Discrete-State Variational Autoencoders for Joint Discovery and Factorization of Relations (TACL Paper)

Diego Marcheggiani and Ivan Titov

University of Amsterdam

Universiteit van Amsterdam

Relation Extraction

Given two entities, predict the semantic relation that holds between them

Chomsky embarked on a program of study at UPenn

Relation Extraction

Given two entities, predict the semantic relation that holds between them

$$\underbrace{\textit{Chomsky}}_{e_1}\ \textit{embarked on a program of study at}\ \underbrace{\textit{UPenn}}_{e_2}$$

Relation Extraction

Given two entities, predict the semantic relation that holds between them

$$\underbrace{\textit{Chomsky}}_{e_1}\ \underbrace{\textit{embarked on a program of study at}}_{studied\_at}\ \underbrace{\textit{UPenn}}_{e_2}$$

Motivation

- Much of previous work has focused on (distantly-) supervised methods:
  - supervision is not available for many domains
  - knowledge bases are often incomplete

Riedel et al. (2010); Surdeanu et al. (2012)

In this work we do unsupervised relation extraction

Motivation

- Existing work on unsupervised modeling used restricted features and restrictive modeling assumptions.

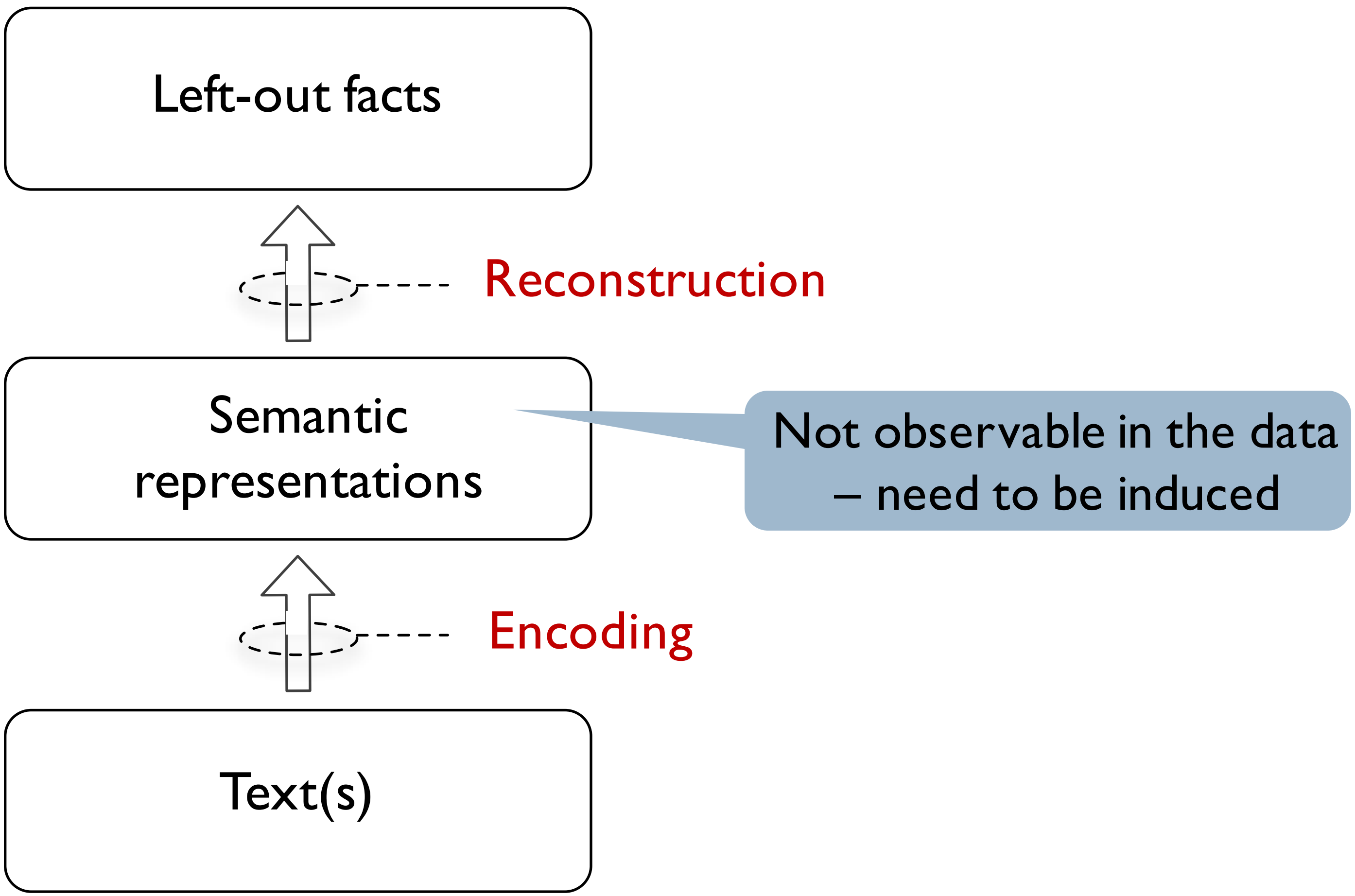
Lin and Pantel (2001);
Yao et al. (2011);
Yao et al. (2012)

We define an unsupervised feature-rich model

Outline

- Framework: reconstruction error minimization
- Instantiation: our model for relation discovery
- Empirical evaluation: experiments on NYT corpus

General framework



Instead of using annotated data, induce representations beneficial for inferring left-out facts

Unsupervised setting

$\underbrace{\textbf{\textit{Chomsky}}}_{e_1}$ *embarked on a program of study at* $\underbrace{\textbf{\textit{UPenn}}}_{e_2}$

$\underbrace{\textbf{\textit{Barak Obama}}}_{e_1}$ *studied at* $\underbrace{\textbf{\textit{Harvard}}}_{e_2}$

$\underbrace{\textbf{\textit{Iggy Pop}}}_{e_1}$ *has lived in* $\underbrace{\textbf{\textit{Berlin}}}_{e_2}$ *during the 70's*

Unsupervised setting

$\underbrace{\textbf{\textit{Chomsky}}}_{e_1}\ \underbrace{\textit{embarked on a program of study at}}_{studied\_at}\ \underbrace{\textbf{\textit{UPenn}}}_{e_2}$

$\underbrace{\textbf{\textit{Barak Obama}}}_{e_1}\ \underbrace{\textit{studied at}}_{studied\_at}\ \underbrace{\textbf{\textit{Harvard}}}_{e_2}$

