Deep Learning for Unsupervised Relation Extraction

Étienne Simon

5 july 2022

ISIR, Sorbonne Université

Pr. Alexandre Allauzen
Université Paris-Dauphine PSL, ESPCI

Dr. Benoît Favre
Aix-Marseille Université

Rapporteurs

Pr. Pascale Sébillot
IRISA, INSA Rennes

Pr. Xavier Tannier
Sorbonne Université

Examinateurs

Dr. Benjamin Piwowarski
CNRS, Sorbonne Université

Dr. Vincent Guigue
Sorbonne Université

Directeurs

$\Longrightarrow$ An "otter" entity exists.

Historical Context


$\Longrightarrow$ An "otter" entity exists.


$\Longrightarrow$ An "*inside of*" relation exists.


$\Longrightarrow$ An "otter" entity exists.


$\Longrightarrow$ An "*inside of*" relation exists.

Structuralism: interrelations are keys to our understanding of the world.

Historical Context


$\Longrightarrow$ An "otter" entity exists.


$\Longrightarrow$ An "*inside of*" relation exists.

Structuralism: interrelations are keys to our understanding of the world.

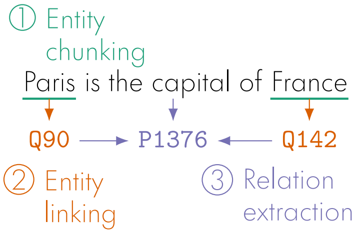
Realism

Nominalism

Entities and relations are unproductive concepts.
They only capture synonymy.

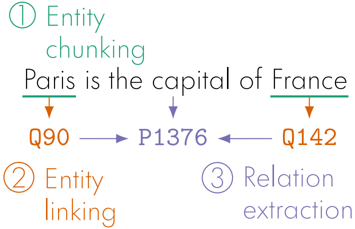
Information Extraction

Maps between two symbolic representations (text and knowledge bases).
Knowledge bases are set of facts:
(entity, *relation*, entity)



Information Extraction

Maps between two symbolic representations (text and knowledge bases).
Knowledge bases are set of facts:
(entity, *relation*, entity)



Symbolic Representations

symbol ↔ concept
e.g.: one-hot vector, text (Paris is the capital of France),
knowledge base (Paris^{Q90}, *capital*^{P1376}, France^{Q142})

Distributed Representations

concept → several units; unit → part of several concepts
e.g.: embeddings, neural network activations

Megrez$_{e_1}^{\texttt{Q850779}}$ is a star in the northern circumpolar constellation of Ursa Major$_{e_2}^{\texttt{Q10460}}$. → e_1 *part of constellation* e_2

Posidonius$_{e_1}^{\texttt{Q185770}}$ was a Greek philosopher, astronomer, historian, mathematician, and teacher native to Apamea, Syria$_{e_2}^{\texttt{Q617550}}$. → e_1 *born in* e_2

Hipparchus$_{e_1}^{\texttt{Q159905}}$ was born in Nicaea, Bithynia$_{e_2}^{\texttt{Q739037}}$, and probably died on the island of Rhodes, Greece. → e_1 *born in* e_2

In an unsupervised fashion.
Two kind of approaches: clustering and similarity function.

Megrez$_{e_1}^{\texttt{Q850779}}$ is a star in the northern circumpolar constellation of Ursa Major$_{e_2}^{\texttt{Q10460}}$. → e_1 *part of constellation* e_2

Posidonius$_{e_1}^{\texttt{Q185770}}$ was a Greek philosopher, astronomer, historian, mathematician, and teacher native to Apamea, Syria$_{e_2}^{\texttt{Q617550}}$. → e_1 *born in* e_2

Hipparchus$_{e_1}^{\texttt{Q159905}}$ was born in Nicaea, Bithynia$_{e_2}^{\texttt{Q739037}}$, and probably died on the island of Rhodes, Greece. → e_1 *born in* e_2

Megrez$_{e_1}^{\texttt{Q850779}}$ is a star in the northern circumpolar constellation of Ursa Major$_{e_2}^{\texttt{Q10460}}$. → e_1 *part of constellation* e_2

Posidonius$_{e_1}^{\texttt{Q185770}}$ was a Greek philosopher, astronomer, historian, mathematician, and teacher native to Apamea, Syria$_{e_2}^{\texttt{Q617550}}$. → e_1 *born in* e_2

Hipparchus$_{e_1}^{\texttt{Q159905}}$ was born in Nicaea, Bithynia$_{e_2}^{\texttt{Q739037}}$, and probably died on the island of Rhodes, Greece. → e_1 *born in* e_2

Megrez$_{e_1}^{\text{Q850779}}$ is a star in the northern circumpolar constellation of Ursa Major$_{e_2}^{\text{Q10460}}$. → e_1 part of constellation e_2

Posidonius$_{e_1}^{\text{Q185770}}$ was a Greek philosopher, astronomer, historian, mathematician, and teacher native to Apamea, Syria$_{e_2}^{\text{Q617550}}$. → e_1 born in e_2

Hipparchus$_{e_1}^{\text{Q159905}}$ was born in Nicaea, Bithynia$_{e_2}^{\text{Q739037}}$, and probably died on the island of Rhodes, Greece. → e_1 born in e_2

Same cluster $\iff$ Same relation

Induced clusters need not be labeled with a relation.

$\underline{\text{Megrez}}_{e_1}^{\text{Q850779}}$ is a star in the northern circumpolar constellation of $\underline{\text{Ursa Major}}_{e_2}^{\text{Q10460}}$.

→ e_1 part of constellation e_2

$\underline{\text{Posidonius}}_{e_1}^{\text{Q185770}}$ was a Greek philosopher, astronomer, historian, mathematician, and teacher native to $\underline{\text{Apamea, Syria}}_{e_2}^{\text{Q617550}}$.

→ e_1 born in e_2

$\underline{\text{Hipparchus}}_{e_1}^{\text{Q159905}}$ was born in $\underline{\text{Nicaea, Bithynia}}_{e_2}^{\text{Q739037}}$, and probably died on the island of Rhodes, Greece.

→ e_1 born in e_2

Same cluster $\iff$ Same relation

Induced clusters need **not** be labeled with a relation.

Clustering Metrics

B^3	Similar to standard F_1
V-measure	Entropic F_1
ARI	Pair of samples consistency

Megrez$_{e_1}^{\texttt{Q850779}}$ is a star in the northern circum-polar constellation of Ursa Major$_{e_2}^{\texttt{Q10460}}$. → e_1 *part of constellation* e_2

Posidonius$_{e_1}^{\texttt{Q185770}}$ was a Greek philosopher, astronomer, historian, mathematician, and teacher native to Apamea, Syria$_{e_2}^{\texttt{Q617550}}$. → e_1 *born in* e_2

Hipparchus$_{e_1}^{\texttt{Q159905}}$ was born in Nicaea, Bithynia$_{e_2}^{\texttt{Q739037}}$, and probably died on the island of Rhodes, Greece. → e_1 *born in* e_2

$\underline{\text{Megrez}}_{e_1}^{\texttt{Q850779}}$ is a star in the northern circumpolar constellation of $\underline{\text{Ursa Major}}_{e_2}^{\texttt{Q10460}}$.

$\underline{\text{Posidonius}}_{e_1}^{\texttt{Q185770}}$ was a Greek philosopher, astronomer, historian, mathematician, and teacher native to $\underline{\text{Apamea, Syria}}_{e_2}^{\texttt{Q617550}}$.

$\underline{\text{Hipparchus}}_{e_1}^{\texttt{Q159905}}$ was born in $\underline{\text{Nicaea, Bithynia}}_{e_2}^{\texttt{Q739037}}$, and probably died on the island of Rhodes, Greece.

Megrez$_{e_1}^{\text{Q850779}}$ is a star in the northern circumpolar constellation of Ursa Major$_{e_2}^{\text{Q10460}}$. } x_1

Posidonius$_{e_1}^{\text{Q185770}}$ was a Greek philosopher, astronomer, historian, mathematician, and teacher native to Apamea, Syria$_{e_2}^{\text{Q617550}}$. } x_2

Hipparchus$_{e_1}^{\text{Q159905}}$ was born in Nicaea, Bithynia$_{e_2}^{\text{Q739037}}$, and probably died on the island of Rhodes, Greece. } x_3

Learn a similarity function
$\text{sim}\colon \mathcal{D} \times \mathcal{D} \to \mathbb{R}$

$$\text{sim}(x_1, x_2) < \text{sim}(x_2, x_3)$$
$$\text{sim}(x_1, x_3) < \text{sim}(x_2, x_3)$$

5 way 1 shot: given 1 query and 5 candidates, which of the candidates is most similar to the query?
Evaluated using accuracy.

