Provenance for Natural Language Queries

Daniel Deutch Nave Frost Amir Gilad

Tel Aviv University

August 2017

Presented by Amir Gilad

Outline

Motivation

NL Query

Return the organization of authors who published papers in database conferences after 2005

Motivation

NL Query

Return the organization of authors who published papers in database conferences after 2005

Formal Query

```
query(oname) :- org(oid, oname), conf(cid, cname),
pub(wid, cid, ptitle, pyear), author(aid, aname, oid),
domainConf(cid, did), domain(did, dname),

writes(aid, wid), dname = 'Databases', pyear > 2005
```

Motivation

NL Query

Return the organization of authors who published papers in database conferences after 2005

Formal Query

```
query(oname) :- org(oid, oname), conf(cid, cname),
pub(wid, cid, ptitle, pyear), author(aid, aname, oid),
domainConf(cid, did), domain(did, dname),
writes(aid, wid), dname = 'Databases', pyear > 2005
```

Result

Tel Aviv University (TAU)

Motivation

NL Query

Return the organization of authors who published papers in database conferences after 2005

Formal Query

```
query(oname) :- org(oid, oname), conf(cid, cname),
pub(wid, cid, ptitle, pyear), author(aid, aname, oid),
domainConf(cid, did), domain(did, dname),

writes(aid, wid), dname = 'Databases', pyear > 2005
```

Result

Tel Aviv University (TAU) (why?)

Motivation

NL Query

Return the organization of authors who published papers in database conferences after 2005

Formal Query

```
query(oname) :- org(oid, oname), conf(cid, cname),
pub(wid, cid, ptitle, pyear), author(aid, aname, oid),
domainConf(cid, did), domain(did, dname),

writes(aid, wid), dname = 'Databases', pyear > 2005
```

What We Have - Provenance

```
(oname,TAU)·(aname,Tova M.)·(ptitle,OASSIS...)·(cname,SIGMOD)·(pyear,14')+
(oname,TAU)·(aname,Tova M.)·(ptitle,Querying...)·(cname,VLDB)·(pyear,06')+
(oname,TAU)·(aname,Tova M.)· (ptitle,Monitoring...)·(cname,VLDB)·(pyear,07')+
(oname,TAU)·(aname,Slava N.)·(ptitle,OASSIS...)·(cname,SIGMOD)·(pyear,14')+
(oname,TAU)·(aname,Tova M.)·(ptitle,A sample...)·(cname,SIGMOD)·(pyear,14')+
...
```

Motivation

NL Query

Return the organization of authors who published papers in database conferences after 2005

Formal Query

```
query(oname) :- org(oid, oname), conf(cid, cname),
pub(wid, cid, ptitle, pyear), author(aid, aname, oid),
domainConf(cid, did), domain(did, dname),
writes(aid, wid), dname = 'Databases', pyear > 2005
```

What We Want - Explanations

TAU is the organization of 43 authors who published 170 papers in 31 conferences in 2006 - 2015

Solution Overview

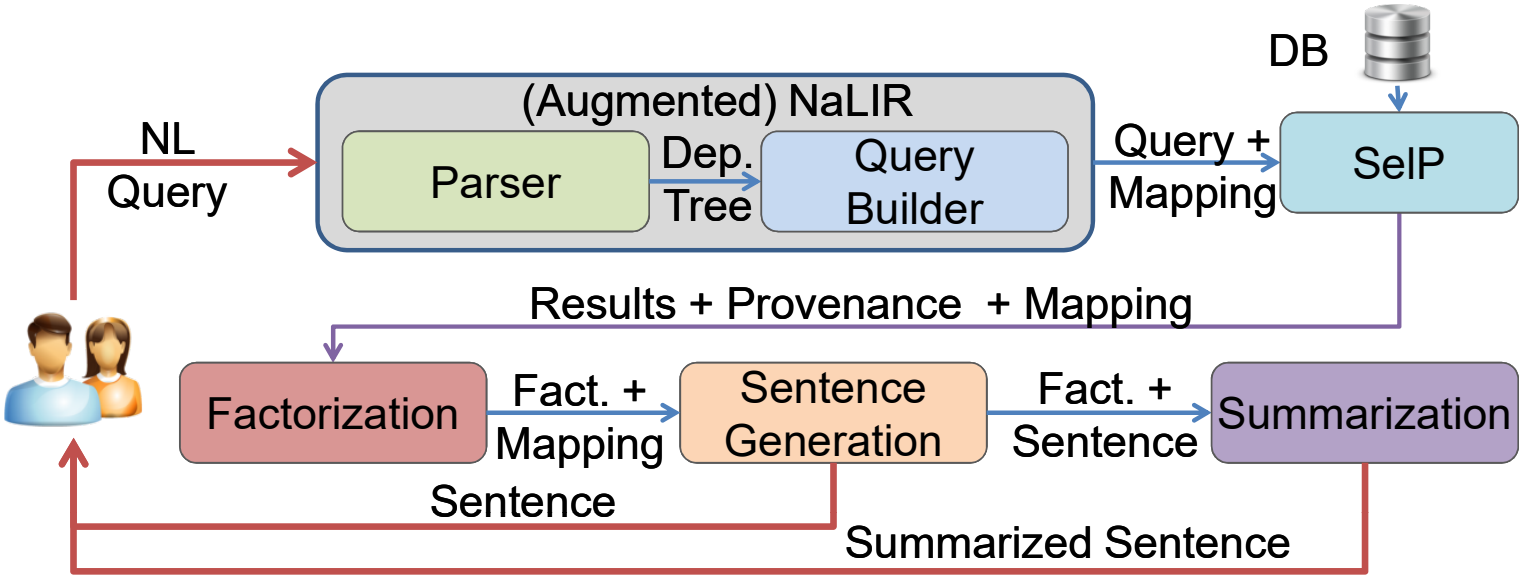
Solution

- Use the input question to formulate a detailed NL answer by replacing words with values
  - This is a general idea: showing provenance in a way that corresponds to the standard user interaction
- When a long answer is needed, compact it using algebraic factorization and summarization
  - Here, again, we leverage the structure of the user question

Current Limitations

- Only conjunctive queries are supported
- Some aspects of the solution are limited to a specific NLIDB
  - But the general idea is not

Framework



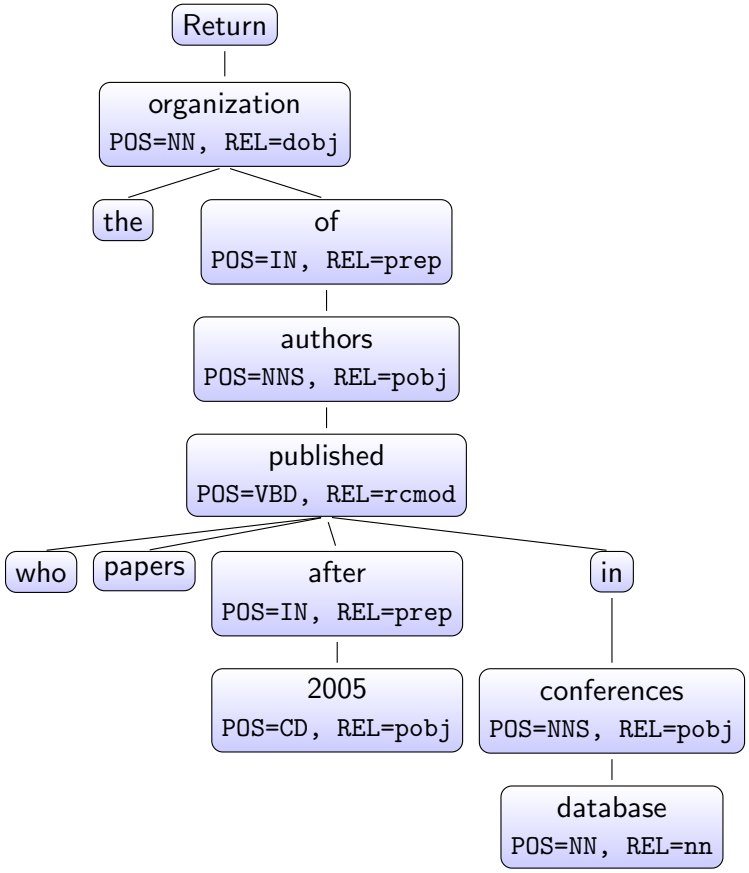
- Augment NaLIR [Fei Li, Jagadish, 15']
- Use a provenance-aware engine - SelP [Deutch et al., 15']
- Store the provenance and mappings
- Translate results and provenance to NL using factorization and summarization

Outline

Mappings

Return the organization of authors who published papers in database conferences after 2005