$\underbrace{\textbf{\textit{Iggy Pop}}}_{e_1}\ \underbrace{\textit{has lived in}}_{has\_lived}\ \underbrace{\textbf{\textit{Berlin}}}_{e_2}$ *during the 70's*

Unsupervised setting

$\underbrace{\textbf{\textit{Chomsky}}}_{e_1}\ \underbrace{\textit{embarked on a program of study at}}_{01}\ \underbrace{\textbf{\textit{UPenn}}}_{e_2}$

$\underbrace{\textbf{\textit{Barak Obama}}}_{e_1}\ \underbrace{\textit{studied at}}_{01}\ \underbrace{\textbf{\textit{Harvard}}}_{e_2}$

$\underbrace{\textbf{\textit{Iggy Pop}}}_{e_1}\ \underbrace{\textit{has lived in}}_{07}\ \underbrace{\textbf{\textit{Berlin}}}_{e_2}\ \textit{during the 70's}$

Arguments reconstruction

Chomsky *embarked on a program of study at* ***UPenn***

Not observable in the data

Studied_at (e1: Chomsky, e2:UPenn)

Arguments reconstruction

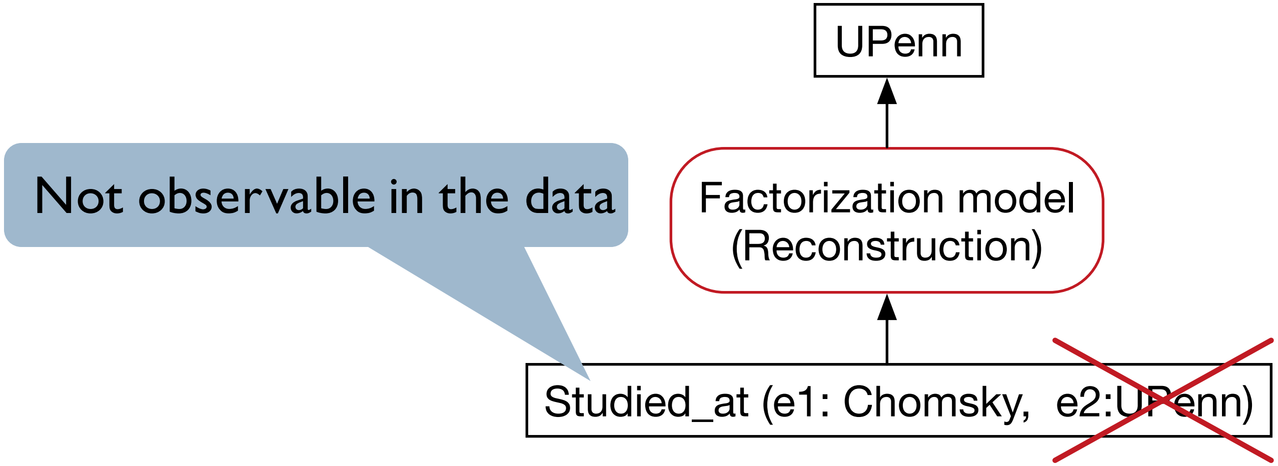
Chomsky** embarked on a program of study at **UPenn

Not observable in the data

Studied_at (e1: Chomsky, ~~e2:UPenn~~)

