustering
  - Canonicalization
- Experiment
  - Evaluation Metrics
  - Results of Entity Clustering
  - Results of Relation Clustering
- Conclusion

# Information Extraction (IE)

Natural Languages



Muhammad Yunus

From Wikipedia, the free encyclopedia

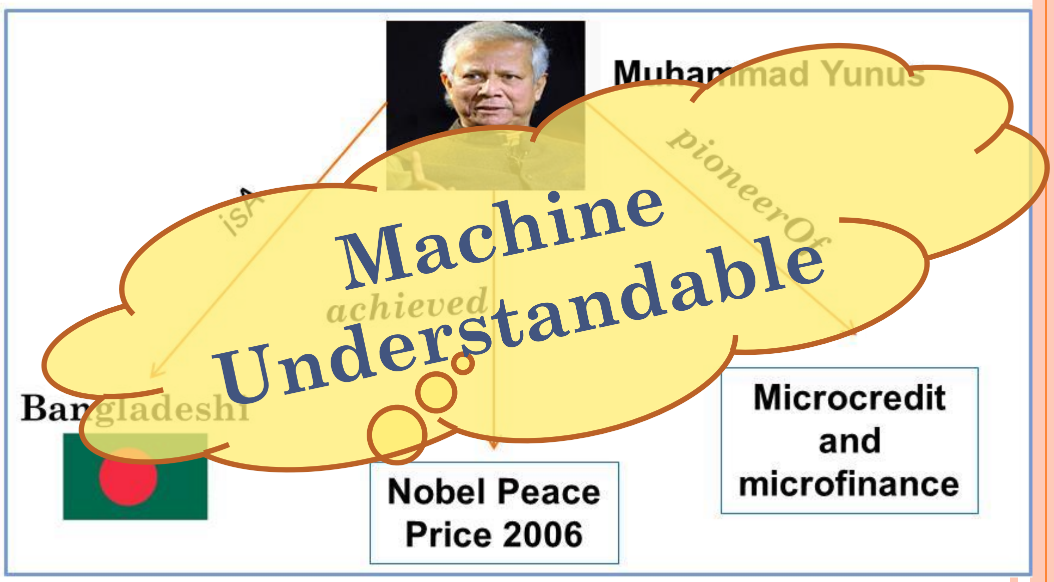
*For other people with similar names, see Mohammad Yunus (disambiguation).*

**Muhammad Yunus** (Bengali: মুহাম্মদ ইউনূস; born 28 June 1940) is a Bangladeshi social entrepreneur, banker, economist and civil society leader who was awarded the Nobel Peace Prize for founding the Grameen Bank and pioneering the concepts of microcredit and microfinance. These loans are given to entrepreneurs too poor to qualify for traditional bank loans. In 2006, Yunus and the Grameen Bank were jointly awarded the Nobel Peace Prize "for their efforts through microcredit to create economic and social development from below". The Norwegian Nobel Committee noted that "lasting peace can not be achieved unless large population groups find ways in which to break out of poverty" and that "across cultures and civilizations, Yunus and Grameen Bank have shown that even the poorest of the poor can work to bring about their own development." Yunus has received several other national and international honours. He received the United States Presidential Medal of Freedom in 2009 and the Congressional Gold Medal in 2010[3]

Muhammad Yunus

Information Extraction

Knowledge Bases



# Popular Knowledge Bases

# Types of IE

| Closed IE | Open IE |
|---|---|
| Domain is known beforehand | Coverage is much more bigger than Closed IE. |
| Applied to semi- structured sources | Applied  to natural language text |
| High precision, canonicalized | Dirty, non-canonicalized |
| Known schema. | Extraction of shema free facts |
| YAGO, Freebase, DBpedia | ReVerb |

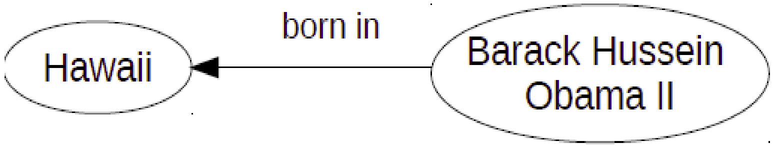
# Presentation Outline

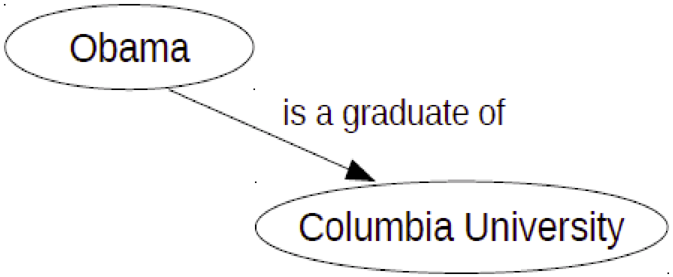
- **Motivation**
  - Information Extraction
  - **Problems in Open Knowledge Bases**
  - Contribution
- Canonicalization of Noun Phrases
  - Mention
  - Clustering
  - Similarity Functions
- Canonicalization of Verbal Phrases
  - Procedure
  - Rule Mining
  - Phrase Clustering
  - Canonicalization
- Experiment
  - Evaluation Metrics
  - Results of Entity Clustering
  - Results of Relation Clustering
- Conclusion

# Problems of Open Knowledge Bases

### Barack Hussein Obama II

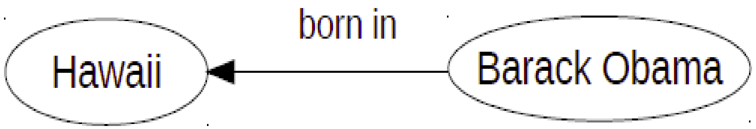
Born in Honolulu, Hawaii. Obama is a graduate of Columbia University and Harvard Law School.

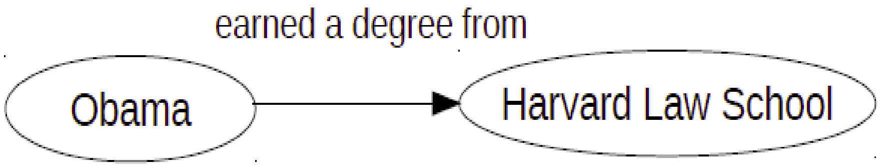




### Barack Obama

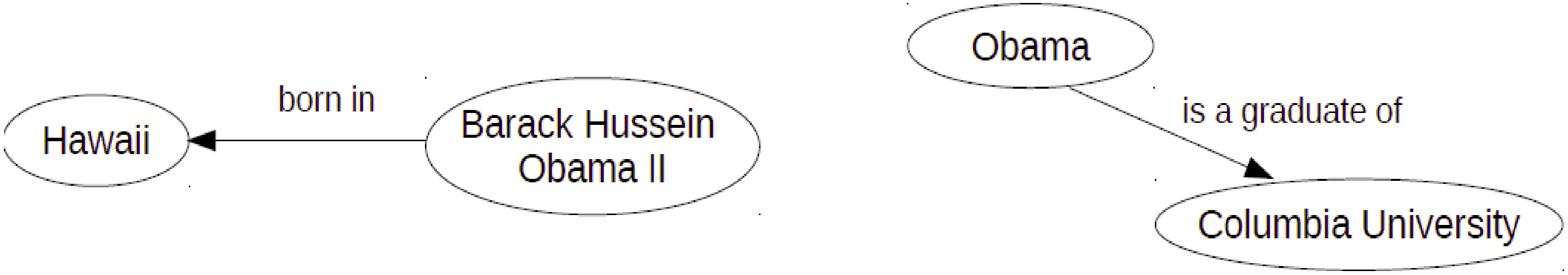
Born in Hawaii. Obama earned a degree from Harvard Law School.



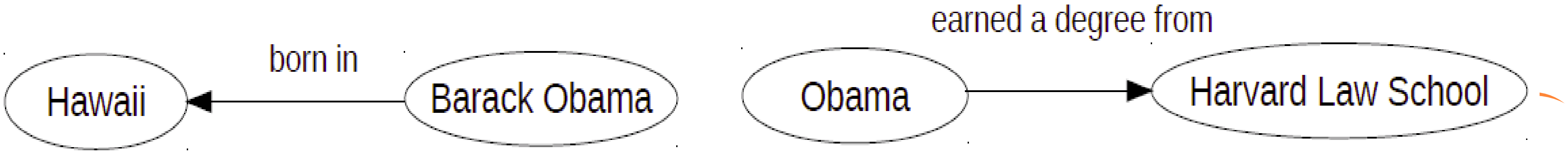


# Problems of Open Knowledge Bases

Barak Hussein Obama II earned a degree from ?