Étienne Simon, Vincent Guigue, Benjamin Piwowarski. **"Unsupervised Information Extraction: Regularizing Discriminative Approaches with Relation Distribution Losses"** ACL 2019 Part 1

Étienne Simon, Vincent Guigue, Benjamin Piwowarski. **"Graph-Based Unsupervised Relation Extraction"** Work in progress Part 2

Étienne Simon, Vincent Guigue, Benjamin Piwowarski. **"Unsupervised Information Extraction: Regularizing Discriminative Approaches with Relation Distribution Losses"** ACL 2019

- Introduce relation distribution losses
- First to train a deep RE classifier without supervision
- Improve over then SOTA

Part 1

Étienne Simon, Vincent Guigue, Benjamin Piwowarski. **"Graph-Based Unsupervised Relation Extraction"** Work in progress

Part 2

Étienne Simon, Vincent Guigue, Benjamin Piwowarski. **"Unsupervised Information Extraction: Regularizing Discriminative Approaches with Relation Distribution Losses"** ACL 2019

- Introduce relation distribution losses
- First to train a deep RE classifier without supervision
- Improve over then SOTA

Part 1

Étienne Simon, Vincent Guigue, Benjamin Piwowarski. **"Graph-Based Unsupervised Relation Extraction"** Work in progress

- Evaluate the quantity of topological information available
- Explicitly exploit aggregate setup for unsupervised RE
- Draw parallels between WL isomorphism test and unsupervised RE

Part 2

Regularizing Discriminative Models

Megrez$_{e_1}^{\text{Q850779}}$ is a star in the northern circumpolar constellation of Ursa Major$_{e_2}^{\text{Q10460}}$. → e_1 part of constellation e_2

Posidonius$_{e_1}^{\text{Q185770}}$ was a Greek philosopher, astronomer, historian, mathematician, and teacher native to Apamea, Syria$_{e_2}^{\text{Q617550}}$. → e_1 born in e_2

Hipparchus$_{e_1}^{\text{Q159905}}$ was born in Nicaea, Bithynia$_{e_2}^{\text{Q739037}}$, and probably died on the island of Rhodes, Greece. → e_1 born in e_2

Same cluster $\iff$ Same relation

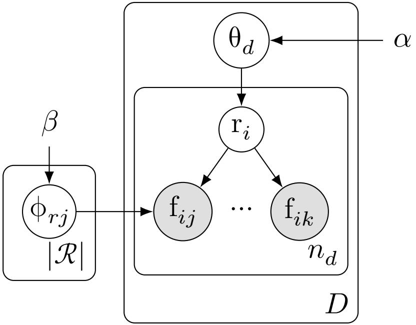
Induced clusters need **not** be labeled with a relation.

Evaluated using clustering metrics similar to standard F_1/precision/recall.

1. Related work
2. Limitation: can't train deep classifier
3. Model details
4. Analysis of limitation
5. Proposed solution
6. Results

Related Work: RelLDA (generative, 2011)

An LDA-like model:



θ_d distribution of relations in document d

r_i conveyed relation

ϕ_{rj} associate features to relations

f_i features:

1. bag of words of the infix;
2. surface form of the entities;
3. lemma words on the dependency path;
4. POS of the infix words;

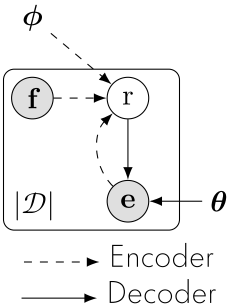
...

Assume $\mathscr{H}_{\text{BICLIQUE}}$: $\forall r \in \mathcal{R} : \exists A, B \subseteq \mathcal{E} : r \bullet \breve{r} = A^2 \wedge \breve{r} \bullet r = B^2$

Problem: Makes large independance assumptions.

Related Work: Marcheggiani (discriminative, 2016)

A conditional β-VAE:



Autoencode the entities $\mathbf{e}$ given the sentence features $\mathbf{f}$.

$$\mathcal{L}_{\text{VAE}}(\boldsymbol{\theta}, \boldsymbol{\phi}) = \mathcal{L}_{\text{reconstruction}}(\boldsymbol{\theta}, \boldsymbol{\phi}) + \mathcal{L}_{\text{VAE REG}}(\boldsymbol{\phi})$$

$$\mathcal{L}_{\text{VAE REG}}(\boldsymbol{\phi}) = \mathrm{D}_{\text{KL}}(Q(\mathrm{r} \mid \mathbf{e}; \boldsymbol{\phi}) \parallel \mathcal{U}(\mathcal{R}))$$

Assume $\mathscr{H}_{\text{UNIFORM}}$: All relations occur with equal frequency.

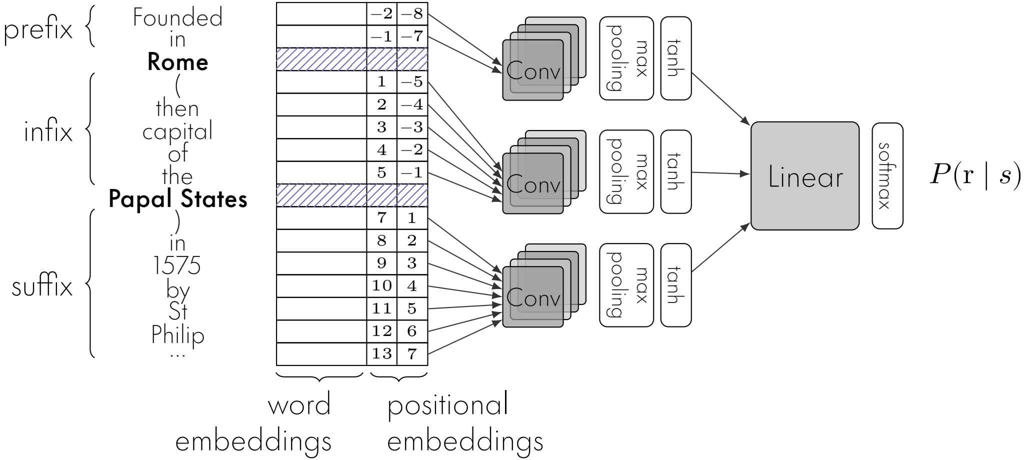
$$\forall r \in \mathcal{R}\colon P(r) = \frac{1}{|\mathcal{R}|}$$

Assume $\mathscr{H}_{1 \to 1}$: All relations are bijective.

$$\forall r \in \mathcal{R}\colon r \bullet \breve{r} \cup \boldsymbol{I} = \breve{r} \bullet r \cup \boldsymbol{I} = \boldsymbol{I}$$

Problem: Still uses hand designed features.

Supervised (old) SOTA: PCNN



Zeng et al. "Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks" EMNLP 2015

Experimental Setup

We introduced:

- 2 metrics (V-measure, ARI)
- 2 datasets (T-REx es)

B^3	Similar to standard F_1
V-measure	Entropic F_1
ARI	Pair of samples consistency

Model		B^3			V-measure			ARI
Classifier	Reg.	F_1	Prec.	Rec.	F_1	Hom.	Comp.	
rel-LDA		29.1	24.8	35.2	30.0	26.1	35.1	13.3
rel-LDA1		36.9	30.4	47.0	37.4	31.9	45.1	24.2
Linear	$\mathcal{L}_{\text{VAE REG}}$	35.2	23.8	67.1	27.0	18.6	49.6	18.7
PCNN	$\mathcal{L}_{\text{VAE REG}}$	27.6	24.3	31.9	24.7	21.2	29.6	15.7

Experimental Setup

We introduced:

- 2 metrics (V-measure, ARI)
- 2 datasets (T-REXes)

B^3 Similar to standard F_1
V-measure Entropic F_1
ARI Pair of samples consistency

Model: Classifier	Model: Reg.	B^3: F_1	B^3: Prec.	B^3: Rec.	V-measure: F_1	V-measure: Hom.	V-measure: Comp.	ARI
rel-LDA		29.1	24.8	35.2	30.0	26.1	35.1	13.3
rel-LDA1		36.9	30.4	47.0	37.4	31.9	45.1	24.2
Linear	$\mathcal{L}_{\text{VAE REG}}$	35.2	23.8	67.1	27.0	18.6	49.6	18.7
PCNN	$\mathcal{L}_{\text{VAE REG}}$	27.6	24.3	31.9	24.7	21.2	29.6	15.7

Yao et al. "Structured Relation Discovery using Generative Models" EMNLP 2011

Experimental Setup

We introduced:

- 2 metrics (V-measure, ARI)
- 2 datasets (T-RExes)

B^3 Similar to standard F_1

V-measure Entropic F_1

ARI Pair of samples consistency

Model		B^3			V-measure			ARI
Classifier	Reg.	F_1	Prec.	Rec.	F_1	Hom.	Comp.	
rel-LDA		29.1	24.8	35.2	30.0	26.1	35.1	13.3
rel-LDA1		36.9	30.4	47.0	37.4	31.9	45.1	24.2
Linear	$\mathcal{L}_{\text{VAE REG}}$	35.2	23.8	67.1	27.0	18.6	49.6	18.7
PCNN	$\mathcal{L}_{\text{VAE REG}}$	27.6	24.3	31.9	24.7	21.2	29.6	15.7

Marcheggiani and Titov "Discrete-State Variational Autoencoders for Joint Discovery and Factorization of Relations" TACL 2016

Experimental Setup

We introduced:

- 2 metrics (V-measure, ARI)
- 2 datasets (T-REXes)

B^3 Similar to standard F_1

V-measure Entropic F_1

ARI Pair of samples consistency

Model		B^3			V-measure			ARI
Classifier	Reg.	F_1	Prec.	Rec.	F_1	Hom.	Comp.	
rel-LDA		29.1	24.8	35.2	30.0	26.1	35.1	13.3
rel-LDA1		36.9	30.4	47.0	37.4	31.9	45.1	24.2
Linear	$\mathcal{L}_{\text{VAE REG}}$	35.2	23.8	67.1	27.0	18.6	49.6	18.7
PCNN	$\mathcal{L}_{\text{VAE REG}}$	27.6	24.3	31.9	24.7	21.2	29.6	15.7

Problem: Using a deep encoder does not work.

- We introduce a new formalism.
- The encoder and decoder are sub-models performing different tasks.
- The interaction between these two sub-models is problematic.

Fill-in-the-blank Surrogate Task

"The $\underline{\text{sol}}_{e_1}$ was the currency of $\underline{\text{?}}_{e_2}$ between 1863 and 1985."