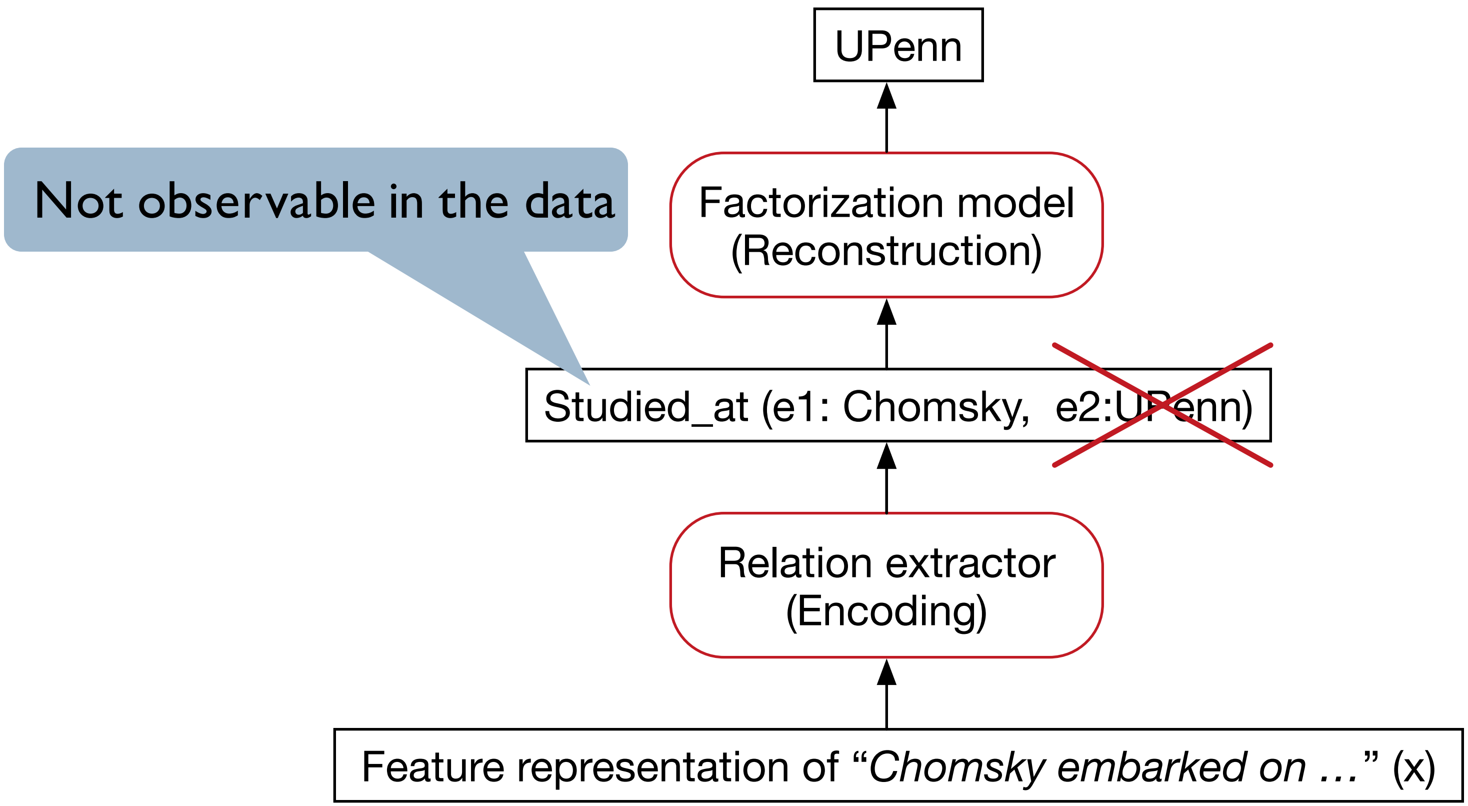
Arguments reconstruction

Chomsky** embarked on a program of study at **UPenn



Relation induction

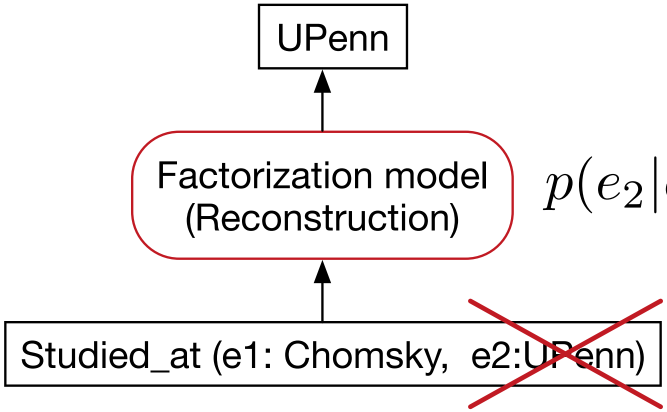
Chomsky *embarked on a program of study at* ***UPenn***



Outline

- Framework: reconstruction error minimization
- Instantiation: our model for relation discovery
- Empirical evaluation: experiments on NYT corpus

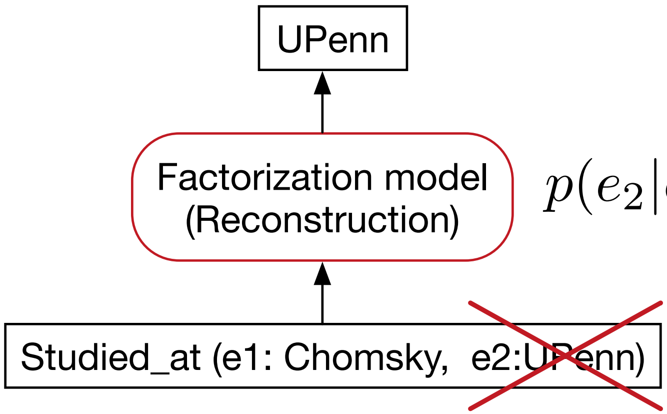
Reconstruction component



$p(e_2|e_1, r, \theta)$

$\mathbf{u}_{e_1}, \mathbf{u}_{e_2} \in \mathbb{R}^d$ - encode semantic properties of entities e_1 *and* e_2

Reconstruction component



$p(e_2|e_1, r, \theta)$

$\mathbf{u}_{e_1}, \mathbf{u}_{e_2} \in \mathbb{R}^d$ - encode semantic properties of entities e_1 *and* e_2

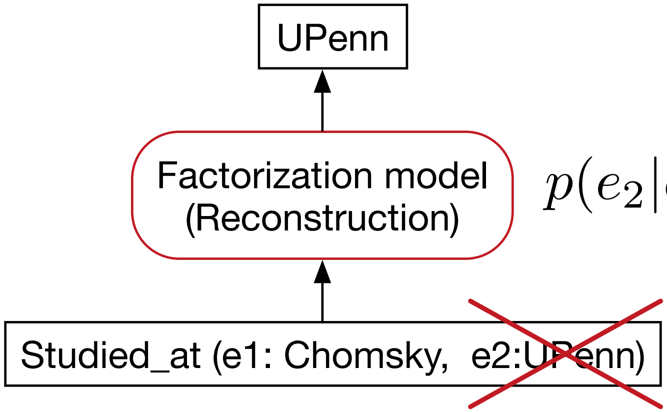
RESCAL factorization

Nickel et al. (2011)

$$\psi^{RS}(e_1, e_2, r, \theta) = \mathbf{u}_{e_1}^T C_r \mathbf{u}_{e_2}$$

encodes interdependencies between entities

Reconstruction component



$$p(e_2|e_1, r, \theta)$$

$\mathbf{u}_{e_1}, \mathbf{u}_{e_2} \in \mathbb{R}^d$ - encode semantic properties of entities e_1 *and* e_2

RESCAL factorization

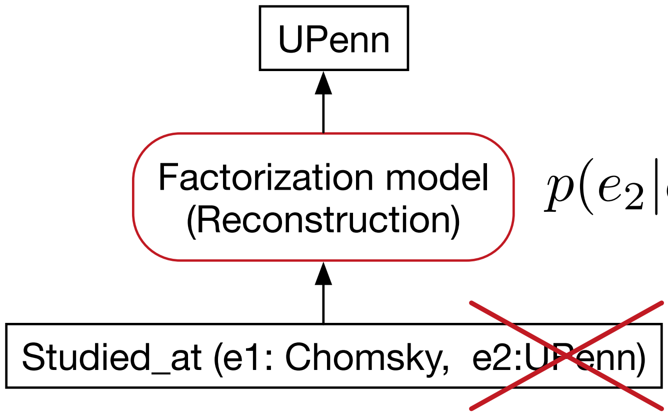
Nickel et al. (2011)

$$\psi^{RS}(e_1, e_2, r, \theta) = \mathbf{u}_{e_1}^T C_r \mathbf{u}_{e_2}$$

encodes interdependencies between entities

The reconstruction model: $$p(e_2|e_1, r, \theta) = \frac{\exp(\psi(e_1, e_2, r, \theta))}{\sum_{e' \in \mathcal{E}} \exp(\psi(e_1, e', r, \theta))}$$