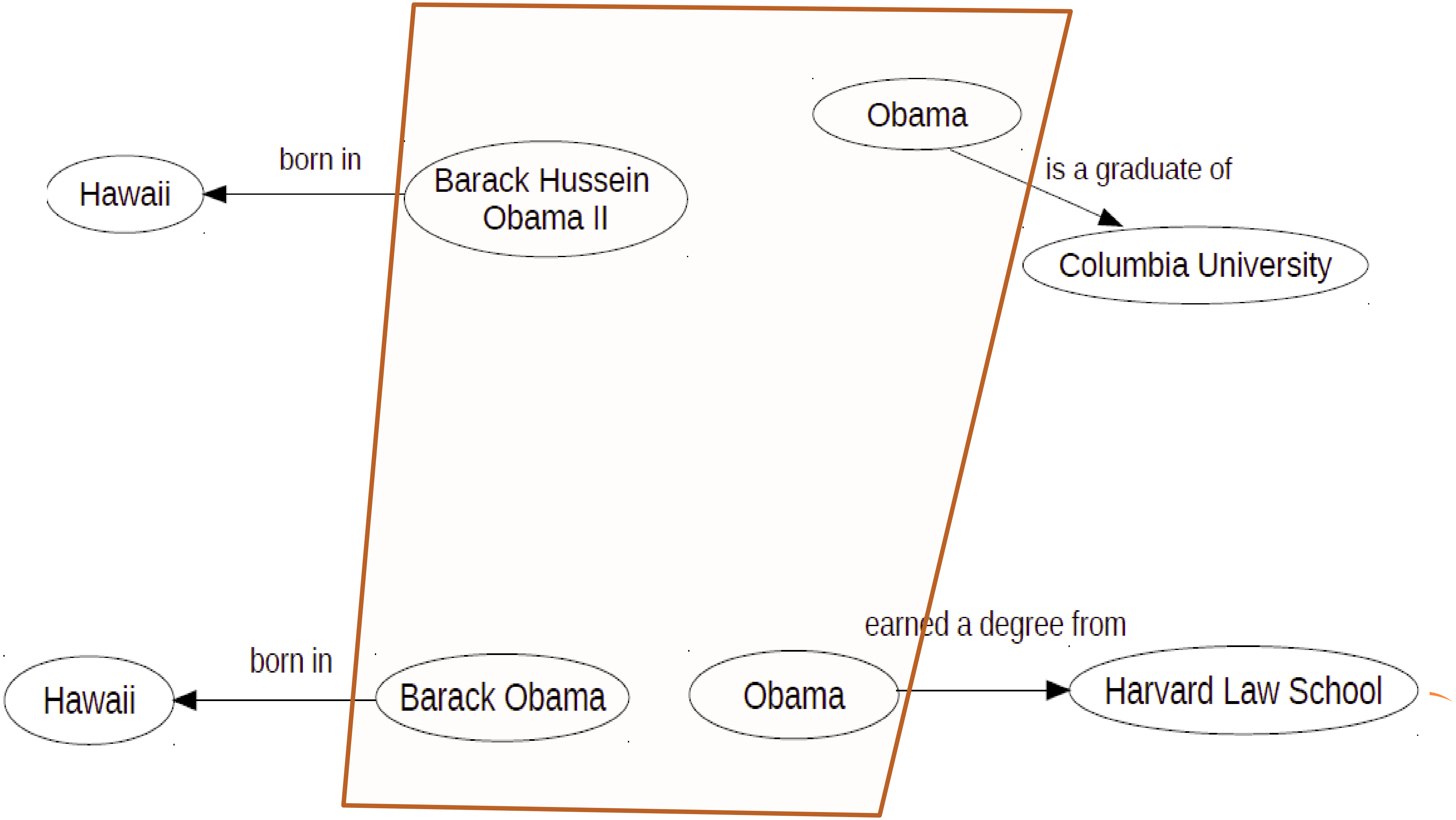


**Not Canonical implies problem for Query Answering**



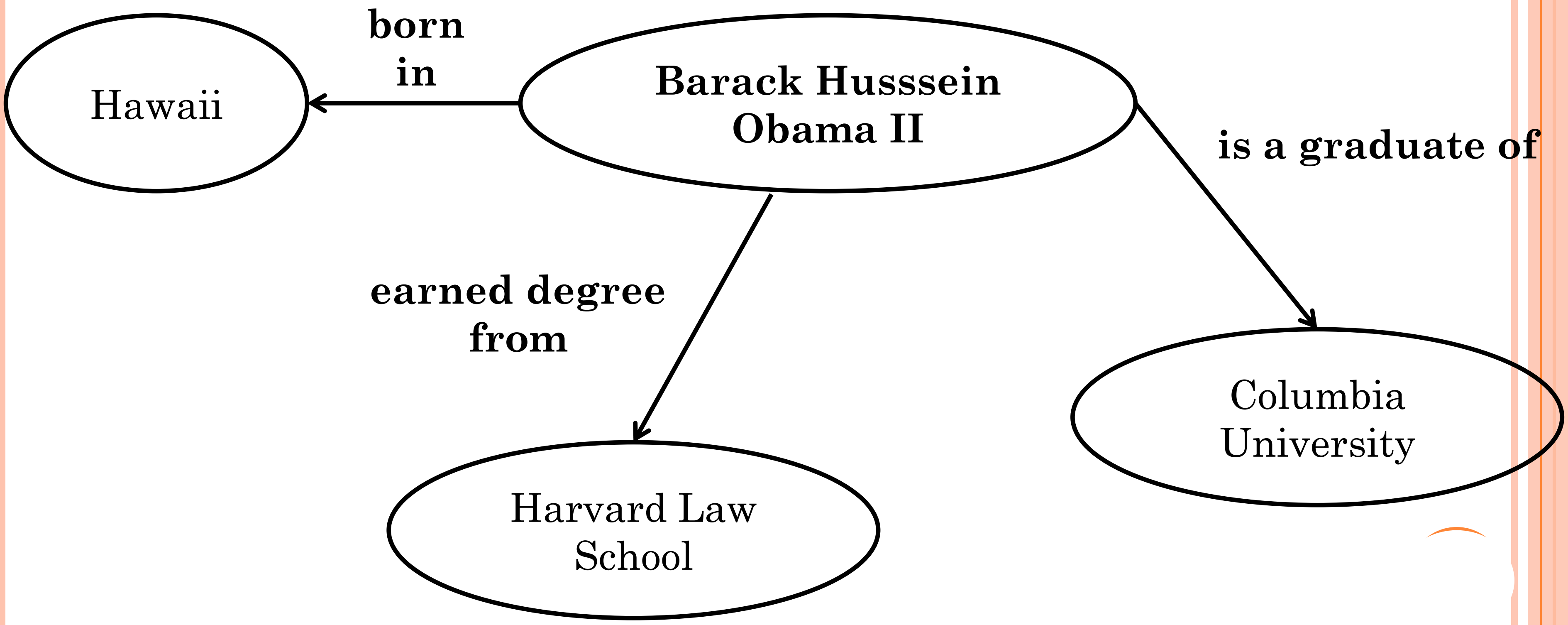
# Problems of Open Knowledge Bases

Barak Hussein Obama II earned a degree from ?



# Problems of Open Knowledge Bases

Barak Hussein Obama II earned a degree from ?

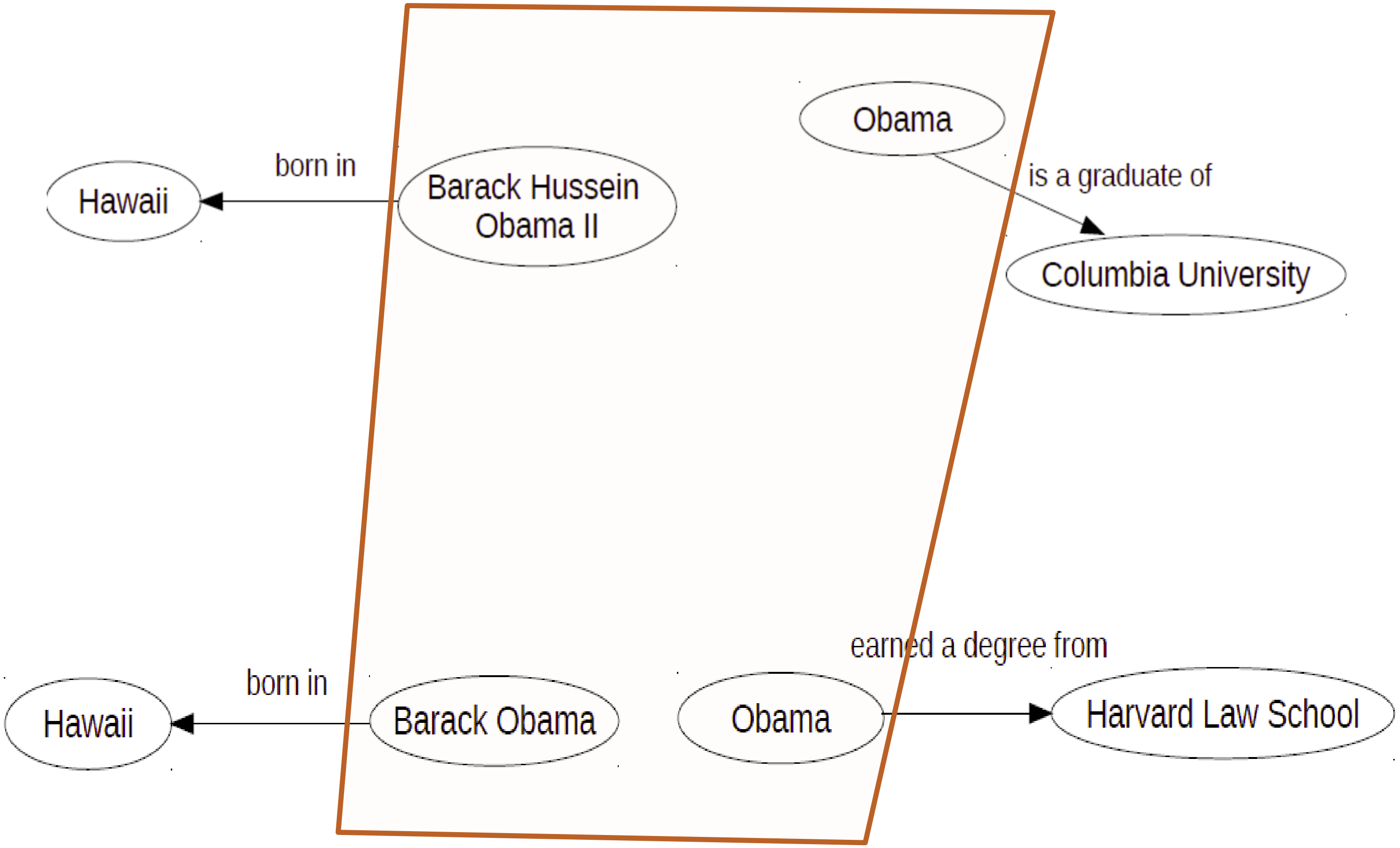


# Presentation Outline

- **Motivation**
  - Information Extraction
  - Problems in Open Knowledge Bases
  - **Contribution**
- Canonicalization of Noun Phrases
  - Mention
  - Clustering
  - Similarity Functions
- Canonicalization of Verbal Phrases
  - Procedure
  - Rule Mining
  - Phrase Clustering
  - Canonicalization
- Experiment
  - Evaluation Metrics
  - Results of Entity Clustering
  - Results of Relation Clustering
- Conclusion

