Simple Large-scale Relation Extraction from Unstructured Text

Christos Christodoulopoulos and Arpit Mittal

Amazon Research Cambridge

amazon

Alexa Question Answering

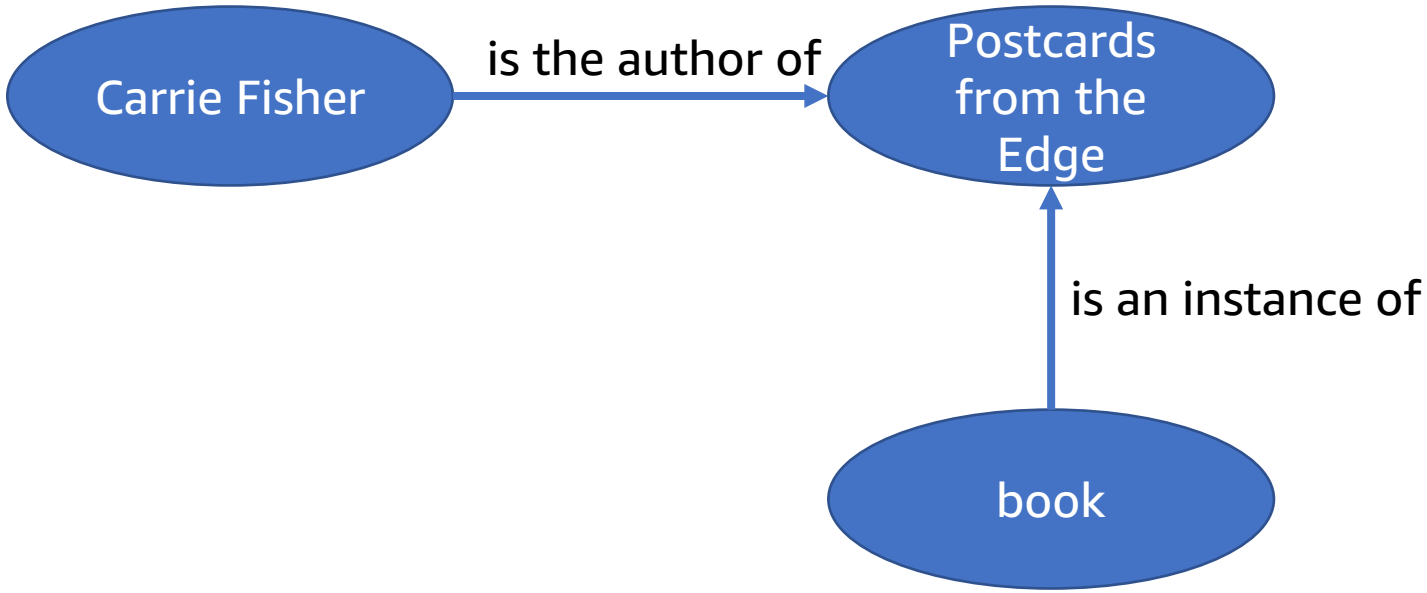
“Alexa, what books did Carrie Fisher write?”

“The books that Carrie Fisher is an author of are Delusions of Grandma, Shockaholic, Surrender the Pink, Postcards from the Edge, The Best Awful There Is and Wishful Drinking.”