"The $\underline{\text{sol}}_{e_1}$ was the currency of $\underline{\ ?\ }_{e_2}$ between 1863 and 1985."

e_{-i} missing entity, e_i remaining entity, s conveying sentence

$$\text{for } i = 1, 2: \qquad \overbrace{P(e_{-i} \mid s, e_i)}^{\text{fill-in-the-blank}}$$

"The $\underline{\text{sol}}_{e_1}$ was the currency of $\underline{\ ?\ }_{e_2}$ between 1863 and 1985."

e_{-i} missing entity, e_i remaining entity, s conveying sentence, r conveyed relation

$$\text{for } i = 1, 2: \quad \overbrace{P(e_{-i} \mid s, e_i)}^{\text{fill-in-the-blank}} \qquad \overbrace{P(e_{-i} \mid r, e_i)}^{\text{entity predictor}}$$

"The $\underline{\text{sol}}_{e_1}$ was the currency of $\underline{\ ?\ }_{e_2}$ between 1863 and 1985."

e_{-i} missing entity, e_i remaining entity, s conveying sentence, r conveyed relation

$$\text{for } i = 1, 2: \qquad \overbrace{P(e_{-i} \mid s, e_i)}^{\text{fill-in-the-blank}} = \sum_{r \in \mathcal{R}} \overbrace{P(r \mid s)}^{\text{classifier}} \; \overbrace{P(e_{-i} \mid r, e_i)}^{\text{entity predictor}}$$

Assume $\mathscr{H}_{\text{BLANKABLE}}$: The relation can be predicted from the text surrounding the two entities alone.

"The $\underline{\text{sol}}_{e_1}$ was the currency of $\underline{\ ?\ }_{e_2}$ between 1863 and 1985."

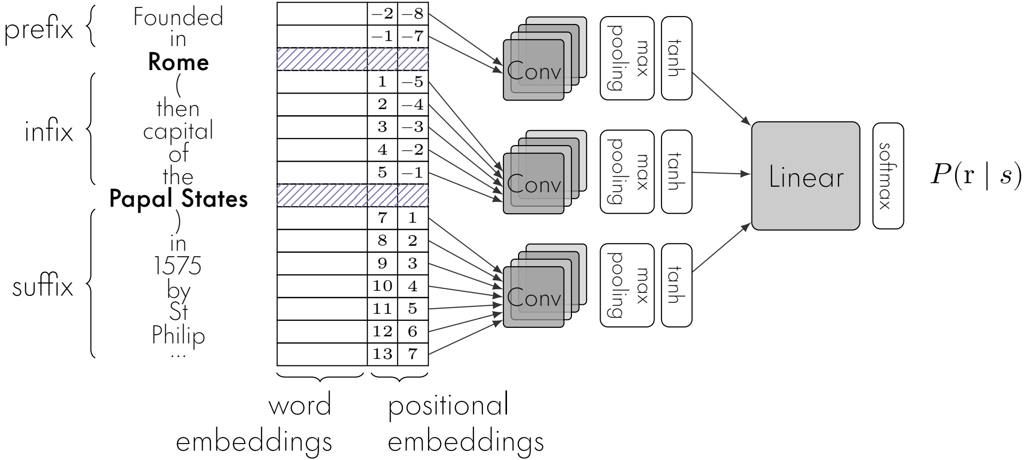
e_{-i} missing entity, e_i remaining entity, s conveying sentence, r conveyed relation

$$\text{for } i = 1, 2: \qquad \overbrace{P(e_{-i} \mid s, e_i)}^{\text{fill-in-the-blank}} = \sum_{r \in \mathcal{R}} \overbrace{P(r \mid s)}^{\text{classifier}} \overbrace{P(e_{-i} \mid r, e_i)}^{\text{entity predictor}}$$

Assume $\mathscr{H}_{\text{BLANKABLE}}$: The relation can be predicted from the text surrounding the two entities alone.

1. Train a fill-in-the-blank model on an unsupervised dataset.
2. Throw away the entity predictor.
3. Use the classifier on new samples.

Relation Classifier: PCNN



$$\overbrace{P(e_{-i} \mid s, e_i)}^{\text{fill-in-the-blank}} = \sum_{r \in \mathcal{R}} \overbrace{P(r \mid s)}^{\text{classifier}} \overbrace{P(e_{-i} \mid r, e_i)}^{\text{entity predictor}}$$

Hybrid (Marcheggiani and Titov 2016)

$$\psi(e_1, r, e_2) = \psi_{\text{SP}}(e_1, r, e_2) + \psi_{\text{RESCAL}}(e_1, r, e_2)$$

$$P(e_1 \mid r, e_2) = \frac{\exp \psi(e_1, r, e_2)}{\sum_{e' \in \mathcal{E}} \exp \psi(e', r, e_2)}$$

Selectional Preferences

$$\psi_{\text{SP}}(e_1, r, e_2) = \boldsymbol{u}_{e_1}^{\top} \boldsymbol{a}_r + \boldsymbol{u}_{e_2}^{\top} \boldsymbol{b}_r$$

$\boldsymbol{U} \in \mathbb{R}^{\mathcal{E} \times d}$ entity embeddings
$\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{\mathcal{R} \times d}$ relation embeddings

RESCAL

$$\psi_{\text{RESCAL}}(e_1, r, e_2) = \boldsymbol{u}_{e_1}^{\top} \boldsymbol{C}_r \boldsymbol{u}_{e_2}$$

$\boldsymbol{U} \in \mathbb{R}^{\mathcal{E} \times d}$ entity embeddings
$\boldsymbol{C} \in \mathbb{R}^{\mathcal{R} \times d \times d}$ relation embeddings

$$\overbrace{P(e_{-i} \mid s, e_i)}^{\text{fill-in-the-blank}} = \sum_{r \in \mathcal{R}} \overbrace{P(r \mid s)}^{\text{classifier}} \overbrace{P(e_{-i} \mid r, e_i)}^{\text{entity predictor}}$$

Negative Sampling Approximation

$$\mathcal{L}_{\mathsf{EP}}(\boldsymbol{\theta}, \boldsymbol{\phi}) = \mathop{\mathbb{E}}_{\substack{(\mathrm{s},\mathrm{e}_1,\mathrm{e}_2)\sim\mathcal{U}(\mathcal{D}) \\ \mathrm{r}\sim\mathsf{PCNN}(\mathrm{s};\boldsymbol{\phi})}} \Bigg[- \log \sigma\left(\psi(\mathrm{e}_1, \mathrm{r}, \mathrm{e}_2; \boldsymbol{\theta})\right) - \sum_{j=1}^{k} \mathop{\mathbb{E}}_{\mathrm{e}'\sim\mathcal{U}_{\mathcal{D}}(\mathcal{E})} \left[\log \sigma\left(-\psi(\mathrm{e}_1, \mathrm{r}, \mathrm{e}'; \boldsymbol{\theta})\right)\right] - \sum_{j=1}^{k} \mathop{\mathbb{E}}_{\mathrm{e}'\sim\mathcal{U}_{\mathcal{D}}(\mathcal{E})} \left[\log \sigma\left(-\psi(\mathrm{e}', \mathrm{r}, \mathrm{e}_2; \boldsymbol{\theta})\right)\right] \Bigg]$$

$$\mathcal{L}_{\mathsf{EP}}(\boldsymbol{\theta}, \boldsymbol{\phi}) = \underset{\substack{(\mathrm{s},\mathrm{e}_1,\mathrm{e}_2)\sim\mathcal{U}(\mathcal{D}) \\ \mathrm{r}\sim\mathsf{PCNN}(\mathrm{s};\boldsymbol{\phi})}}{\mathbb{E}} \Bigg[-\log\sigma\left(\psi(\mathrm{e}_1, \mathrm{r}, \mathrm{e}_2; \boldsymbol{\theta})\right) - \sum_{j=1}^{k} \underset{\mathrm{e}'\sim\mathcal{U}_{\mathcal{D}}(\mathcal{E})}{\mathbb{E}} \left[\log\sigma\left(-\psi(\mathrm{e}_1, \mathrm{r}, \mathrm{e}'; \boldsymbol{\theta})\right)\right] - \sum_{j=1}^{k} \underset{\mathrm{e}'\sim\mathcal{U}_{\mathcal{D}}(\mathcal{E})}{\mathbb{E}} \left[\log\sigma\left(-\psi(\mathrm{e}', \mathrm{r}, \mathrm{e}_2; \boldsymbol{\theta})\right)\right] \Bigg]$$

1. Take a sample uniformly from the dataset.

$$\mathcal{L}_{\mathsf{EP}}(\boldsymbol{\theta}, \boldsymbol{\phi}) = \underset{\substack{(\mathrm{s},\mathrm{e}_1,\mathrm{e}_2)\sim\mathcal{U}(\mathcal{D}) \\ \mathrm{r}\sim\mathsf{PCNN}(\mathrm{s};\boldsymbol{\phi})}}{\mathbb{E}} \Bigg[-\log\sigma\left(\psi(\mathrm{e}_1, \mathrm{r}, \mathrm{e}_2; \boldsymbol{\theta})\right) - \sum_{j=1}^{k} \underset{\mathrm{e}'\sim\mathcal{U}_{\mathcal{D}}(\mathcal{E})}{\mathbb{E}} \left[\log\sigma\left(-\psi(\mathrm{e}_1, \mathrm{r}, \mathrm{e}'; \boldsymbol{\theta})\right)\right] - \sum_{j=1}^{k} \underset{\mathrm{e}'\sim\mathcal{U}_{\mathcal{D}}(\mathcal{E})}{\mathbb{E}} \left[\log\sigma\left(-\psi(\mathrm{e}', \mathrm{r}, \mathrm{e}_2; \boldsymbol{\theta})\right)\right] \Bigg]$$