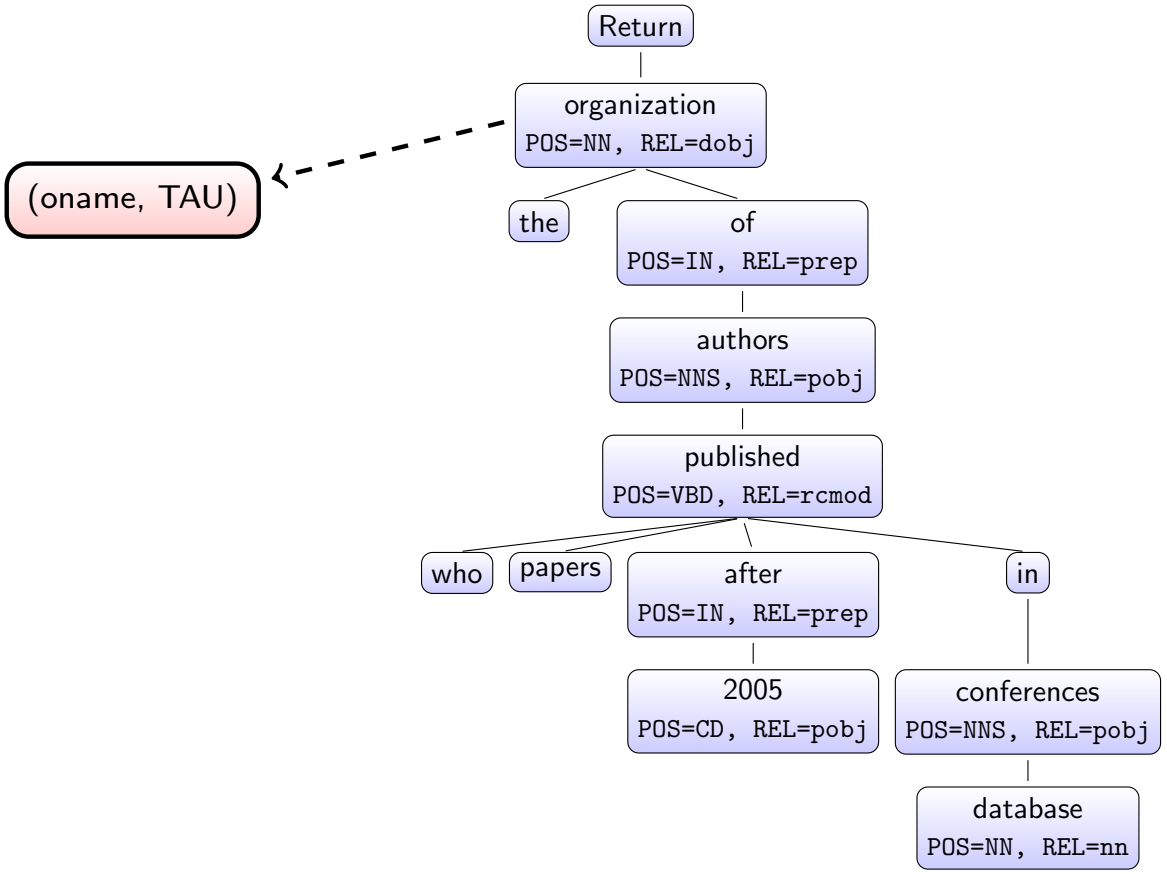
query(oname) :- org(oid, oname), conf(cid, cname), pub(wid, cid, ptitle, pyear), author(aid, aname, oid), domainConf(cid, did), domain(did, dname), writes(aid, wid), dname = 'Databases', pyear > 2005



Mappings

Return the **organization** *of authors who published papers in database conferences after 2005*

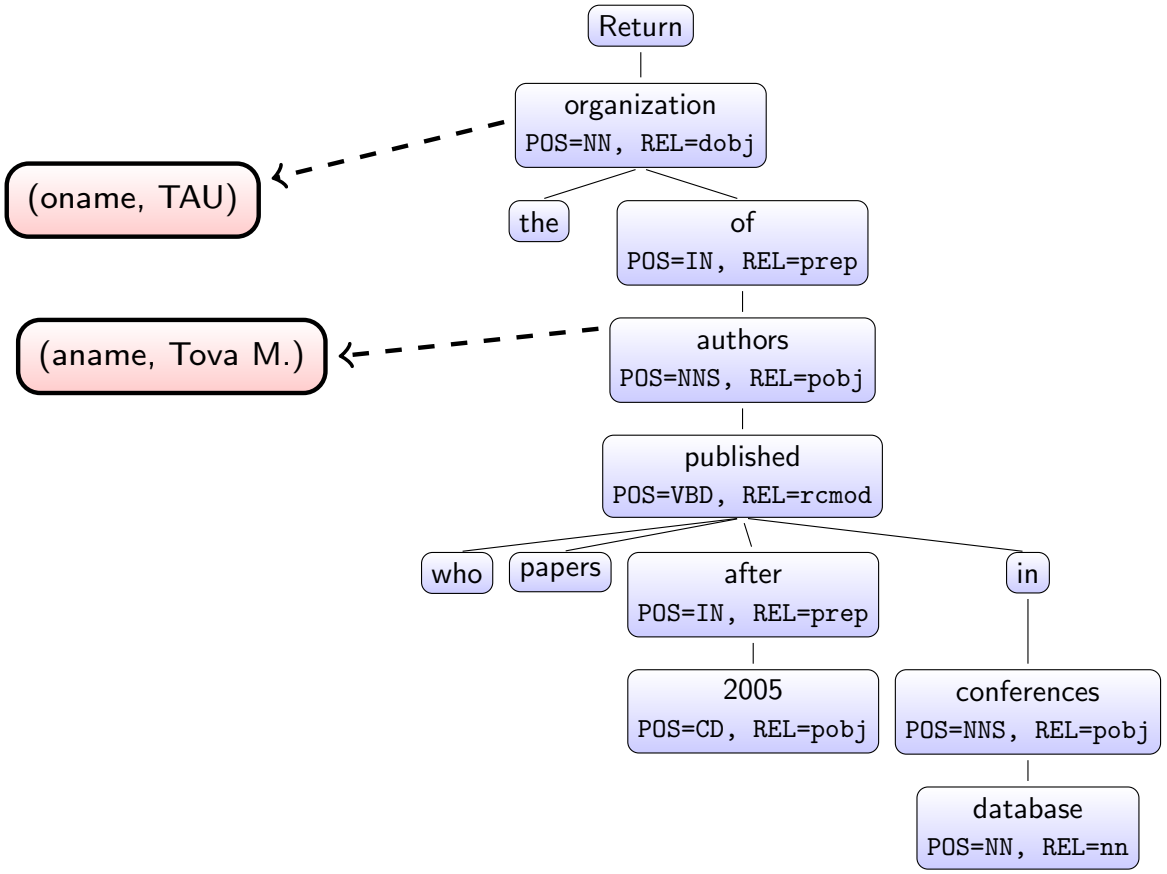
*query(***oname***) :- org(oid,* ***oname****), conf(cid, cname), pub(wid, cid, ptitle, pyear), author(aid, aname, oid), domainConf(cid, did), domain(did, dname), writes(aid, wid), dname = 'Databases', pyear > 2005*



Mappings

Return the organization of **authors** *who published papers in database conferences after 2005*

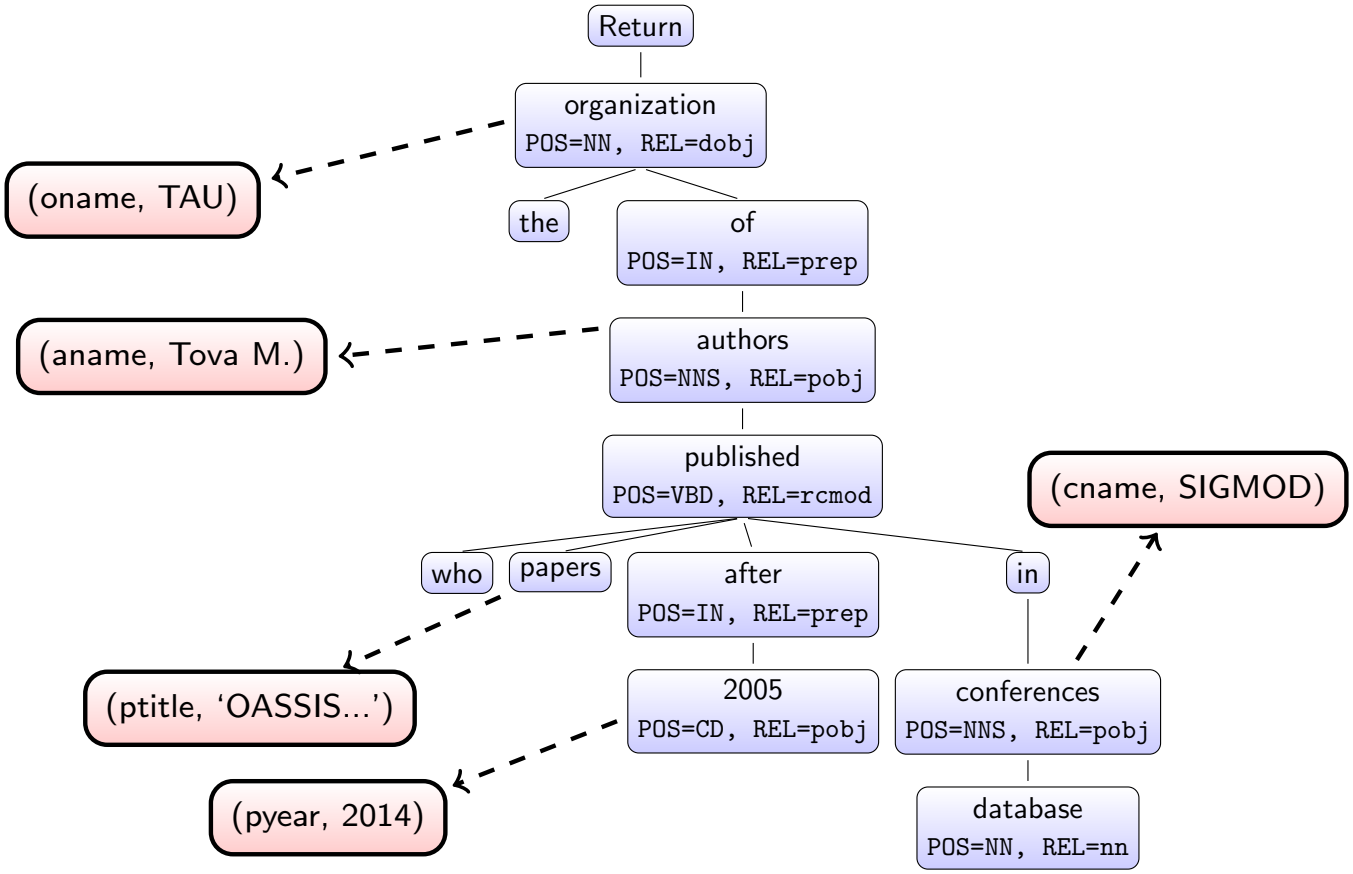
query(oname) :- org(oid, oname), conf(cid, cname), pub(wid, cid, ptitle, pyear), author(aid, **aname**, *oid), domainConf(cid, did), domain(did, dname), writes(aid, wid), dname = 'Databases', pyear > 2005*