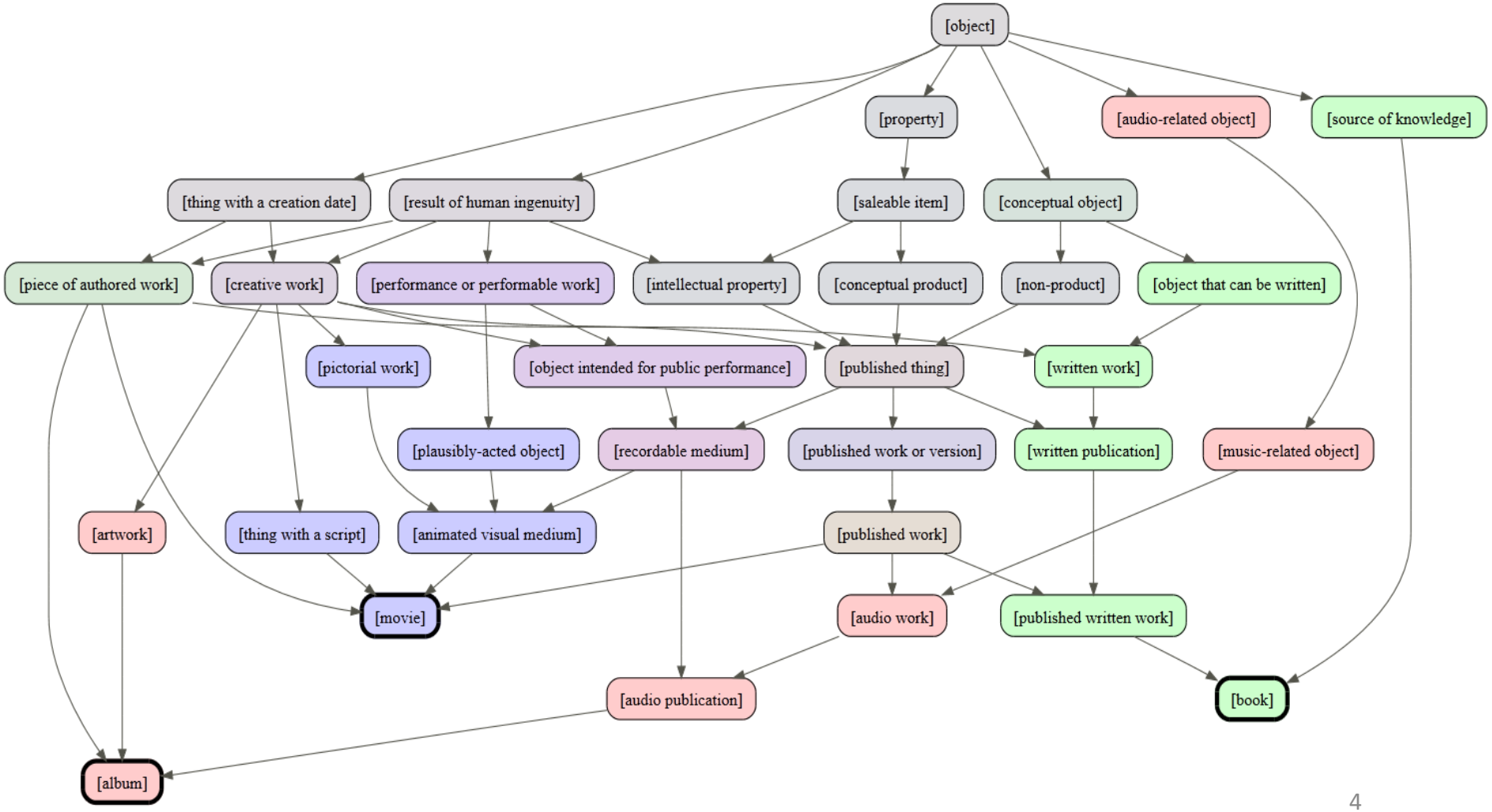
amazon

Alexa Knowledge Base

Named relations between entities



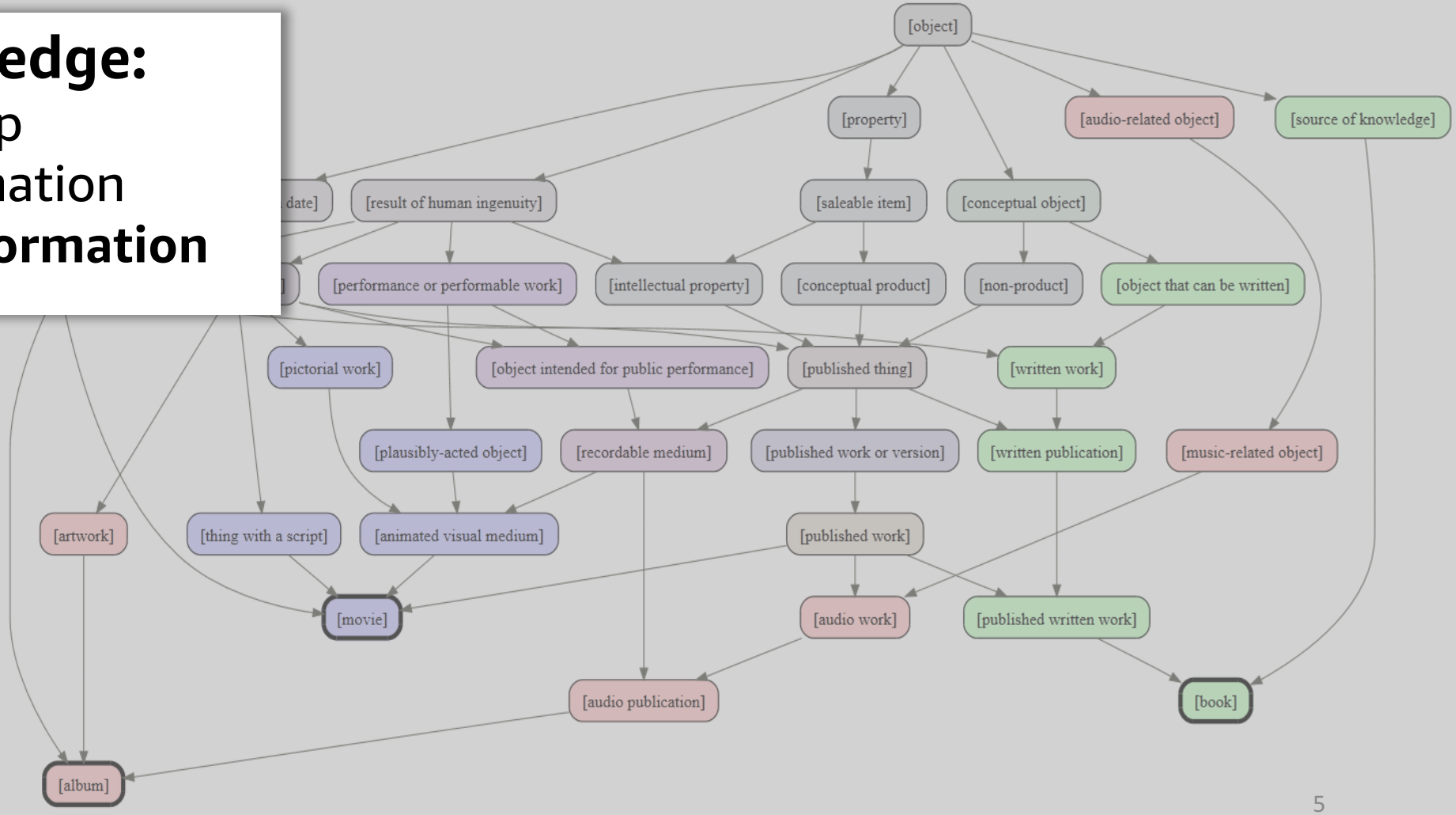
Alexa Knowledge Base



Alexa Knowledge Base

Sources of knowledge:

1. Human authorship
2. Structured information
3. **Unstructured information**



Knowledge from Unstructured Text

The Goal:

Carrie Fisher wrote several semi-autobiographical novels, including Postcards from the Edge.

Knowledge from Unstructured Text

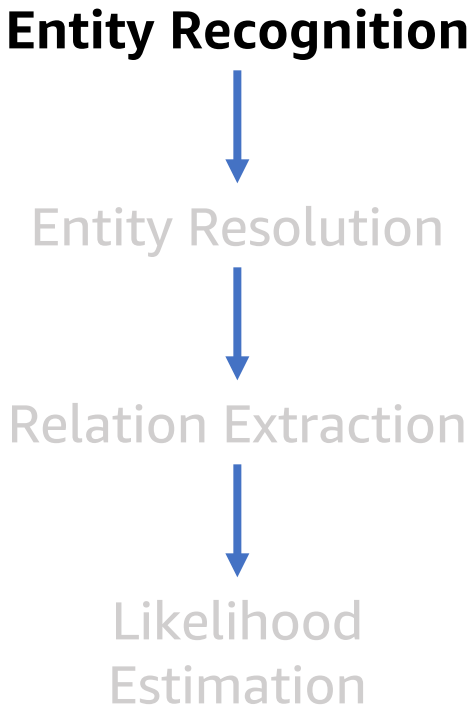
The Goal:

Carrie Fisher wrote several semi-autobiographical novels, including **Postcards from the Edge**.

Knowledge from Unstructured Text

The Goal:

Carrie Fisher wrote several semi-autobiographical novels, including **Postcards from the Edge**.

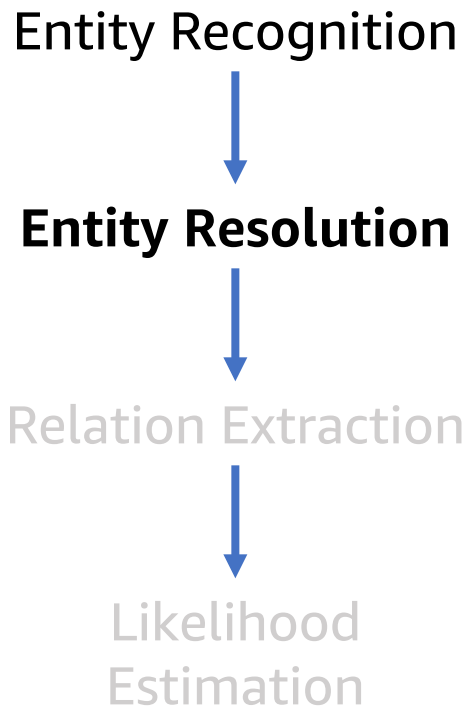


Knowledge from Unstructured Text

The Goal:

Carrie Fisher wrote several semi-autobiographical novels, including **Postcards from the Edge**.

