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

Diego Marcheggiani and Ivan Titov

University of 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

Chomsky embarked on a program of study at UPenn

\( e_1 \quad e_2 \)
Relation Extraction

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

Chomsky embarked on a program of study at UPenn

$e_1$ studied_at $e_2$
Motivation

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

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

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

**Barak Obama** studied at **Harvard**

**Iggy Pop** has lived in **Berlin** during the 70’s
Unsupervised setting

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

\[
e_1 \quad \text{studied_at} \quad e_2
\]

**Barak Obama** studied at **Harvard**

\[
e_1 \quad \text{studied_at} \quad e_2
\]

**Iggy Pop** has lived in **Berlin** during the 70’s

\[
e_1 \quad \text{has_lived} \quad e_2
\]
Unsupervised setting

Chomsky embarked on a program of study at UPenn

Barak Obama studied at Harvard

Iggy Pop has lived in Berlin 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)
```
Chomsky embarked on a program of study at UPenn

Arguments reconstruction

Not observable in the data
Relation induction

Chomsky embarked on a program of study at UPenn

Not observable in the data
Outline

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

\[ p(e_2 | e_1, r, \theta) \]

\[ u_{e_1}, u_{e_2} \in \mathbb{R}^d \quad - \text{encode semantic properties of entities } e_1 \text{ and } e_2 \]
Reconstruction component

Factorization model (Reconstruction)

\[ p(e_2|e_1, r, \theta) \]

\[ u_{e_1}, u_{e_2} \in \mathbb{R}^d \] - encode semantic properties of entities \( e_1 \) and \( e_2 \)

RESCAL factorization

Nickel et al. (2011)

\[ \psi^{RS}(e_1, e_2, r, \theta) = u_{e_1}^T C_r u_{e_2} \]

encodes interdependencies between entities
Reconstruction component

\[ p(e_2 | e_1, r, \theta) \]

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

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:

\[ p(e_2 | e_1, r, \theta) = \frac{\exp(\psi(e_1, e_2, r, \theta))}{\sum_{e' \in \mathcal{E}} \exp(\psi(e_1, e', r, \theta))} \]
Selectional preferences

\[
\psi^{SP}(e_1, e_2, r, \theta) = \sum_{i=1}^{2} u_{ei}^T c_{ir}
\]

Séaghdha (2010)
Reconstruction component

Selectional preferences

\[ \psi^{SP}(e_1, e_2, r, \theta) = \sum_{i=1}^{2} u_{e_i}^T c_{ir} \]

Séaghdha (2010)

Hybrid

\[ \psi^{HY}(e_1, e_2, r, \theta) = u_{e_1}^T C_r u_{e_2} + \sum_{i=1}^{2} u_{e_i}^T c_{ir} \]

combines RESCAL model and selectional preferences

scores each entity independently
The relation extraction model:

\[ q(r | x, w) = \frac{\exp(w^T g(r, x))}{\sum_{r' \in R} \exp(w^T g(r', x))} \]
For each sentence, we optimize the entity prediction quality while marginalizing over relations:

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

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

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

\[H(q)\]
For each sentence, we optimize the entity prediction quality while marginalizing over relations:

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

\[
E_q[\log p(e_i | e_{-i}, r, \theta)] - H(q)
\]

Variational lower bound on the pseudo-likelihood

Kingma and Welling (2014)
Joint learning

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

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

\[
E_{q}[\log p(e_i | e_{-i}, r, \theta)]
\]

\[H(q)\]

Not very tractable in this exact form:

- negative sampling (as, e.g., in Mikolov et al '13) instead of 'softmax'

Kingma and Welling (2014)
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)

- **DIRT**: 0.283
- **Rel-LDA**: 0.296
- **RESCAL**: 0.345
- **Select. Pref.**: 0.334
- **Hybrid**: 0.358

Clusters:
- **Clustering baseline**
- **Generative baseline**
Results (F1)

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

Generative baseline
Modelling the interdependence of arguments is beneficial.
Modelling the interdependence of arguments is beneficial. 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