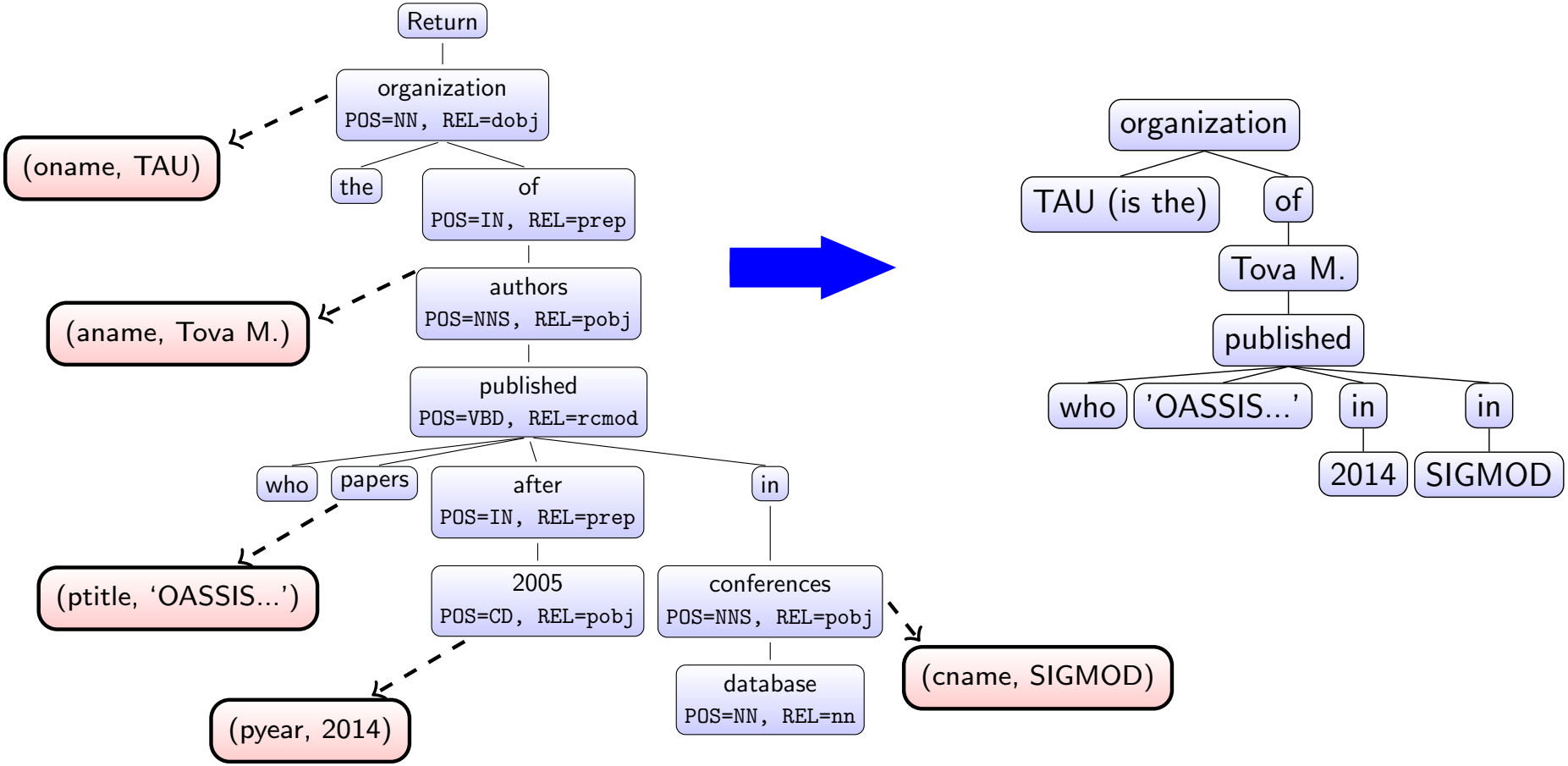
Mappings

Return the organization of authors who published papers in database conferences after 2005

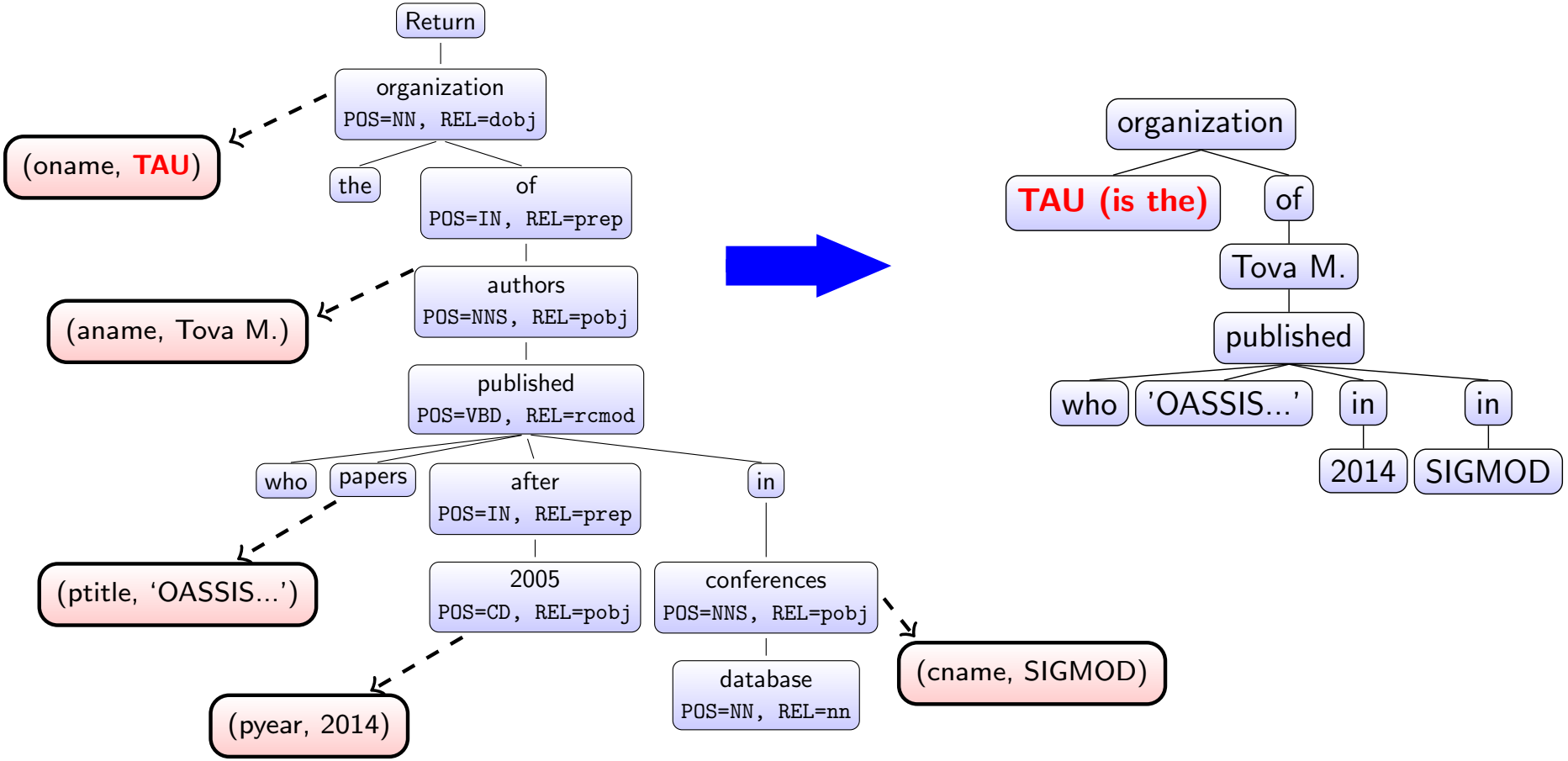
query(oname) :- org(oid, oname), conf(cid, cname), pub(wid, cid, ptitle, pyear), author(aid, aname, oid), domainConf(cid, did), domain(did, dname), writes(aid, wid), dname = 'Databases', pyear > 2005



