Knowledge Graph and Corpus Driven Segmentation and Answer Inference for Telegraphic Entity-seeking Queries

EMNLP 2014

Mandar Joshi

IBM Research

mandarj90@in.ibm.com

Uma Sawant

IIT Bombay, Yahoo Labs

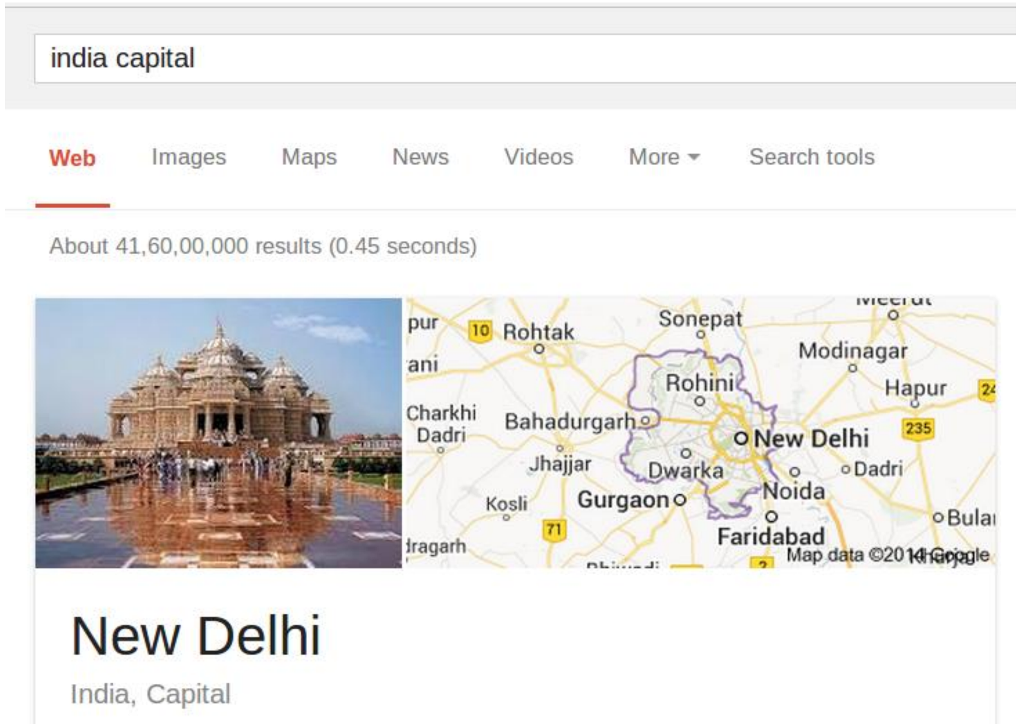
uma@cse.iitb.ac.in

Soumen Chakrabarti

IIT Bombay

soumen@cse.iitb.ac.in

Entity-seeking Telegraphic Queries



- Short
- Unstructured (like natural language questions)
- Expect entities as answers

Challenges

- No reliable syntax clues
  - Free word order
  - No or rare capitalization, quoted phrases
- Ambiguous
  - Multiple interpretations
  - `aamir khan films`
    - Aamir Khan - the Indian actor or British boxer
    - Films - appeared in, directed by, or about
- Previous QA work
  - Convert to structured query
  - Execute on knowledge graph (KG)

WHY DO WE NEED THE CORPUS?

- KG is high precision but incomplete
  - Work in progress
  - Triples can not represent all information
  - Structured – unstructured gap
- Corpus provides recall
- `fastest odi century batsman`

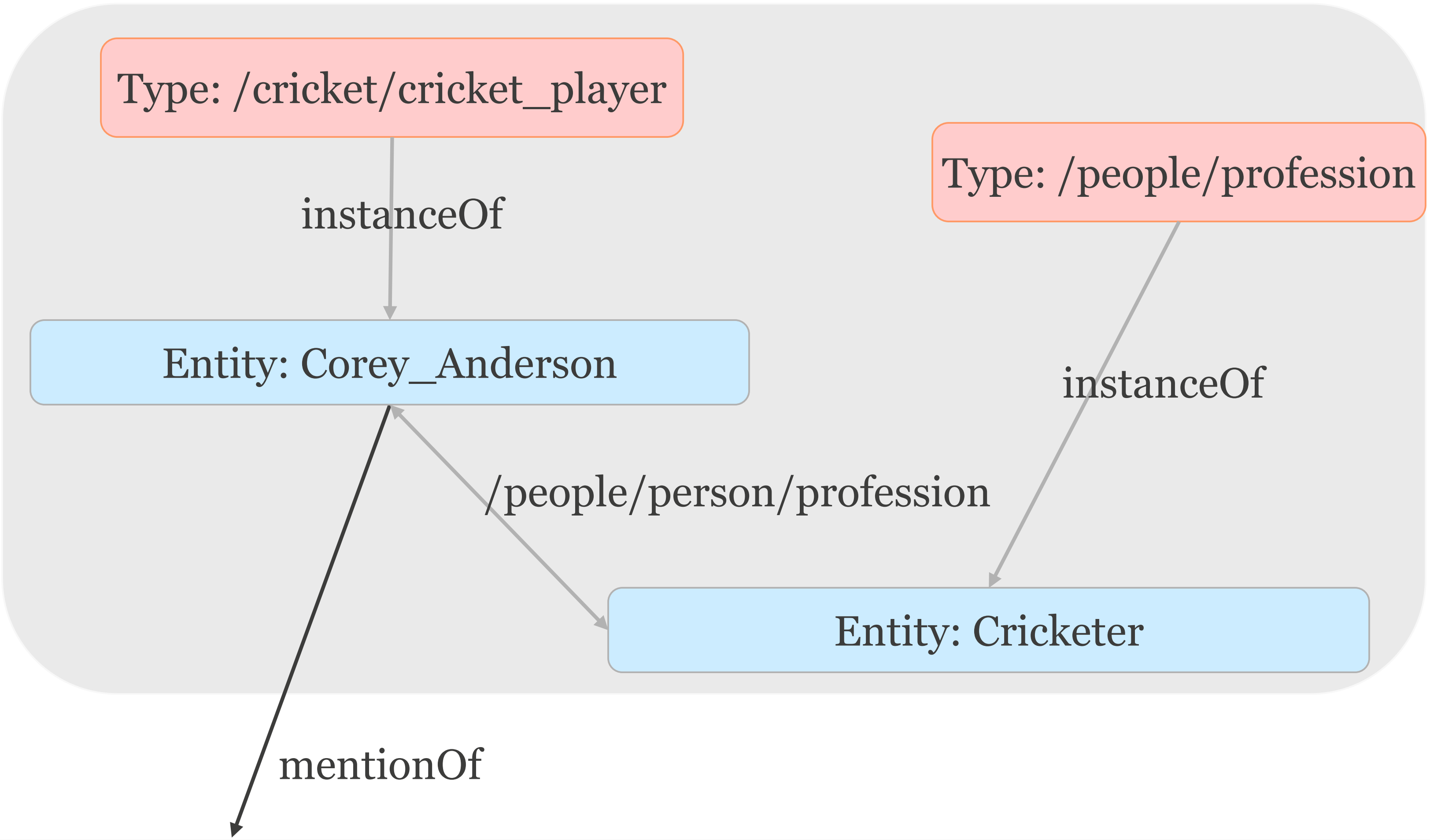
> ... Corey Anderson hits **fastest ODI century**. This was the first time two **batsmen** have hit hundreds in under 50 balls in the same **ODI**.

WHY DO WE NEED THE CORPUS?

- KG is high precision but incomplete
  - Work in progress
  - Triples can not represent all information
  - Structured – unstructured gap
- Corpus provides recall
- `fastest odi century batsman`

> .. Corey Anderson hits **fastest ODI century**. This was the first time two **batsmen** have hit hundreds in under 50 balls in the same **ODI**.

