

Facilitating Human Intervention in Coreference Resolution with Comparative Entity Summaries

Danyun Xu, Gong Cheng, Yuzhong Qu
Nanjing University, China

Presented at ESWC 2014, Crete, Greece

Coreference resolution



TimBL

givenName: "Tim"
surname: "Berners-Lee"
altName: "Tim BL"
type: Scientist
gender: "male"
isDirectorOf: W3C

TBL

name: "Tim Berners-Lee"
type: ComputerScientist
type: RoyalSocietyFellow
sex: "Male"
invented: WWW
founded: WSRI

Wendy

fullName: "Wendy Hall"
type: ComputerScientist
type: RoyalSocietyFellow
sex: "Female"
birthplace: London
founded: WSRI

Methods with humans in the loop (or, coordinating “ings”)

- Active learning
- Crowdsourcing
- Pay-as-you-go

Methods with humans in the loop (or, coordinating “ings”)

- Active learning
- Crowdsourcing
- Pay-as-you-go

Candidate coreferent entities



Methods with humans in the loop (or, coordinating “ings”)

- Active learning
- Crowdsourcing
- Pay-as-you-go

Candidate coreferent entities



Methods with humans in the loop (or, coordinating “ings”)

- Active learning
- Crowdsourcing
- Pay-as-you-go

Candidate coreferent entities



Present entire entity descriptions?

```
publicHomePage : Berners-Lee
sameAs : 512908782
based near : [ ]
  ·longitude : -71.091840
  ·latitude : 42.361860
sameAs : Tim+Berners-Lee
sameAs : 100007
account : User:Timbl
personal mailbox : timbl@w3.org
account : timberners_lee
seeAlso : query?date=All+past+and+future+talks&event=None&ac
one&rdfOnly=yes&submit=Submit
family_name : Berners-Lee
nick : TimBL
sha1sum of a personal mailbox URI name : 965c47c5a70db740721
likes : readfile?fk_files=2372108&pageno=11
assistant : amy
type : Male
account : timbl
title : Sir
homepage : Berners-Lee
homePage : Berners-Lee
uses : findMyLoc
Given name : Timothy
sameAs : 45563
preferredURI : http://www.w3.org/People/Berners-Lee/card#i
label : Tim Berners-Lee
account : timbl
nick : timbl
name : Timothy Berners-Lee
```

```
image : image.php?id=person_7113&checksum=e14a4903654
name : Professor Sir Tim Berners-Lee
workplace homepage : Professor Sir Tim Berners-Lee
hasFamilyName : Berners-Lee
sameAs : ext-7113
past project : EnAKTing the Unbounded Data Web: Challenges
past project : OpenKnowledge
image : image.php?id=person_7113&maxw=250&maxh=300&
hasAppellation : Professor Sir
hasFullName : Professor Sir Tim Berners-Lee
title : Professor Sir
memberOf : EnAKTing the Unbounded Data Web: Challenges
personal mailbox : timbl@ecs.soton.ac.uk
memberOf : OpenKnowledge
hasGivenName : Tim
image : image.php?id=person_7113&square=50&cr=255&cg=2
family_name : Berners-Lee
type : Person
hasRole : 7113
type : Person
hasBiography : <p>Tim Berners-Lee graduated from <a href="h
```

```
<p>Tim Berners-Lee graduated from <a href="h
England, 1976. Whilst there he built his first co
</p> <p>He spent two years with Plessey Tele
urer, working on distributed transaction system
oin D.G Nash Ltd (Ferndown, Dorset, UK), whe
ultitasking operating system.</p> <p>A year a
980) as consultant software engineer at <a href
eneva, Switzerland. Whilst there, he wrote for h
```

Present a compact comparative summary!

givenName: "Tim"

surname: "Berners-Lee"

isDirectorOf: W3C

name: "Tim Berners-Lee"

invented: WWW

Present a compact comparative summary!

Which property-value (PV) pairs are more helpful?