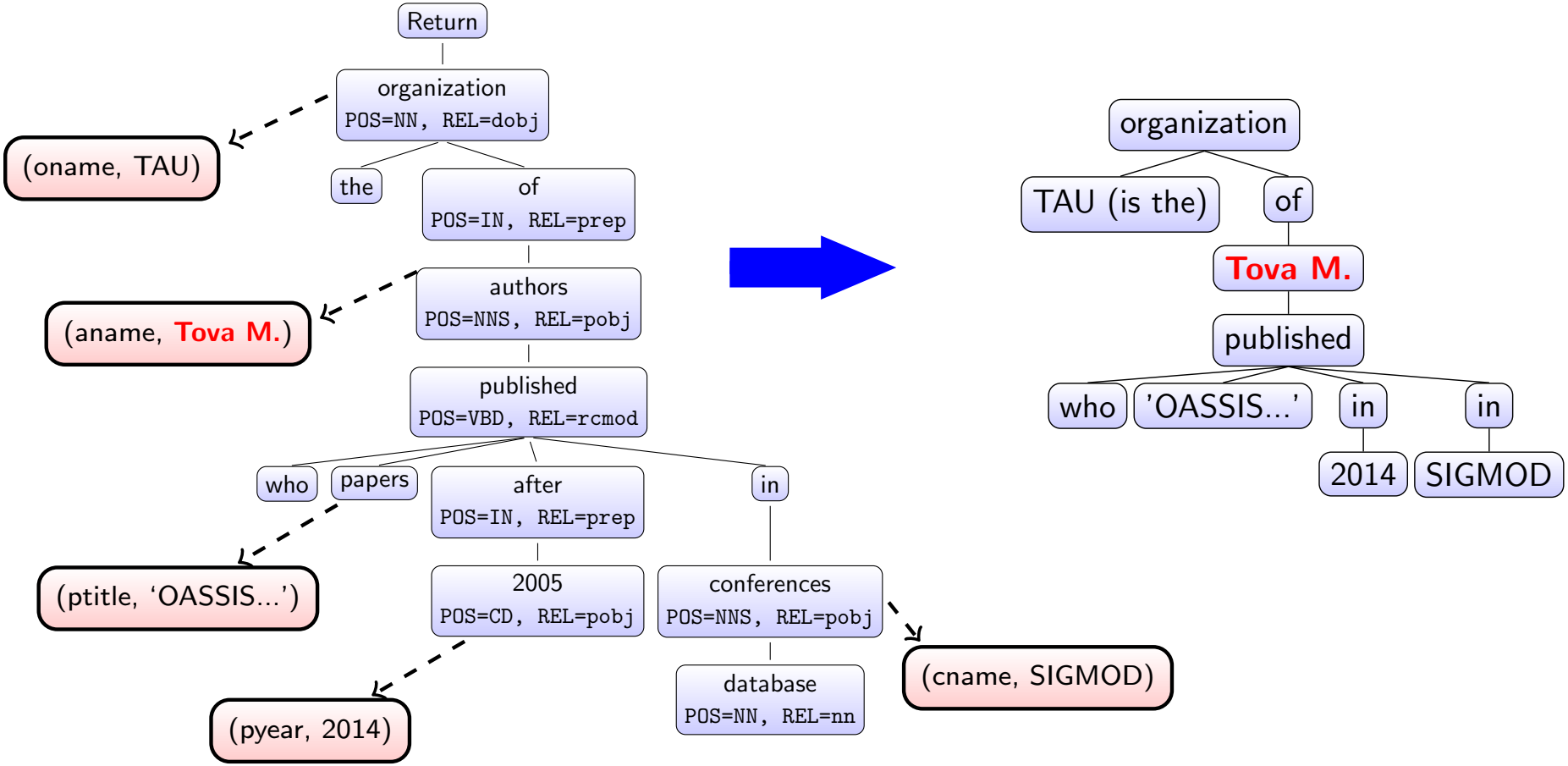
From Mappings to an Answer



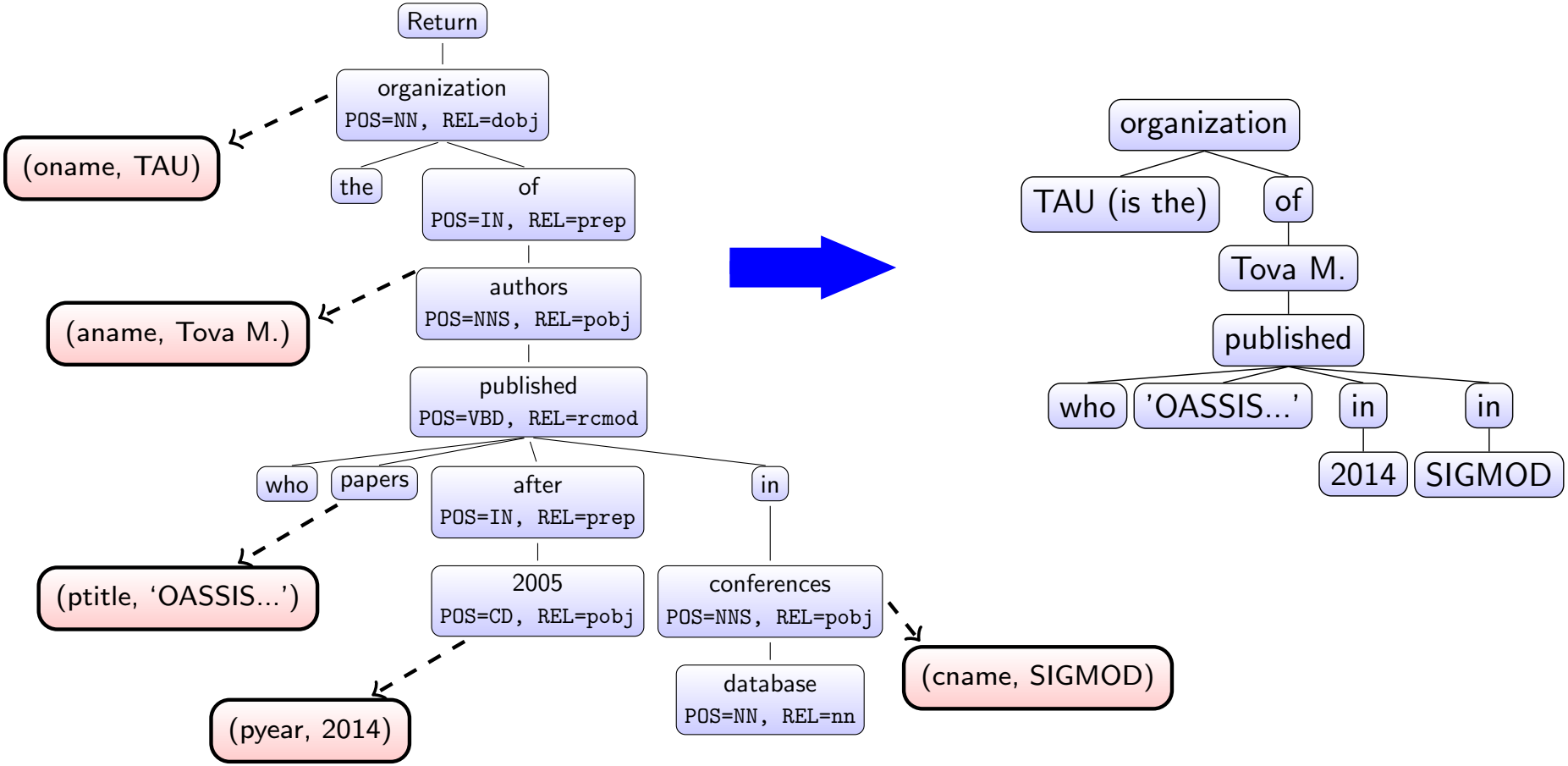
From Mappings to an Answer



From Mappings to an Answer



From Mappings to an Answer



Answer

TAU is the organization of Tova M. who published 'OASSIS...' in SIGMOD in 2014

Outline

Provenance Factorization

Idea

- Use algebraic factorization of the provenance to take-out common values that appear in multiple assignments