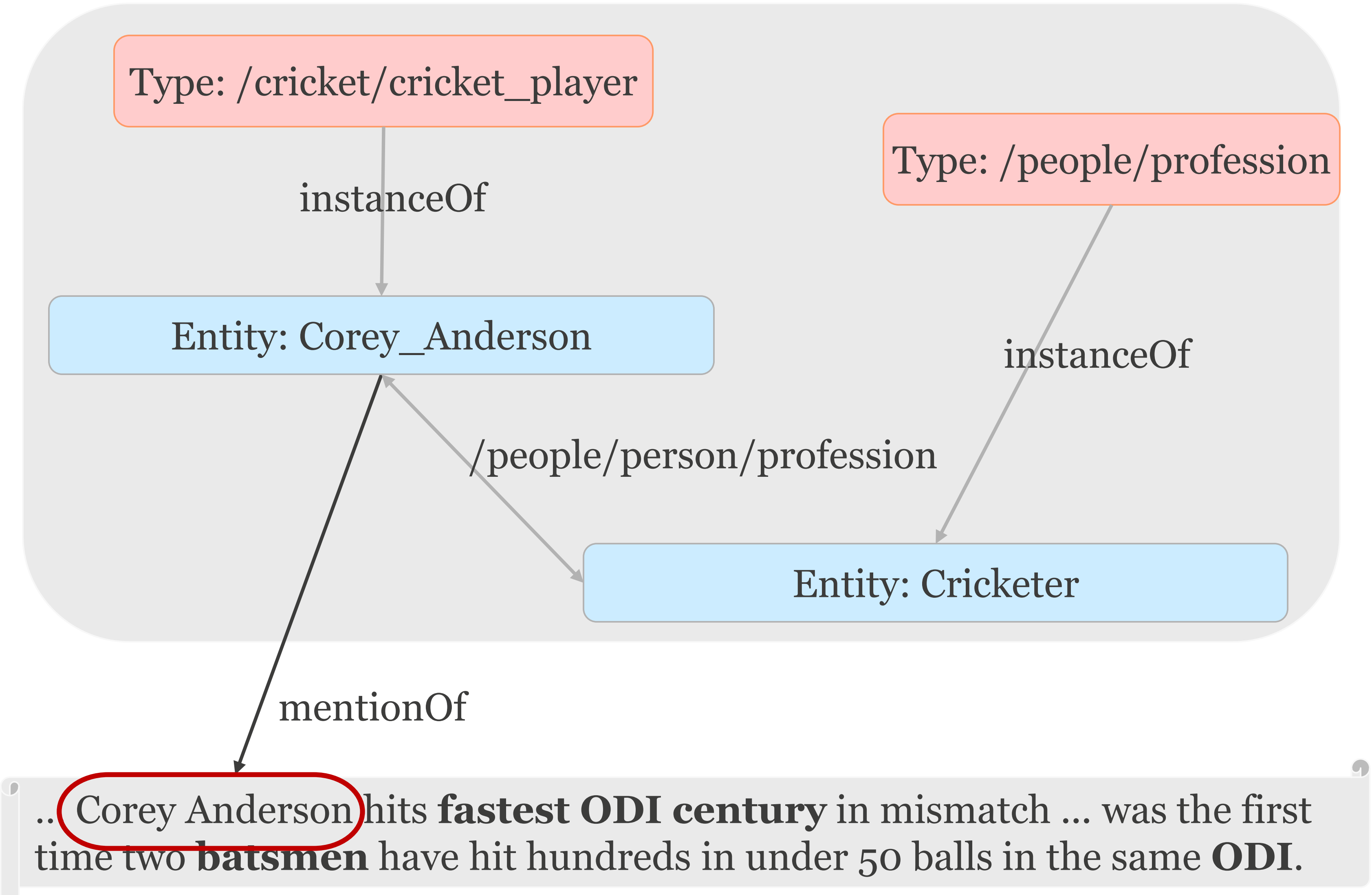
ANNOTATED WEB WITH KNOWLEDGE GRAPH



... Corey Anderson hits **fastest ODI century** in mismatch ... was the first time two **batsmen** have hit hundreds in under 50 balls in the same **ODI**.

Annotated document

Annotated Web with Knowledge Graph



Annotated document

Interpretation via Segmentation

Signals from the Query

- Queries seek answer entities (e_2)
- Contain (query) *entities* (e_1) , *target types* (t_2), *relations* (r), and *selectors* (s).

query	e_1	r	t_2	s
`washington first governor`	washington	governor	governor	first
	washington	-	governor	first
`spider automobile company`	spider	-	automobile company	-
	automobile	company	company	spider

Assignment of tokens to columns for illustration only; not necessarily optimal

Segmentation and Interpretation

- Interpretation = Segmentation + Annotation
- Segmentation of query tokens into 3 partitions
  - Query entity (E_1)
  - Relation and Type (T_2/R)
  - Selectors (S)
- Multiple ways to annotate each partition

Segmentation and Interpretation

- Interpretation = Segmentation + Annotation
- Segmentation of query tokens into 3 partitions
  - Query entity (E_1)
  - Relation and Type (T_2/R)
  - Selectors (S)
- Multiple ways to annotate each partition

washington first governor

SEGMENTATION AND INTERPRETATION

- Interpretation = Segmentation + Annotation
- Segmentation of query tokens into 3 partitions
  - Query entity (E_1)
  - Relation and Type (T_2/R)
  - Selectors (S)
- Multiple ways to annotate each partition

washington first governor

E_1 partition S partition T_2/R partition

Segmentation and Interpretation

- Interpretation = Segmentation + Annotation
- Segmentation of query tokens into 3 partitions
  - Query entity (E_1)
  - Relation and Type (T_2/R)
  - Selectors (S)
- Multiple ways to annotate each partition

washington first governor

E_1 partition S partition T_2/R partition

1. Washington (State)
2. Washington_D.C. (City)

Segmentation and Interpretation

- Interpretation = Segmentation + Annotation
- Segmentation of query tokens into 3 partitions
  - Query entity (E_1)
  - Relation and Type (T_2/R)
  - Selectors (S)
- Multiple ways to annotate each partition

washington first governor

E_1 partition

S partition

T_2/R partition

1. Washington (State)
2. Washington_D.C. (City)

r: governorOf t2: us_state_governor
r: null t2: us_state_governor

Combining KG and Corpus Evidence

Combining KG and Corpus Evidence

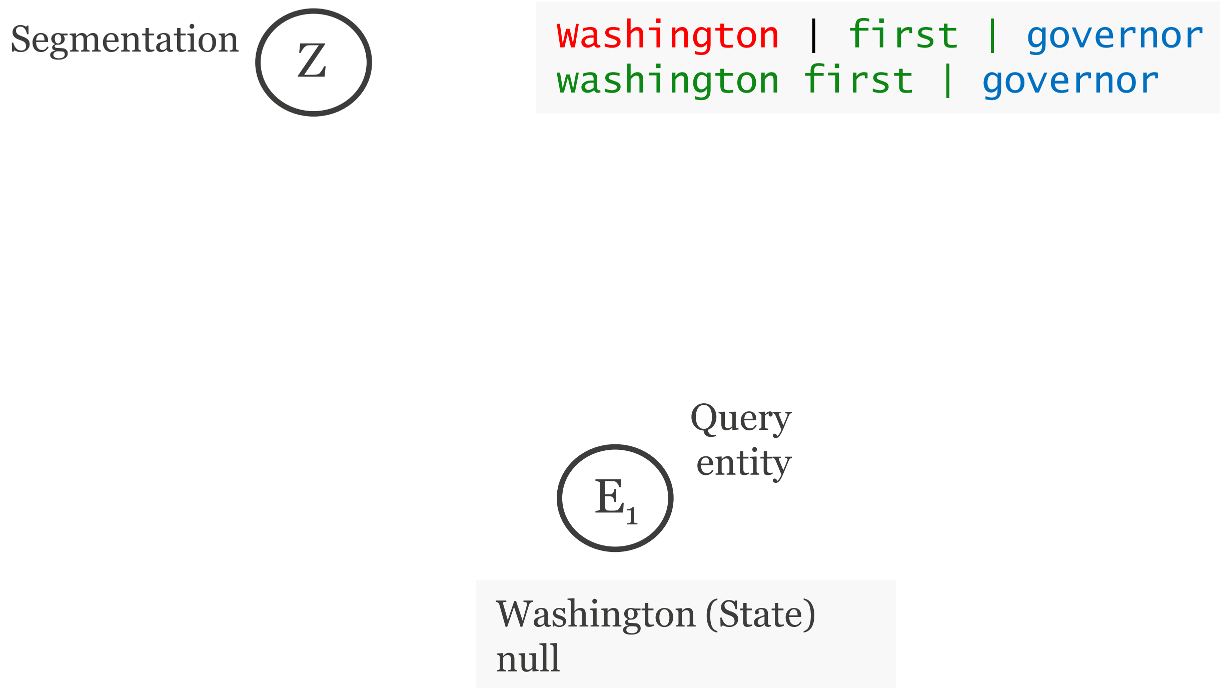
Segmentation Z

Combining KG and Corpus Evidence

Segmentation Z

```
Washington | first | governor
washington first | governor
```