Reconstruction component



$$p(e_2|e_1, r, \theta)$$

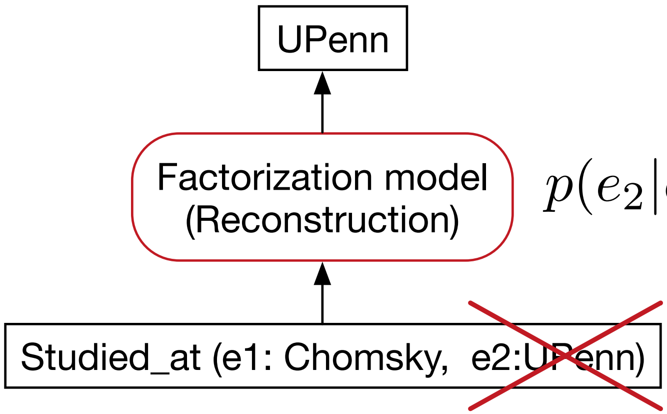
scores each entity independently

Selectional preferences

Séaghdha (2010)

$$\psi^{SP}(e_1, e_2, r, \theta) = \sum_{i=1}^{2} \mathbf{u}_{e_i}^T \mathbf{c}_{ir}$$

Reconstruction component



$$p(e_2|e_1, r, \theta)$$

scores each entity independently

Selectional preferences

Séaghdha (2010)

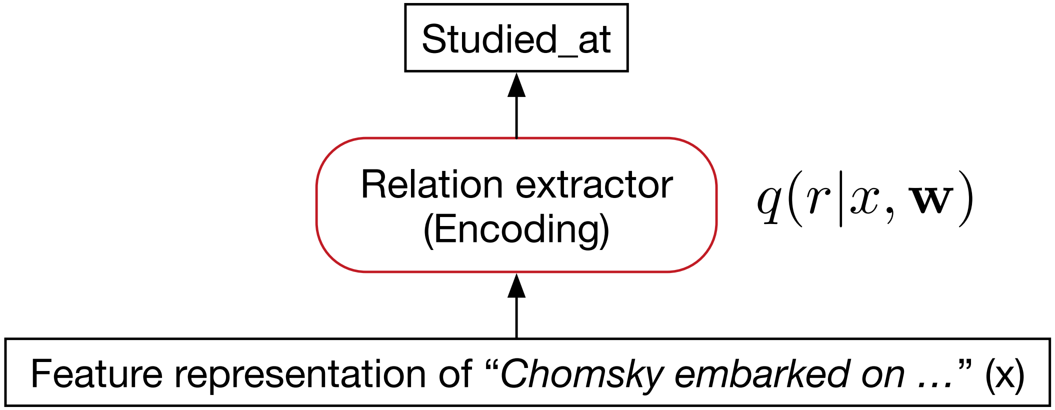
$$\psi^{SP}(e_1, e_2, r, \theta) = \sum_{i=1}^{2} \mathbf{u}_{e_i}^T \mathbf{c}_{ir}$$

Hybrid

$$\psi^{HY}(e_1, e_2, r, \theta) = \overbrace{\mathbf{u}_{e_1}^T C_r \mathbf{u}_{e_2}}^{\psi^{RS}} + \overbrace{\sum_{i=1}^{2} \mathbf{u}_{e_i}^T \mathbf{c}_{ir}}^{\psi^{SP}}$$

combines RESCAL model
and selectional preferences

Encoding component

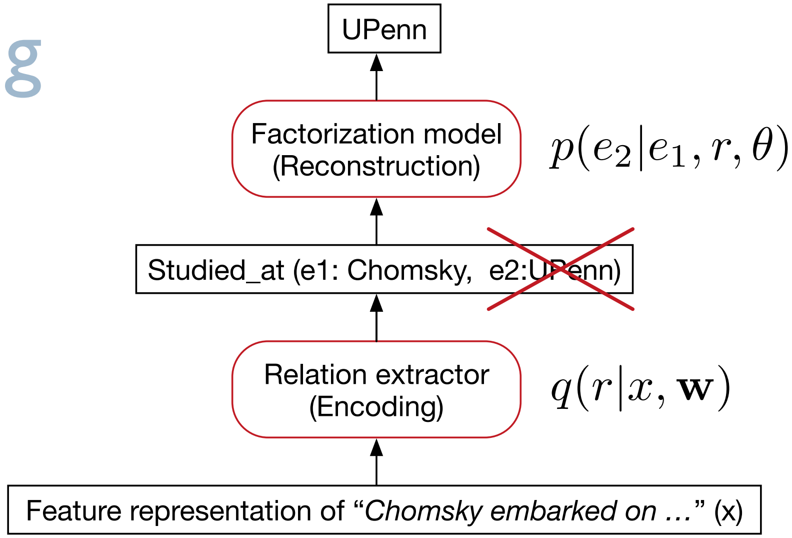


A feature-rich representation

The relation extraction model:

$$q(r|x, \mathbf{w}) = \frac{\exp(\mathbf{w}^T \mathbf{g}(r, x)))}{\sum_{r' \in \mathcal{R}} \exp(\mathbf{w}^T \mathbf{g}(r', x))}$$

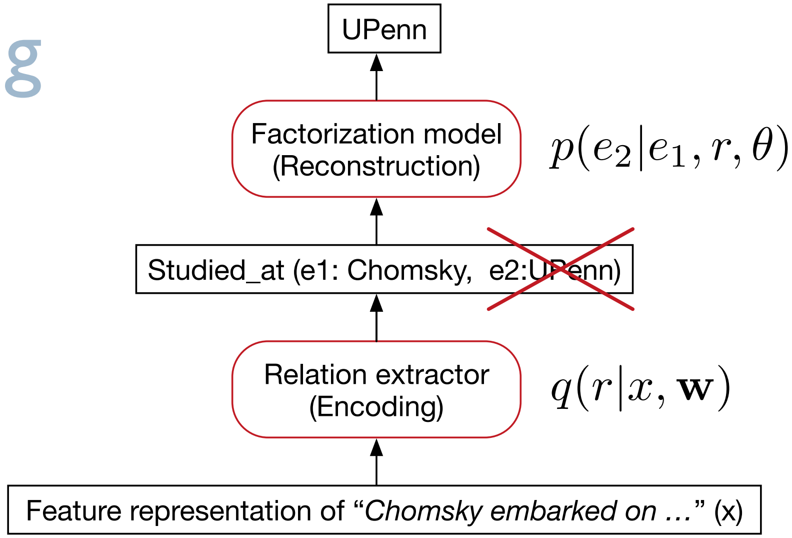
Joint learning



- For each sentence, we optimize the entity prediction quality while marginalizing over relations:

$$\sum_{i=1}^{2} \sum_{r \in \mathcal{R}} q(r|x, \mathbf{w}) \log p(e_i|e_{-i}, r, \theta)$$