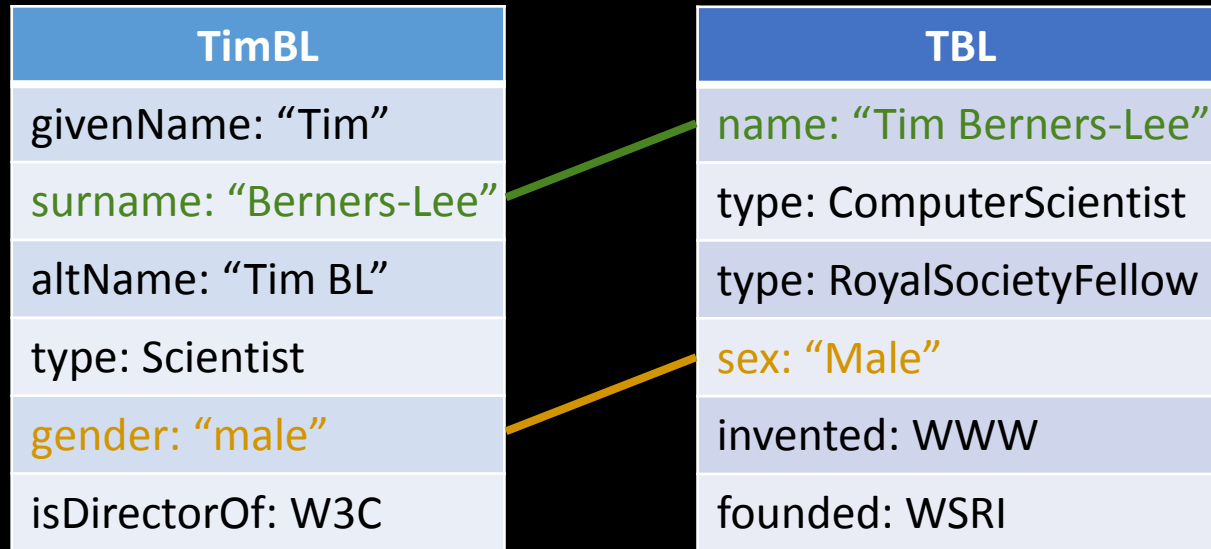
```
nameAs: 512908762
location: []
  -longitude: -71.091840
  -latitude: 42.361990
nameAs: Tim+Berners-Lee
nameAs: 100007
account: User:Timbl
personal mailbox: timbl@w3.org
account: timbers_lee
webPage: query?date=All+past+and+future+talks&event=Name&
  name=Tim+Berners-Lee&submit=Submit
family name: Berners-Lee
nick: Tim
platform: a personal home page for the World Wide Web
feed: rss/feeds%_file=2372108&page=11
assistant: amy
sex: Male
account: timbl
title: Sir
homepage: Berners-Lee
homePage: Berners-Lee
sex: Indeterminate
given name: Timothy
nameAs: 45563
preferredURL: http://www.w3.org/People/Berners-Lee/card#h
label: Tim Berners-Lee
account: timbl
nick: timbl
name: Timothy Berners-Lee
```

```
image: image.php?id=person_7113&maxw=250&maxh=300&
name: Professor Sir Tim Berners-Lee
workplace homepage: Professor Sir Tim Berners-Lee
familyName: Berners-Lee
nameAs: ext-7113
last project: EnAKTing the Unbounded Data Web: Challenges
last project: OpenKnowledge
image: image.php?id=person_7113&maxw=250&maxh=300&
appellation: Professor Sir
fullName: Professor Sir Tim Berners-Lee
title: Professor Sir
last project: EnAKTing the Unbounded Data Web: Challenges
last project: OpenKnowledge
memberOf: OpenKnowledge
givenName: Tim
image: image.php?id=person_7113&square=50&crop=255&crop=
type: Person
role: 7113
type: Person
biography: <p>Tim Berners-Lee graduated from <a href="
  England, 1976. Whilst there he built his first computer
</p> <p>He spent two years with Plessey Telecommunications
  urer, working on distributed transaction systems for
  on D G Nash Ltd (Ferndown, Dorset, UK), whilst also
  ultitasking operating systems. </p> <p>A year or so later
  980) as consultant software engineer at <a href="http://www.
  erve.ch>Switzerland. Whilst there, he wrote for
```