1. Take a sample uniformly from the dataset.
2. Sample a relation r from the output of the PCNN classifier.

$$\mathcal{L}_{\mathsf{EP}}(\boldsymbol{\theta}, \boldsymbol{\phi}) = \mathop{\mathbb{E}}_{\substack{(\mathrm{s},\mathrm{e}_1,\mathrm{e}_2)\sim\mathcal{U}(\mathcal{D}) \\ \mathrm{r}\sim\mathsf{PCNN}(\mathrm{s};\boldsymbol{\phi})}} \Bigg[-\log\sigma\left(\psi(\mathrm{e}_1, \mathrm{r}, \mathrm{e}_2; \boldsymbol{\theta})\right) - \sum_{j=1}^{k} \mathop{\mathbb{E}}_{\mathrm{e}'\sim\mathcal{U}_{\mathcal{D}}(\mathcal{E})} \left[\log\sigma\left(-\psi(\mathrm{e}_1, \mathrm{r}, \mathrm{e}'; \boldsymbol{\theta})\right)\right] - \sum_{j=1}^{k} \mathop{\mathbb{E}}_{\mathrm{e}'\sim\mathcal{U}_{\mathcal{D}}(\mathcal{E})} \left[\log\sigma\left(-\psi(\mathrm{e}', \mathrm{r}, \mathrm{e}_2; \boldsymbol{\theta})\right)\right] \Bigg]$$

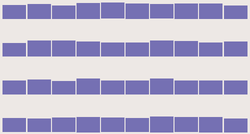
1. Take a sample uniformly from the dataset.
2. Sample a relation r from the output of the PCNN classifier.
3. Increase the energy of this fact.

$$\mathcal{L}_{\mathsf{EP}}(\boldsymbol{\theta}, \boldsymbol{\phi}) = \mathop{\mathbb{E}}_{\substack{(\mathrm{s}, \mathrm{e}_1, \mathrm{e}_2) \sim \mathcal{U}(\mathcal{D}) \\ \mathrm{r} \sim \mathsf{PCNN}(\mathrm{s}; \boldsymbol{\phi})}} \Bigg[-\log \sigma\left(\psi(\mathrm{e}_1, \mathrm{r}, \mathrm{e}_2; \boldsymbol{\theta})\right) - \sum_{j=1}^{k} \mathop{\mathbb{E}}_{\mathrm{e}' \sim \mathcal{U}_{\mathcal{D}}(\mathcal{E})} \left[\log \sigma\left(-\psi(\mathrm{e}_1, \mathrm{r}, \mathrm{e}'; \boldsymbol{\theta})\right)\right] - \sum_{j=1}^{k} \mathop{\mathbb{E}}_{\mathrm{e}' \sim \mathcal{U}_{\mathcal{D}}(\mathcal{E})} \left[\log \sigma\left(-\psi(\mathrm{e}', \mathrm{r}, \mathrm{e}_2; \boldsymbol{\theta})\right)\right] \Bigg]$$

1. Take a sample uniformly from the dataset.
2. Sample a relation r from the output of the PCNN classifier.
3. Increase the energy of this fact.
4. Decrease the energy of negative facts. ($\mathscr{H}_{1 \to 1}$)

Degenerate distributions

$$P(\mathrm{r} \mid s_1) =$$
$$P(\mathrm{r} \mid s_2) =$$
$$P(\mathrm{r} \mid s_3) =$$
$$P(\mathrm{r} \mid s_4) =$$
$$\vdots$$



$$P(\mathrm{r} \mid s_1) =$$
$$P(\mathrm{r} \mid s_2) =$$
$$P(\mathrm{r} \mid s_3) =$$
$$P(\mathrm{r} \mid s_4) =$$
$$\vdots$$

Desired distribution

$$P(\mathrm{r} \mid s_1) =$$
$$P(\mathrm{r} \mid s_2) =$$
$$P(\mathrm{r} \mid s_3) =$$
$$P(\mathrm{r} \mid s_4) =$$
$$\vdots$$

VAE Model Reminder (Marcheggiani)

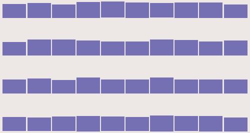
$$\overbrace{P(e_{-i} \mid s, e_i)}^{\text{fill-in-the-blank}} = \sum_{r \in \mathcal{R}} \overbrace{P(r \mid s)}^{\text{classifier}} \overbrace{P(e_{-i} \mid r, e_i)}^{\text{entity predictor}}$$

$$\mathcal{L}_{\text{VAE REG}}(\phi) = \mathrm{D}_{\text{KL}}(Q(\mathrm{r} \mid \mathbf{e}; \phi) \parallel \mathcal{U}(\mathcal{R}))$$

Source of Low Scores

Degenerate distributions

$$P(\mathrm{r} \mid s_1) = \quad P(\mathrm{r} \mid s_2) = \quad P(\mathrm{r} \mid s_3) = \quad P(\mathrm{r} \mid s_4) =$$



$$P(\mathrm{r} \mid s_1) = \quad P(\mathrm{r} \mid s_2) = \quad P(\mathrm{r} \mid s_3) = \quad P(\mathrm{r} \mid s_4) =$$

Problem: Marcheggiani's model cannot handle deep encoder.

Desired distribution

$$P(\mathrm{r} \mid s_1) = \quad P(\mathrm{r} \mid s_2) = \quad P(\mathrm{r} \mid s_3) = \quad P(\mathrm{r} \mid s_4) = \quad \vdots$$

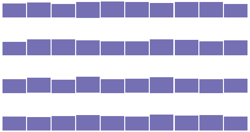
VAE Model Reminder (Marcheggiani)

$$\overbrace{P(e_{-i} \mid s, e_i)}^{\text{fill-in-the-blank}} = \sum_{r \in \mathcal{R}} \overbrace{P(r \mid s)}^{\text{classifier}} \overbrace{P(e_{-i} \mid r, e_i)}^{\text{entity predictor}}$$

$$\mathcal{L}_{\text{VAE REG}}(\phi) = \mathrm{D}_{\text{KL}}(Q(\mathrm{r} \mid \mathbf{e}; \phi) \parallel \mathcal{U}(\mathcal{R}))$$

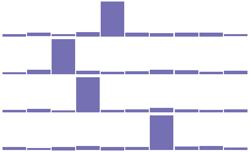
Degenerate distributions:

$$P(\mathrm{r} \mid s_1) =$$
$$P(\mathrm{r} \mid s_2) =$$
$$P(\mathrm{r} \mid s_3) =$$
$$P(\mathrm{r} \mid s_4) =$$
$$\vdots$$



Desired distributions:

$$P(\mathrm{r} \mid s_1) =$$
$$P(\mathrm{r} \mid s_2) =$$
$$P(\mathrm{r} \mid s_3) =$$
$$P(\mathrm{r} \mid s_4) =$$
$$\vdots$$



Ensure Confidence

$$\mathcal{L}_{\mathrm{S}}(\boldsymbol{\phi}) = \mathop{\mathbb{E}}_{(\mathrm{s},\mathbf{e})\sim\mathcal{U}(\mathcal{D})} [\mathrm{H}(\mathrm{R} \mid \mathrm{s}, \mathbf{e}; \boldsymbol{\phi})]$$

The entropy of the relation distribution must be low for each sample.

Distribution Distance Loss

Degenerate distributions:

$P(\mathrm{r} \mid s_1) =$

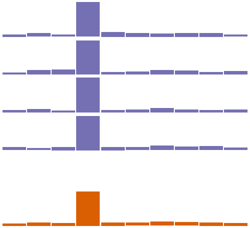
$P(\mathrm{r} \mid s_2) =$

$P(\mathrm{r} \mid s_3) =$

$P(\mathrm{r} \mid s_4) =$

$\vdots$

expectation =



Desired distributions:

$P(\mathrm{r} \mid s_1) =$

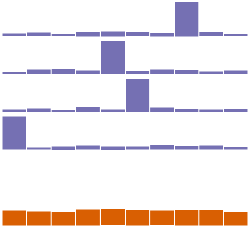
$P(\mathrm{r} \mid s_2) =$

$P(\mathrm{r} \mid s_3) =$

$P(\mathrm{r} \mid s_4) =$

$\vdots$

expectation =



Ensure Diversity

$$\mathcal{L}_{\mathsf{D}}(\boldsymbol{\phi}) = \mathrm{D}_{\mathsf{KL}}(P(\mathrm{R} \mid \boldsymbol{\phi}) \parallel \mathcal{U}(\mathcal{R}))$$

At the level of the dataset (or mini-batch) the distribution of relations must be uniform.

Quantitative Results: NYT+FB

Model		B^3			V-measure			ARI
Classifier	Reg.	F_1	Prec.	Rec.	F_1	Hom.	Comp.	
rel-LDA		29.1	24.8	35.2	30.0	26.1	35.1	13.3
rel-LDA1		36.9	30.4	47.0	37.4	31.9	45.1	24.2
Linear	$\mathcal{L}_{\text{VAE REG}}$	35.2	23.8	67.1	27.0	18.6	49.6	18.7
PCNN	$\mathcal{L}_{\text{VAE REG}}$	27.6	24.3	31.9	24.7	21.2	29.6	15.7
Linear	$\mathcal{L}_S + \mathcal{L}_D$	37.5	31.1	47.4	**38.7**	32.6	47.8	27.6
PCNN	$\mathcal{L}_S + \mathcal{L}_D$	**39.4**	32.2	50.7	38.3	32.2	47.2	**33.8**
BERTcoder	$\mathcal{L}_S + \mathcal{L}_D$	41.5	34.6	51.8	39.9	33.9	48.5	35.1
BERTcoder	SelfORE	*49.1*	47.3	51.1	*46.6*	45.7	47.6	*40.3*

Quantitative Results: NYT+FB

Model		B^3			V-measure			ARI
Classifier	Reg.	F_1	Prec.	Rec.	F_1	Hom.	Comp.	
rel-LDA		29.1	24.8	35.2	30.0	26.1	35.1	13.3
rel-LDA1		36.9	30.4	47.0	37.4	31.9	45.1	24.2
Linear	$\mathcal{L}_{\text{VAE REG}}$	35.2	23.8	67.1	27.0	18.6	49.6	18.7
PCNN	$\mathcal{L}_{\text{VAE REG}}$	27.6	24.3	31.9	24.7	21.2	29.6	15.7
Linear	$\mathcal{L}_{\text{S}} + \mathcal{L}_{\text{D}}$	37.5	31.1	47.4	**38.7**	32.6	47.8	27.6
PCNN	$\mathcal{L}_{\text{S}} + \mathcal{L}_{\text{D}}$	**39.4**	32.2	50.7	38.3	32.2	47.2	**33.8**
BERTcoder	$\mathcal{L}_{\text{S}} + \mathcal{L}_{\text{D}}$	41.5	34.6	51.8	39.9	33.9	48.5	35.1
BERTcoder	SelfORE	***49.1***	47.3	51.1	***46.6***	45.7	47.6	***40.3***