Combining KG and Corpus Evidence



Combining KG and Corpus Evidence

Segmentation Z

Washington | first | governor
washington first | governor

Target type T_2

us_state_governor
governor_general

Query entity E_1

Washington (State)
null

Combining KG and Corpus Evidence

Segmentation Z

```
Washington | first | governor
washington first | governor
```

Target type T_2

us_state_governor
governor_general

R Relation

governorOf
null

E_1 Query entity

Washington (State)
null

Combining KG and Corpus Evidence

Segmentation Z

```
Washington | first | governor
washington first | governor
```

Target type T_2

us_state_governor
governor_general

R Relation

governorOf
null

E_1 Query entity

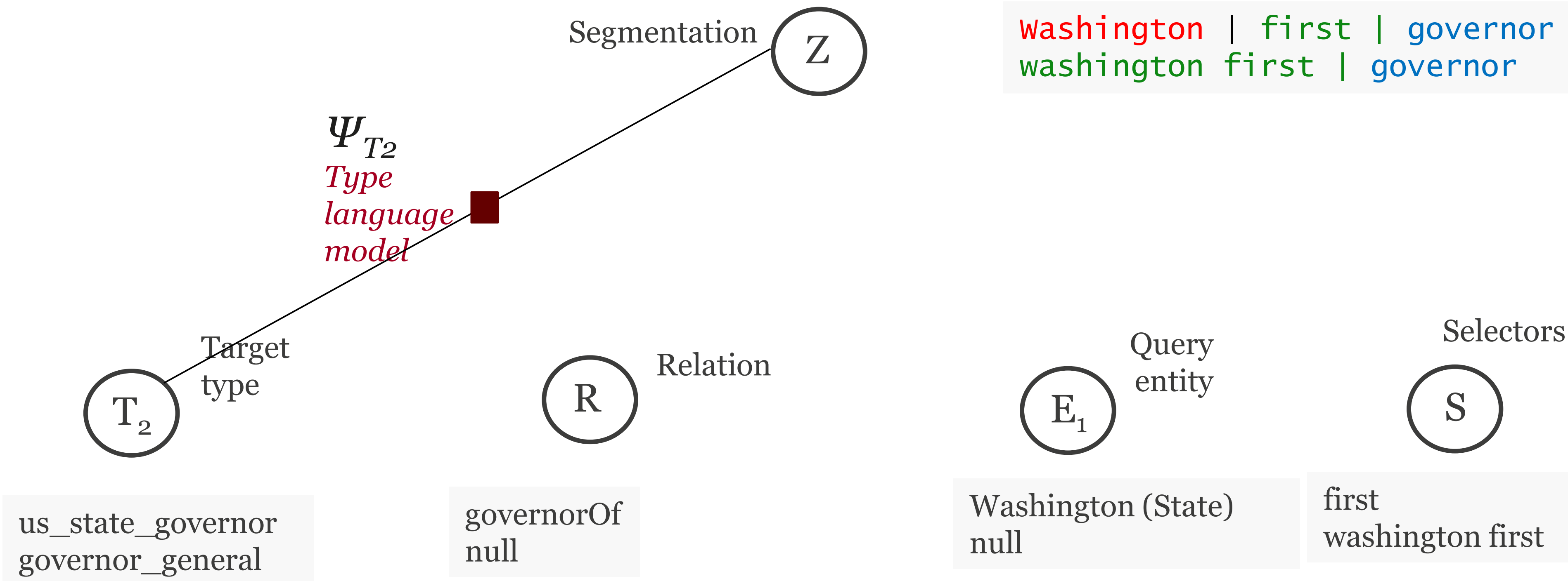
Washington (State)
null

Selectors

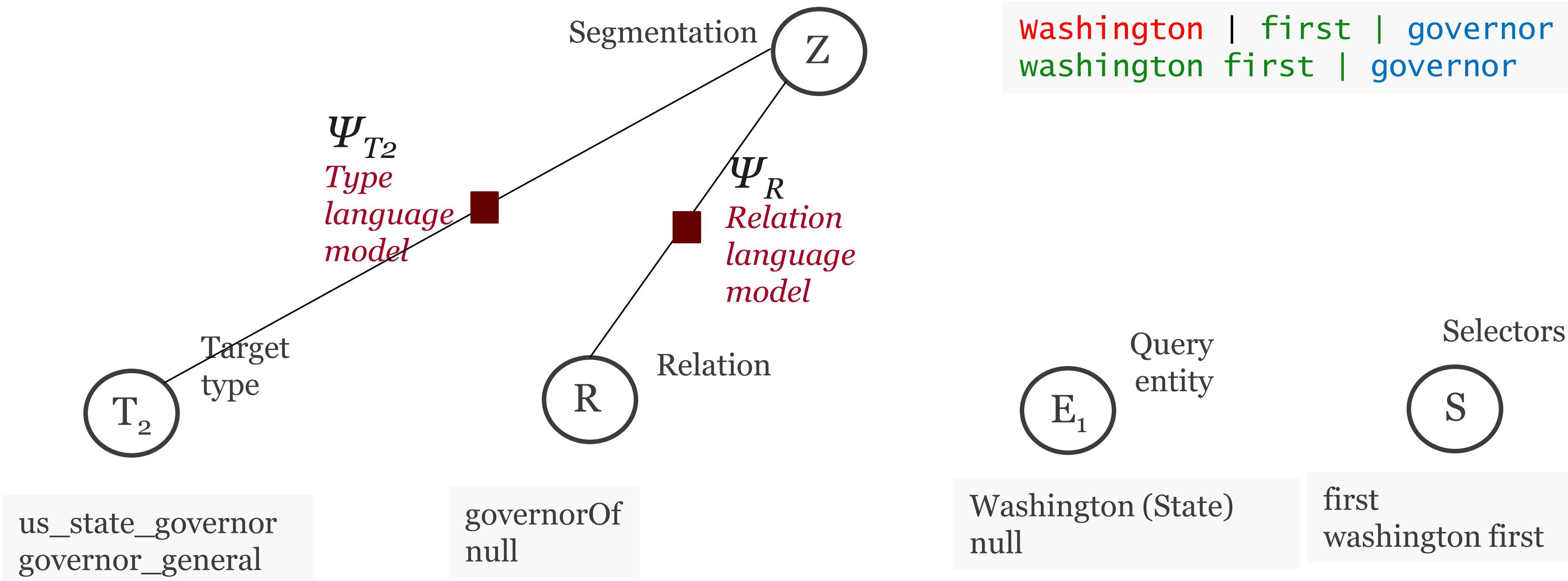
S

first
washington first

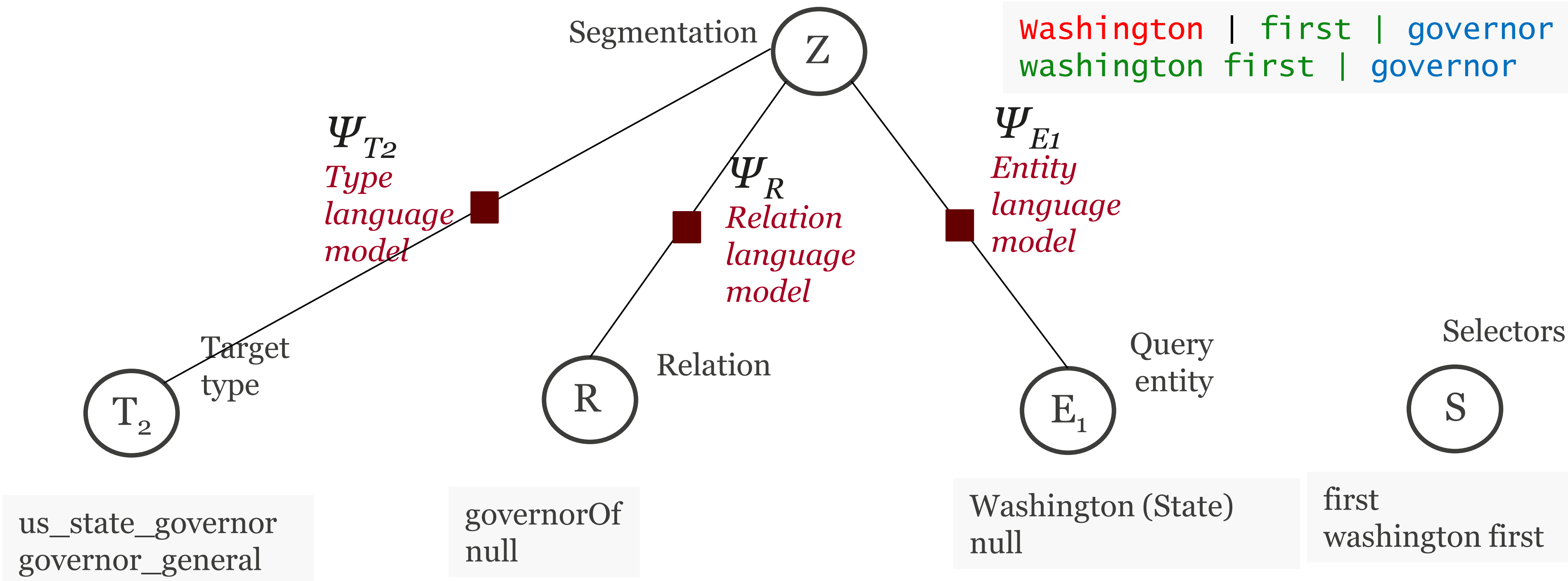
Combining KG and Corpus Evidence



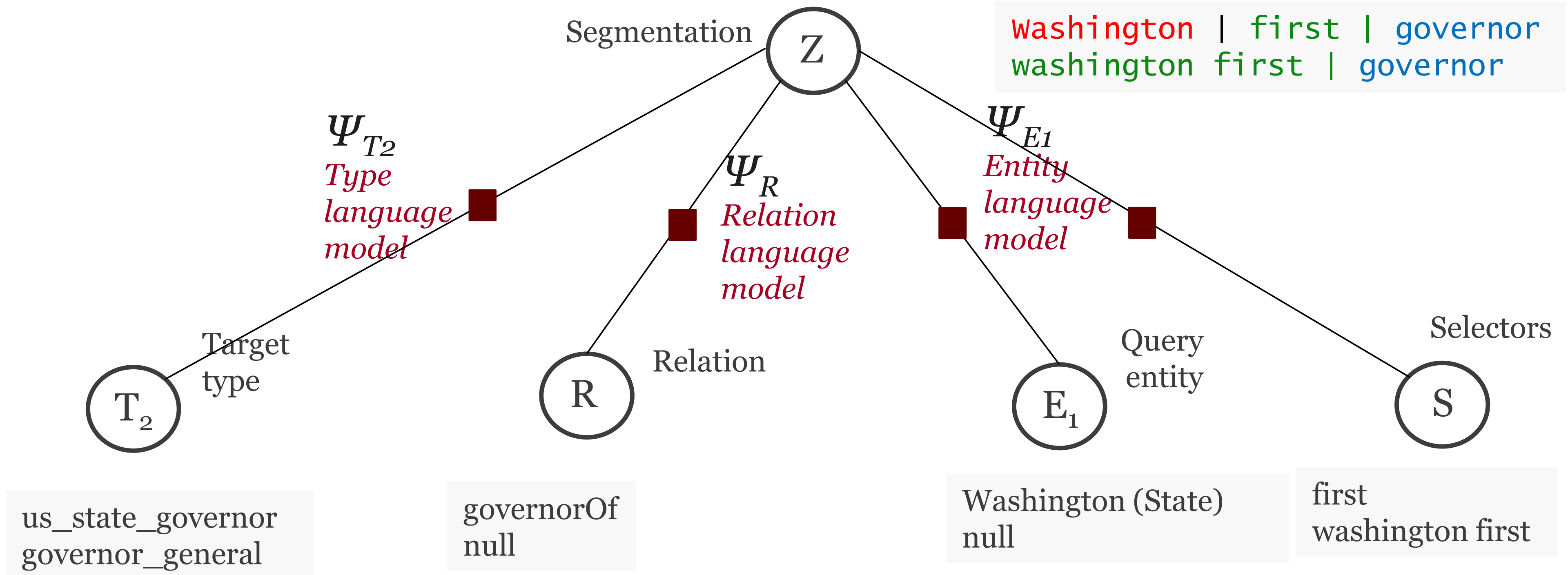