Four aspects of a good comparative summary

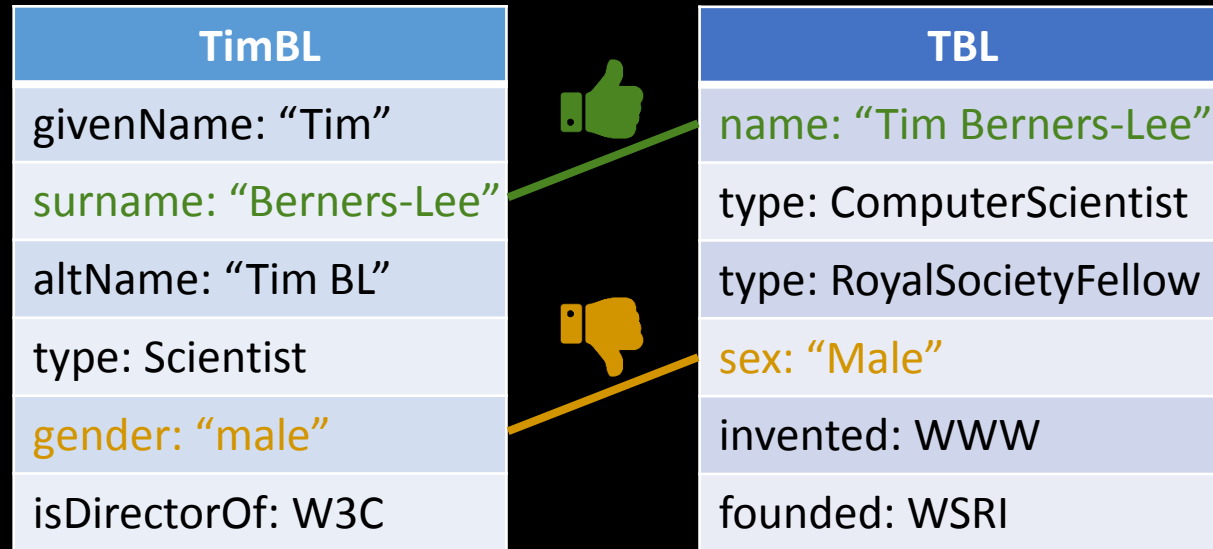
1. Reflecting commonality
2. Reflecting difference
3. Providing information on identity
4. Providing diverse information

1. Commonality



- Common PV pairs = comparable properties + similar values

1. Commonality



- Common PV pairs = comparable properties + similar values
- More helpful properties = more like an Inverse Functional Property (IFP)

1. Commonality (details)

- Comparability between properties
 - Learned from known coreferent entities

Comparable properties = Properties having similar values

- String similarity

1. Commonality (details)

- Comparability between properties
 - Learned from known coreferent entities

Comparable properties = Properties having similar values

- String similarity
- Similarity between values
 - String similarity

1. Commonality (details)

- Comparability between properties
 - Learned from known coreferent entities

Comparable properties = Properties having similar values

- String similarity
- Similarity between values
 - String similarity
- Likeness to an IFP
 - Estimated based on the data set

$$\textit{Likeness} = \frac{\textit{Number of distinct values}}{\textit{Number of all values}}$$

1. Commonality (weakness)

- Only reflecting commonality can be misleading.

TBL
name: "Tim Berners-Lee"
type: ComputerScientist
type: RoyalSocietyFellow
sex: "Male"
invented: WWW
founded: WSRI

Wendy
fullName: "Wendy Hall"
type: ComputerScientist
type: RoyalSocietyFellow
sex: "Female"
birthplace: London
founded: WSRI

2. Difference

TBL	Wendy
name: "Tim Berners-Lee"	fullName: "Wendy Hall"
type: ComputerScientist	type: ComputerScientist
type: RoyalSocietyFellow	type: RoyalSocietyFellow
sex: "Male"	sex: "Female"
invented: WWW	birthplace: London
founded: WSRI	founded: WSRI

- Different PV pairs = comparable properties + dissimilar values

2. Difference

TBL	Wendy
name: "Tim Berners-Lee"	fullName: "Wendy Hall"
type: ComputerScientist	type: ComputerScientist
type: RoyalSocietyFellow	type: RoyalSocietyFellow
sex: "Male"	sex: "Female"
invented: WWW	birthplace: London
founded: WSRI	founded: WSRI