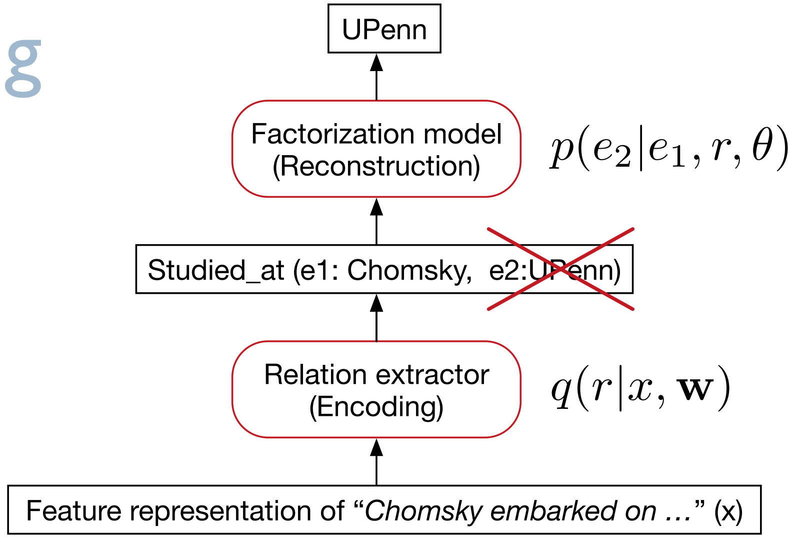
Joint learning



- For each sentence, we optimize the entity prediction quality while marginalizing over relations:

$$\sum_{i=1}^{2} \sum_{r \in \mathcal{R}} q(r|x, \mathbf{w}) \log p(e_i|e_{-i}, r, \theta) \underbrace{- \sum_{r \in \mathcal{R}} q(r|x, \mathbf{w}) \log q(r|x, \mathbf{w})}_{H(q)}$$

Joint learning



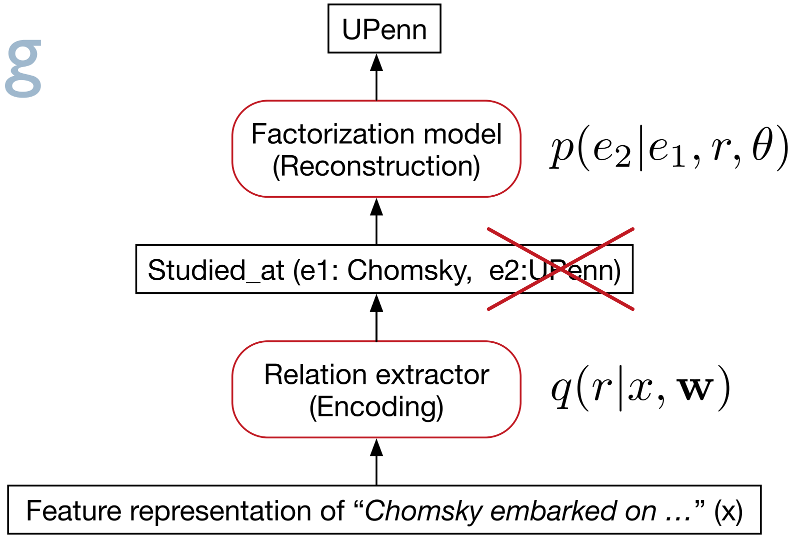
- For each sentence, we optimize the entity prediction quality while marginalizing over relations:

Kingma and Welling (2014)

$$\sum_{i=1}^{2} \underbrace{\sum_{r\in\mathcal{R}} q(r|x,\mathbf{w}) \log p(e_i|e_{-i}, r, \theta)}_{E_q[\log p(e_i|e_{-i}, r, \theta)]} - \underbrace{\sum_{r\in\mathcal{R}} q(r|x,\mathbf{w}) \log q(r|x,\mathbf{w})}_{H(q)}$$

Variational lower bound on the pseudo-likelihood

Joint learning



- For each sentence, we optimize the entity prediction quality while marginalizing over relations:

Kingma and Welling (2014)

$$\sum_{i=1}^{2} \underbrace{\sum_{r \in \mathcal{R}} q(r|x, \mathbf{w}) \log p(e_i|e_{-i}, r, \theta)}_{E_q[\log p(e_i|e_{-i}, r, \theta)]} - \underbrace{\sum_{r \in \mathcal{R}} q(r|x, \mathbf{w}) \log q(r|x, \mathbf{w})}_{H(q)}$$

Variational lower bound on the pseudo-likelihood

- Not very tractable in this exact form:
  - negative sampling (as, e.g., in Mikolov et al '13) instead of 'softmax'