Combining KG and Corpus Evidence



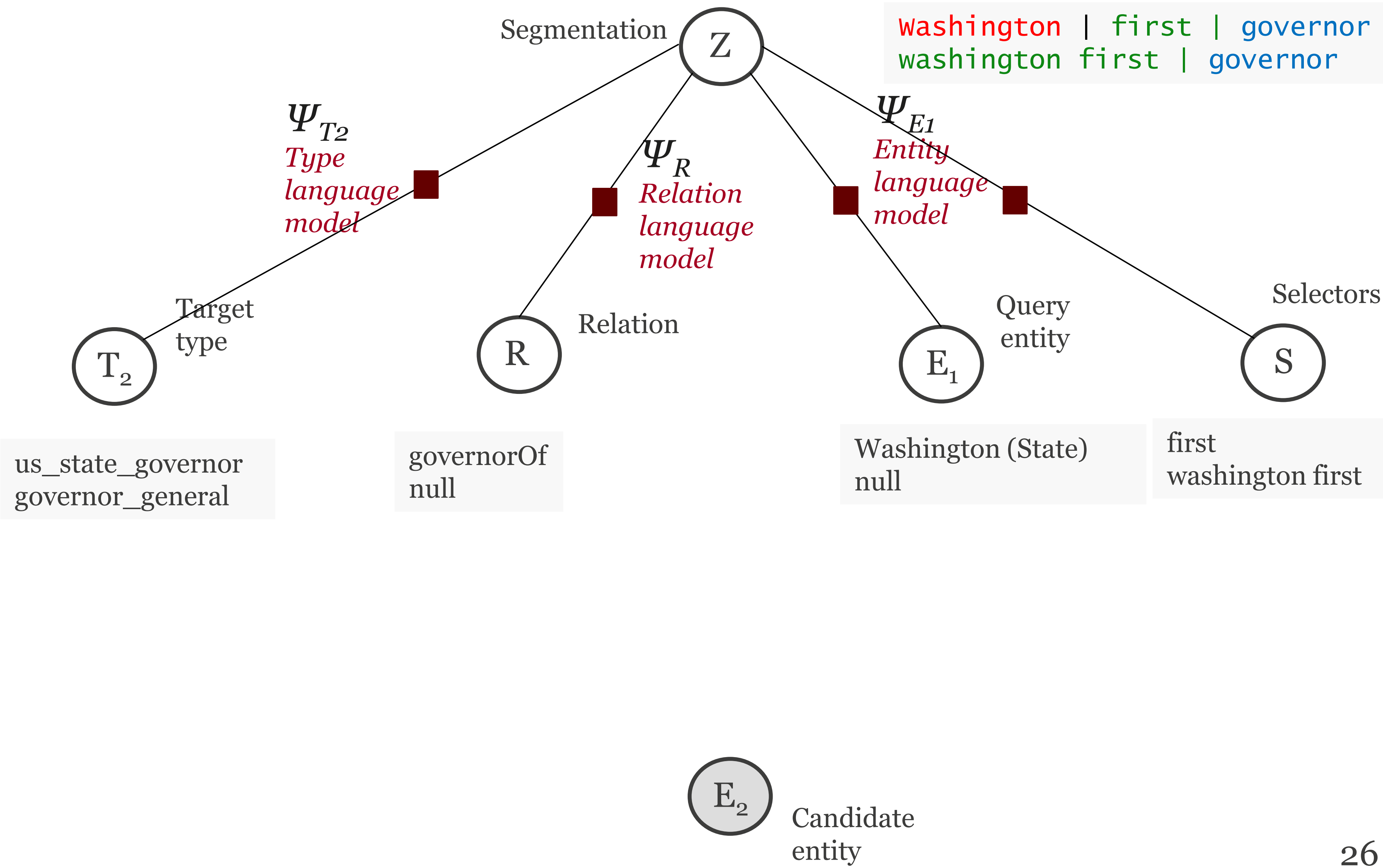
Combining KG and Corpus Evidence



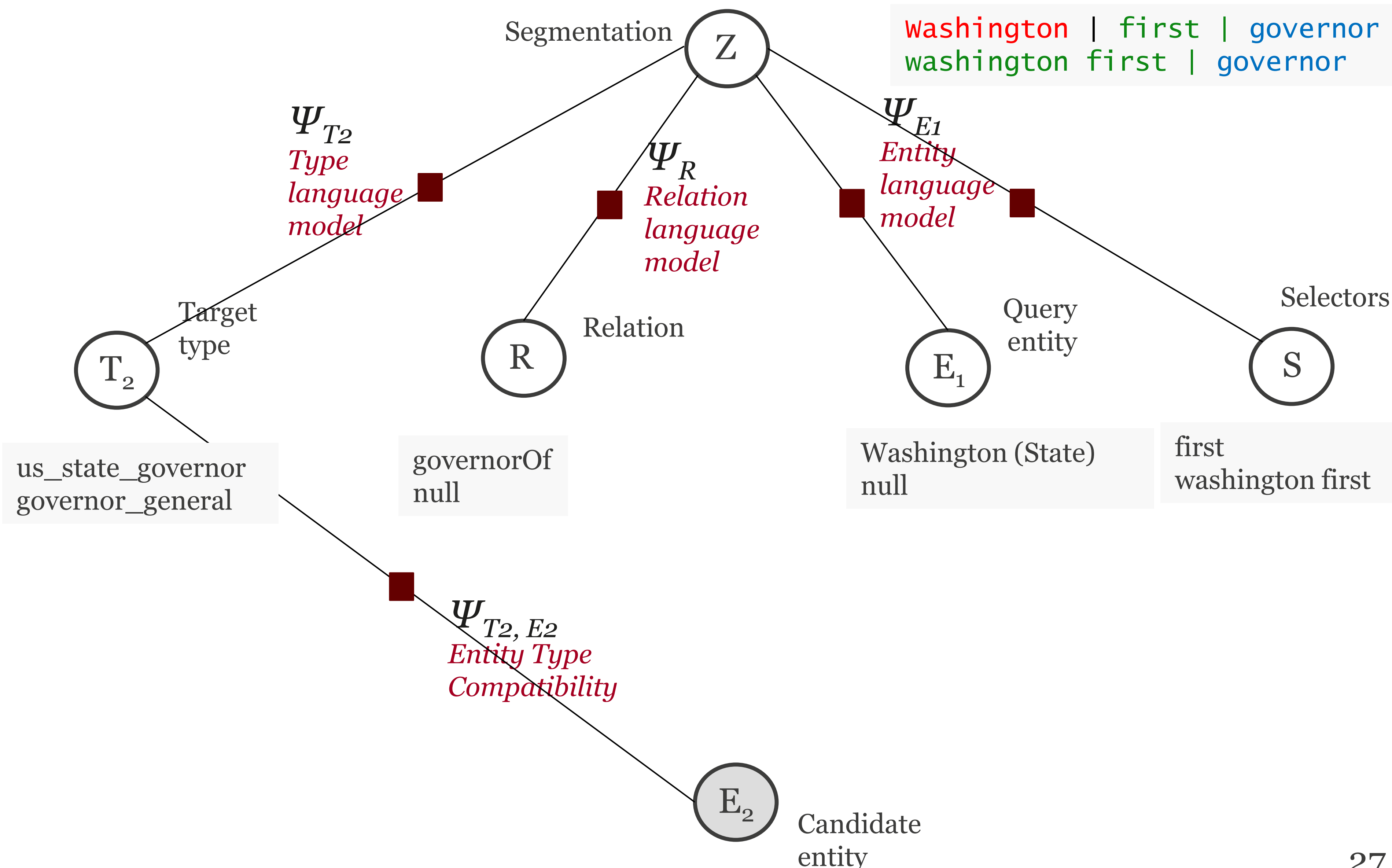
Combining KG and Corpus Evidence



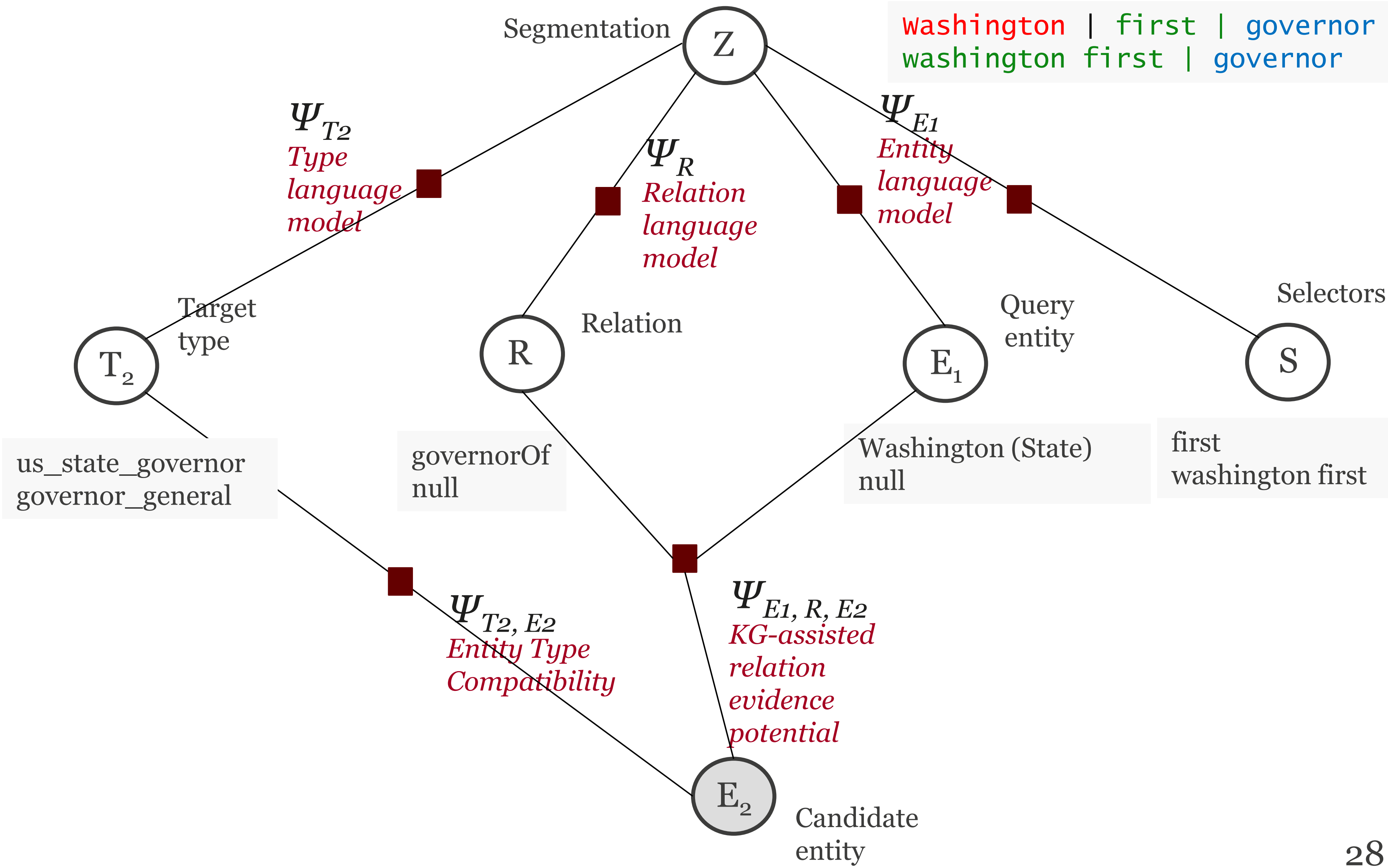
Combining KG and Corpus Evidence



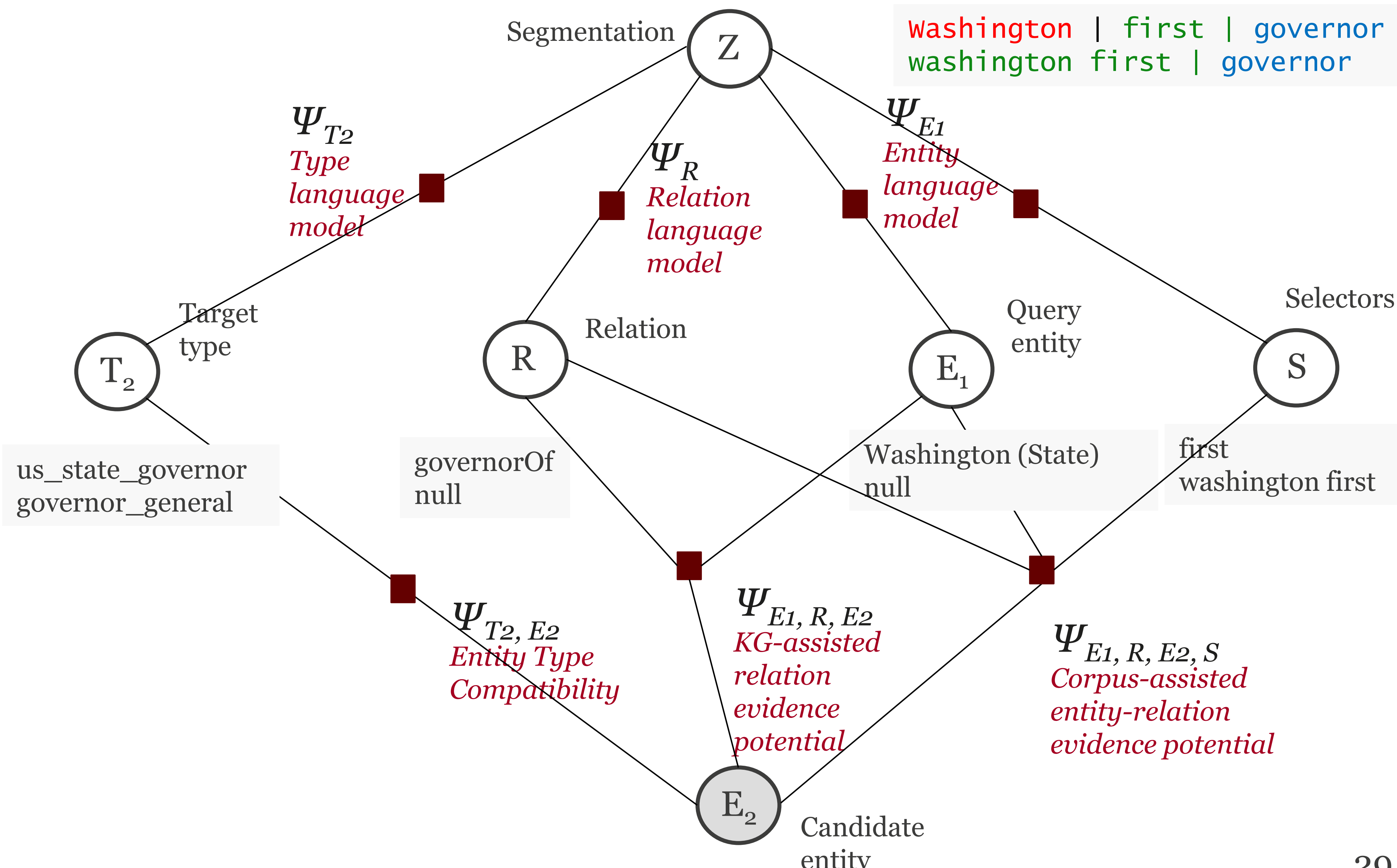
Combining KG and Corpus Evidence



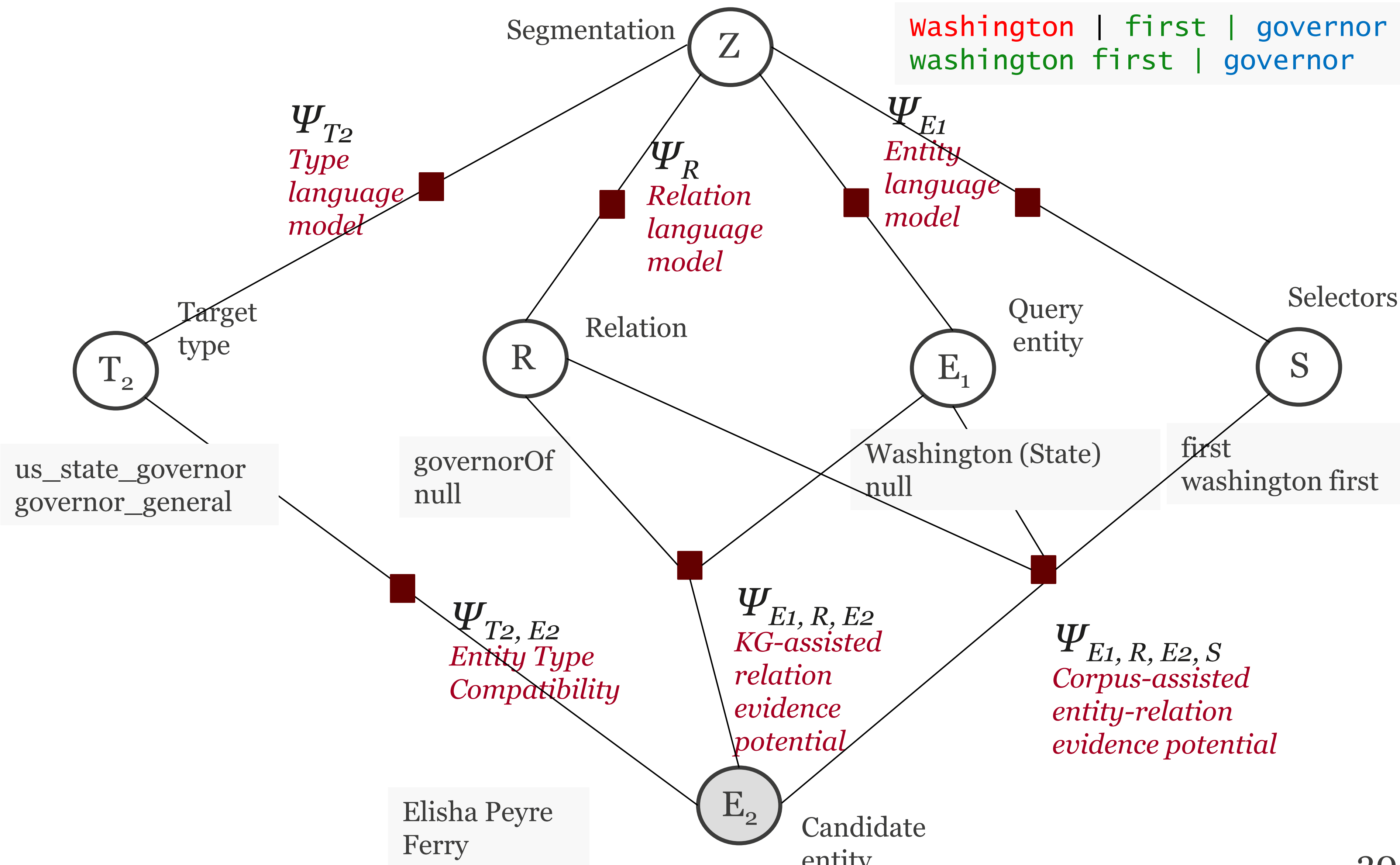
Combining KG and Corpus Evidence



Combining KG and Corpus Evidence



Combining KG and Corpus Evidence



From query to Answer Entity

- Generate interpretations
- Retrieve snippets for each interpretation
- Construct candidate answer entities (e_2) set
  - Top *k* from corpus based on snippet frequency
  - By KG links that are in interpretations set
- Inference

$$score(e_2) = \max_{z,t_2,r,e_1} \Psi_{T_2}(q,z,t_2)\ \Psi_R(q,z,r)\ \Psi_{E_1}(q,z,e_1)\ \Psi_{E_2,T_2}(e_2,t_2)\ \Psi_{E_1,R,E_2,S}(e_1,r,e_2,s)\ \Psi_{E_1,R,E_2}(e_1,r,e_2)$$

From query to Answer Entity

- Generate interpretations