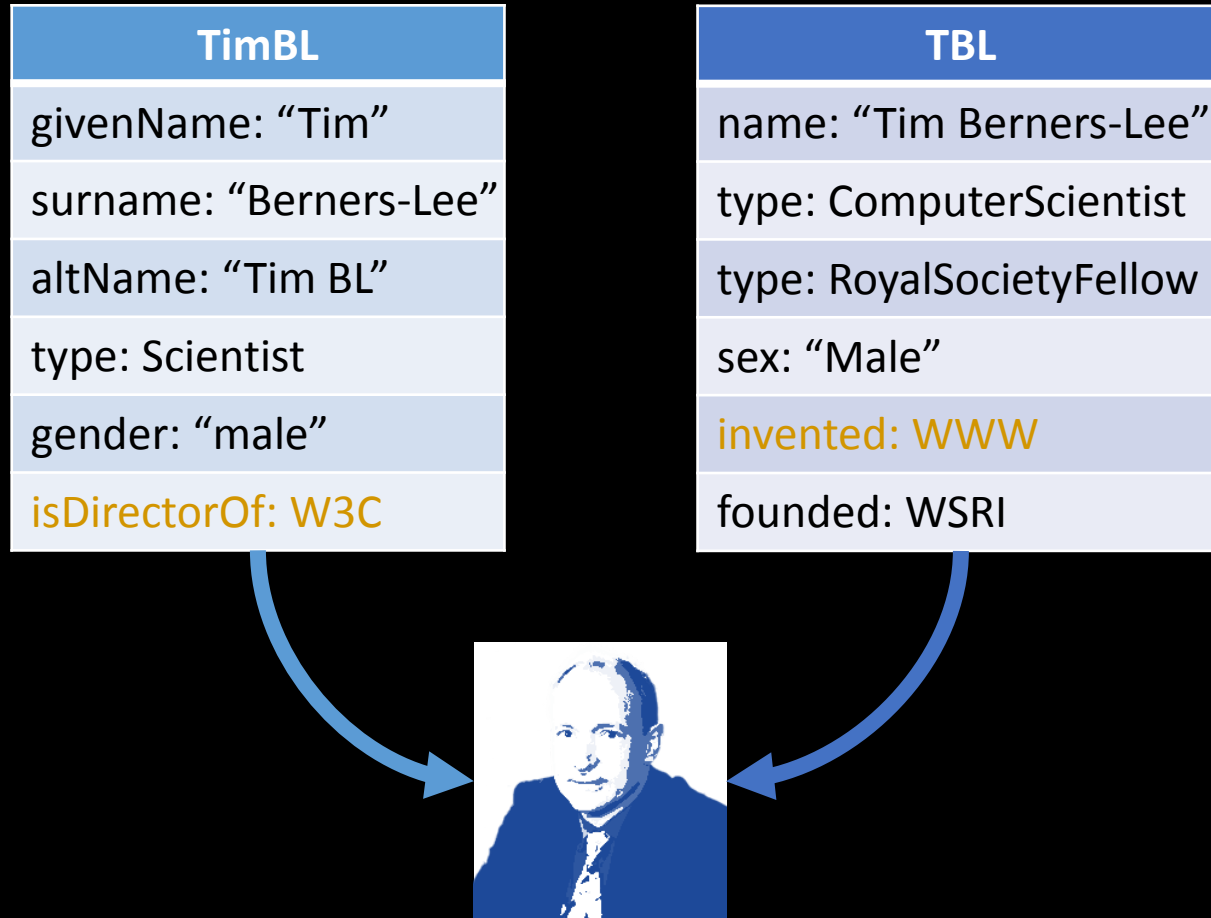
- Different PV pairs = comparable properties + dissimilar values
- **More helpful properties = more like a Functional Property (FP)**

2. Difference (details)

- Comparability between properties
 - Learned from known coreferent entities
 - String similarity
- Dissimilarity between values
 - String similarity
- Likeness to a FP
 - Estimated based on the data set

$$\textit{Likeness} = \frac{\textit{Number of distinct subjects}}{\textit{Number of all subjects}}$$