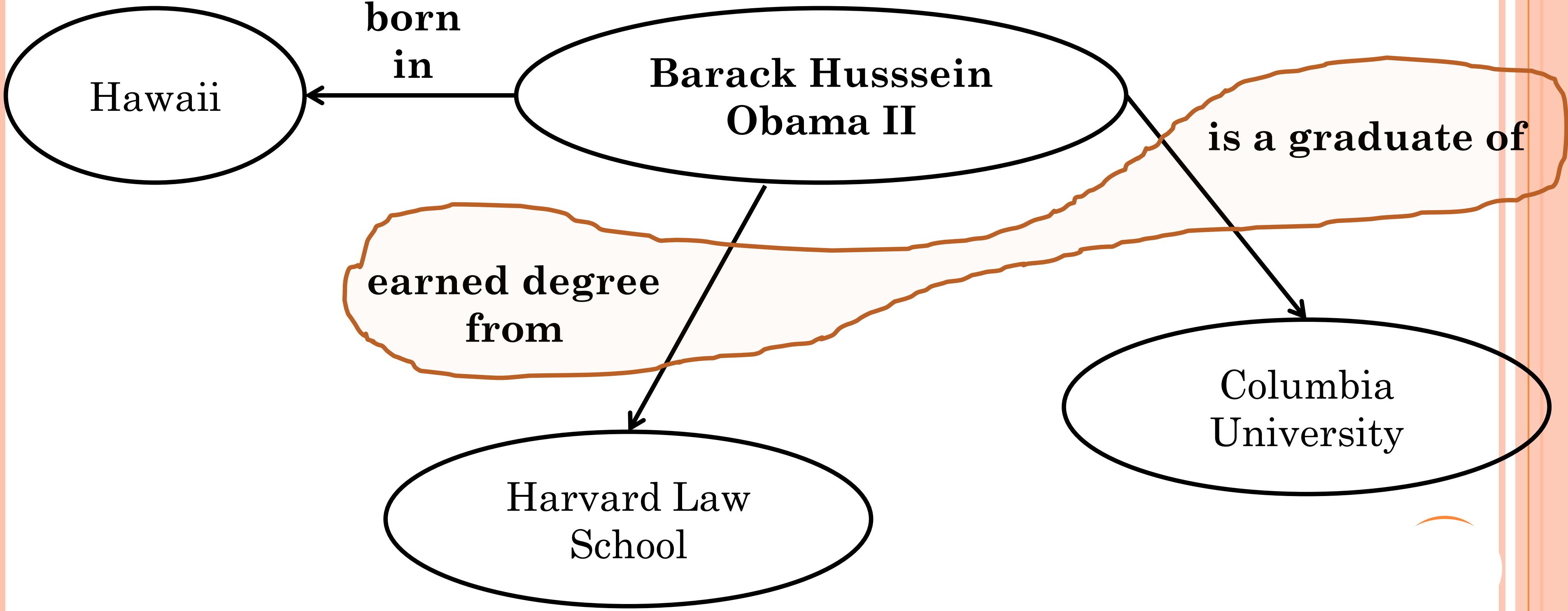
# CONTRIBUTION

❖ To Canonicalize entities (subject and objects)
- clustering technique with simple blocking and similarity function



# CONTRIBUTION

- To identify synonym verbal phrases (predicates)
  - use of rule mining

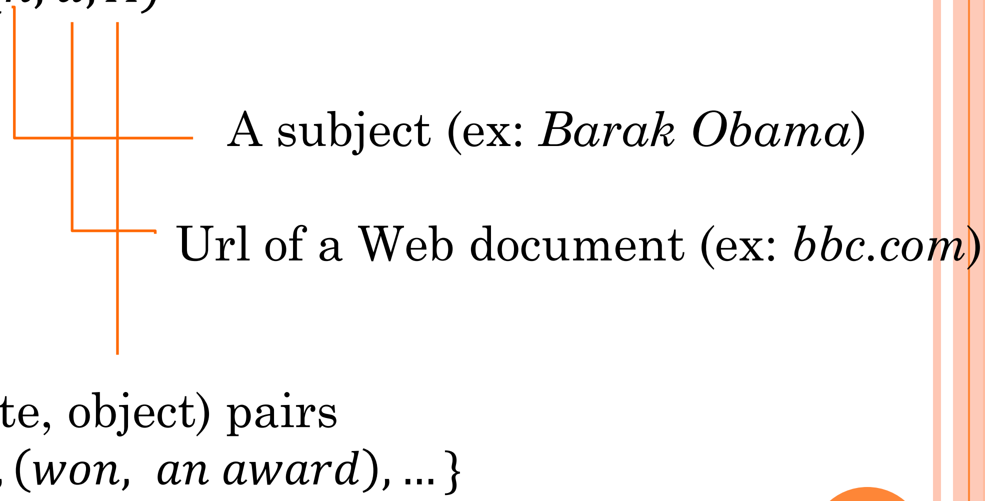


# Presentation Outline

- Motivation
  - Information Extraction
  - Problems in Open Knowledge Bases
  - Contribution
- **Canonicalization of Noun Phrases**
  - **Mention**
  - Clustering
  - Similarity Functions
- Canonicalization of Verbal Phrases
  - Procedure
  - Rule Mining
  - Phrase Clustering
  - Canonicalization
- Experiment
  - Evaluation Metrics
  - Results of Entity Clustering
  - Results of Relation Clustering
- Conclusion

# MENTIONS

- One-sense-per-category assumption
  - Same string in two different pages may have different meanings.
  - Same subject in same Web page refers to the same entity
- Mention defines the profile of a noun phrase *(n)* in a particular Web source *(u)*.
- Represented as a triple $m = (n, u, A)$



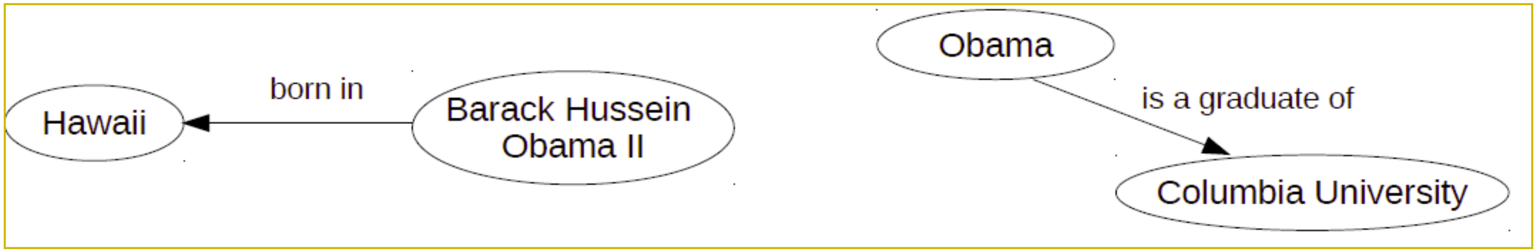
A subject (ex: *Barak Obama*)

Url of a Web document (ex: *bbc.com*)