Provenance

```
[TAU]·[Tova M.]·[OASSIS...]·[SIGMOD]·[2014]+
[TAU]·[Tova M.]·[Querying...]·[VLDB]·[2006]+
[TAU]·[Tova M.]· [Monitoring..]·[VLDB]·[2007]+
[TAU]·[Slava N.]·[OASSIS...]·[SIGMOD]·[2014]+
[TAU]·[Tova M.]·[A sample...]·[SIGMOD]·[2014]
```

Provenance Factorization

Idea

- Use algebraic factorization of the provenance to take-out common values that appear in multiple assignments

Provenance

```
[TAU]·[Tova M.]·[OASSIS...]·[SIGMOD]·[2014]+
[TAU]·[Tova M.]·[Querying...]·[VLDB]·[2006]+
[TAU]·[Tova M.]· [Monitoring..]·[VLDB]·[2007]+
[TAU]·[Slava N.]·[OASSIS...]·[SIGMOD]·[2014]+
[TAU]·[Tova M.]·[A sample...]·[SIGMOD]·[2014]
```

Two Different Factorizations

```
[TAU] ·
 ([SIGMOD] · [2014] ·
 ([OASSIS...] ·
 ([Tova M.] + [Slava N.]))
  + [Tova M.] · [A Sample...])
  + [VLDB] · [Tova M.] ·
 ([2006] · [Querying...]
  + [2007] · [Monitoring...])
```

```
[TAU] ·
([Tova M.] ·
([VLDB] ·
 ([2006] · [Querying...]
 + [2007] · [Monitoring...]))
+ [SIGMOD] · [2014] ·
 ([OASSIS...] + [A Sample...]))
+ [Slava N.] · [OASSIS...] · [SIGMOD] · [2014])
```

Provenance Factorization

Idea

- Use algebraic factorization of the provenance to take-out common values that appear in multiple assignments

Provenance

```
[TAU]·[Tova M.]·[OASSIS...]·[SIGMOD]·[2014]+
[TAU]·[Tova M.]·[O...]·[VLDB]·[2006]+
[TAU]·[Tova M.]·[...]·[VLDB]·[2007]+
[TAU]·[Slava N.]...IGMOD]·[2014]+
[TAU]·[Tova M.]...SIGMOD]·[2014]
```

Shorter means better?

Two Different Factorizations

```
[TAU] ·
 ([SIGMOD] · [2014] ·
 ([OASSIS...] ·
 ([Tova M.] + [Slava N.]))
  + [Tova M.] · [A Sample...])
  + [VLDB] · [Tova M.] ·
 ([2006] · [Querying...]
  + [2007] · [Monitoring...])
```

```
[TAU] ·
([Tova M.] ·
([VLDB] ·
 ([2006] · [Querying...]
 + [2007] · [Monitoring...]))
+ [SIGMOD] · [2014] ·
 ([OASSIS...] + [A Sample...]))
+ [Slava N.] · [OASSIS...] · [SIGMOD] · [2014])
```

T-Compatibility

NL Query

Return the organization of authors who published papers in database conferences after 2005

Shortest Factorization

```
[TAU] ·
 ([SIGMOD] · [2014] ·
 ([OASSIS...] ·
 ([Tova M.] + [Slava N.]))
  + [Tova M.] · [A Sample...])
  + [VLDB] · [Tova M.] ·
 ([2006] · [Querying...]
  + [2007] · [Monitoring...])
```

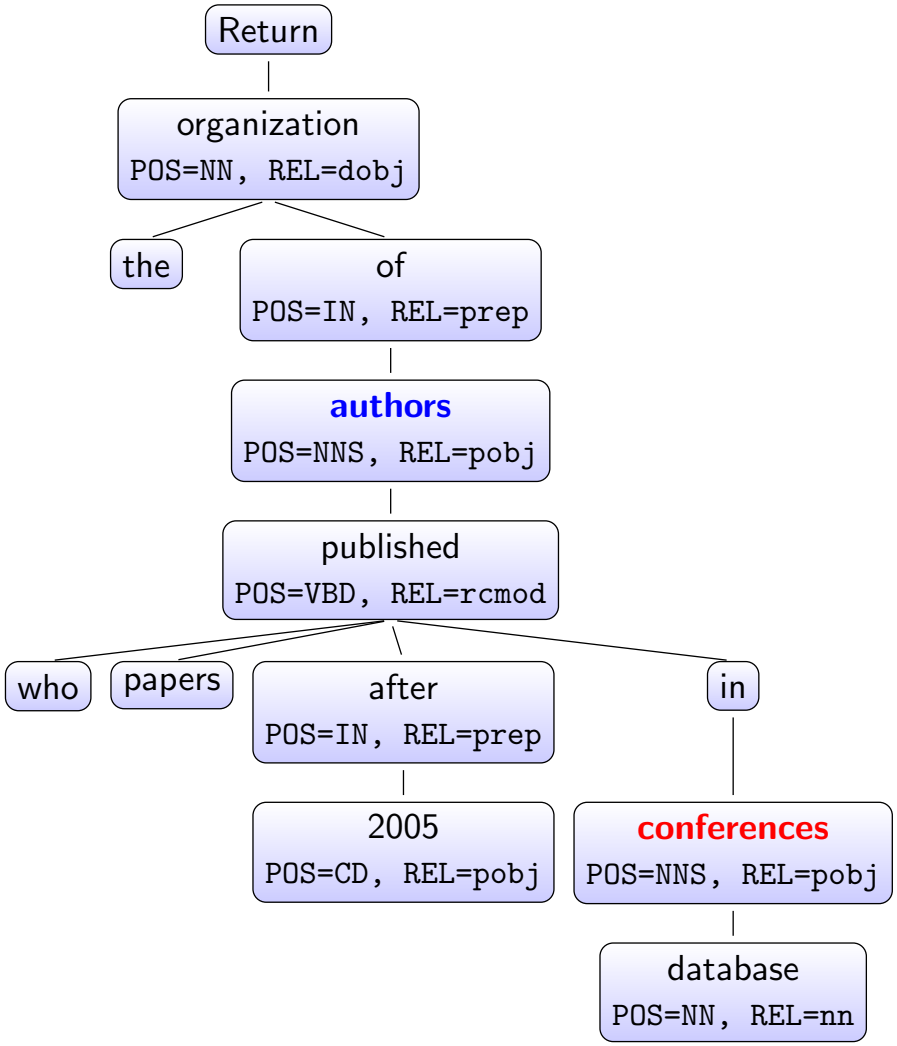
As a Sentence

TAU is the organization **of authors who published in**
SIGMOD 2014
'OASSIS...' **which was published by**
Tova M. and Slava N.
and Tova M. published 'A sample...'
and Tova M. published in VLDB
'Querying...' in 2014
and 'Monitoring...' in 2007.

T-Compatibility

Shortest Factorization

```
[TAU] ·
 ([SIGMOD] · [2014] ·
 ([OASSIS...] ·
 ([Tova M.] + [Slava N.]))
  + [Tova M.] · [A Sample...])
  + [VLDB] · [Tova M.] ·
 ([2006] · [Querying...]
  + [2007] · [Monitoring...])
```



$conferences \leq_T authors$ but $conferences \not\leq_{f_{bad}} authors$

T-Compatibility

NL Query

Return the organization of authors who published papers in database conferences after 2005

Longer, T-Compatible Factorization

```
[TAU] ·
([Tova M.] ·
([VLDB] ·
 ([2006] · [Querying...]
 + [2007] · [Monitoring...]))
+ [SIGMOD] · [2014] ·
 ([OASSIS...] + [A Sample...]))
+ [Slava N.] · [OASSIS...] · [SIGMOD] · [2014])
```

As a Sentence

```
TAU is the organization of
 Tova M. who published
 in VLDB
 'Querying...'  in 2006 and
 'Monitoring...'  in 2007
 and in SIGMOD in 2014
 'OASSIS...'  and 'A sample...'
 and Slava N. who published
 'OASSIS...'  in SIGMOD in 2014.
```

Factorization Algorithm

Proposition

- Obtaining a minimal T-compatible factorization is coNP-hard

Factorization Algorithm

Proposition

- Obtaining a minimal T-compatible factorization is coNP-hard

Algorithm

- Factorize greedily: traverse the dependency tree level-by-level
- For every level with mapped words, factorize their corresponding values in the provenance
- Prioritize which values to take-out in each level by frequency