3. Information on identity

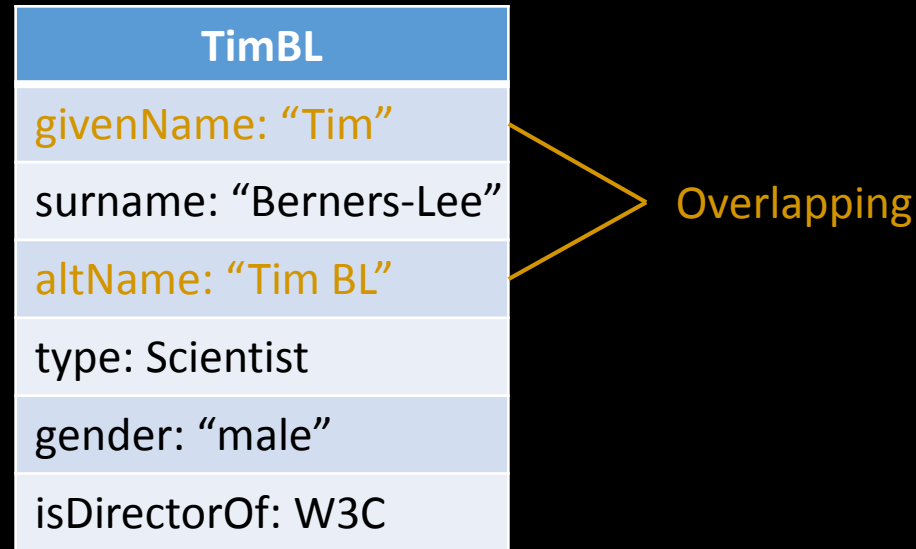


3. Information on identity (details)

- Information on identity
 - Estimated based on the data set

$$\text{information} = 1 - \frac{\log(\text{Number of entities having this PV pair})}{\log(\text{Number of all entities})}$$

4. Diversity of information



- Overlapping PV pairs = similar properties or similar values

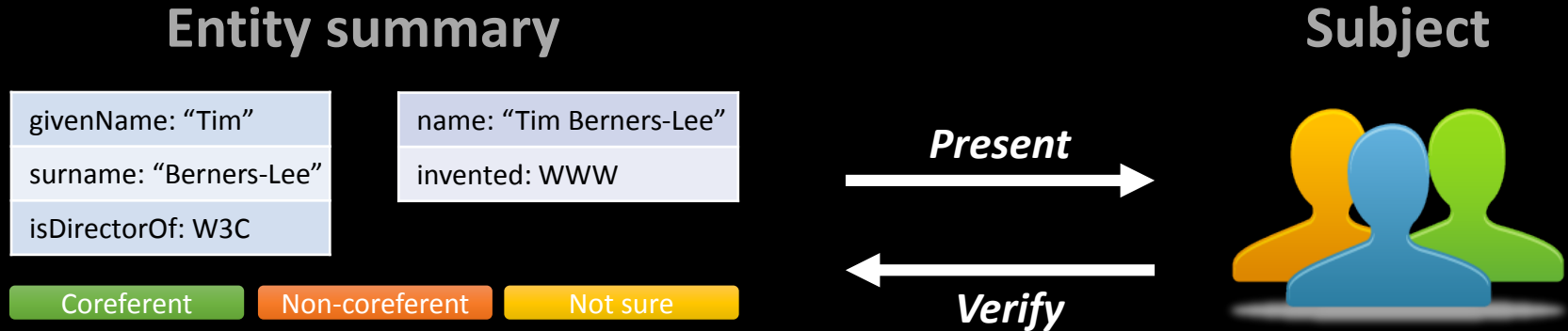
To find an optimal summary
(or, to find the most helpful PV pairs)

- Maximize
 - Commonality
 - Difference
 - Information on identity
 - Diversity of information
- Subject to
 - A length limit

To find an optimal summary
(or, to find the most helpful PV pairs)

- Maximize
 - Commonality
 - Difference
 - Information on identity
 - Diversity of information
- Subject to
 - A length limit
- Formulated as a binary quadratic knapsack problem
- Solved by GRASP-based local search

Evaluation method



- 4 approaches to be blindly tested
- 20 subjects (university students)
- 24 random tasks for each subject
 - 4 approaches * (3 positive cases + 3 negative cases)
 - Sorted in random order