[carrie fisher] [postcards from the edge]

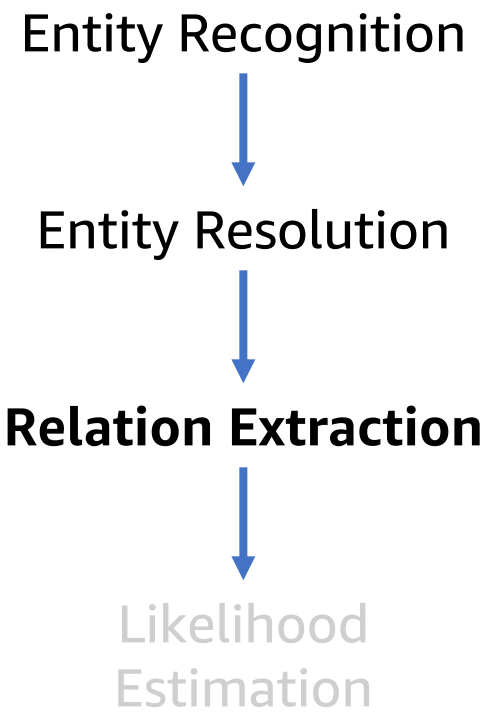


Knowledge from Unstructured Text

The Goal:

Carrie Fisher **wrote** several semi-autobiographical novels, including **Postcards from the Edge**.

[carrie fisher] [is the author of] [postcards from the edge]



Knowledge from Unstructured Text

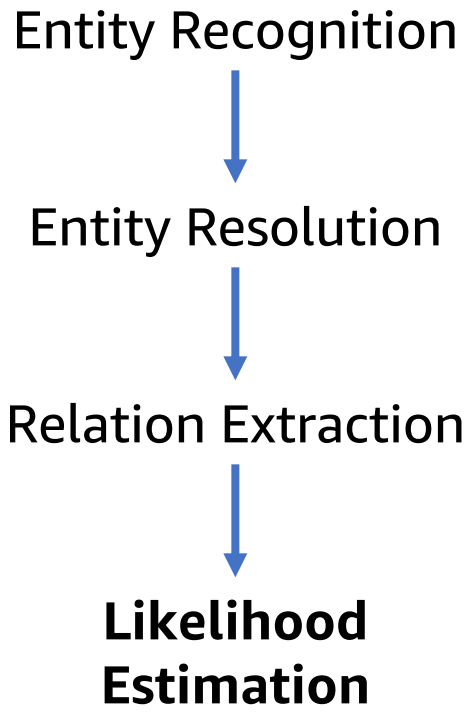
The Goal:

Carrie Fisher **wrote** several semi-autobiographical novels, including **Postcards from the Edge**.

[carrie fisher] [is the author of] [postcards from the edge]

Ontological constraints
Entity embeddings
Distributional information

98% likelihood



Learning approaches for RE

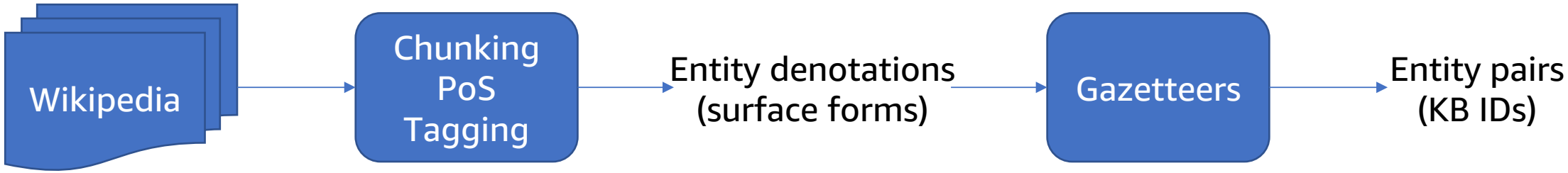
- Rule-based
- Fully supervised
- Unsupervised
- **Distant/weakly supervised**
  - Snow, Jurafsky, Ng, 2005
  - Main assumption: if two entities are linked by a relation, any sentence containing both sentences is *likely* to express that relation
    - [steven spielberg] [is the director of] [saving private ryan]
    - "Spielberg's film Saving Private Ryan is based on…"

Learning approaches for RE

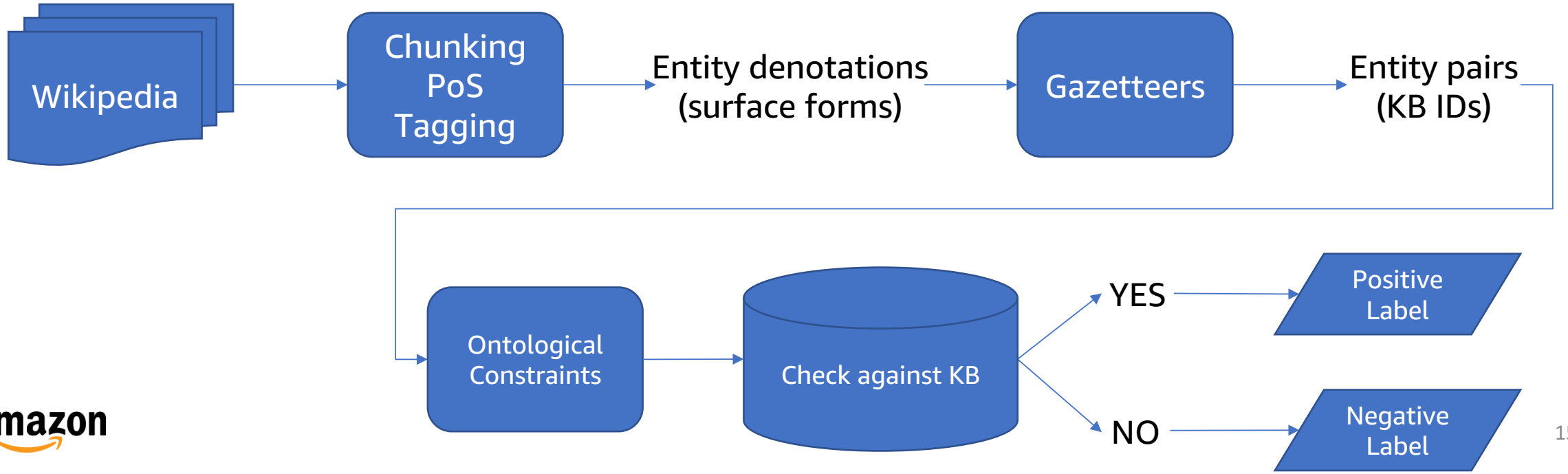
- Rule-based
- Fully supervised
- Unsupervised
- **Distant/weakly supervised**
  - Snow, Jurafsky, Ng, 2005
  - Main assumption: if two entities are linked by a relation, any sentence containing both sentences is *likely* to express that relation
    - [steven spielberg] [is the director of] [saving private ryan]
    - "Spielberg's film Saving Private Ryan is based on…"