- Retrieve snippets for each interpretation
- Construct candidate answer entities (e_2) set
  - Top *k* from corpus based on snippet frequency
  - By KG links that are in interpretations set
- Inference

query – signals compatibility

$$score(e_2) = \max_{z,t_2,r,e_1} \Psi_{T_2}(q,z,t_2)\ \Psi_R(q,z,r)\ \Psi_{E_1}(q,z,e_1)$$

$$\Psi_{E_2,T_2}(e_2,t_2)$$

$$\Psi_{E_1,R,E_2,S}(e_1,r,e_2,s)\ \Psi_{E_1,R,E_2}(e_1,r,e_2)$$

From query to Answer Entity

- Generate interpretations
- Retrieve snippets for each interpretation
- Construct candidate answer entities (e_2) set
  - Top *k* from corpus based on snippet frequency
  - By KG links that are in interpretations set
- Inference

query – signals compatibility

$$score(e_2) = \max_{z,t_2,r,e_1} \Psi_{T_2}(q,z,t_2)\ \Psi_R(q,z,r)\ \Psi_{E_1}(q,z,e_1)$$

$$\Psi_{E_2,T_2}(e_2,t_2)$$

e2-t2 compatibility

$$\Psi_{E_1,R,E_2,S}(e_1,r,e_2,s)\ \Psi_{E_1,R,E_2}(e_1,r,e_2)$$

From query to Answer Entity

- Generate interpretations
- Retrieve snippets for each interpretation
- Construct candidate answer entities (e_2) set
  - Top *k* from corpus based on snippet frequency
  - By KG links that are in interpretations set
- Inference

query – signals compatibility

$$score(e_2) = \max_{z,t_2,r,e_1} \quad \Psi_{T_2}(q,z,t_2)\ \Psi_R(q,z,r)\ \Psi_{E_1}(q,z,e_1)$$

$$\Psi_{E_2,T_2}(e_2,t_2)$$

e2-t2 compatibility

$$\Psi_{E_1,R,E_2,S}(e_1,r,e_2,s)\ \Psi_{E_1,R,E_2}(e_1,r,e_2)$$

evidence from KG and corpus

Relation and Type Models

- Objective: To map relation (or type) mentions in query to Freebase relation (or types)
- Relation Language Model (Ψ_R)
  - Use annotated ClueWeb09 + Freebase triples
  - Locate Freebase relation endpoints in corpus
  - Extract dependency path words between entities
  - Maintain co-occurrence counts of <words, rel>
  - Assumption: Co-occurrence implies relation
- Type Language Model (Ψ_{T2})
  - Smoothed Dirichlet language model using Freebase type names