Data sets and tasks

- Data sets

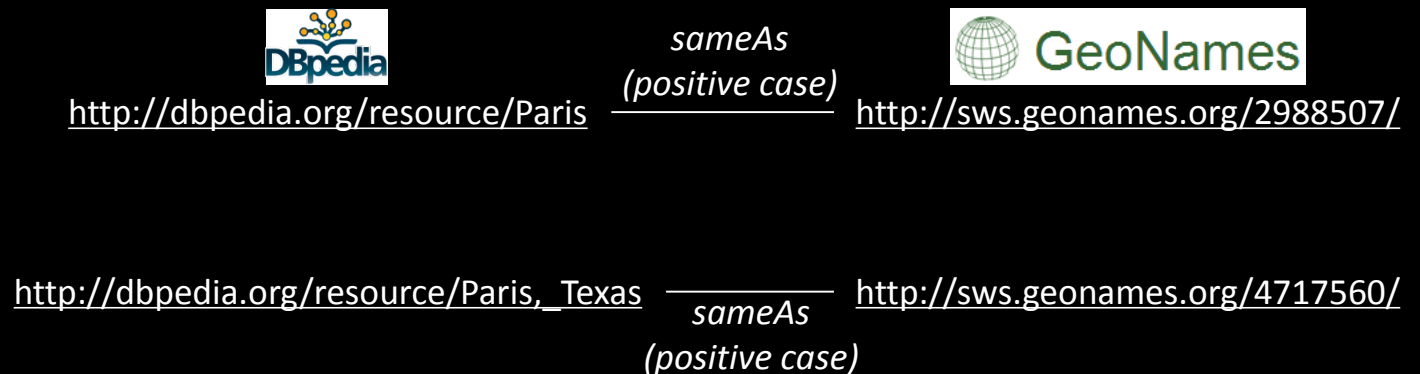


Data sets and tasks

- Data sets



- Tasks

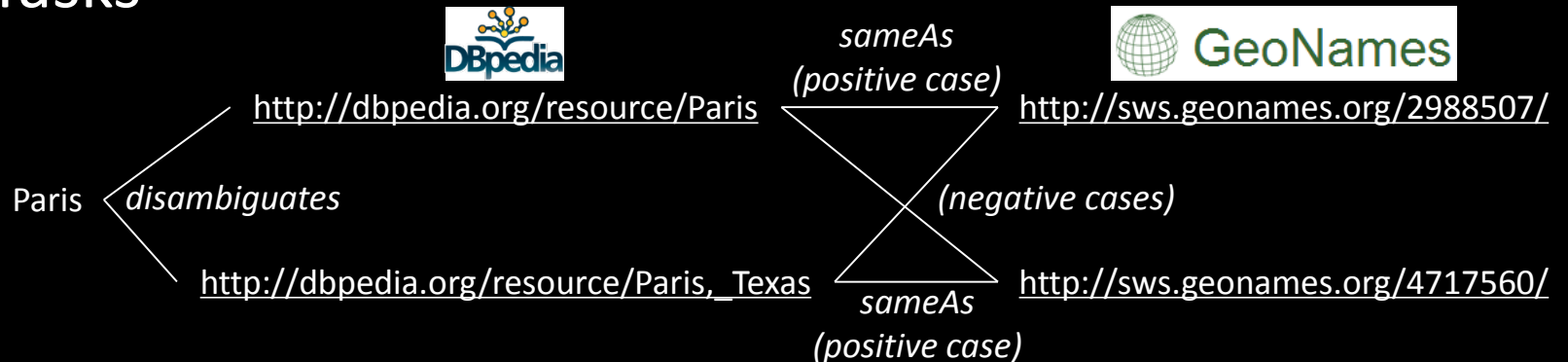


Data sets and tasks

- Data sets



- Tasks

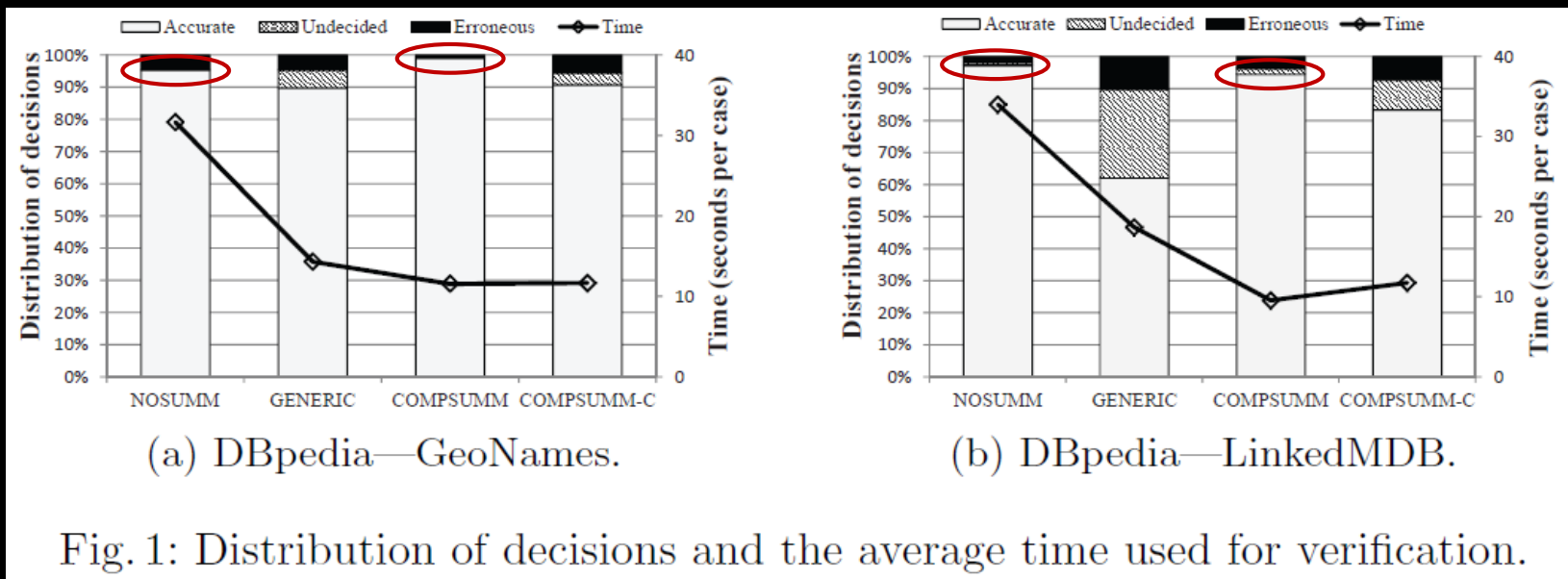


Approaches

Approach	Description
NOSUMM	Present entire entity descriptions