Christodoulopoulos and Mittal (*under review*)

Distant supervision label generation



Distant supervision label generation

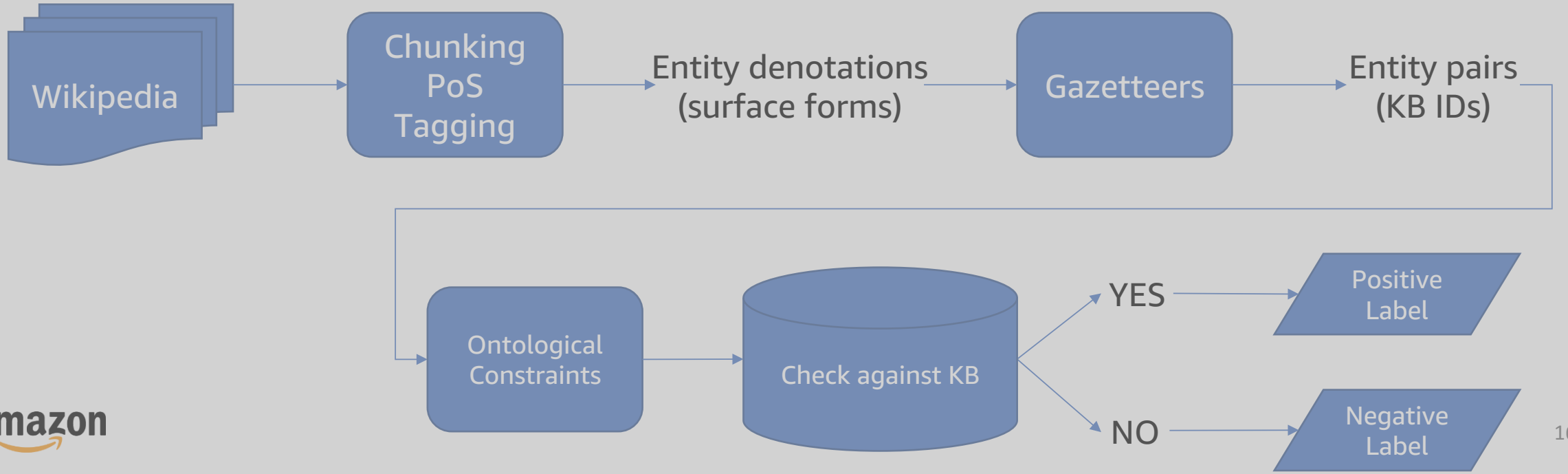


Distant supervision label generation

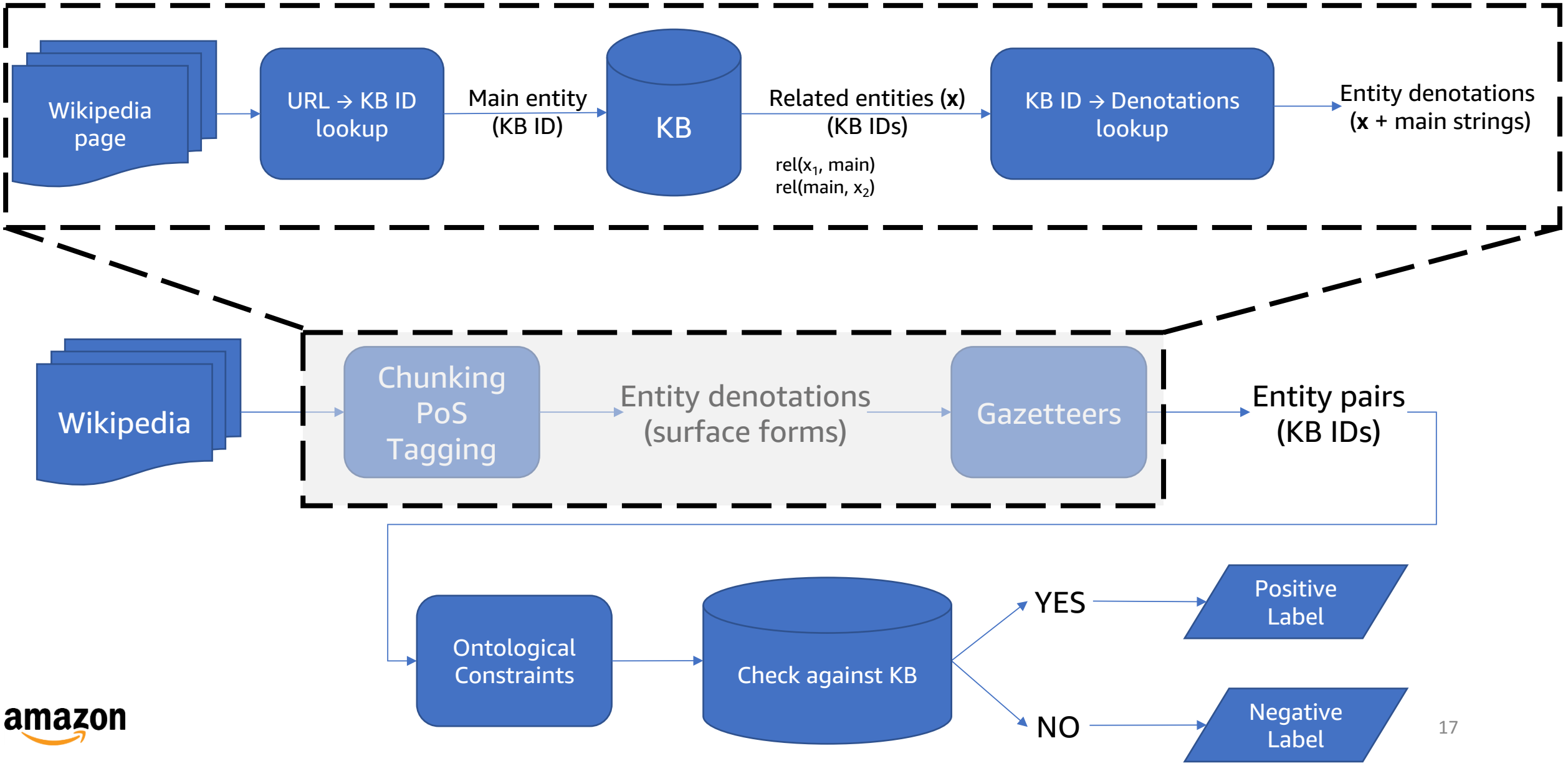
His studies were interrupted by army service and at the *end* of the *war* he was forced to return. . .
`[the second world war] [is an instance of] [cause of death]`