Factorization Algorithm

Proposition

- Obtaining a minimal T-compatible factorization is coNP-hard

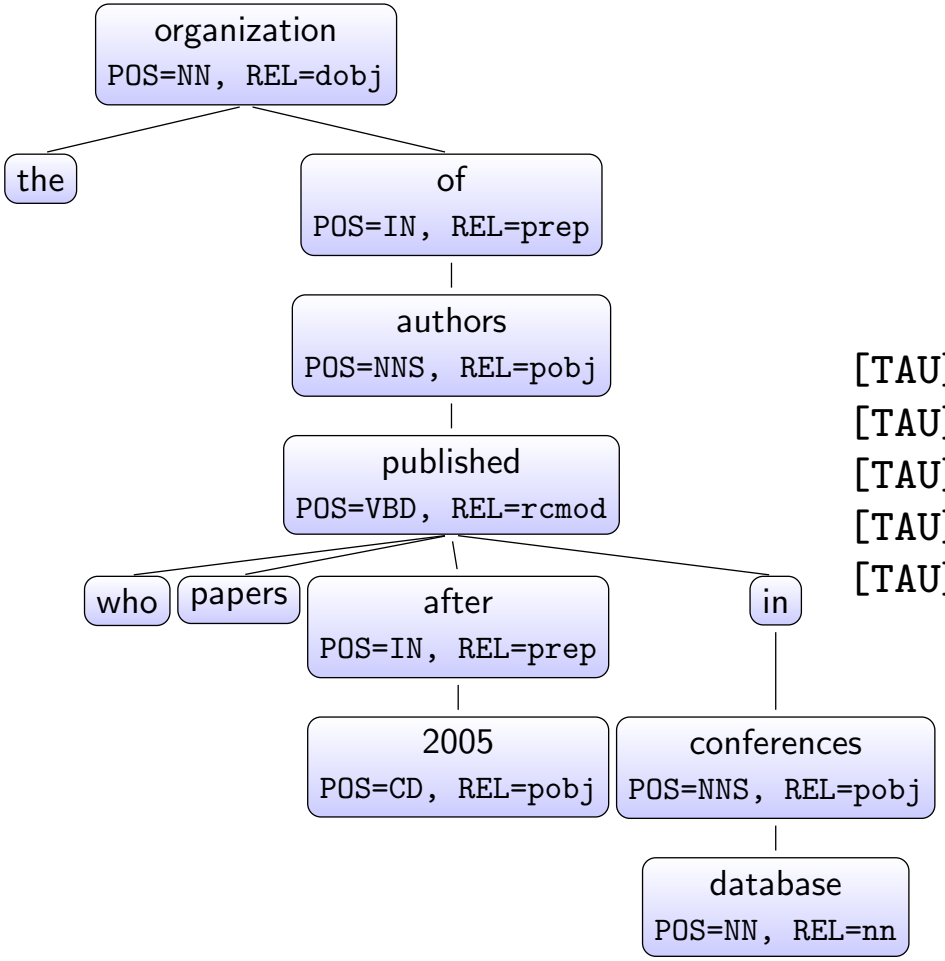
Algorithm

- Factorize greedily: traverse the dependency tree level-by-level
- For every level with mapped words, factorize their corresponding values in the provenance
- Prioritize which values to take-out in each level by frequency

Complexity

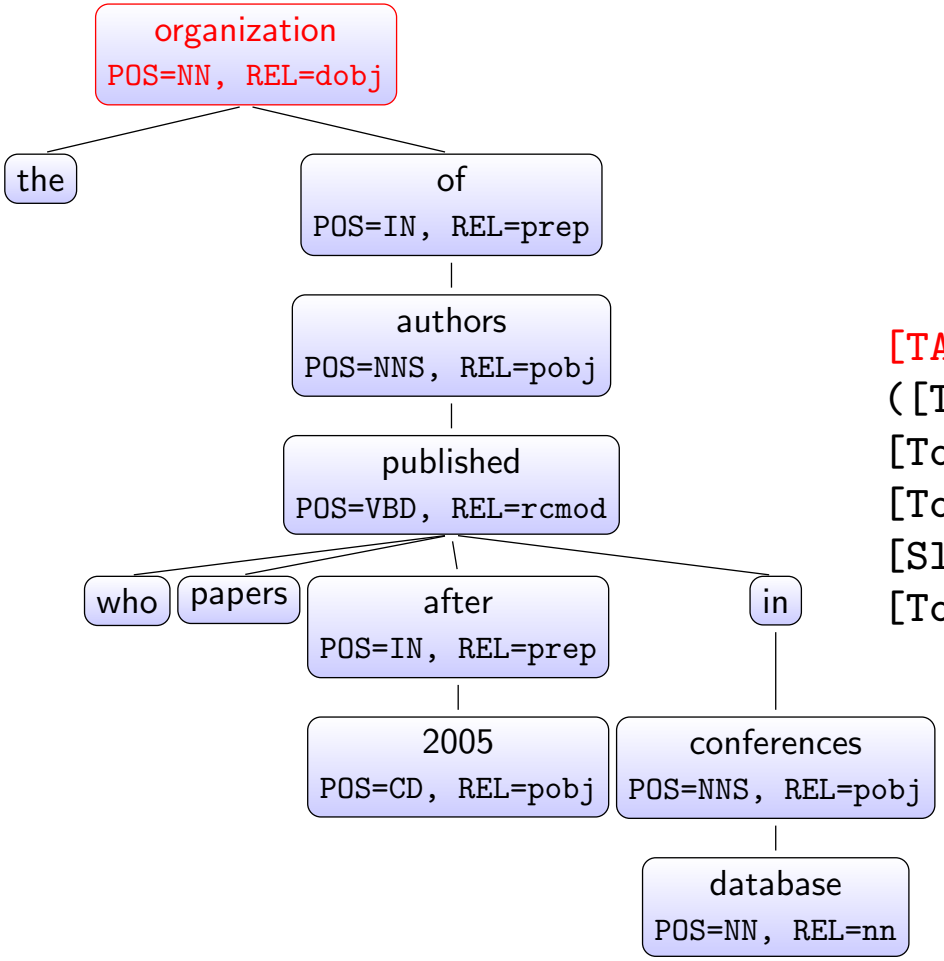
- $\mathcal{O}(n^2 \cdot \log n)$: recursively traverse the dependency tree and sort the variables at each layer by their frequency in $\mathcal{O}(n \cdot \log n)$

Factorization Example



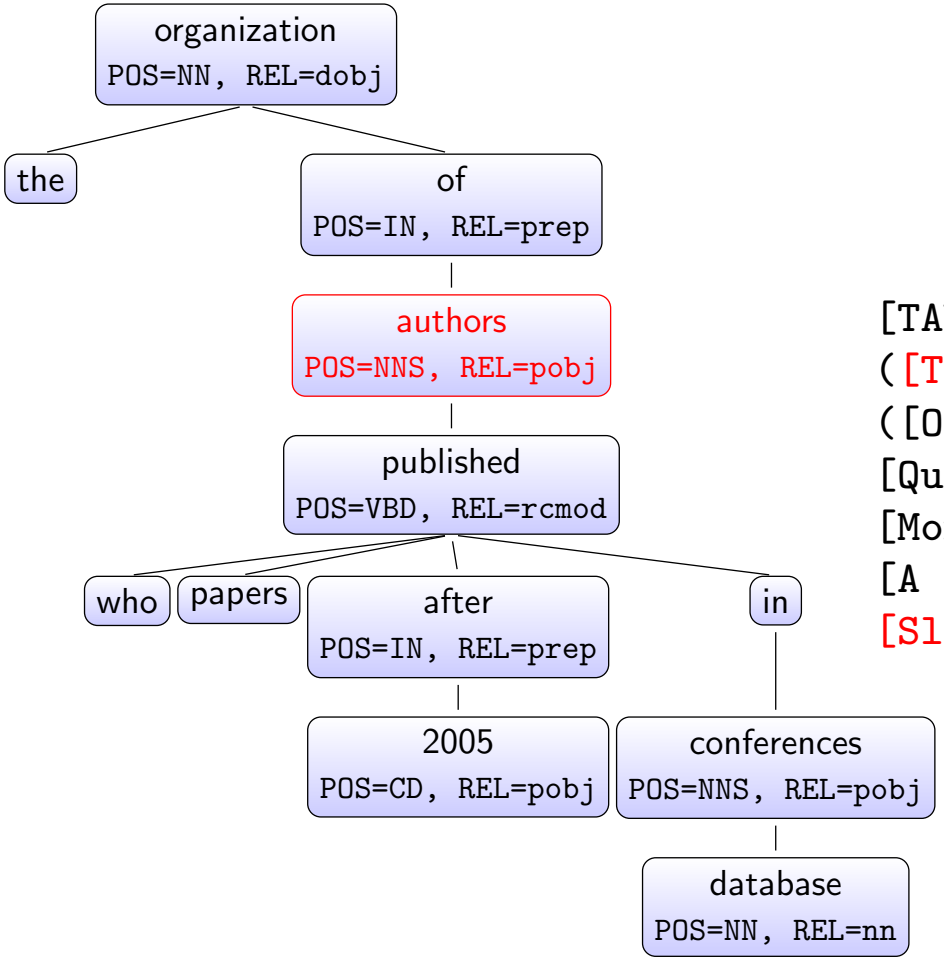
[TAU]·[Tova M.]·[OASSIS...]·[SIGMOD]·[2014]+
[TAU]·[Tova M.]·[Querying...]·[VLDB]·[2006]+
[TAU]·[Tova M.]· [Monitoring..]·[VLDB]·[2007]+
[TAU]·[Slava N.]·[OASSIS...]·[SIGMOD]·[2014]+
[TAU]·[Tova M.]·[A sample...]·[SIGMOD]·[2014]