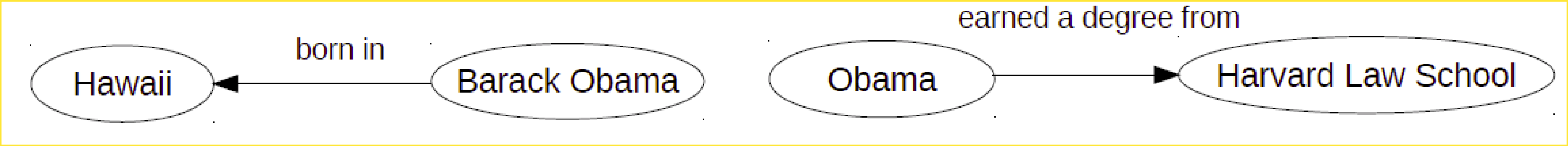
Set of (predicate, object) pairs

ex: $\{(born\ in, Hawii), (won,\ an\ award), \ldots\}$

# MENTIONS

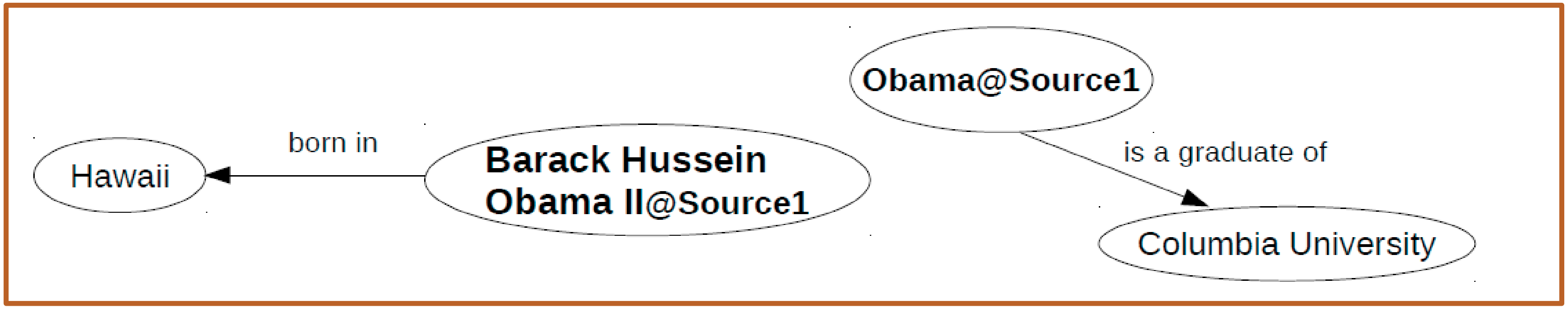


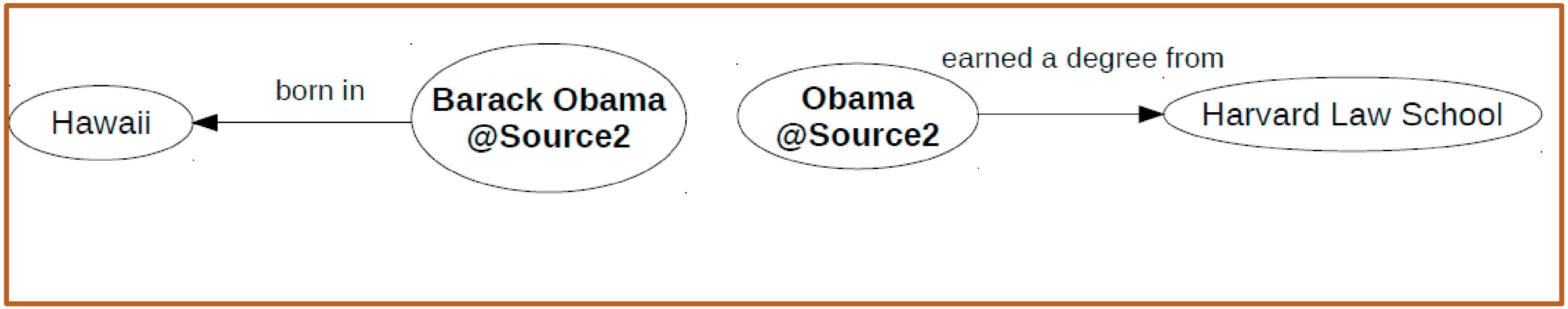
Source 1



Source 2

# MENTIONS

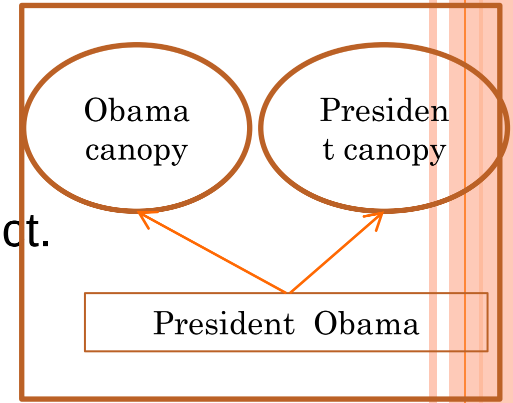

Obama@Source1
born in
is a graduate of
Hawaii
Barack Hussein Obama II@Source1
Columbia University


earned a degree from
born in
Hawaii
Barack Obama @Source2
Obama @Source2
Harvard Law School

# Presentation Outline

- Motivation
  - Information Extraction
  - Problems in Open Knowledge Bases
  - Contribution
- **Canonicalization of Noun Phrases**
  - Mention
  - **Clustering**
  - Similarity Functions
- Canonicalization of Verbal Phrases
  - Procedure
  - Rule Mining
  - Phrase Clustering
  - Canonicalization
- Experiment
  - Evaluation Metrics
  - Results of Entity Clustering
  - Results of Relation Clustering
- Conclusion

# CLUSTERING



Obama canopy

Presiden t canopy

President Obama

- Objectives
  - Partition the set of mentions.
  - Mentions of same partition refer to same real world object.
- Hierarchical Agglomerative Clustering (HAC)
  - High computational complexity ($O(N^3)$)
- Assign each mention to one or several groups, called canopies
  - Based on words of subject
  - Same subject may be in different canopy

    $(\boldsymbol{Mumbai}, is\ located\ in,\ the\ Republic\ of\ india)$

    $(\boldsymbol{Bombay}, is\ a\ city\ in,\ india)$
  - Two mentions $m_1 = (n_1, m_1, A_1)\ and\ m_2 = (n_2, m_2, A_2)$ are in same canopy if

    (1) their subjects share a non-stopword or

    (2) two objects across mentions share a word.
- Apply HAC on each canopy.

# Presentation Outline

- Motivation
  - Information Extraction
  - Problems in Open Knowledge Bases
  - Contribution
- **Canonicalization of Noun Phrases**
  - Mention
  - Clustering
  - **Similarity Functions**
- Canonicalization of Verbal Phrases
  - Procedure
  - Rule Mining
  - Phrase Clustering
  - Canonicalization
- Experiment
  - Evaluation Metrics
  - Results of Entity Clustering
  - Results of Relation Clustering
- Conclusion

# Similarity Functions

- Similarity between two mentions $m = (n, u, A)$ and $m' = (n', u', A')$
- Using Simple features and Corpus features
- Simple features (using KB triples)
  - Attribute Overlap :

    - $f_{attr}(m, m') \coloneqq jacard(A, A')$ $[jacard(S, S') = \frac{|S \cap S'|}{|S \cup S'|}]$

    - $(p, o) \in A$ and $(p', o') \in A'$ are equal if $p = p'$ and $o = o'$

  - String Similarity :

    - $f_{strsm}(m, m') \coloneqq jarowinkler\ (n, n')$

  - String Identity : special case of string similarity

    - $f_{strid}(m, m') = \begin{cases} 1, & if\ n = n' \\ 0, & else \end{cases}$

# Similarity Functions

- IDF Token Overlap :

  -$f_{itol} = \frac{\sum_{w \in w(n) \cap w(n')} \log(1+df(w))^{-1}}{\sum_{w \in w(n) \cup w(n')} \log(1+df(w))^{-1}}$

  - $w(.) \; - set\ of\ words\ of\ a\ string$

  - $df(w) - the\ frequency\ of\ the\ word\ in$

  $the\ subjects\ and\ objects\ of\ the\ OpenIE\ triples$

- Corpus features

- Word Overlap :

  - $f_{wol}(m, m') = jaccard(t(u), t(u'))$

  -$t(.) - is\ the\ set\ of\ top\ 100\ words\ on\ a\ page\ (ranked\ by\ TF - IDF)$

- Entity Overlap : word may be ambiguous

  - $f_{eol}(m, m') = jaccard(e(u), e(u'))$

  -e$(u) - is\ the\ set\ linked\ Freebase\ entities\ on\ the\ page\ u$.

# Similarity Functions

- Type Overlap :

  - $f_{tol}(m,m') = jaccard(types(\pi_{pred}(A), \pi_{pred}(A'))$

- Combined Feature

  -$f_{ml}(m,m') = \frac{1}{1+ e^{-f_{sim}(m,m')}}$

  -$f_{sim}(m,m') = c_o + \sum_{i=1}^{N} c_i f_i\,(m,m')$

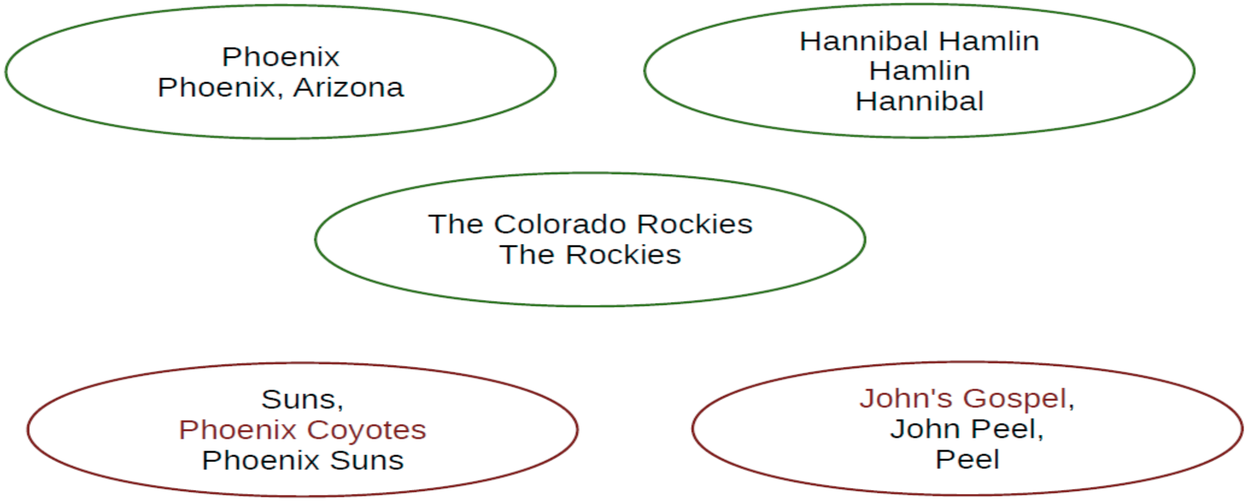
  -$f_i - is\ the\ similarity\ function$