Quantitative Results: NYT+FB

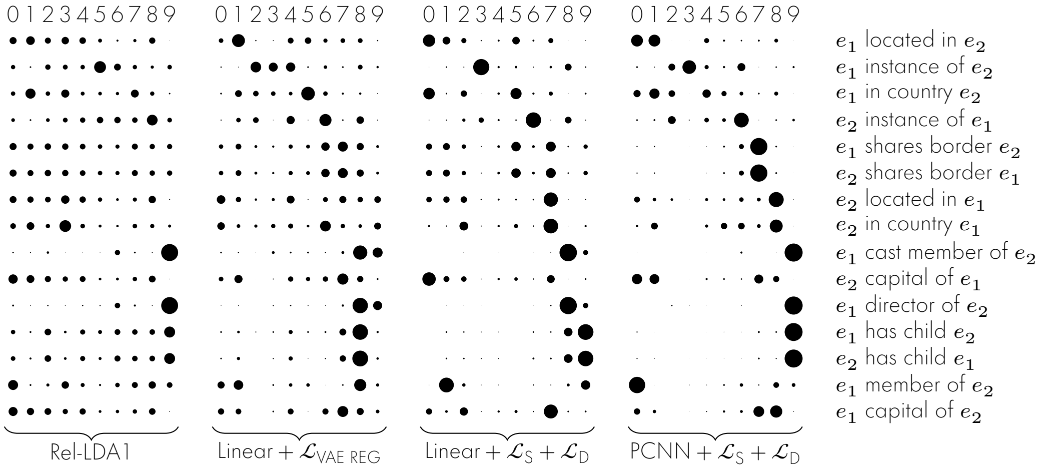
Model		B^3			V-measure			ARI
Classifier	Reg.	F_1	Prec.	Rec.	F_1	Hom.	Comp.	
rel-LDA		29.1	24.8	35.2	30.0	26.1	35.1	13.3
rel-LDA1		36.9	30.4	47.0	37.4	31.9	45.1	24.2
Linear	$\mathcal{L}_{\text{VAE REG}}$	35.2	23.8	67.1	27.0	18.6	49.6	18.7
PCNN	$\mathcal{L}_{\text{VAE REG}}$	27.6	24.3	31.9	24.7	21.2	29.6	15.7
Linear	$\mathcal{L}_S + \mathcal{L}_D$	37.5	31.1	47.4	**38.7**	32.6	47.8	27.6
PCNN	$\mathcal{L}_S + \mathcal{L}_D$	**39.4**	32.2	50.7	38.3	32.2	47.2	**33.8**
BERTcoder	$\mathcal{L}_S + \mathcal{L}_D$	41.5	34.6	51.8	39.9	33.9	48.5	35.1
BERTcoder	SelfORE	*49.1*	47.3	51.1	*46.6*	45.7	47.6	*40.3*

Quantitative Results: NYT+FB

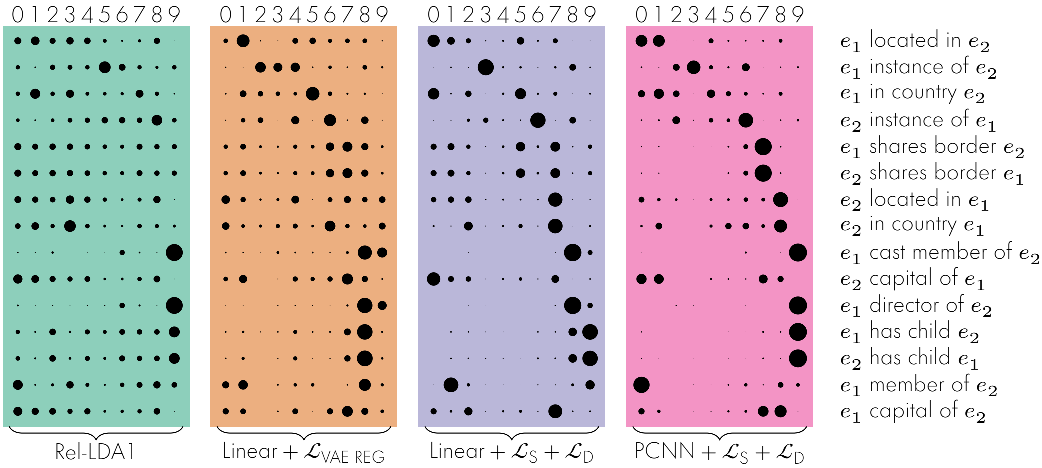
Model		B^3			V-measure			ARI
Classifier	Reg.	F_1	Prec.	Rec.	F_1	Hom.	Comp.	
rel-LDA		29.1	24.8	35.2	30.0	26.1	35.1	13.3
rel-LDA1		36.9	30.4	47.0	37.4	31.9	45.1	24.2
Linear	$\mathcal{L}_{\text{VAE REG}}$	35.2	23.8	67.1	27.0	18.6	49.6	18.7
PCNN	$\mathcal{L}_{\text{VAE REG}}$	27.6	24.3	31.9	24.7	21.2	29.6	15.7
Linear	$\mathcal{L}_S + \mathcal{L}_D$	37.5	31.1	47.4	**38.7**	32.6	47.8	27.6
PCNN	$\mathcal{L}_S + \mathcal{L}_D$	**39.4**	32.2	50.7	38.3	32.2	47.2	**33.8**
BERTcoder	$\mathcal{L}_S + \mathcal{L}_D$	41.5	34.6	51.8	39.9	33.9	48.5	35.1
BERTcoder	SelfORE	***49.1***	47.3	51.1	***46.6***	45.7	47.6	***40.3***

Hu et al. "SelfORE: Self-supervised Relational Feature Learning for Open Relation Extraction" EMNLP 2020