In the *intro* to the *song*, Fred Durst makes reference to. . .
`[intro 15367][is an instance of] [song]`

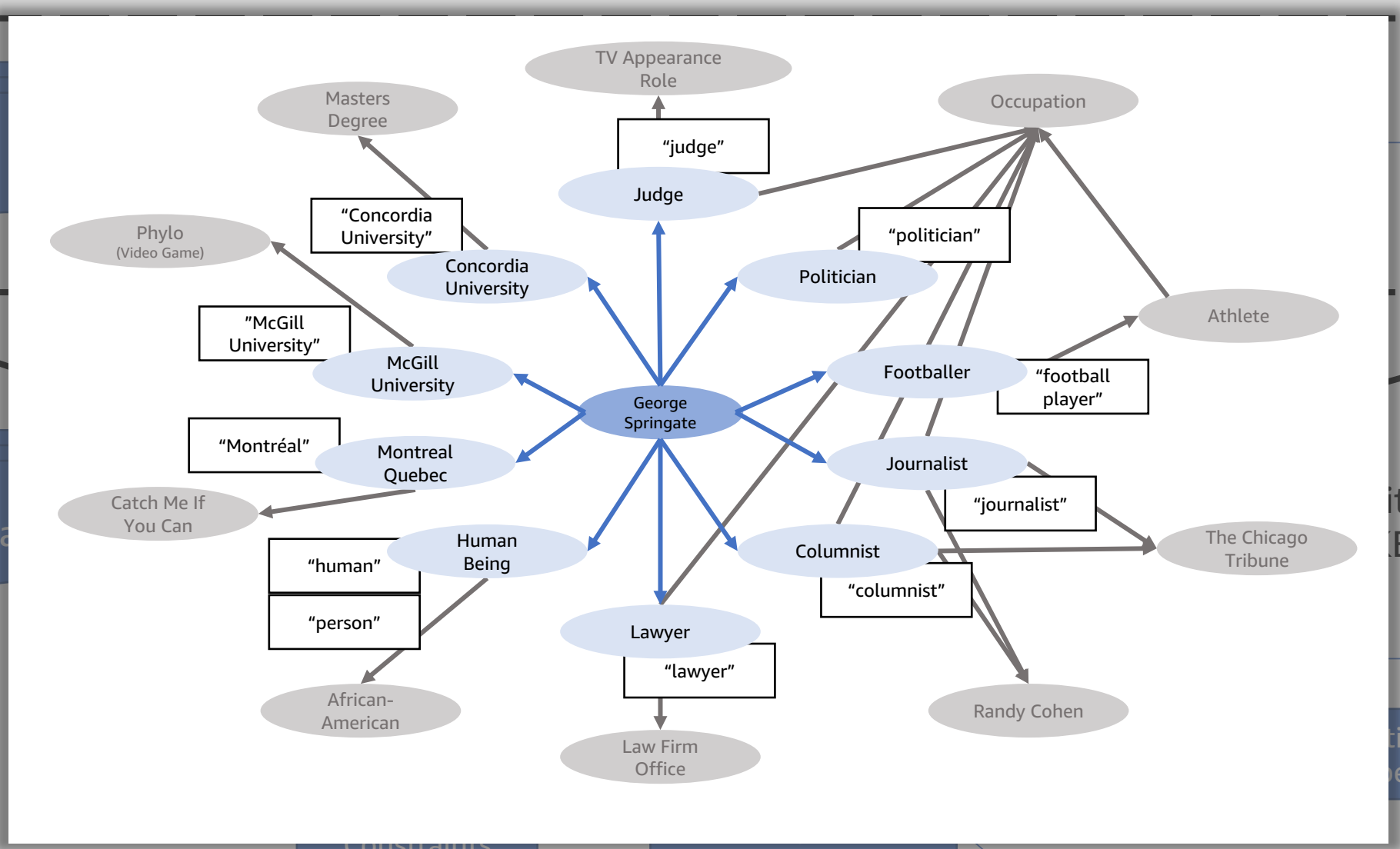
Turner also released one *album* and several *singles* under the moniker Repeat.
`[the singles the 2011 album] [is an instance of] [album]`



Distant supervision label generation



Distant supervision label generation



Wikipedia page

Wikipedia

Entity denotations (x + main strings)

ity pairs
B IDs)

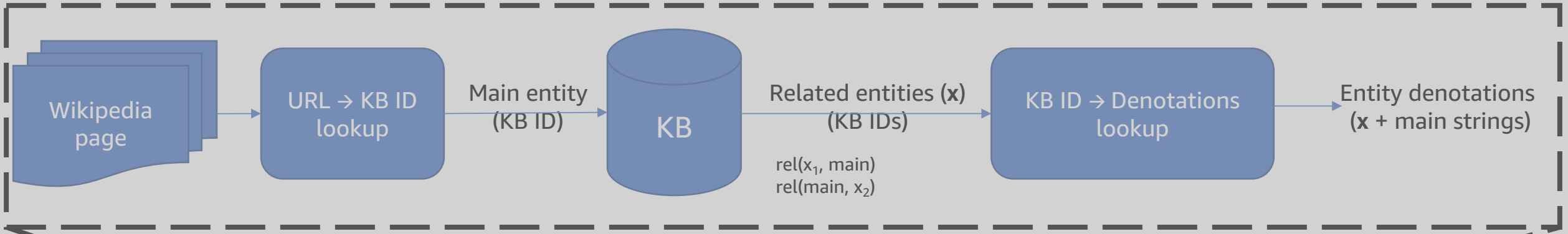
Constraints

(Bloom filters)

NO

Negative Label

Distant supervision label generation



Call Your Girlfriend was written by Robyn, Alexander Kronlund and Klas Åhlund, with the latter producing the *song*.

`[call your girlfriend 3] [is an instance of] [song]`

Forget Her is a *song* by Jeff Buckley.