Corpus Potential

- Estimates support to e_1-r-e_2-s in corpus
- Snippet retrieval and scoring
- Snippets scored using RankSVM
- Partial list of features
  - #snippets with distance(e_2 , e_1) < k (k = 5, 10)
  - #snippets with distance(e_2 , r) < k (k = 3, 6)
  - #snippets with relation r = ⊥
  - #snippets with relation phrases as prepositions
  - #snippets covering fraction of query IDF > k (k = 0.2, 0.4, 0.6, 0.8)

Latent Variable Discriminative Training (LVDT)

- Constraints are formulated using the best scoring interpretation
- Training

$$\max_{q,z,e_1,t_2,r} w \cdot \phi(q,z,e_1,t_2,r,e_2^+) + \xi$$
$$\geq 1 + \max_{q,z,e_1,t_2,r} w \cdot \phi(q,z,e_1,t_2,r,e_2^-)$$

- Inference

$$\max_{q,z,e_1,t_2,r} w \cdot \phi(q,z,e_1,t_2,r,e_2),$$

- q, e_2 are observed; e_1, t_2, r and z are latent
- Non-convex formulation

Experiments

TEST BED

- Freebase entity, type and relation knowledge graph
  - ~29 million entities
  - 14000 types
  - 2000 selected relation

- Annotated corpus
  - Clueweb09B Web corpus having 50 million pages
  - Google (FACC1), ~ 13 annotations per page
  - Text and Entity Index

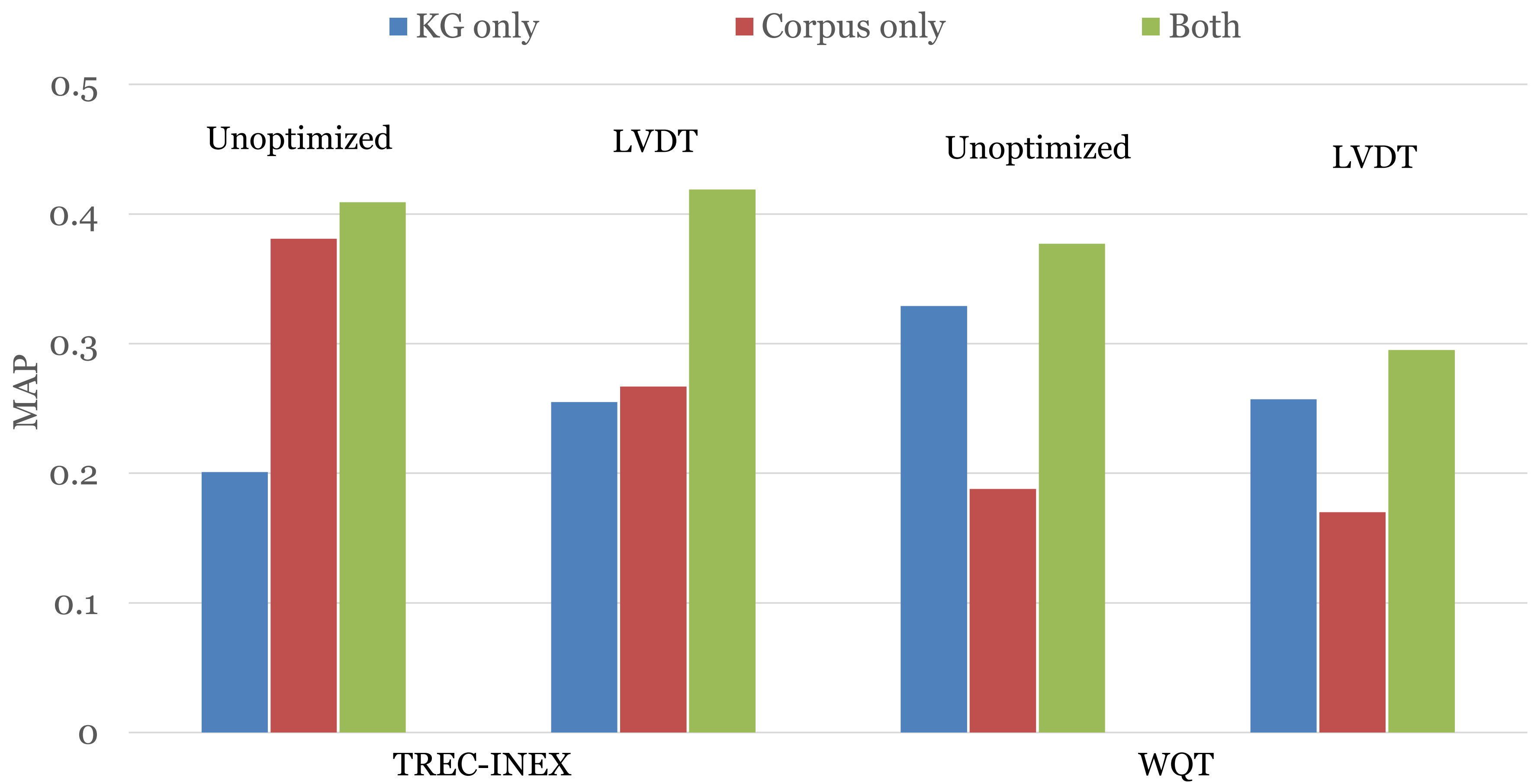
TEST BED

- Query sets
  - TREC-INEX: 700 entity search queries
  - WQT: Subset of ~800 queries from WebQuestions (WQ) natural language query set [1], manually converted to telegraphic form
  - Available at http://bit.ly/Spva49

TREC-INEX	WQT
Has type and/or relation hints	Has mostly relation hints
Answers from KG and corpus collected by volunteers	Answers from KG only collected by turkers.
Answer evidence from corpus (+ KG)	Answer evidence from KG

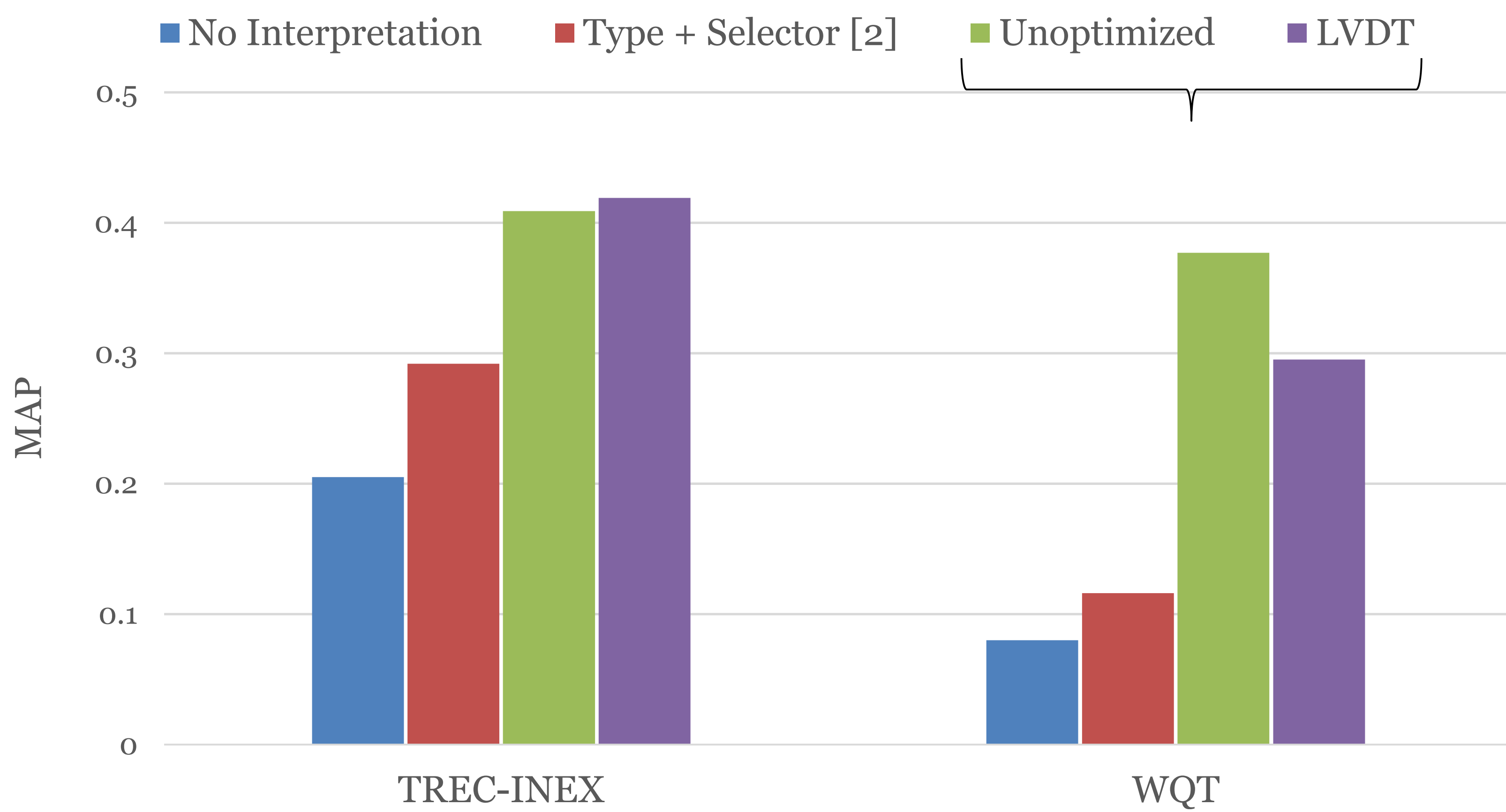
[1] Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on Freebase from question-answer pairs. In Empirical Methods in Natural Language Processing (EMNLP).