Factorization Example



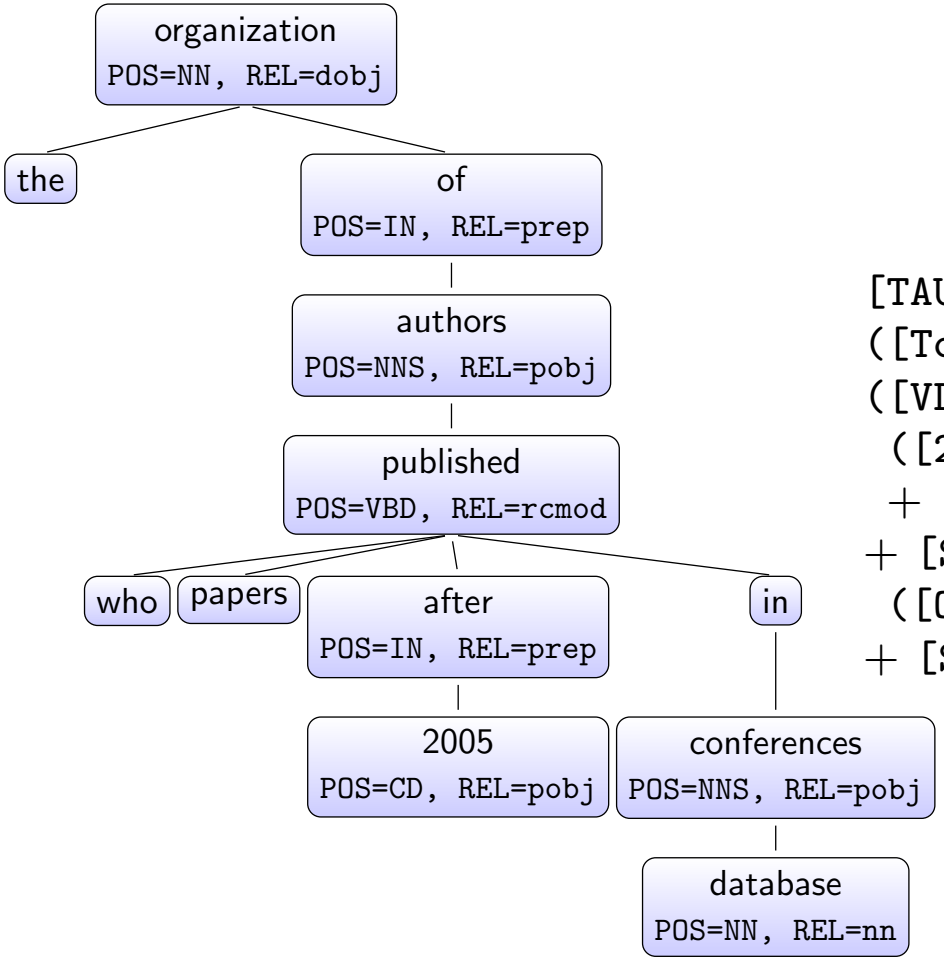
[TAU] ·
([Tova M.]·[OASSIS...]·[SIGMOD]·[2014]+
[Tova M.]·[Querying...]·[VLDB]·[2006]+
[Tova M.]· [Monitoring..]·[VLDB]·[2007]+
[Slava N.]·[OASSIS...]·[SIGMOD]·[2014]+
[Tova M.]·[A sample...]·[SIGMOD]·[2014])

Factorization Example



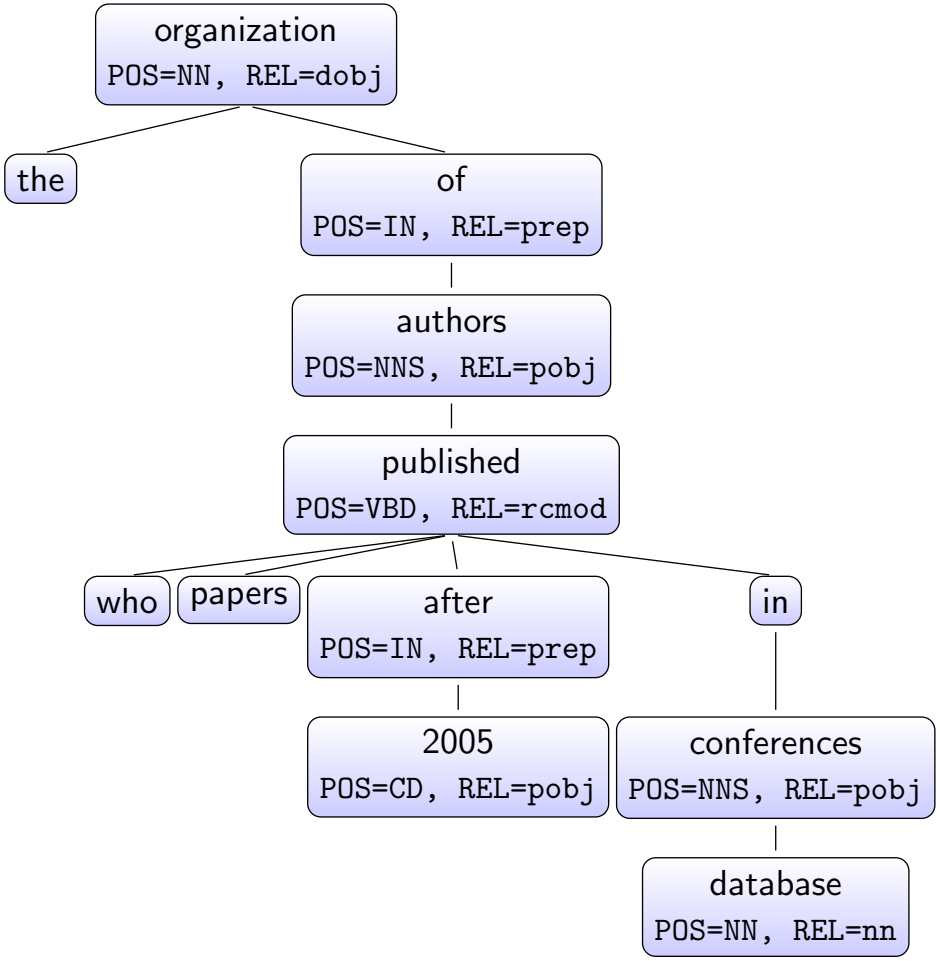
[TAU] ·
([Tova M.] ·
([OASSIS...]·[SIGMOD]·[2014]+
[Querying...]·[VLDB]·[2006]+
[Monitoring..]·[VLDB]·[2007]+
[A sample...]·[SIGMOD]·[2014])+
[Slava N.] · [OASSIS...] · [SIGMOD] · [2014])

Factorization Example




```
[TAU] ·
([Tova M.] ·
([VLDB] ·
 ([2006] · [Querying...]
 + [2007] · [Monitoring...]))
+ [SIGMOD] · [2014] ·
 ([OASSIS...] + [A Sample...]))
+ [Slava N.] · [OASSIS...] · [SIGMOD] · [2014])
```

Factorization Example




```
TAU is the organization of
 Tova M. who published
 in VLDB
 'Querying...'  in 2006 and
 'Monitoring...'  in 2007
 and in SIGMOD in 2014
 'OASSIS...'  and 'A sample...'
 and Slava N. who published
 'OASSIS...'  in SIGMOD in 2014.
```

Outline

Summarization

Two Levels of Summarization

```
[TAU] ·
  A { ([Tova M.] ·
        B { ([VLDB] ·
              ([2006] · [Querying...]
              + [2007] · [Monitoring...]))
            + [SIGMOD] · [2014] ·
              ([OASSIS...] + [A Sample...]))  } B
      + [Slava N.] · [OASSIS...] · [SIGMOD] · [2014]) } A
```

Shorter Summarized Answer Based on A

TAU is the organization of 2 authors who published 4 papers in 2 conferences in 2006 - 2014

More Detailed Summarized Answer Based on B

TAU is the organization of Tova M. who published 4 papers in 2 conferences in 2006 - 2014 and Slava N. who published 'OASSIS...' in SIGMOD in 2014.