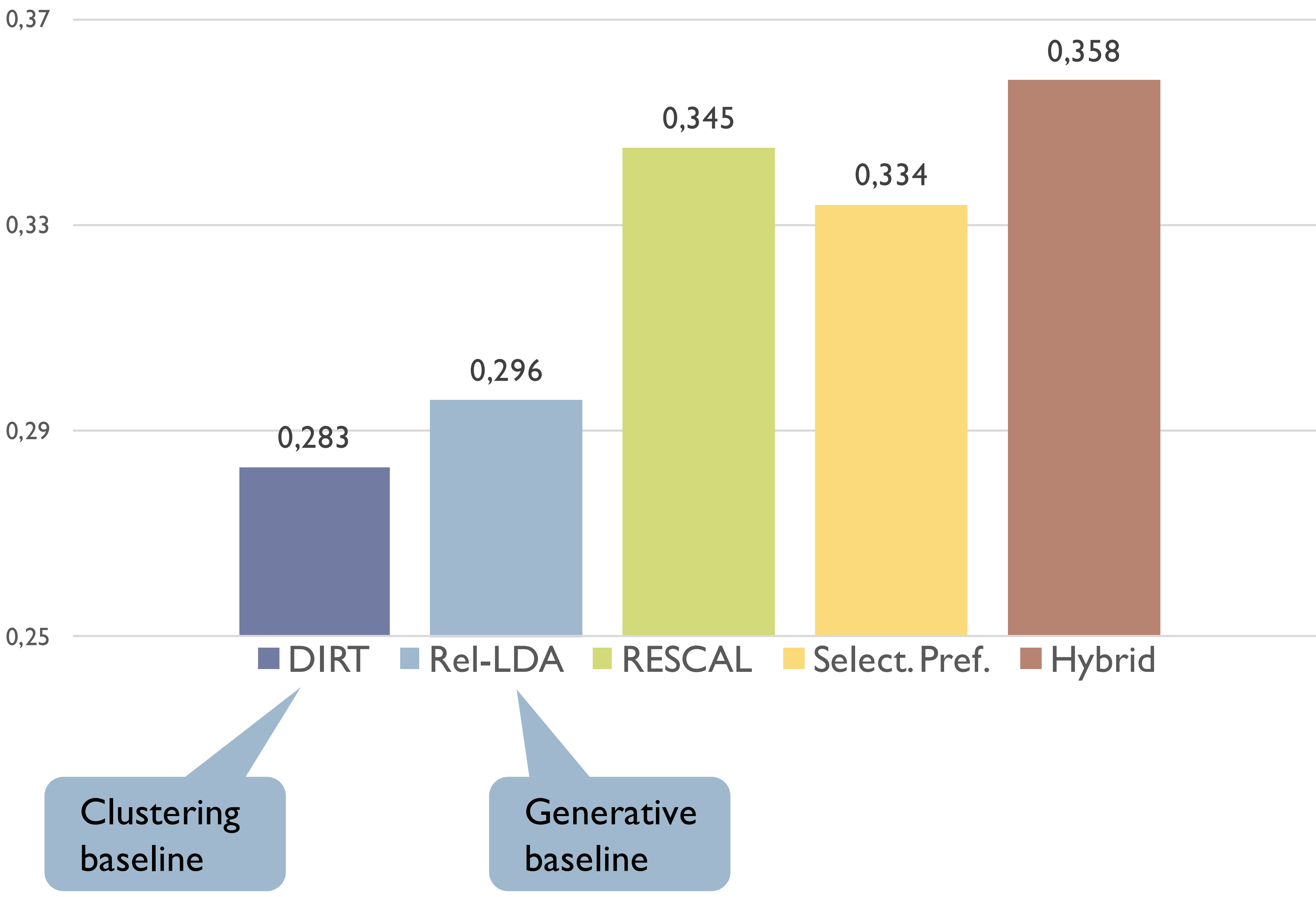
Outline

- Framework: reconstruction error minimization
- Instantiation: our model for relation discovery
- Empirical evaluation: experiments on NYT corpus

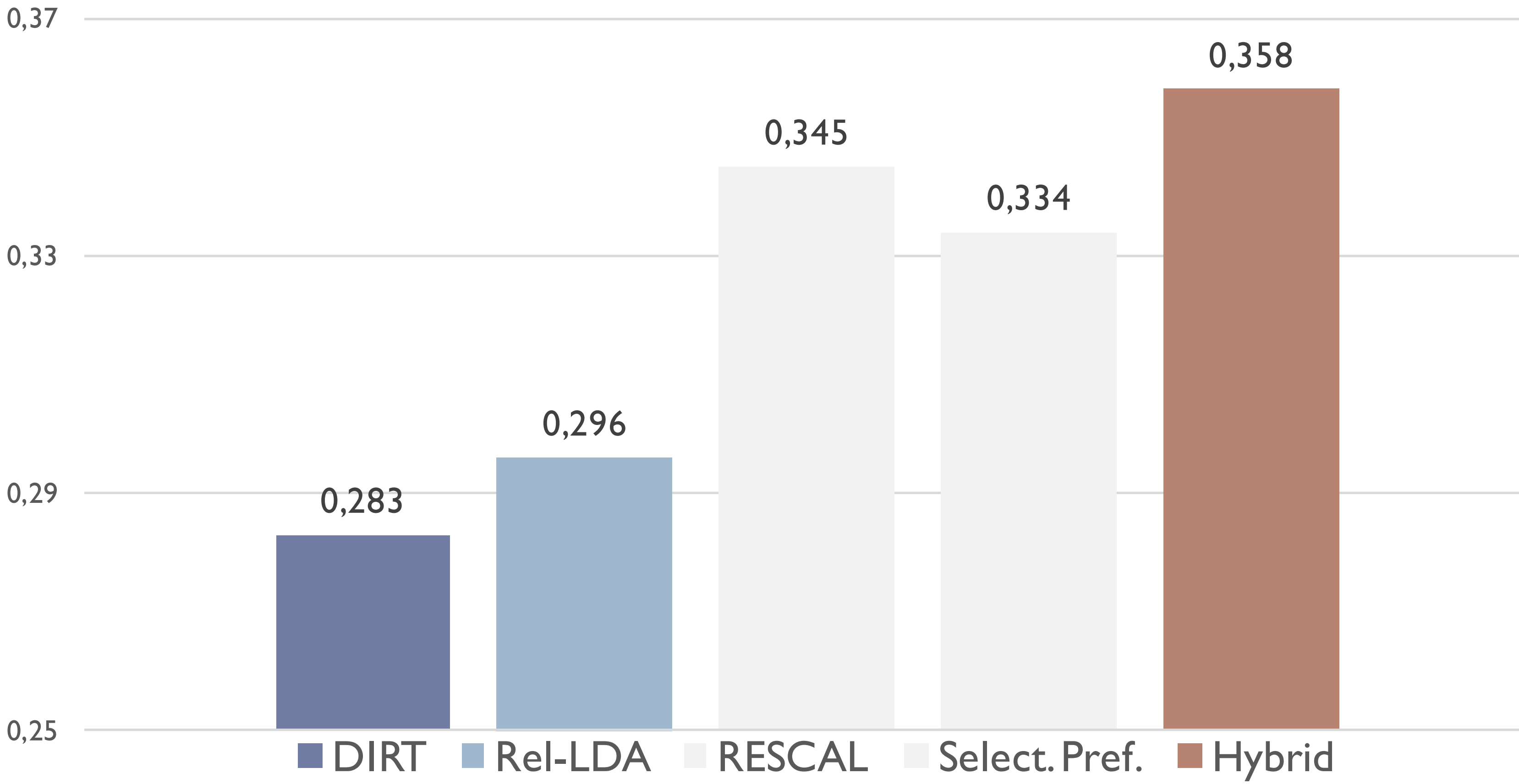
Experimental setup

- Data:
  - New York Times corpus (~2 million examples) aligned with Freebase relations (only for evaluation)
- Baseline:
  - Rel-LDA, state-of-the-art generative model for unsupervised relation discovery (Yao et al. (2011))
  - DIRT, agglomerative clustering baseline (Lin and Pantel (2001))
- Evaluation:
  - F1 of the B-Cube measure