  - $c_i - determined\ by\ traning\ a\ logistic\ regression\ classifier.$

  -Simple ML includes the similartiy functions of simple features

  -Full ML includes the similarity functions of simple and corpus features

# Similarity Functions



Phoenix
Phoenix, Arizona

Hannibal Hamlin
Hamlin
Hannibal

The Colorado Rockies
The Rockies

Suns,
Phoenix Coyotes
Phoenix Suns

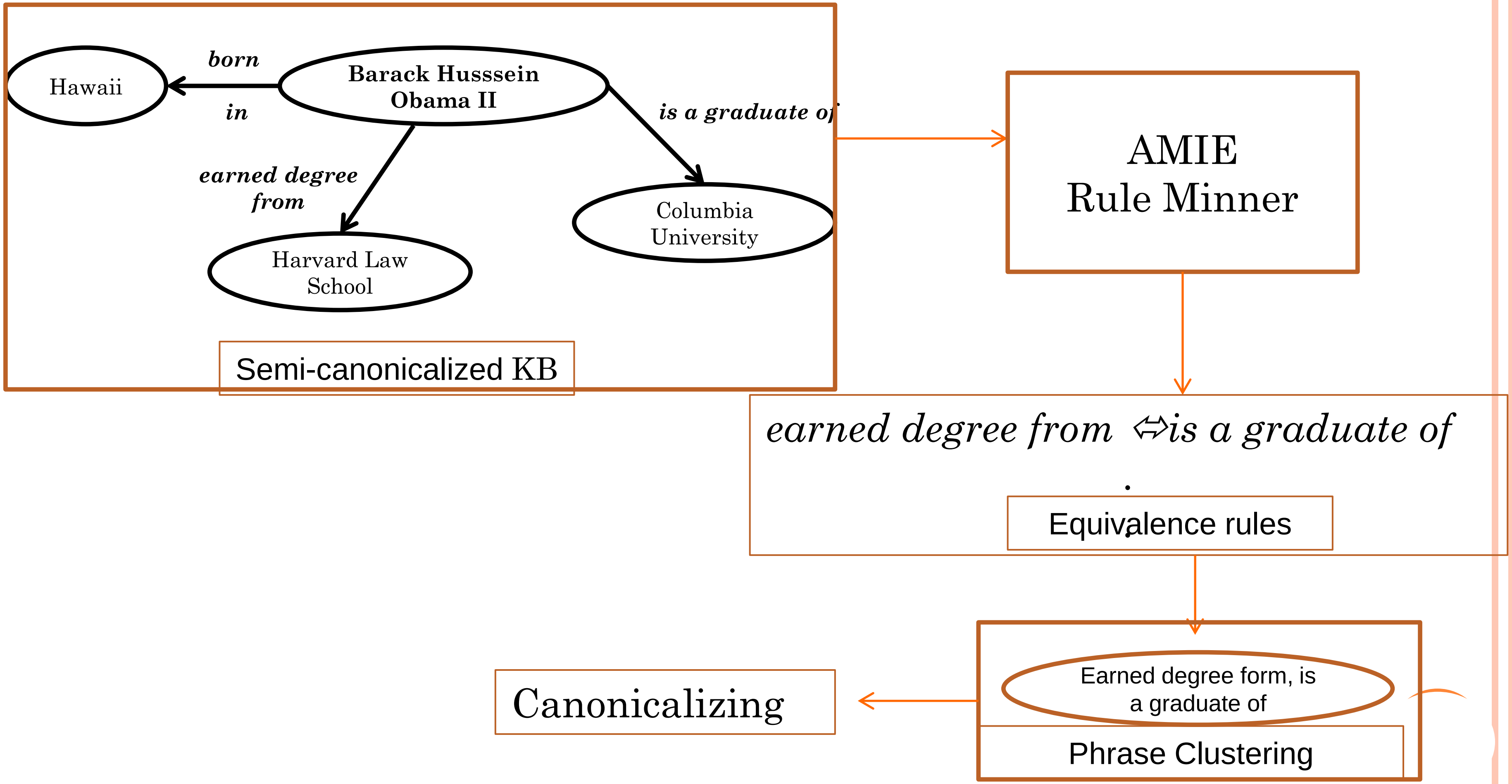
John's Gospel,
John Peel,
Peel

- Similarity between two clusters is calculated using the single linkage criterion
- Canonicalization : procedure of selecting a representive noun phrase for a cluster
    - noun pharase that have highest frequenceies in different web sources
    - If there is a tie, selectes the longest noun phrase

# Presentation Outline

- Motivation
  - Information Extraction
  - Problems in Open Knowledge Bases
  - Contribution
- Canonicalization of Noun Phrases
  - Mention
  - Clustering
  - Similarity Functions
- **Canonicalization of Verbal Phrases**
  - **Procedure**
  - Rule Mining
  - Phrase Clustering
  - Canonicalization
- Experiment
  - Evaluation Metrics
  - Results of Entity Clustering
  - Results of Relation Clustering
- Conclusion

# Canonicalizing verbal phrases



Semi-canonicalized KB

AMIE
Rule Minner

*earned degree from ⇔is a graduate of*

Equivalence rules

Earned degree form, is a graduate of

Phrase Clustering

Canonicalizing

# SEMI-CANONICALIZED KB

- Canonicalized subjects and objects of OpenIE triples
- Two ways to canonicalize subjects and objects
  - Using the noun phrase clustering technique
  - Mappings to Freebase
    - Consider the subset of ReVerb triples whose subjects are linked to Freebase
    - String matching can be used to canonicalize object

# Presentation Outline

- Motivation
  - Information Extraction
  - Problems in Open Knowledge Bases
  - Contribution
- Canonicalization of Noun Phrases
  - Mention
  - Clustering
  - Similarity Functions
- **Canonicalization of Verbal Phrases**
  - Procedure
  - **Rule Mining**
  - Phrase Clustering
  - Canonicalization
- Experiment
  - Evaluation Metrics
  - Results of Entity Clustering
  - Results of Relation Clustering
- Conclusion

# Rule Mining : AMIE algorithm

- Open IE relations $r = graduate\ of$ and $r' = earned\ degree\ from$
- Objective is to discover

$$\forall x, y: r(x, y) \Leftrightarrow r'(x, y) \quad \text{i.e } r \sqsubset r'\ and\ r' \sqsubset r$$

- Not all triples in $r$ are in $r'$
- Sparse relation may contain same subject and predicate may not reflect equivalance.
  - *'s stepdaughter ⊏ married?*

$$(Woody\ Allen,\ married,\ Soon - Yi\ Previn)$$

$$(Woody\ Allen, 's\ stepdaughter, Soon - Yi\ Previn)\ )$$

# Rule Mining : AMIE algorithm

- AMIE stands for Association rule mining under incomplete evidence.
- Learns Horn rules

$$\underbrace{marriedTo(x,z) \wedge livesIn(z,y)}_{\text{Body}} \Rightarrow \underbrace{livesIn(x,y)}_{\text{Head}}$$

- Subsumption relation can be expressed as a Horn rule in the form of

$$r(x,y) \Rightarrow r'(x,y)$$

- Support of the rule

$$supp\left(\vec{B} \Rightarrow r(x,y)\right) \coloneqq \#(x,y): \exists_{z_1,\ldots\ldots\ldots z_m}: \vec{B} \wedge r(x,y)$$

  - $\#(x,y): A$ is the number of pairs $(x,y)$ that fulfills $A$.

- Confidence of the rule

$$pcaconf\left(\vec{B} \Rightarrow r(x,y)\right) \coloneqq \frac{supp\left(\vec{B} \Rightarrow r(x,y)\right)}{\#(x,y'): \exists_{z_1,\ldots\ldots\ldots z_m}: \vec{B} \wedge r(x,y')}$$

# Presentation Outline

- Motivation
  - Information Extraction
  - Problems in Open Knowledge Bases
  - Contribution
- Canonicalization of Noun Phrases
  - Mention
  - Clustering
  - Similarity Functions
- **Canonicalization of Verbal Phrases**
  - Procedure
  - Rule Mining
  - **Phrase Clustering**
  - Canonicalization
- Experiment
  - Evaluation Metrics
  - Results of Entity Clustering
  - Results of Relation Clustering
- Conclusion

# Phrase Clustering

- The output of AMIE is a set of equivalence rules

*is a graduate of* ⇔ *earned degree from*
*graduated from* ⇔ *is a graduate of*
*belongs to* ⇔ *is part of*
*belongs to* ⇔ *created*

- Equivalence relation is trasitive.
- Itteratively merging of equivalance mappings with at least one verbhal phrase in common makes clusters.

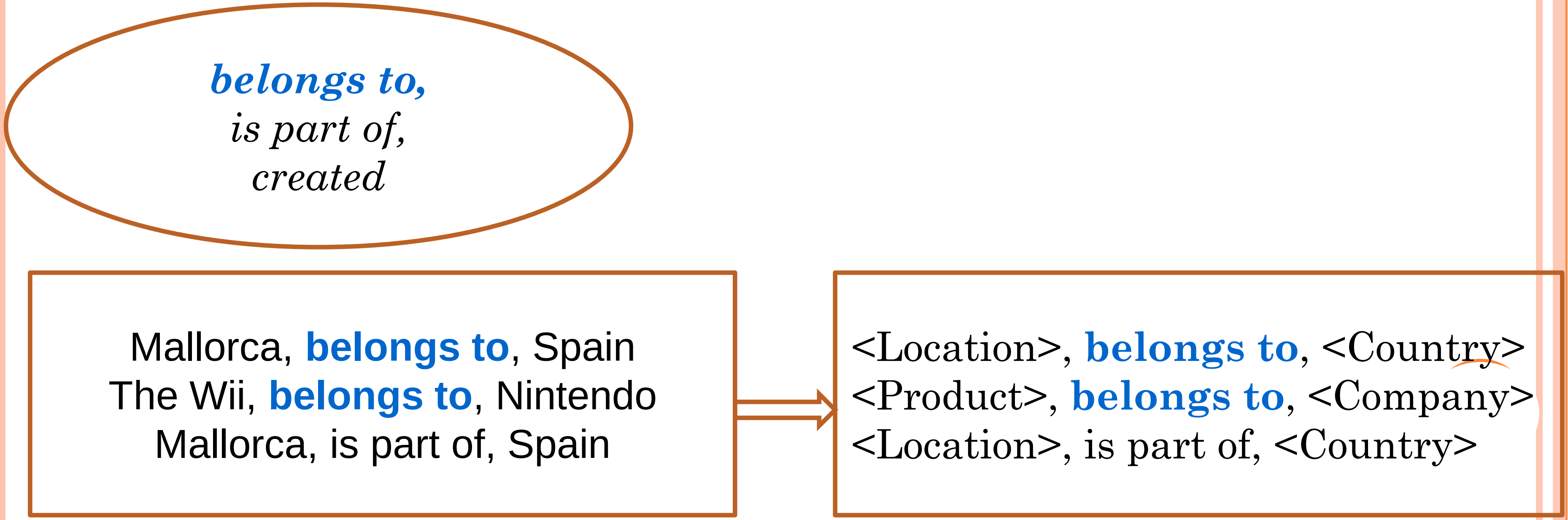
*is a graduate of ,*
*earned degree from,*
*graduated from*

*belongs to,*
*is part of,*
*created*

# PHRASE CLUSTERING

- ❖ Verbal phrase can covey different meanings.
- ❖ Entities are augmented with types (only Linked KB)