GENERIC	<ul style="list-style-type: none">• Information on identity [3]• Diversity of information
COMPSUMM	<ul style="list-style-type: none">• Commonality• Difference• Information on identity• Diversity of information
COMPSUMM-C	<ul style="list-style-type: none">• Commonality• Difference• Information on identity• Diversity of information

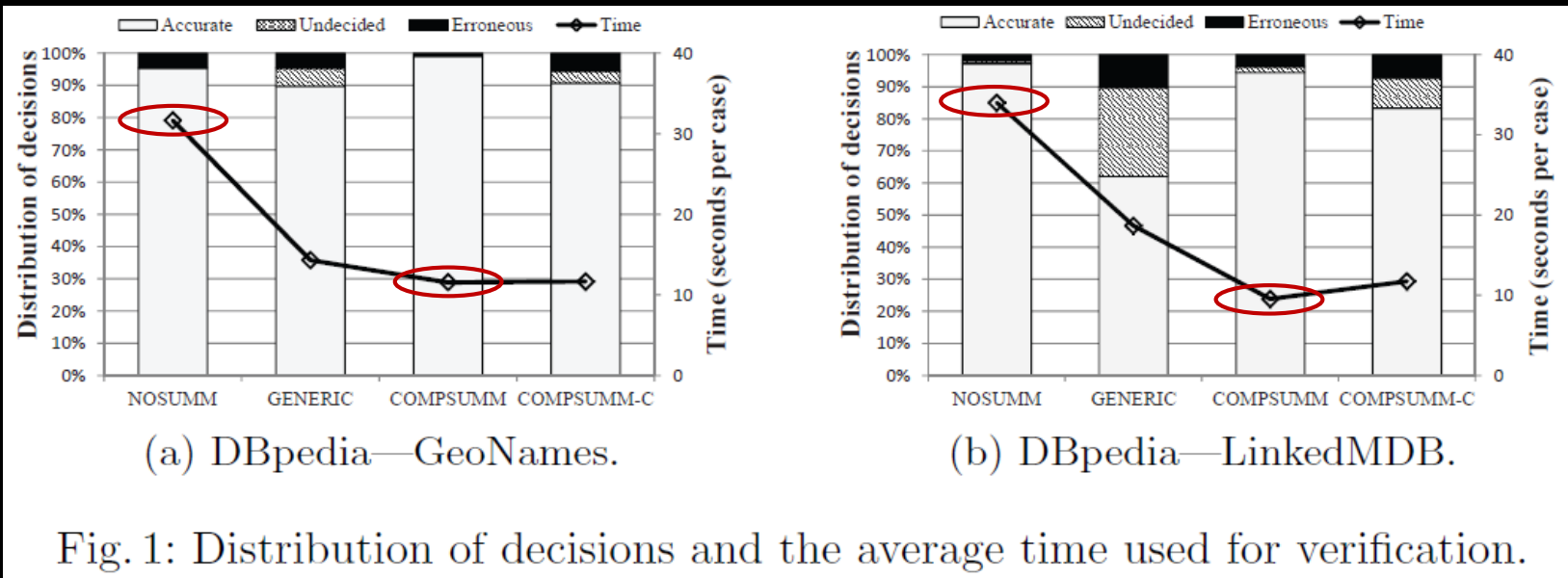
Results (1)