`[forget her] [is an instance of] [song]`

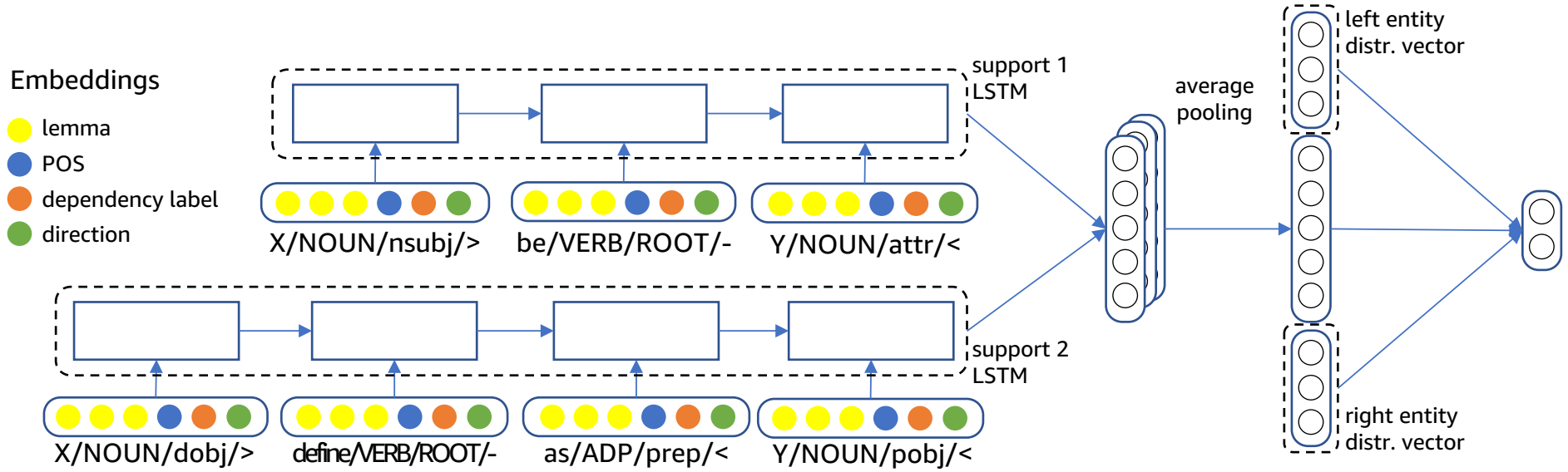
The *Subei Mongol Autonomous County* is an autonomous *county* within the prefecture-level city of Jiuquan in the northwestern Chinese province of Gansu.

`[subei mongol autonomous county] [is an instance of] [chinese county]`

amazon

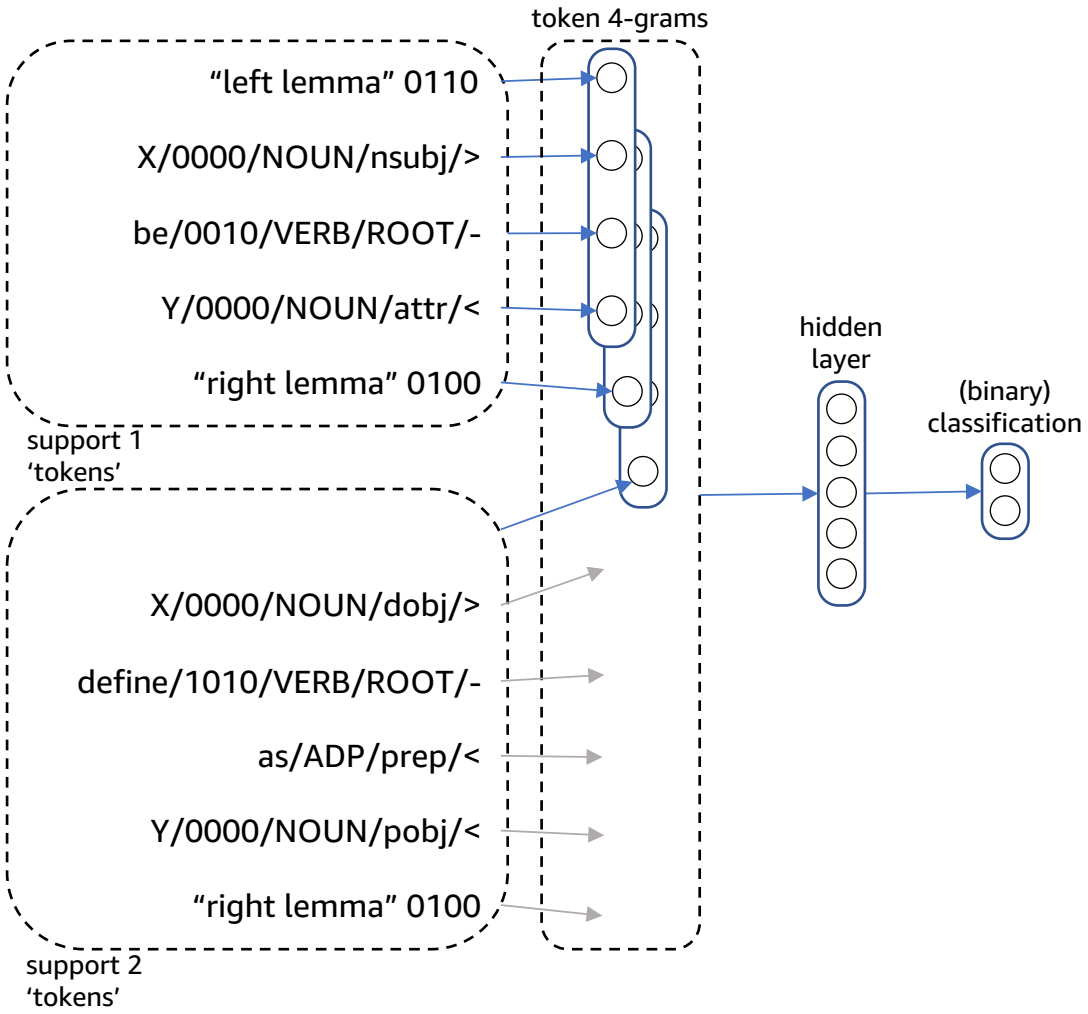
Relation extraction

- HypeNET (Shwartz and Goldberg, 2016)
- Hyponyms [is an instance of] only
  - LexNET extends to multiple relations



Relation extraction

- fastText (Joulin et al., 2016)
- Linear model
  - One hidden layer
  - Rank constraint



Results

HypeNET **equally good** as the much simpler fastText with the **same input features**.

Alexa KB

Relation	HypeNET	fastText
[is an instance of]	94.29 (0.21)	94.31 (0.03)
[is the birthplace of]	85.57 (0.26)	**87.63** (0.01)
[applies to]	81.98 (1.78)	**86.17** (0.01)

amazon

Results

HypeNET **equally good** as the much simpler fastText with the **same input features**.