*is a graduate of ⇔ earned degree from*
*graduated from ⇔ is a graduate of*
*belongs to (Location, Country) ⇔ is part of(Location, Country)*
*belongs to(Product, Company) ⇔ created(Product, Company)*



***belongs to,***
*is part of,*
*created*

Mallorca, **belongs to**, Spain
The Wii, **belongs to**, Nintendo
Mallorca, is part of, Spain

<Location>, **belongs to**, <Country>
<Product>, **belongs to**, <Company>
<Location>, is part of, <Country>

# Presentation Outline

- Motivation
  - Information Extraction
  - Problems in Open Knowledge Bases
  - Contribution
- Canonicalization of Noun Phrases
  - Mention
  - Clustering
  - Similarity Functions
- **Canonicalization of Verbal Phrases**
  - Procedure
  - Rule Mining
  - Phrase Clustering
  - **Canonicalization**
- Experiment
  - Evaluation Metrics
  - Results of Entity Clustering
  - Results of Relation Clustering
- Conclusion

# CANONICALIZATION

- Canonicalize verbal phrase by mapping them to Freebase relations.
- ROSA (Rules for Ontology Schema Alignment ) approace
- Restrict to the subset of triples whose subjects and objects are linked to Freebase
- AMIE is used to mine the rule

$$vp(x,y) \Leftrightarrow fr(x,y)$$

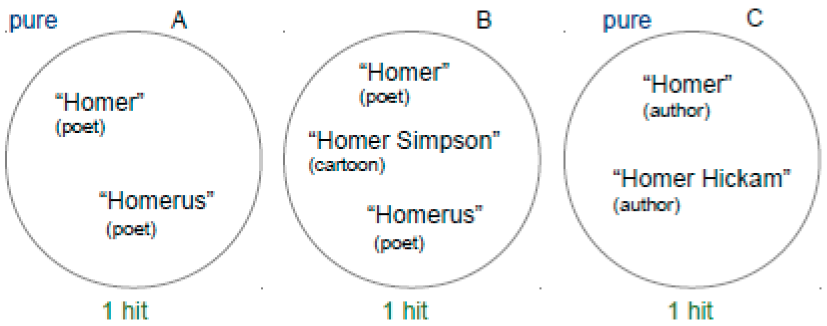
$vp$ is the verbal phrase; $x\ and\ y$ are the Freebase entity, $fr$ is the freebase relation

be the birth place of,
be the hometown of
**f:location.location.people_born_here**

# PRESENTATION OUTLINE

- Motivation
  - Information Extraction
  - Problems in Open Knowledge Bases
  - Contribution
- Canonicalization of Noun Phrases
  - Mention
  - Clustering
  - Similarity Functions
- Canonicalization of Verbal Phrases
  - Procedure
  - Rule Mining
  - Phrase Clustering
  - Canonicalization
- **Experiment**
  - **Evaluation Metrics**
  - Results of Entity Clustering
  - Results of Relation Clustering
- Conclusion

# EVALUATION METRICS



|M|=7 mentions

|E|=3 Freebase entities

|C|= 3 clusters

- Macro-analysis :
- Precision : measures the faction of mentions in the cluster linked to the same entity.

  - $precision_{macro}(C, E) = \frac{|c \in C : \exists_{=1}\ e \in E : e \supseteq c|}{|C|}$

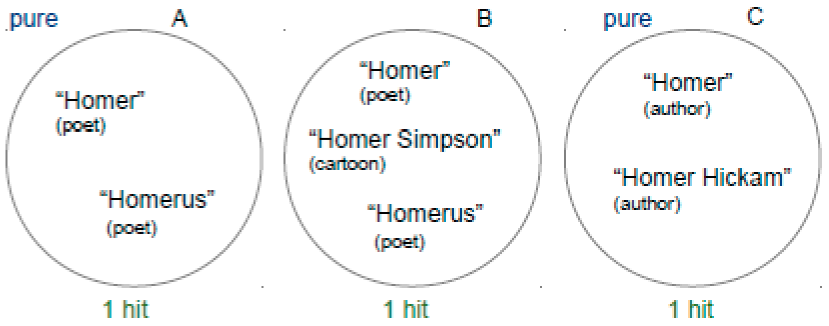
    Example : (1+1)/3=2/3

- Recall : measures the fraction of Freebase entities that is assigned to unique cluster

  - $Recall_{macro}(C, E) = precision_{macro}(C, E)$

    - Example : (1+1)/3 =2/3

# EVALUATION METRICS



|M|=7 mentions

|E|=3 Freebase entities

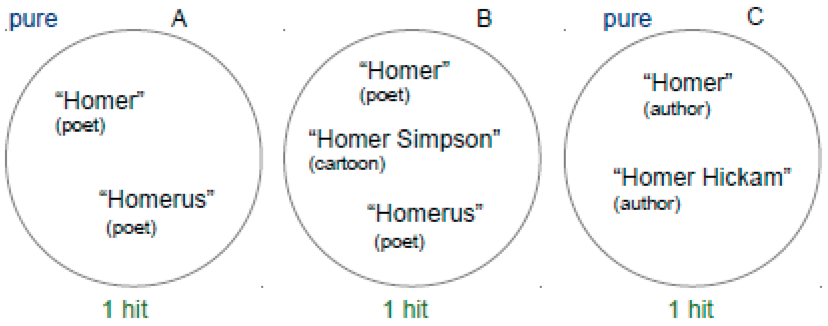
|C|= 3 clusters

- ❖ Micro-analysis
- ❖ Precision: Most frequent Freebase entity of the mention in a cluster is the correct entity.
  - ➢ $precision_{micro}(C,E) = \frac{1}{N}\sum_{c \in C} max_{e \in E}|c \cap e|$

    Example : (2+2+2)/7=6/7
- ❖ Recall :
  - ➢ $Recall_{macro}(C,E) = precision_{macro}(C,E)$

    Example: (2+2+1)/7=5/7

# EVALUATION METRICS



|M|=7 mentions

|E|=3 Freebase entities

|C|= 3 clusters

- ❖ Pairwise-analysis : Two mention from same cluster produce a hit if they refer to same Freebase entity
- ❖ Precision:
  - ➢ $precision_{pairwise}(C,E) = \frac{\sum_{c \in C} \#hits_c}{\sum_{c \in C} \#pairs_c}$

    Example: 3/5
- ❖ Recall:
  - ➢ $recall_{pairwise}(C,E) = \frac{\sum_{c \in C} \#hits_c}{\sum_{e \in E} \#pairs_c}$

# Presentation Outline

- Motivation
    - Information Extraction
    - Problems in Open Knowledge Bases
    - Contribution
- Canonicalization of Noun Phrases
    - Mention
    - Clustering
    - Similarity Functions
- Canonicalization of Verbal Phrases
    - Procedure
    - Rule Mining
    - Phrase Clustering
    - Canonicalization
- **Experiment**
    - Evaluation Metrics
    - **Results of Entity Clustering**
    - Results of Relation Clustering
- Conclusion

# DATASET

- Triples are extracted from ClueWeb09
  - ReVerb extracts 3M triples and subjects of 1.5M triples are linked to Freebase

    -Base Dataset : 150 Freebase entities and 8.5K mentions

    -Ambiguous Dataset: 446 Freebase entities and 34K mentions

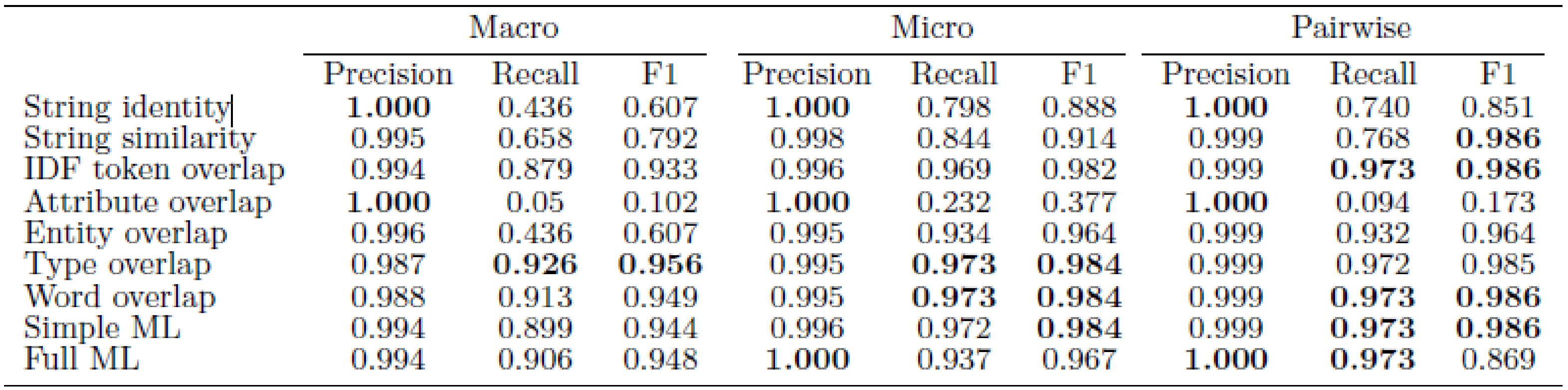
    -Hardware platform : Intel Corei7 with 16GB of RAM

    -Clustering uses HAC with caopies and without canopies

    -Baseline : String identity
  - NELL extractor (Concept Resolver) extracts 57K triples