Results (F1)


0,37
0,33
0,29
0,25
0,283
0,296
0,345
0,334
0,358
DIRT
Rel-LDA
RESCAL
Select. Pref.
Hybrid
Clustering baseline
Generative baseline

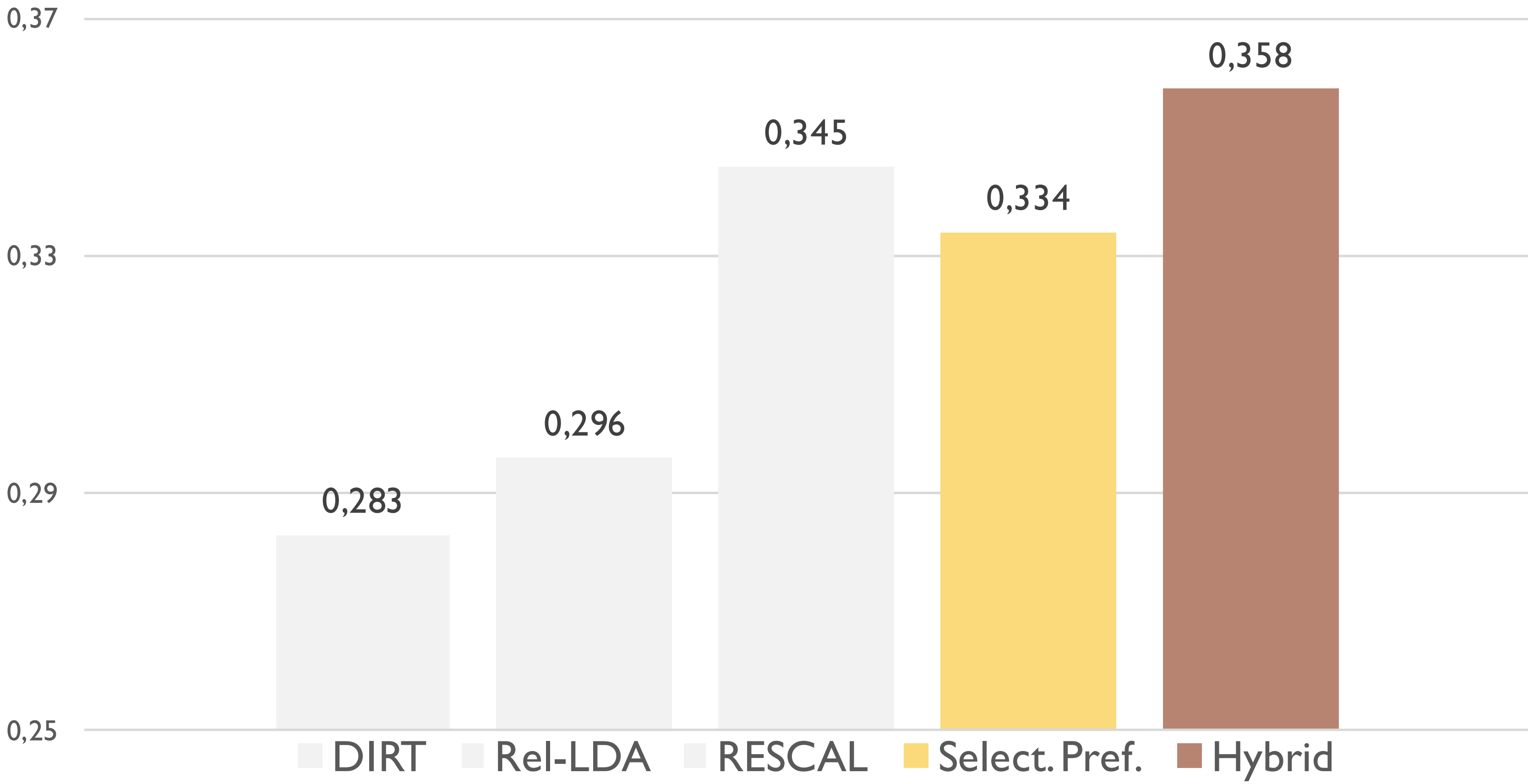
Results (F1)



Generative baseline

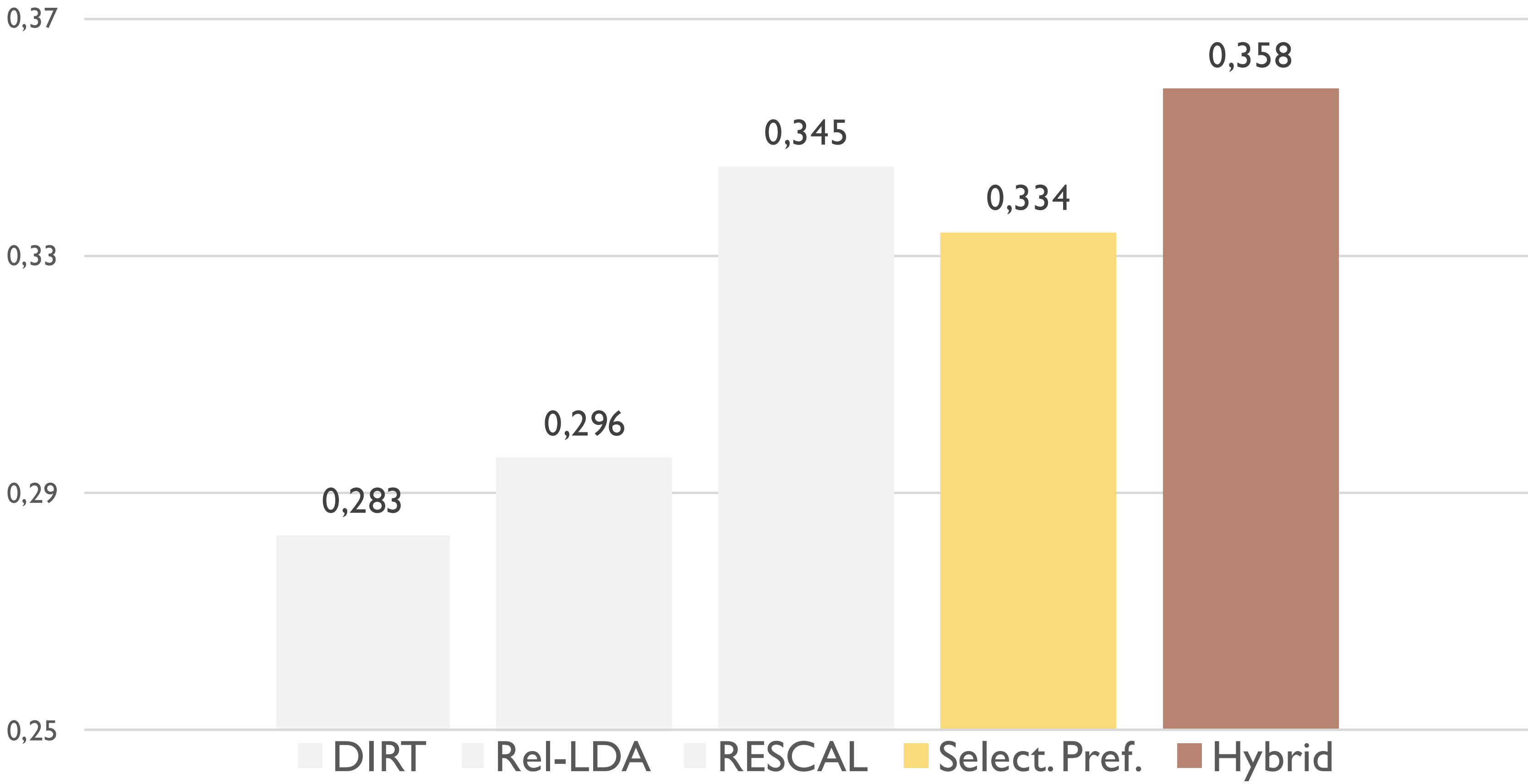
Best model 6.2% more accurate than the Rel-LDA baseline.