Outline

Sample Use-Cases

Q: Return the authors who published papers in VLDB before 2016 and after 2007

- **A:** Tova M. published 16 papers in VLDB in 2008 - 2015

Q: Return the authors who published papers in database conferences

- **A:** Tova M. published 134 papers in 18 conferences

Q: Return the organization of authors who published papers in database conferences after 2005

- **A:** TAU is the organization of 43 authors who published 170 papers in 31 conferences in 2006 - 2015

Sample Use-Cases

Q: Return the authors who published papers in VLDB before 2016 and after 2007

- **A:** Tova M. published 16 papers in VLDB in 2008 - 2015

Q: Return the authors who published papers in database conferences

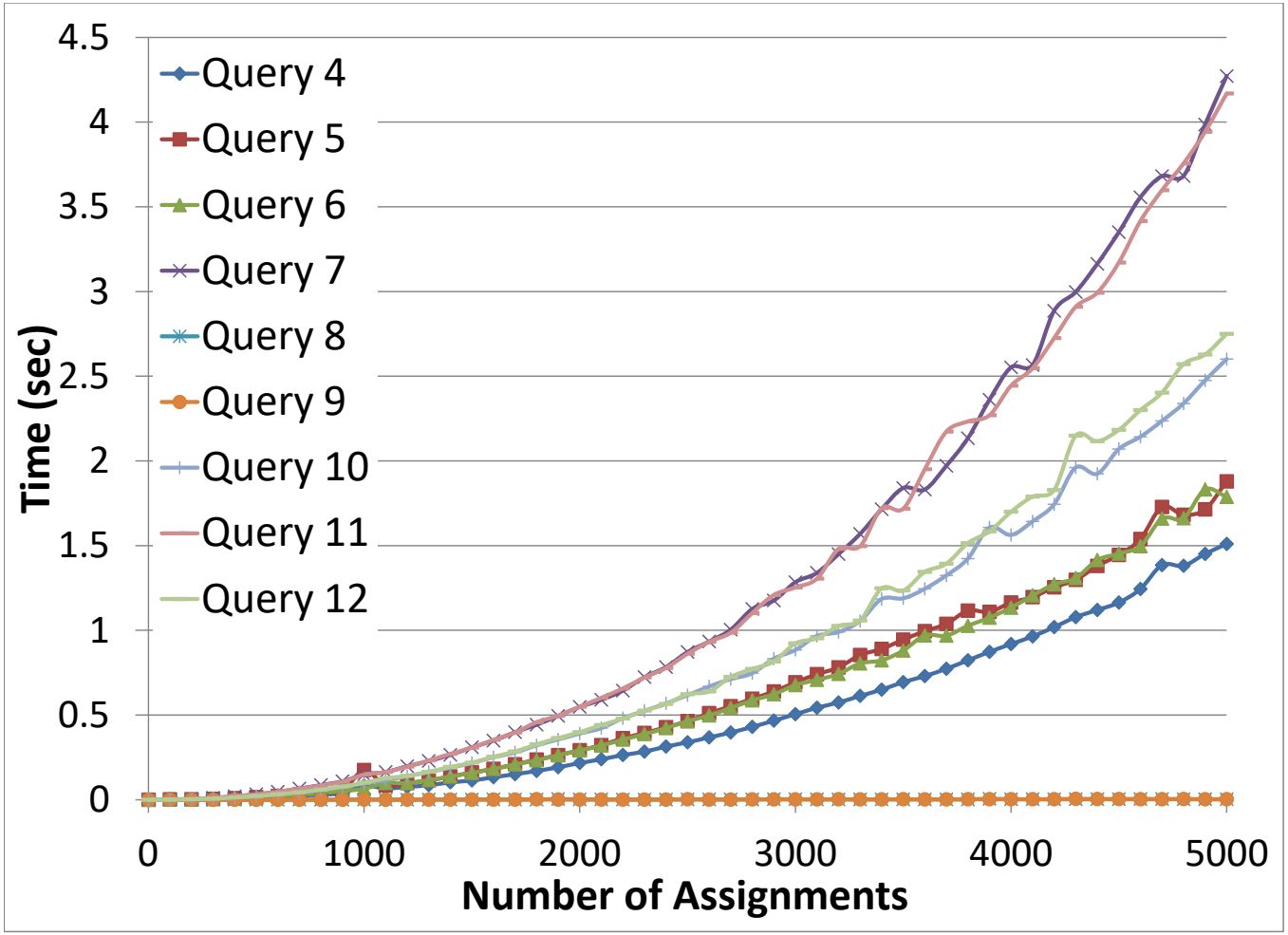
- **A:** Tova M. published 134 papers in 18 conferences

Q: Return the organization of authors who published papers in database conferences after 2005

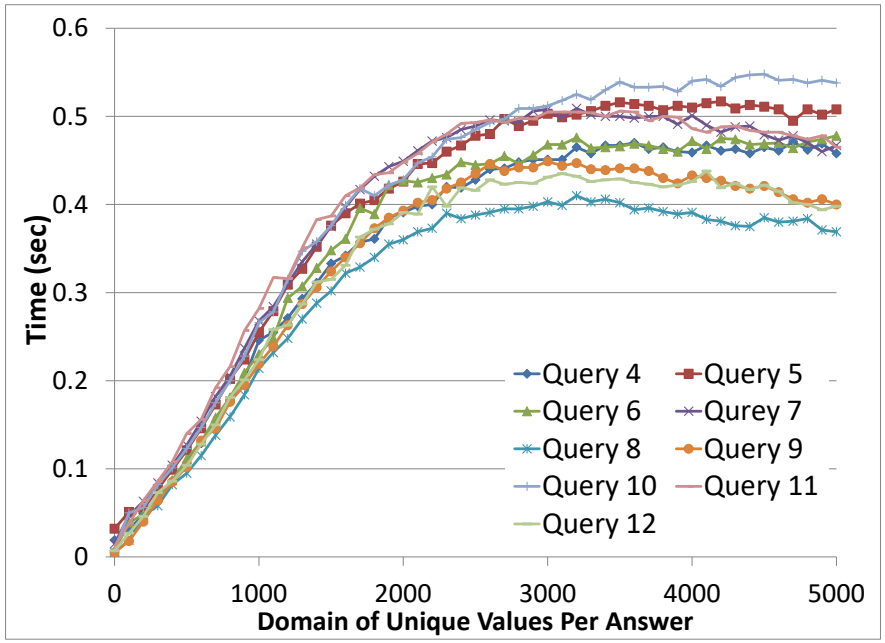
- **A:** TAU is the organization of 43 authors who published 170 papers in 31 conferences in 2006 - 2015

Sample Scalability Results

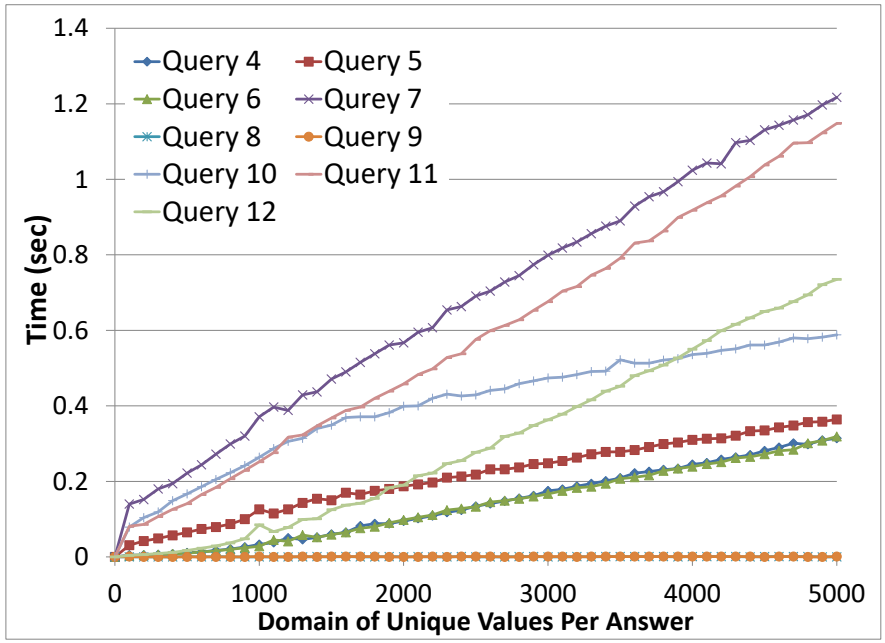
Computation time as a function of the number of assignments.
Overhead of only 16% w.r.t evaluation time.



Breakdown of Computation Time



(a) Factorization time



(b) Sentence gen. time

Summarization time was negligible.

Outline

Related Work

NL Interfaces:

- Formulate the NL query and present the answers, e.g., [Fei Li et al., 15'], [Song et al., 15']
- Present the answers in NL based on the schema [Franconi et al., 14']
- Explain the query in NL [Koutrika et al., 10']

Provenance:

- Showing the provenance in graph form, e.g., [Ailamaki et al., 98'], [Davidson et al., 08']
- Allowing user control over granularity [Cohen-Boulakia et al., 08']
- Provenance factorization and Summarization, e.g., [Chapman et al., 08'], [Olteanu et al., 12']

Summary

Main Contributions:

- First to formulate the *provenance* of output tuples in NL
- Employing both factorization and summarization to make provenance more understandable
- Devising a criterion for provenance factorization that accounts for its *presentation in NL*

Summary

Main Contributions:

- First to formulate the *provenance* of output tuples in NL
- Employing both factorization and summarization to make provenance more understandable
- Devising a criterion for provenance factorization that accounts for its *presentation in NL*

Future Work:

- Extend the solution to UCQs, aggregation, nested queries, ...
- Support more provenance models
- Generalize the requirements from NL interfaces

Thank You

Questions?