Alexa KB

Relation	HypeNET	fastText
[is an instance of]	94.29 (0.21)	94.31 (0.03)
[is the birthplace of]	85.57 (0.26)	**87.63** (0.01)
[applies to]	81.98 (1.78)	**86.17** (0.01)

Wikidata

Relation	HypeNET	fastText
instance of (P31)	93.90 (0.21)	**96.44** (0.01)
birthplace of (P19)	92.06 (0.90)	**93.05** (0.07)
part of (P527)	48.73 (2.59)	**72.87** (0.16)

amazon

Results

HypeNET **equally good** as the much simpler fastText with the **same input features**.

MaxEnt results show that **features alone are not enough.**
Need to create higher-dimensional representations of discrete features.

Alexa KB

Relation	HypeNET	fastText	MaxEnt
[is an instance of]	94.29 (0.21)	94.31 (0.03)	83.93
[is the birthplace of]	85.57 (0.26)	**87.63** (0.01)	80.83
[applies to]	81.98 (1.78)	**86.17** (0.01)	65.27

Wikidata

Relation	HypeNET	fastText	MaxEnt
instance of (P31)	93.90 (0.21)	**96.44** (0.01)	58.45
birthplace of (P19)	92.06 (0.90)	**93.05** (0.07)	66.72
part of (P527)	48.73 (2.59)	**72.87** (0.16)	45.13

amazon

Summary

- New method for entity resolution
  - Page-specific gazetteers
- Features are important
  - HypeNET vs fastText
- Feature representation is important
  - fastText vs MaxEnt

Future directions

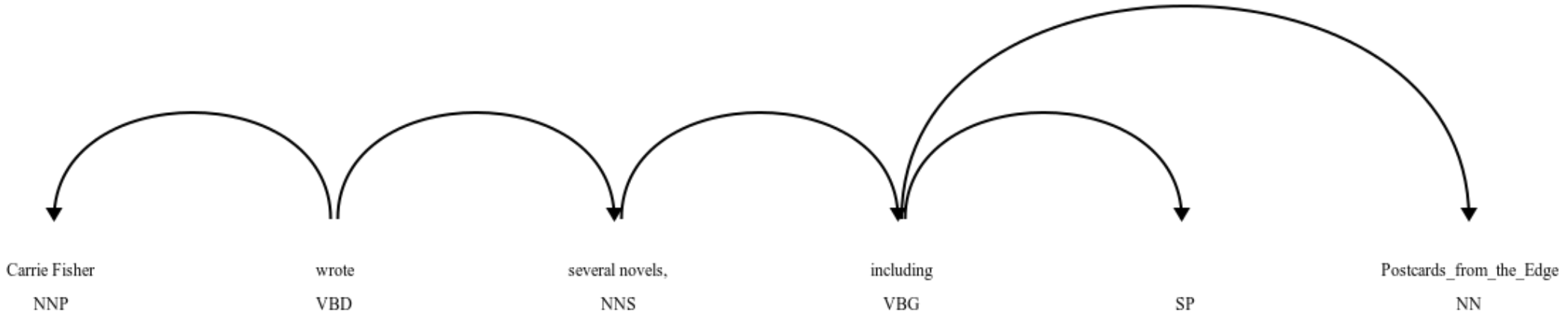
- Enhanced entity recognition
- Use of human annotation for seeding supervision
- Expanding to multiple sources of text
- Coverage of multiple languages

Thanks!

amazon

Dependency parsing for RE

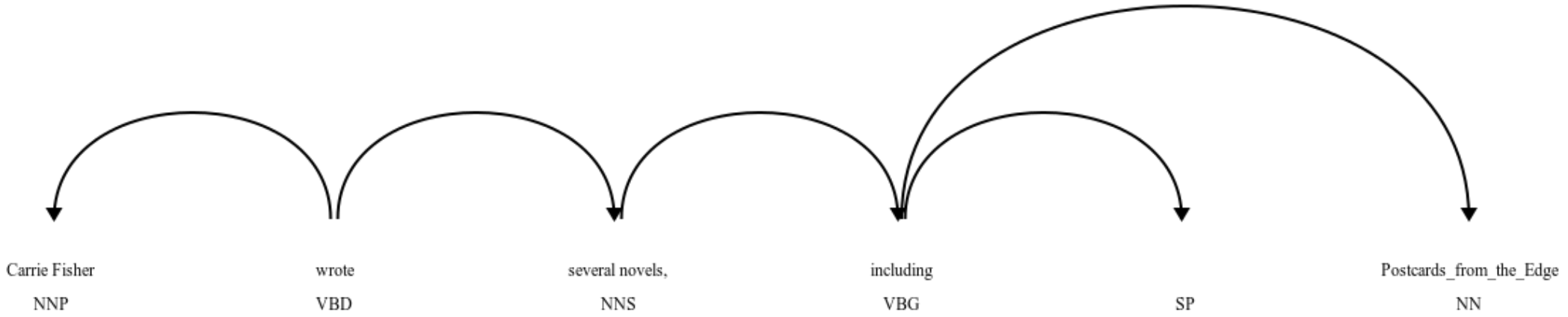
Carrie Fisher wrote several semi-autobiographical novels, including **Postcards from the Edge**.



(shortest) path between entities: **X** → wrote → several → including → **Y**

Dependency parsing for RE

Carrie Fisher wrote several semi-autobiographical novels, including **Postcards from the Edge**.