Results (F1)



Modelling the interdependence of arguments is beneficial.

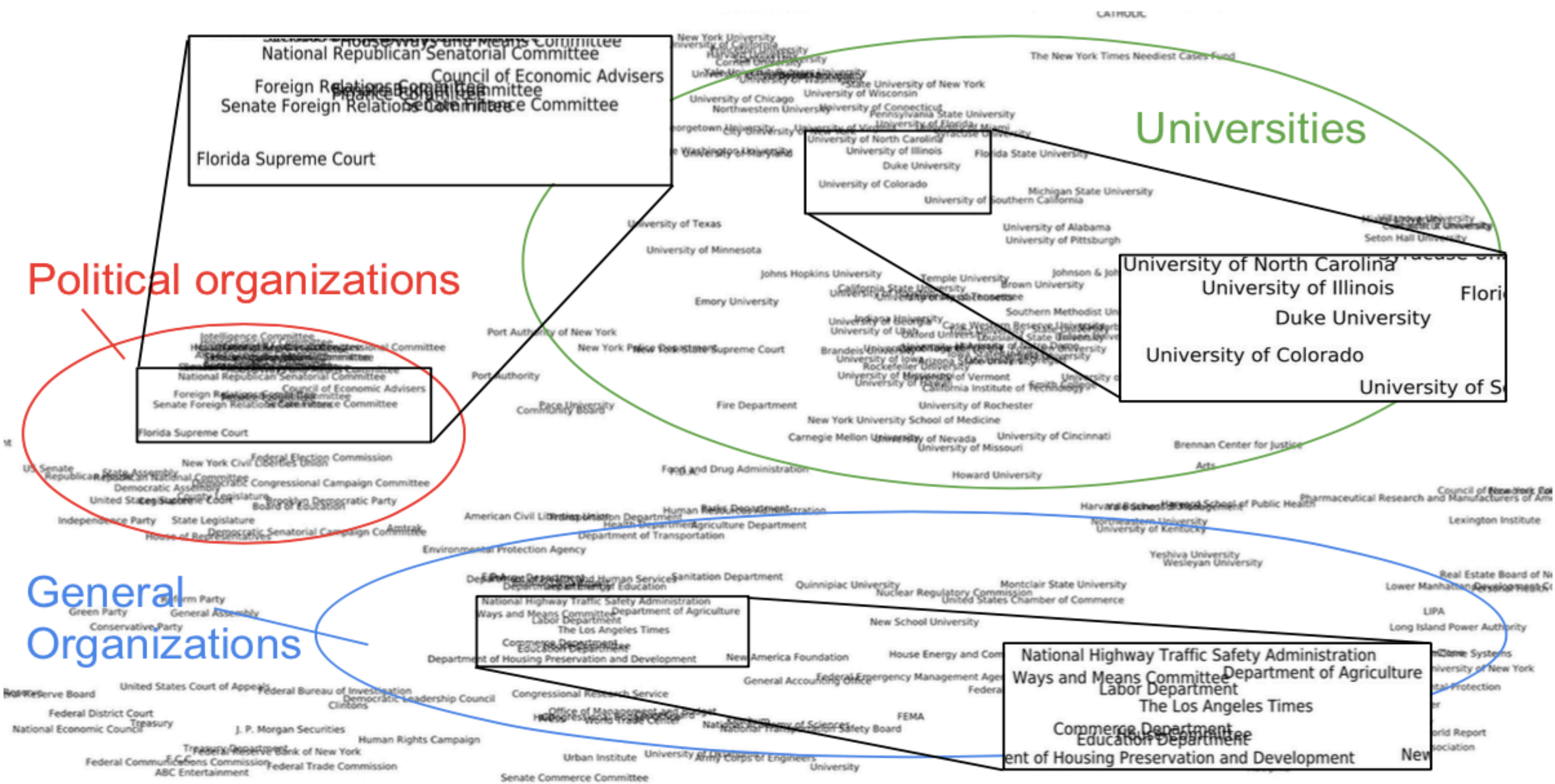
Results (F1)



ndence of

Our model discovers relations not present in Freebase

Qualitative evaluation



Conclusions

- Discrete-state autoencoder for relation extraction
  - Unsupervised
  - Feature-rich
- What's next?
  - Semi-supervised relation extraction with distant supervision
  - Frame-semantic parsing with this framework

Thank you!

Code available at:

github.com/diegma/relation-autoencoder

Funding:

NWO VIDI grant

Google Focused Award on Natural Language Understanding