Synergy Between KG and Corpus



Corpus and knowledge graph help each other to deliver better performance

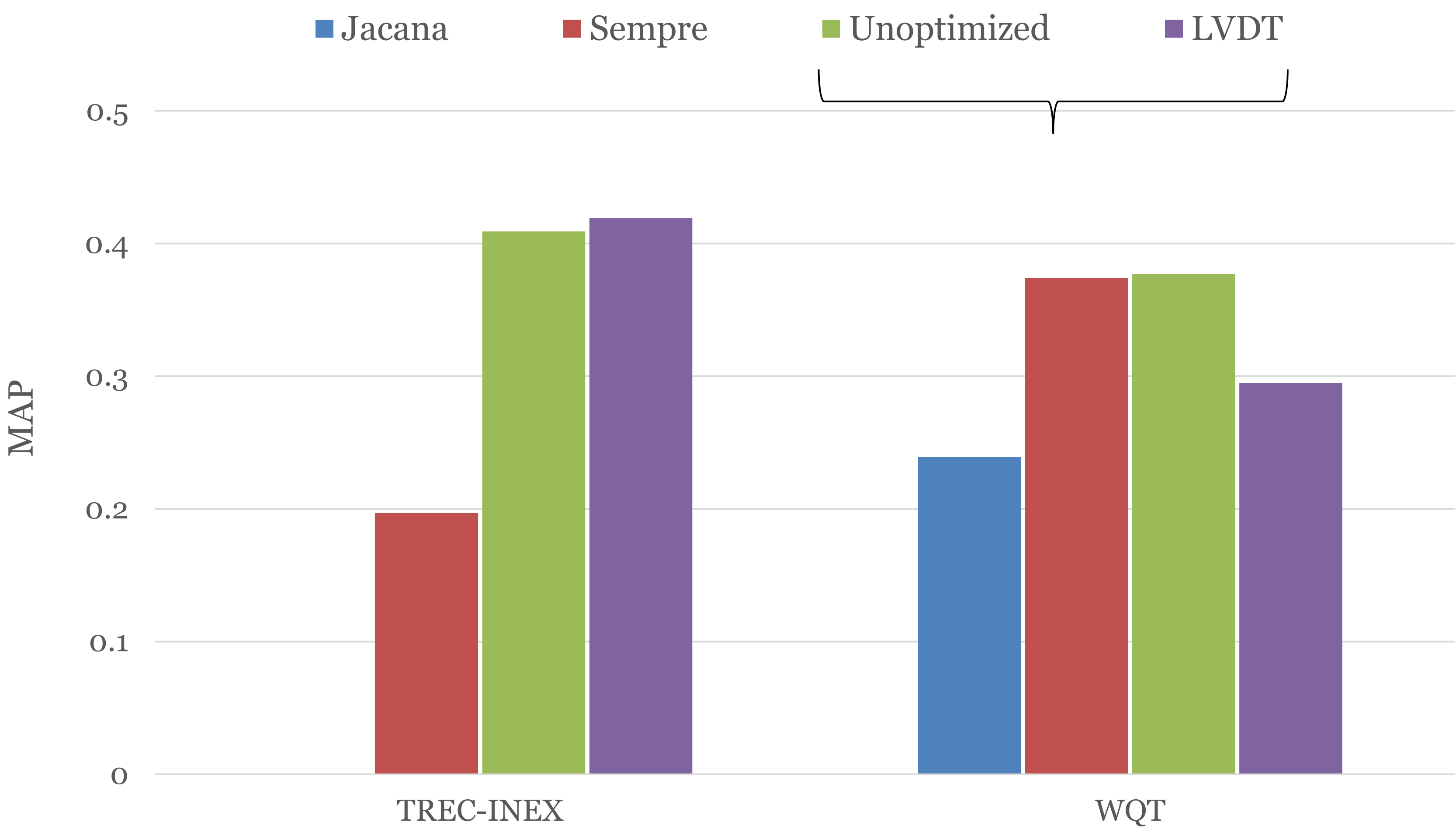
QUERY TEMPLATE COMPARISON



Entity-relation-type-selector template provides yields better accuracy than type-selector template

[2] Uma Sawant and Soumen Chakrabarti. 2013. Learning joint query interpretation and response ranking. In WWW Conference, Brazil.

COMPARISON WITH SEMANTIC PARSERS



QUALITATIVE COMPARISON

- Benefits of collective inference
  - `automobile company makes spider`
  - Entity model fails to identify e_1 (Alfa Romeo Spider)
  - Recovery: `automobile company makes spider`

- Limitations
  - Sparse corpus annotations
  - `south africa political system`
  - Few corpus annotations for e_2: Constitutional Republic
  - Can't find appropriate t_2 (/../form_of_government) and r (/location/country/form_of_government)

QUALITATIVE COMPARISON

- Benefits of collective inference
  - `automobile company makes spider`
  - Entity model fails to identify e_1 (Alfa Romeo Spider)
  - Recovery: `automobile company makes spider`

    e_1: Automobile

- Limitations
  - Sparse corpus annotations
  - `south africa political system`
  - Few corpus annotations for e_2: Constitutional Republic
  - Can't find appropriate t_2 (/../form_of_government) and r (/location/country/form_of_government)

Qualitative Comparison

- Benefits of collective inference
  - `automobile company makes spider`
  - Entity model fails to identify e_1 (Alfa Romeo Spider)
  - Recovery: `automobile` `company` `makes spider`

    e_1: Automobile t_2: /../organization r : /business/industry/companies

- Limitations
  - Sparse corpus annotations
  - `south africa political system`
  - Few corpus annotations for e_2: Constitutional Republic
  - Can't find appropriate t_2 (/../form_of_government) and r (/location/country/form_of_government)

Summary

- Query interpretation is rewarding, but non-trivial
- Segmentation based models work well for telegraphic queries
- Entity-relation-type-selector template better than type-selector template
- Knowledge graph and corpus provide complementary benefits

REFERENCES

- S&C: Uma Sawant and Soumen Chakrabarti. 2013. Learning joint query interpretation and response ranking. In WWW Conference, Brazil.
- Sempre: Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on Freebase from question-answer pairs. In Empirical Methods in Natural Language Processing (EMNLP).
- Jacana: Xuchen Yao and Benjamin Van Durme. 2014. Information extraction over structured data: Question answering with Freebase. In ACL Conference. ACL.

DATA

- TREC-INEX and WQT
  - Short URL http://bit.ly/Spva49
  - Long URL https://docs.google.com/spreadsheets/d/1AbKBdFOIXum_NwXeWuboSdeG-y8Ub4_ub8qTjAw4Qug/edit#gid=0
- Project page
  - http://www.cse.iitb.ac.in/~soumen/doc/CSAW/

THANK YOU!
QUESTIONS?