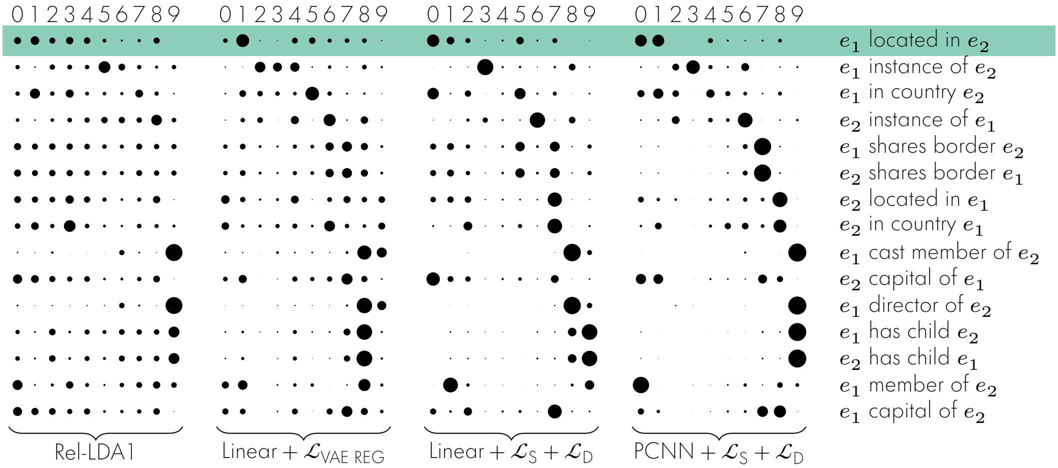
Qualitative Results: Confusion Matrices



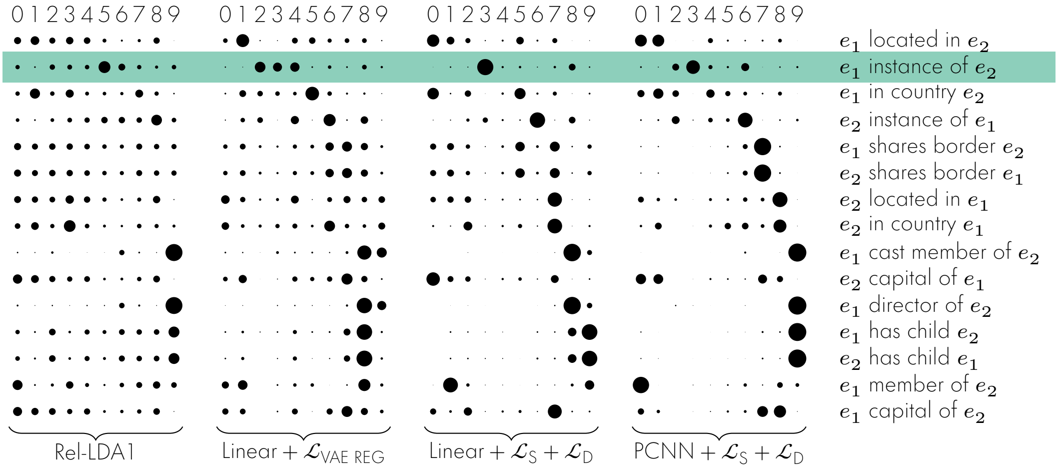
Qualitative Results: Confusion Matrices



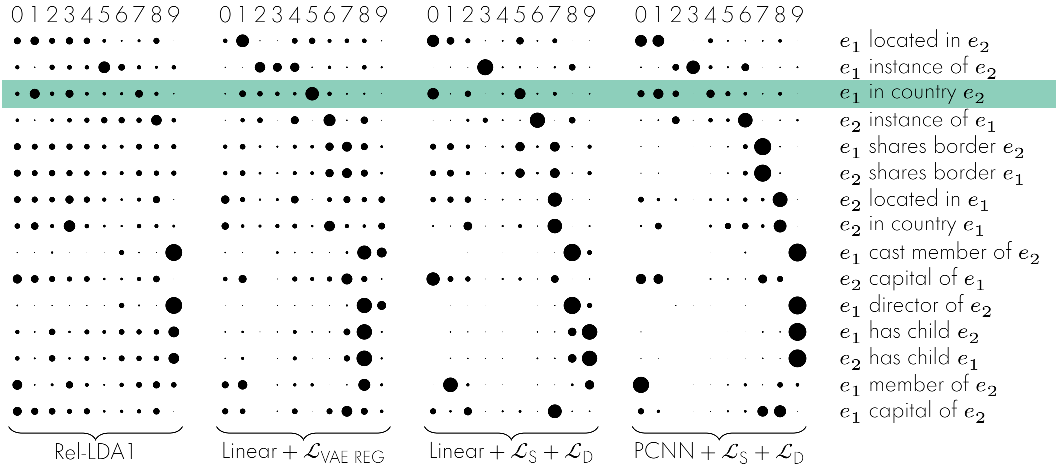
Qualitative Results: Confusion Matrices

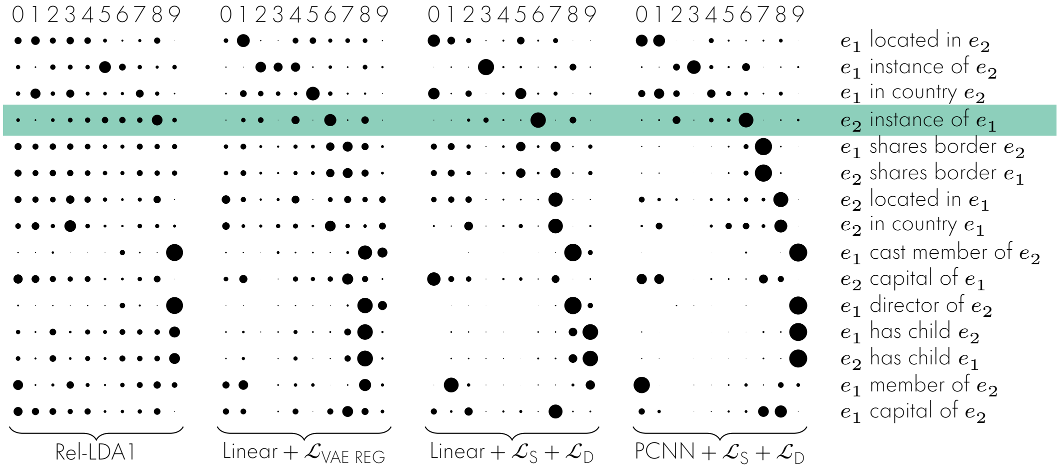


Qualitative Results: Confusion Matrices

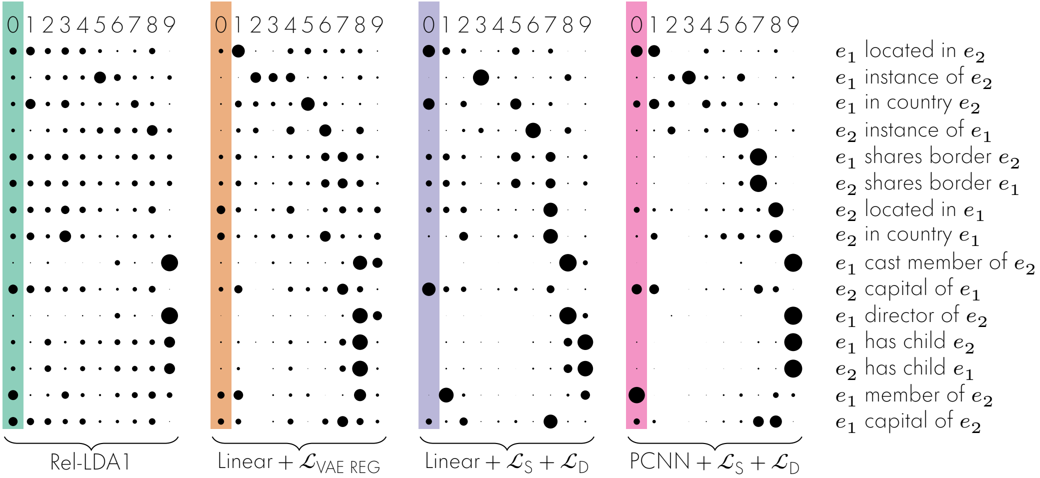


Qualitative Results: Confusion Matrices

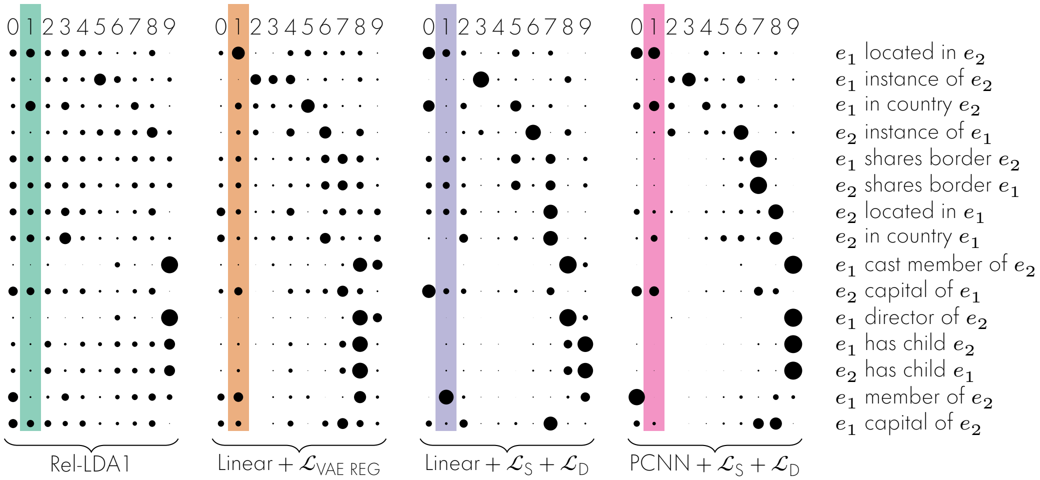



0 1 2 3 4 5 6 7 8 9
0 1 2 3 4 5 6 7 8 9
0 1 2 3 4 5 6 7 8 9
0 1 2 3 4 5 6 7 8 9
e_1 located in e_2
e_1 instance of e_2
e_1 in country e_2
e_2 instance of e_1
e_1 shares border e_2
e_2 shares border e_1
e_2 located in e_1
e_2 in country e_1
e_1 cast member of e_2
e_2 capital of e_1
e_1 director of e_2
e_1 has child e_2
e_2 has child e_1
e_1 member of e_2
e_1 capital of e_2
Rel-LDA1
Linear + $\mathcal{L}_{\text{VAE REG}}$
Linear + $\mathcal{L}_S + \mathcal{L}_D$
PCNN + $\mathcal{L}_S + \mathcal{L}_D$