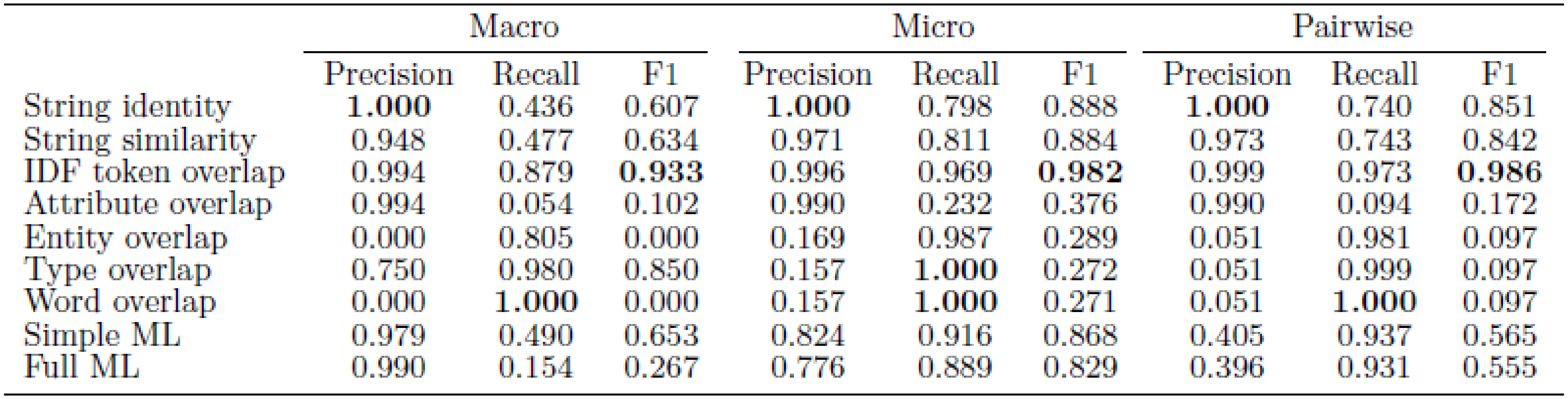
# Results of Entity Clustering

- ❖ Results on Base Dataset : 157 clusters in 54.3 seconds



| | Macro | | | Micro | | | Pairwise | | |
|---|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | F1 | Precision | Recall | F1 | Precision | Recall | F1 |
| String identity | **1.000** | 0.436 | 0.607 | **1.000** | 0.798 | 0.888 | **1.000** | 0.740 | 0.851 |
| String similarity | 0.995 | 0.658 | 0.792 | 0.998 | 0.844 | 0.914 | 0.999 | 0.768 | **0.986** |
| IDF token overlap | 0.994 | 0.879 | 0.933 | 0.996 | 0.969 | 0.982 | 0.999 | **0.973** | **0.986** |
| Attribute overlap | **1.000** | 0.05 | 0.102 | **1.000** | 0.232 | 0.377 | **1.000** | 0.094 | 0.173 |
| Entity overlap | 0.996 | 0.436 | 0.607 | 0.995 | 0.934 | 0.964 | 0.999 | 0.932 | 0.964 |
| Type overlap | 0.987 | **0.926** | **0.956** | 0.995 | **0.973** | **0.984** | 0.999 | 0.972 | 0.985 |
| Word overlap | 0.988 | 0.913 | 0.949 | 0.995 | **0.973** | **0.984** | 0.999 | **0.973** | **0.986** |
| Simple ML | 0.994 | 0.899 | 0.944 | 0.996 | 0.972 | **0.984** | 0.999 | **0.973** | **0.986** |
| Full ML | 0.994 | 0.906 | 0.948 | **1.000** | 0.937 | 0.967 | **1.000** | **0.973** | 0.869 |

Table 1: Precision and recall on ReVerb's *Base* dataset. The highest values in each column are in bold.

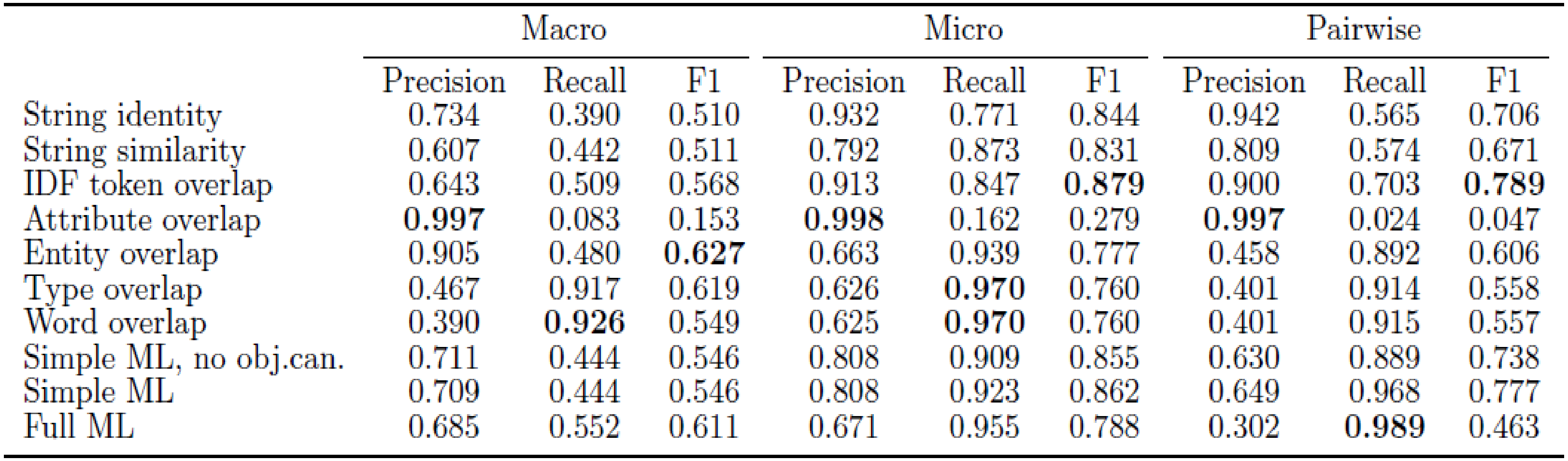


| | Macro | | | Micro | | | Pairwise | | |
|---|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | F1 | Precision | Recall | F1 | Precision | Recall | F1 |
| String identity | **1.000** | 0.436 | 0.607 | **1.000** | 0.798 | 0.888 | **1.000** | 0.740 | 0.851 |
| String similarity | 0.948 | 0.477 | 0.634 | 0.971 | 0.811 | 0.884 | 0.973 | 0.743 | 0.842 |
| IDF token overlap | 0.994 | 0.879 | **0.933** | 0.996 | 0.969 | **0.982** | 0.999 | 0.973 | **0.986** |
| Attribute overlap | 0.994 | 0.054 | 0.102 | 0.990 | 0.232 | 0.376 | 0.990 | 0.094 | 0.172 |
| Entity overlap | 0.000 | 0.805 | 0.000 | 0.169 | 0.987 | 0.289 | 0.051 | 0.981 | 0.097 |
| Type overlap | 0.750 | 0.980 | 0.850 | 0.157 | **1.000** | 0.272 | 0.051 | 0.999 | 0.097 |
| Word overlap | 0.000 | **1.000** | 0.000 | 0.157 | **1.000** | 0.271 | 0.051 | **1.000** | 0.097 |
| Simple ML | 0.979 | 0.490 | 0.653 | 0.824 | 0.916 | 0.868 | 0.405 | 0.937 | 0.565 |
| Full ML | 0.990 | 0.154 | 0.267 | 0.776 | 0.889 | 0.829 | 0.396 | 0.931 | 0.555 |

Table 2: Precision and recall on ReVerb's *Base* dataset, without canopies. Highest values in bold.

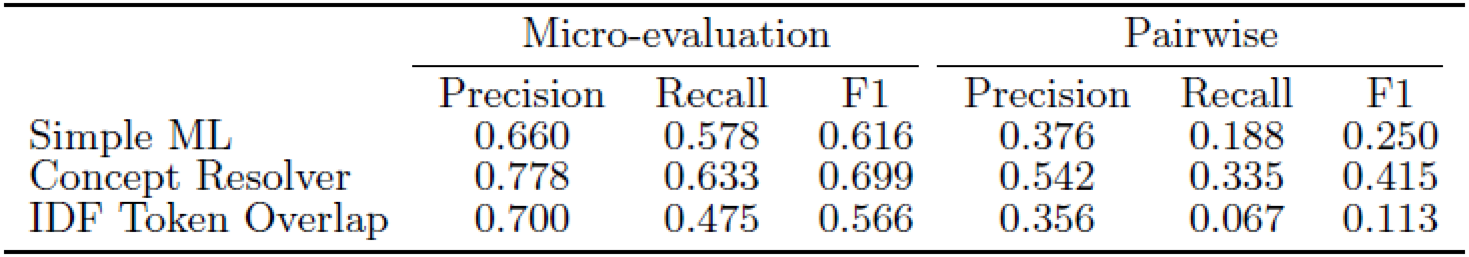
# Results of Entity Clustering

- ❖ Results on Ambiguous Dataset : 823 clusters in 15.04 minutes



| | Macro | | | Micro | | | Pairwise | | |
|---|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | F1 | Precision | Recall | F1 | Precision | Recall | F1 |
| String identity | 0.734 | 0.390 | 0.510 | 0.932 | 0.771 | 0.844 | 0.942 | 0.565 | 0.706 |
| String similarity | 0.607 | 0.442 | 0.511 | 0.792 | 0.873 | 0.831 | 0.809 | 0.574 | 0.671 |
| IDF token overlap | 0.643 | 0.509 | 0.568 | 0.913 | 0.847 | **0.879** | 0.900 | 0.703 | **0.789** |
| Attribute overlap | **0.997** | 0.083 | 0.153 | **0.998** | 0.162 | 0.279 | **0.997** | 0.024 | 0.047 |
| Entity overlap | 0.905 | 0.480 | **0.627** | 0.663 | 0.939 | 0.777 | 0.458 | 0.892 | 0.606 |
| Type overlap | 0.467 | 0.917 | 0.619 | 0.626 | **0.970** | 0.760 | 0.401 | 0.914 | 0.558 |
| Word overlap | 0.390 | **0.926** | 0.549 | 0.625 | **0.970** | 0.760 | 0.401 | 0.915 | 0.557 |
| Simple ML, no obj.can. | 0.711 | 0.444 | 0.546 | 0.808 | 0.909 | 0.855 | 0.630 | 0.889 | 0.738 |
| Simple ML | 0.709 | 0.444 | 0.546 | 0.808 | 0.923 | 0.862 | 0.649 | 0.968 | 0.777 |
| Full ML | 0.685 | 0.552 | 0.611 | 0.671 | 0.955 | 0.788 | 0.302 | **0.989** | 0.463 |

Table 4: Precision and recall on ReVerb's *Ambiguous* dataset. The highest values in each column are in bold.



| | Micro-evaluation | | | Pairwise | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | F1 | Precision | Recall | F1 |
| Simple ML | 0.660 | 0.578 | 0.616 | 0.376 | 0.188 | 0.250 |
| Concept Resolver | 0.778 | 0.633 | 0.699 | 0.542 | 0.335 | 0.415 |
| IDF Token Overlap | 0.700 | 0.475 | 0.566 | 0.356 | 0.067 | 0.113 |

Table 5: Comparison of entity clustering methods on the NELL data.