(shortest) path between entities: **X** → wrote → several → including → **Y**

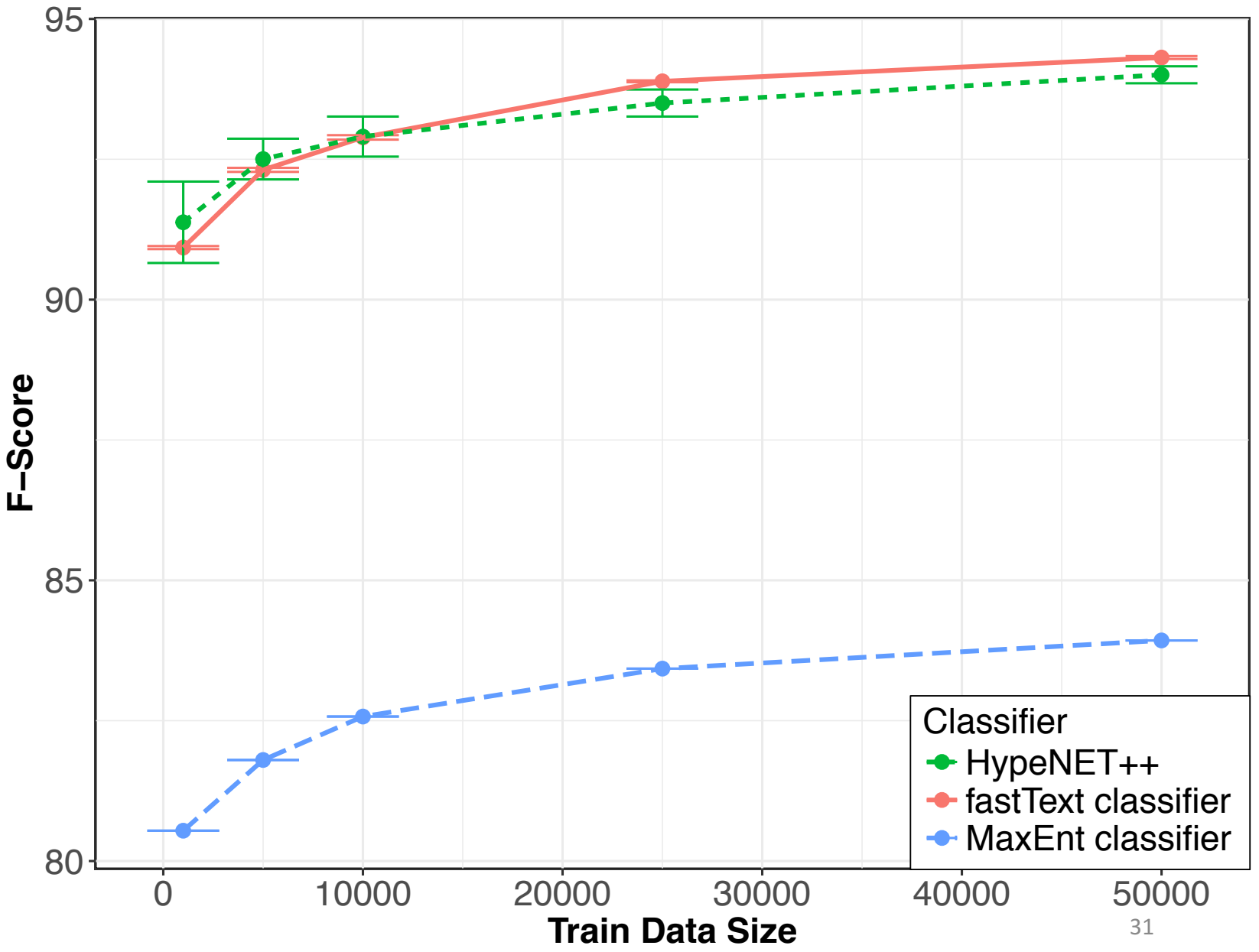
Carrie Fisher, who was friends with Steven Spielberg, wrote several semi-autobiographical novels, including **Postcards from the Edge**.

Results – feature ablation

	instance of	applies to	birthplace of
(1) 5 supports	94.55	86.26	87.56
all supports	94.33	85.92	87.63
(1)-Brown	94.20	85.93	87.51
(1)-lemma	94.17	84.15	86.65
(1)-POS	94.15	85.93	87.71
(1)-dep	93.59	85.42	86.53
(1)-*X*/*Y* entities	93.63	83.89	86.95
X/*Y* only	91.15	74.20	81.15
full sentence	86.70	77.77	87.09

Results

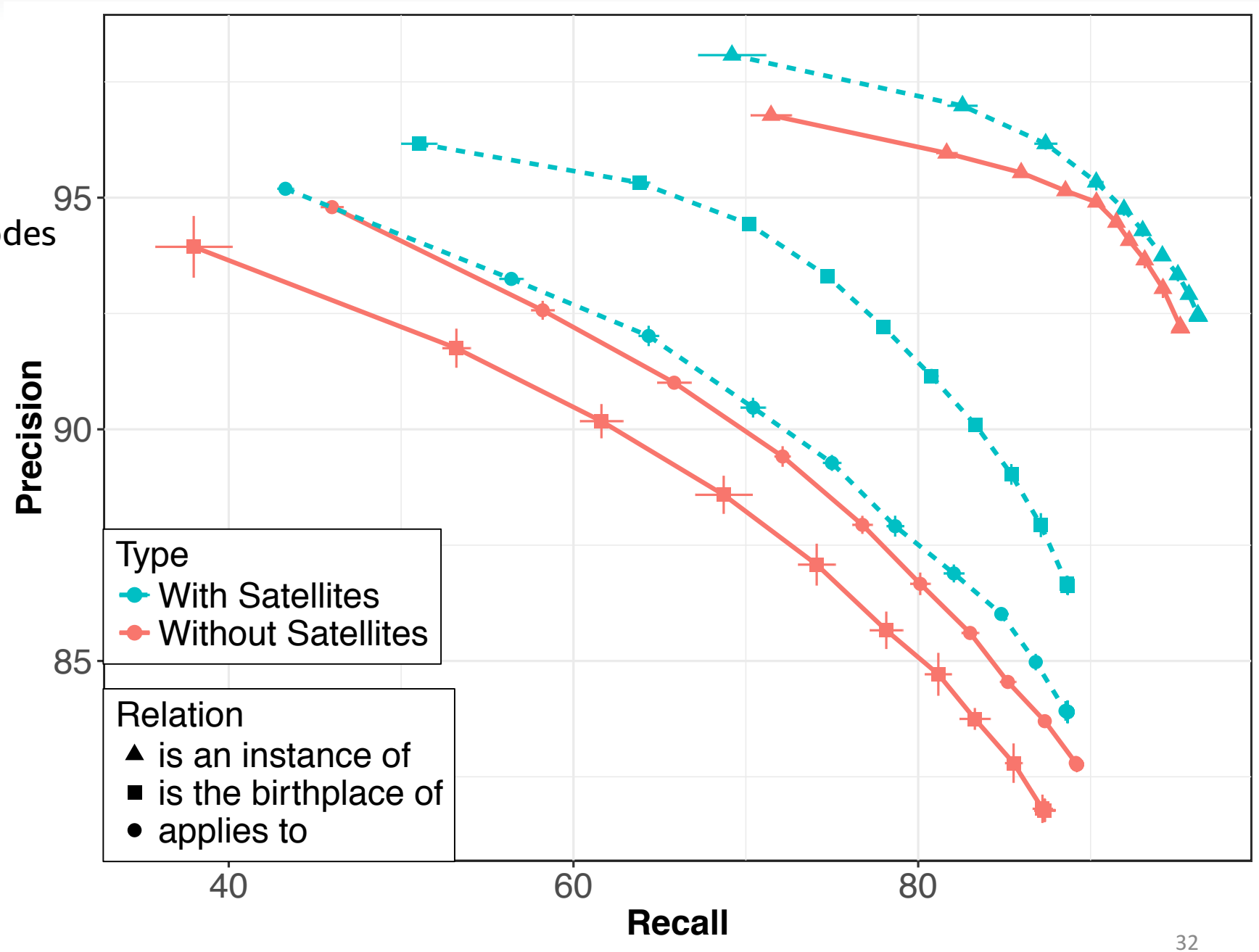
Training data size



amazon

Results

Using dependency satellite nodes



Results

Grouping supports for each entity pair