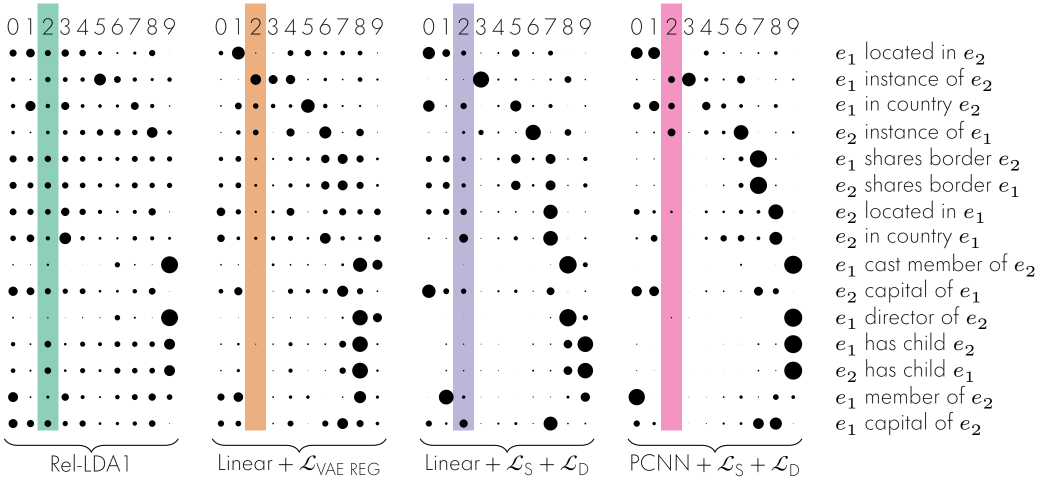
Qualitative Results: Confusion Matrices



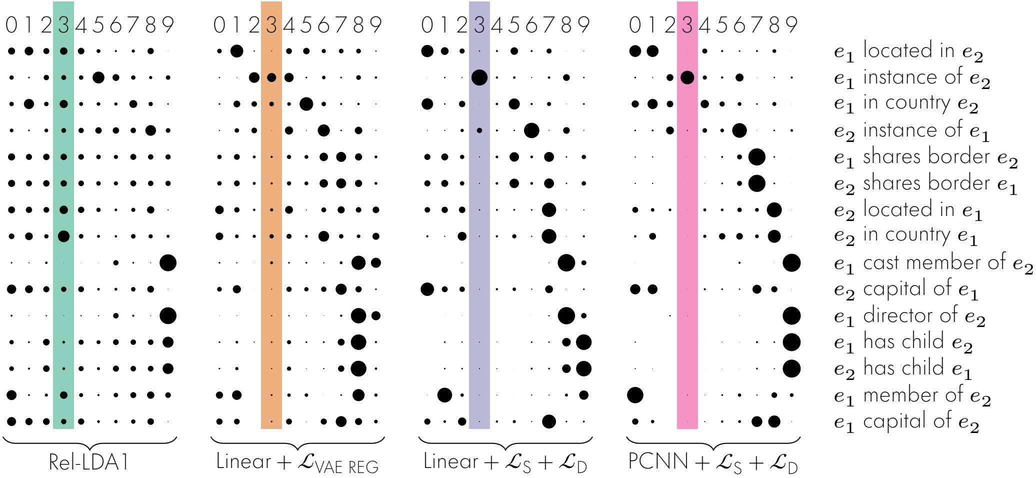
Qualitative Results: Confusion Matrices



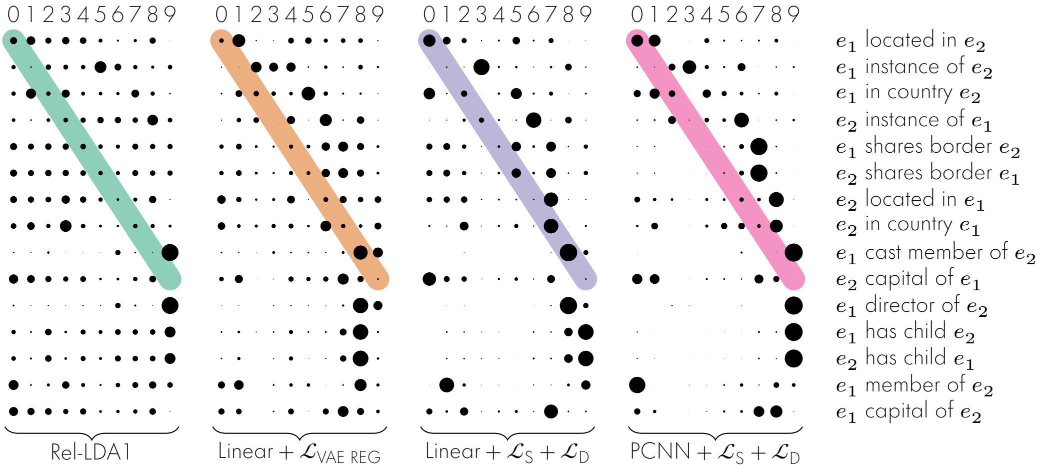
Qualitative Results: Confusion Matrices



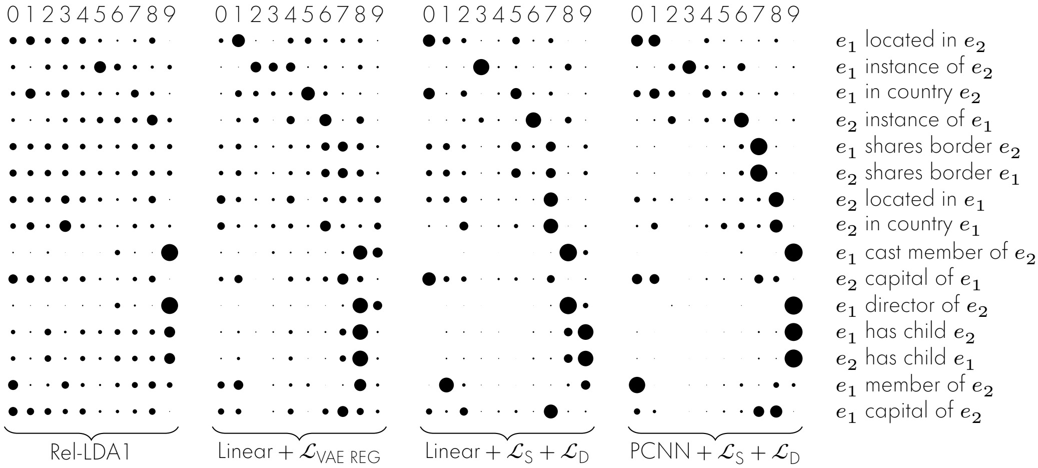
Qualitative Results: Confusion Matrices



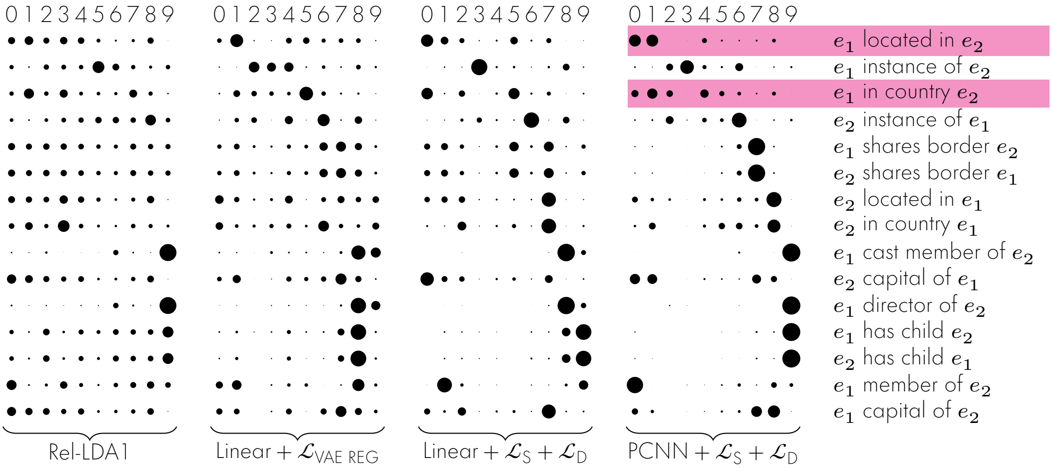
Qualitative Results: Confusion Matrices



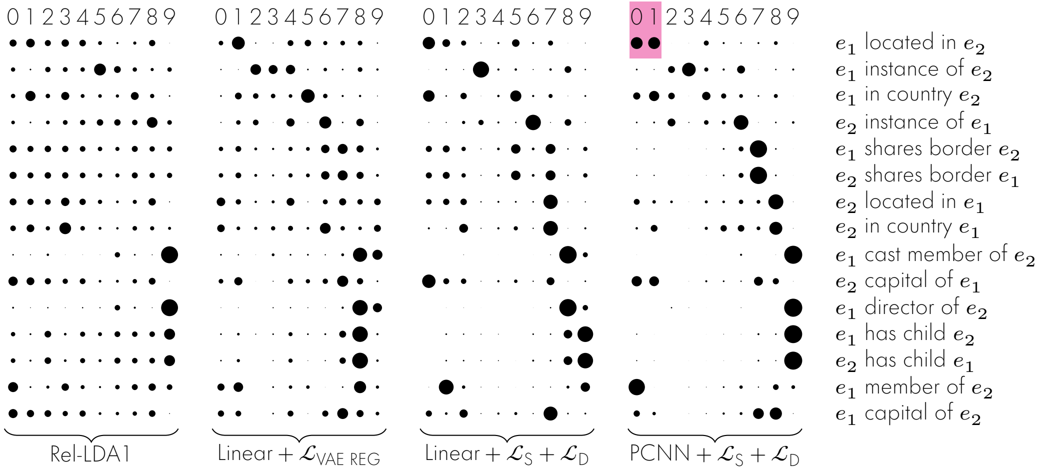
Qualitative Results: Confusion Matrices



Qualitative Results: Confusion Matrices



Qualitative Results: Confusion Matrices



Take-home Message

Selecting good regularizations to enforce modeling hypotheses enables us to train a deep classifier.

Contributions

- Train a PCNN without supervision
- Designed two regularization losses (Skewness, Distribution distance)
- Introduced new datasets (T-REXes)
- Evaluated using additional metrics (V-measure, ARI)

Étienne Simon, Vincent Guigue, Benjamin Piwowarski. **“Unsupervised Information Extraction: Regularizing Discriminative Approaches with Relation Distribution Losses”** ACL 2019

Graph-based Aggregate Extraction

Megrez$_{e_1}^{\text{Q850779}}$ is a star in the northern circumpolar constellation of Ursa Major$_{e_2}^{\text{Q10460}}$. } x_1

Posidonius$_{e_1}^{\text{Q185770}}$ was a Greek philosopher, astronomer, historian, mathematician, and teacher native to Apamea, Syria$_{e_2}^{\text{Q617550}}$. } x_2

Hipparchus$_{e_1}^{\text{Q159905}}$ was born in Nicaea, Bithynia$_{e_2}^{\text{Q739037}}$, and probably died on the island of Rhodes, Greece. } x_3

Learn a similarity function
$\text{sim}\colon \mathcal{D} \times \mathcal{D} \rightarrow \mathbb{R}$

$$\text{sim}(x_1, x_2) < \text{sim}(x_2, x_3)$$
$$\text{sim}(x_1, x_3) < \text{sim}(x_2, x_3)$$

5 way 1 shot: given 1 query and 5 candidates, which of the candidates is most similar to the query?
Evaluated using accuracy.