- ❖ Concept Resolver uses additional inverse and quasi-inverse functions
- ❖ Their relations are canonicalized

# Presentation Outline

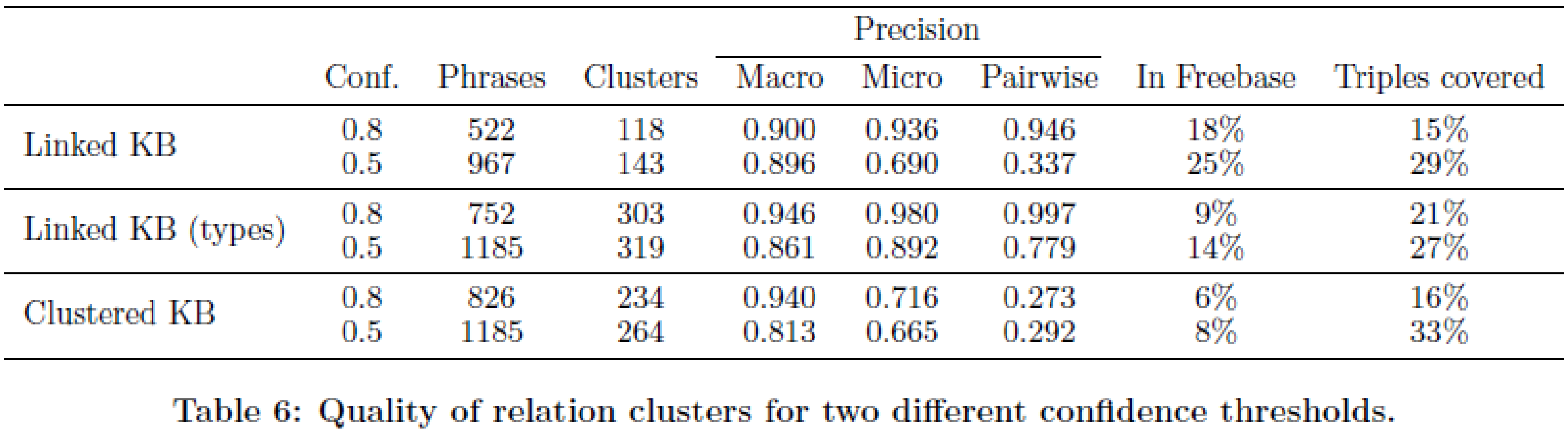
- Motivation
  - Information Extraction
  - Problems in Open Knowledge Bases
  - Contribution
- Canonicalization of Noun Phrases
  - Mention
  - Clustering
  - Similarity Functions
- Canonicalization of Verbal Phrases
  - Procedure
  - Rule Mining
  - Phrase Clustering
  - Canonicalization
- **Experiment**
  - Evaluation Metrics
  - Results of Entity Clustering
  - **Results of Relation Clustering**
- Conclusion

# CONCLUSION

- Linked KB : subject and objects are linked to Freebase
- Clustered KB: canonicalized subject
- Linked KB (type) : Linked KB is augmented with type (to alleviate polsymeous issues)
- AMIE was run using a support threshold of 5
  - Linked KB has 33215 relations
  - Cluster KB has 17259 relations

# Conclusion

- Results: AMIE mined equivalance rules
- 3.5K in Clusterd KB
- 4.3K in Linked KB
- 22K in Linked KB (types)

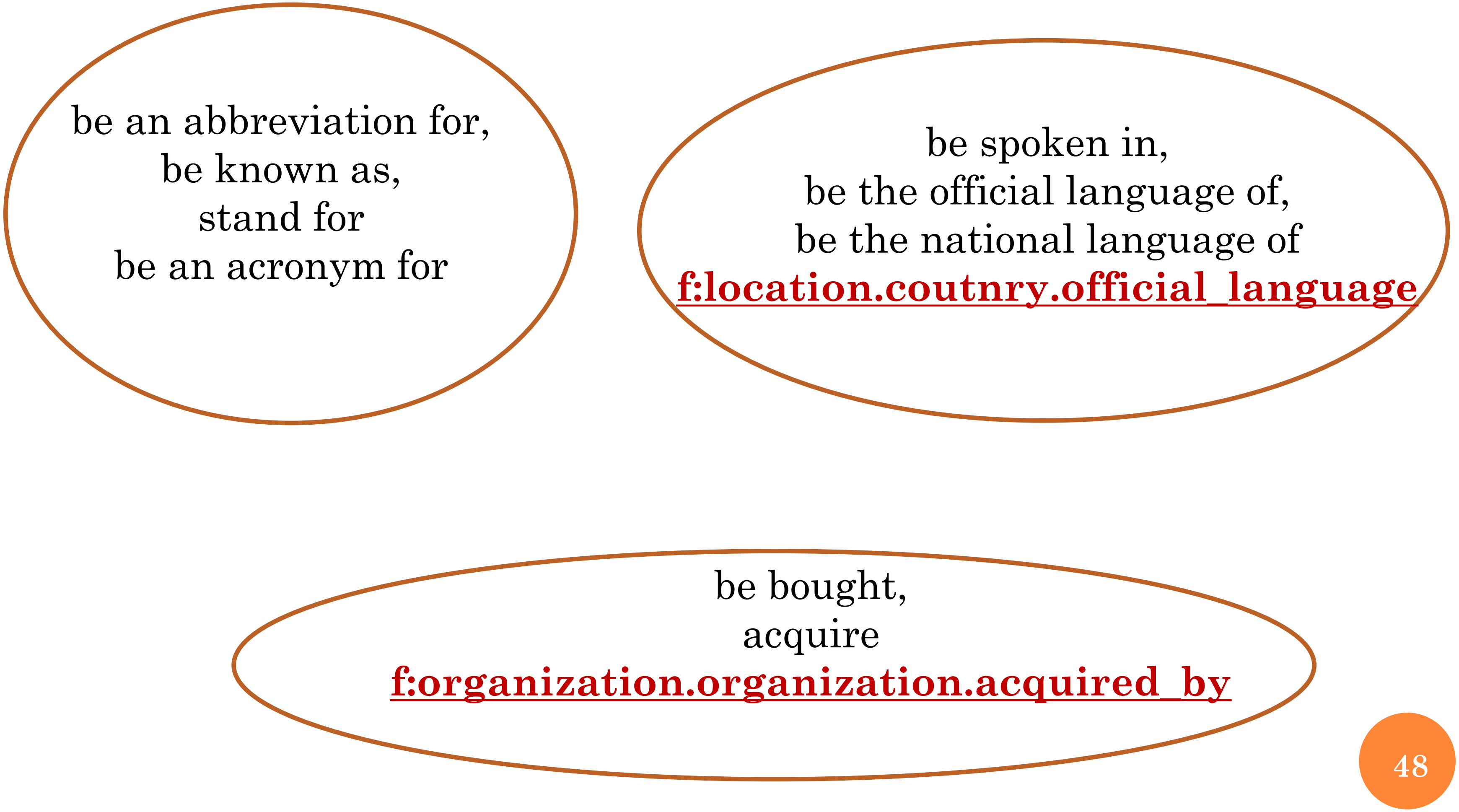


| | Conf. | Phrases | Clusters | Precision<br>Macro | <br>Micro | <br>Pairwise | In Freebase | Triples covered |
|---|---|---|---|---|---|---|---|---|
| Linked KB | 0.8<br>0.5 | 522<br>967 | 118<br>143 | 0.900<br>0.896 | 0.936<br>0.690 | 0.946<br>0.337 | 18%<br>25% | 15%<br>29% |
| Linked KB (types) | 0.8<br>0.5 | 752<br>1185 | 303<br>319 | 0.946<br>0.861 | 0.980<br>0.892 | 0.997<br>0.779 | 9%<br>14% | 21%<br>27% |
| Clustered KB | 0.8<br>0.5 | 826<br>1185 | 234<br>264 | 0.940<br>0.813 | 0.716<br>0.665 | 0.273<br>0.292 | 6%<br>8% | 16%<br>33% |

Table 6: Quality of relation clusters for two different confidence thresholds.

# RESULTS OF RELATION CLUSTERING

- Mapping verbal phrase to Freebase relation



be an abbreviation for,
be known as,
stand for
be an acronym for

be spoken in,
be the official language of,
be the national language of
f:location.coutnry.official_language

be bought,
acquire
f:organization.organization.acquired_by

# Presentation Outline

- Motivation
  - Information Extraction
  - Problems in Open Knowledge Bases
  - Contribution
- Canonicalization of Noun Phrases
  - Mention
  - Clustering
  - Similarity Functions
- Canonicalization of Verbal Phrases
  - Procedure
  - Rule Mining
  - Phrase Clustering
  - Canonicalization
- Experiment
  - Evaluation Metrics
  - Results of Entity Clustering
  - Results of Relation Clustering
- **Conclusion**

# CONCLUSION

- IDF token overlap is better for synonym detection of entity names
- AMIE rule mining is used for canonicalizing verbal phrase.
- Canonicalizing Open KBs can reduct redundancy and ambiguity

Thank you