- Accuracy of verification
 - **COMPSUMM \approx NOSUMM**
 - > COMPSUMM-C
 - > GENERIC



Results (2)

- Efficiency of verification
 - COMPSUMM > NOSUMM (2.7—2.9 times faster)



Take-home messages

- Provide entity summaries for verifying coreference.
 - improves efficiency (2.7—2.9 times faster)
 - without notably affecting accuracy
- Provide comparative (but not just generic) summaries.
- Show both commonality and difference.

Future work

- Present = Summarize + Visualize

Candidate coreferent entities

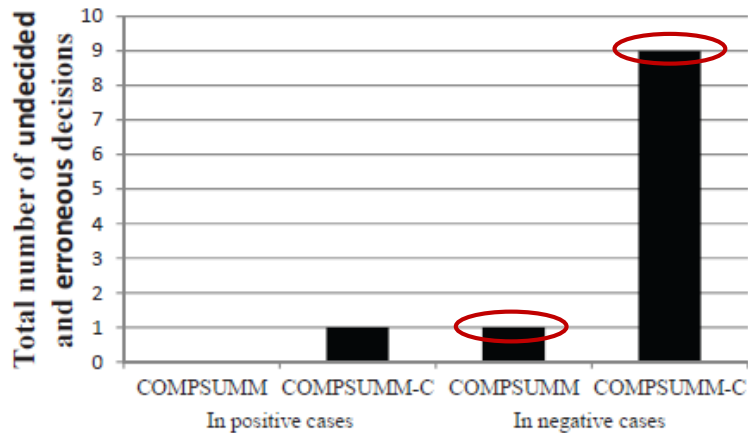


Thanks for your attention

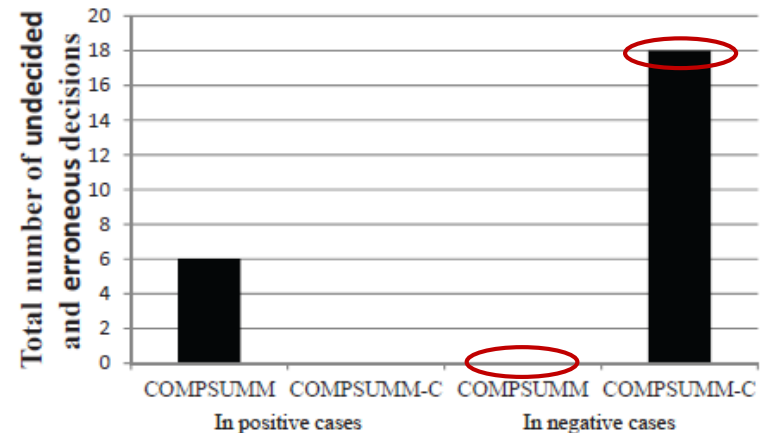


Results (3)

- Erroneous decisions
 - COMPSUMM-C > COMPSUMM (mostly in negative cases)



(a) DBpedia—GeoNames.



(b) DBpedia—LinkedMDB.

Performance testing

- Offline computation
 - Comparability between properties (the learning part)
 - Likeness to an IFP/FP
 - Information on identity

Performance testing

- ~~Offline computation~~
 - ~~Comparability between properties (the learning part)~~
 - ~~Likeness to an IFP/FP~~
 - ~~Information on identity~~
- Online computation
 - Similarity between properties/values
 - Optimization
- Results
 - Places (DBpedia and GeoNames): 24ms per case
 - Films (DBpedia and LinkedMDB): 35ms per case