Sentential approaches: extract sentences' relation independently ($\mathcal{S} \times \mathcal{E}^2 \to \mathcal{R}$)
Aggregate approaches: maps a set of sentences to a set of facts ($2^{\mathcal{S} \times \mathcal{E}^2} \to 2^{\mathcal{E}^2 \times \mathcal{R}}$)

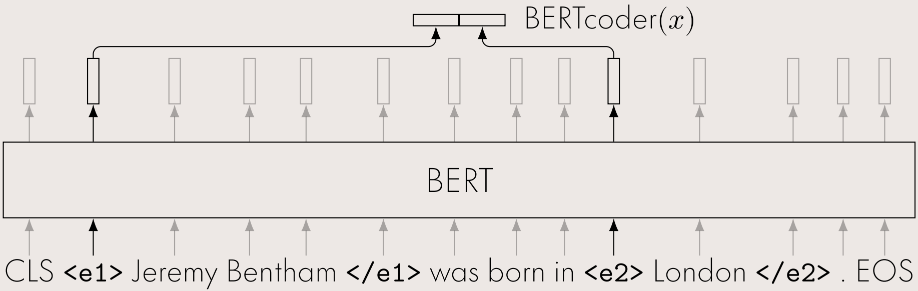
Goal

Exploit dataset-level regularities to leverage additional information

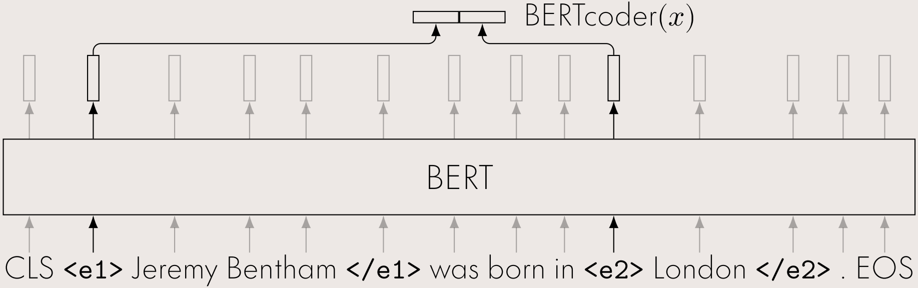
Plan

1. Model datasets as graphs
2. Related relation extraction work only uses **linguistic** similarities
3. Proof that **topological** information can be used
4. How topological features are usually extracted (GCN)
5. How to extract them differently (WL isomorphism test)
6. Experimental results
7. Perspective

BERTcoder (linguistic)



BERTcoder (linguistic)

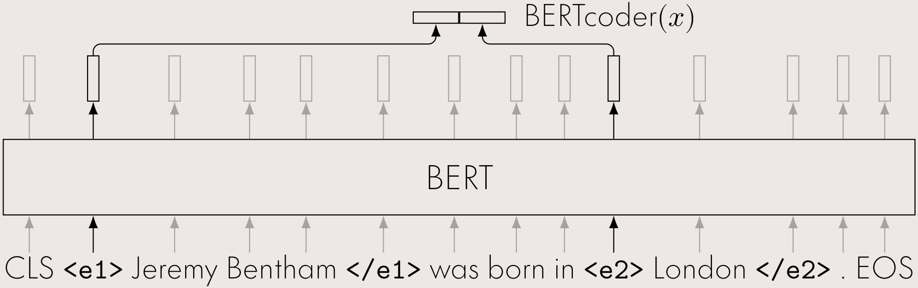


Prediction

Compare samples using:

$\mathrm{sim}(x, x') = \mathrm{sigmoid}($
$\mathrm{BERTcoder}(x)^{\mathsf{T}}\,\mathrm{BERTcoder}(x'))$

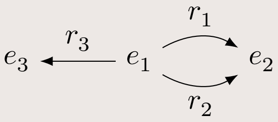
BERTcoder (linguistic)



Prediction

Compare samples using:
$\mathrm{sim}(x, x') = \mathrm{sigmoid}($
$\mathrm{BERTcoder}(x)^\mathsf{T}\,\mathrm{BERTcoder}(x'))$

Hypotheses



MTB assumes:
$r_1 = r_2$ ($\mathscr{H}_{\text{1-ADJACENCY}}$)
$r_3 \neq r_1 \wedge r_3 \neq r_2$ ($\mathscr{H}_{1 \to 1}$)

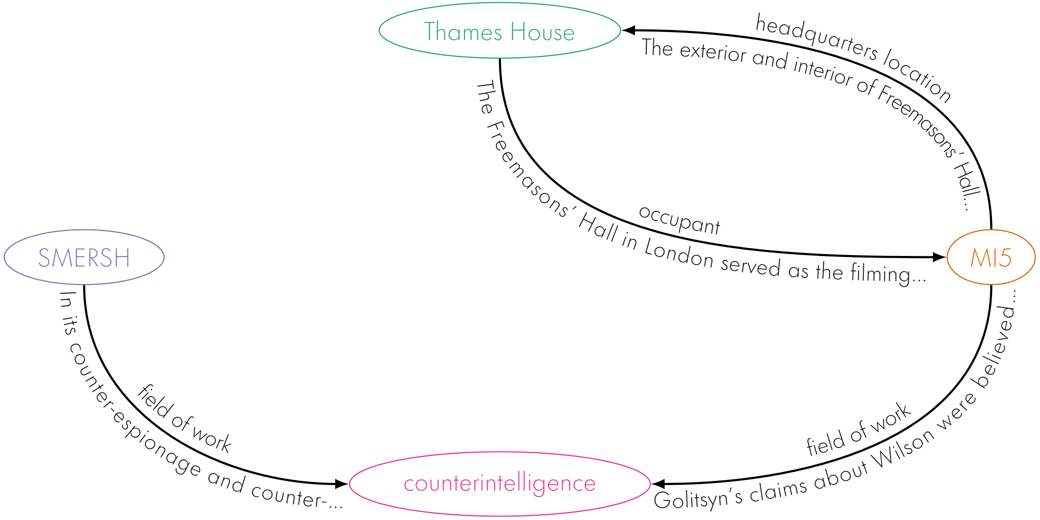
The exterior and interior of Freemasons' Hall continued to be a stand-in for Thames House$_{e_2}$, the headquarters of MI5$_{e_1}$.

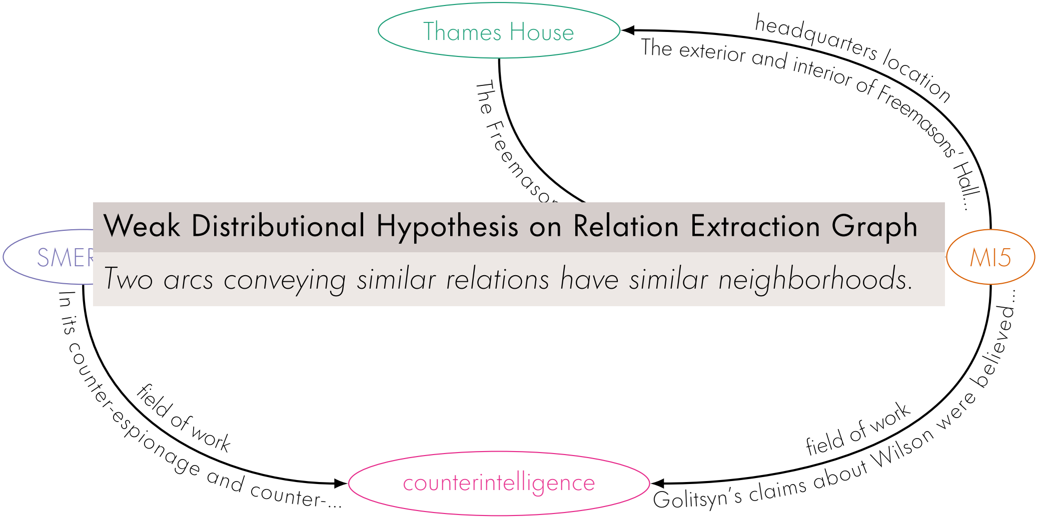
Golitsyn's claims about Wilson were believed in particular by the senior MI5$_{e_1}$ counterintelligence$_{e_2}$ officer Peter Wright.

In its counter-espionage$_{e_2}$ and counter-intelligence roles, SMERSH$_{e_1}$ appears to have been extremely successful throughout World War II.

The Freemasons' Hall in London served as the filming location for Thames House$_{e_1}$, the headquarters for MI5$_{e_2}$.

Encoding Relation Extraction as a Multigraph Problem





Weak Distributional Hypothesis on Relation Extraction Graph

Two arcs conveying similar relations have similar neighborhoods.

Proof of Principle: Counting Paths

Proposition

Given the path $e_1 \xrightarrow{r_1} e_2 \xrightarrow{r_2} e_3 \xrightarrow{r_3} e_4$, we expect $r_1 \not\perp r_2 \not\perp r_3$.

Goal

Compute the mutual information $I(r_2; r_1, r_3)$

Proposition

Given the path $e_1 \xrightarrow{r_1} e_2 \xrightarrow{r_2} e_3 \xrightarrow{r_3} e_4$, we expect $r_1 \not\perp r_2 \not\perp r_3$.

Goal

Compute the mutual information $\mathbf{I}(r_2; r_1, r_3)$

Path Counting Algorithm

We can (slowly) sample walks using power of the adjacency matrix.

1. Sample a walk by chaining neighbors
2. Reject non-path
3. Count the accepted paths weighted by importance

Path Frequency

Frequency	Relation Surface forms	Relation Identifiers
31.696‰	*country* • *diplomatic relation* • $\widehat{\textit{citizen of}}$	**P17** • **P530** • $\widehat{\textbf{P27}}$

Example of path:

Vat Phou —(Vat Phou$_{e_1}$ is a ruined Khmer...)→ Laos —(...the historical relationship between...)→ Japan ←(Sōseki$_{e_1}$ was a Japanese$_{e_2}$ novelist)— Natsume Sōseki

Path Frequency

Frequency	Relation Surface forms	Relation Identifiers
31.696‰	*country* • *diplomatic relation* • $\widehat{\textit{citizen of}}$	**P17** • **P530** • $\widehat{\textbf{P27}}$

Example of path:

Vat Phou —($\underline{\text{Vat Phou}}_{e_1}$ is a ruined Khmer...)→ Laos —(...the historical relationship between...)→ Japan ←($\underline{\text{Sōseki}}_{e_1}$ was a $\underline{\text{Japanese}}_{e_2}$ novelist)— Natsume Sōseki

Summary Statistics

$$\mathrm{I}(\mathrm{r}_2;\mathrm{r}_1,\mathrm{r}_3) = \mathop{\mathbb{E}}_{r_1,r_3}\left[\mathrm{H}_{P(\mathrm{r}_2)}(\mathrm{r}_2 \mid r_1, r_3)\right] - \mathrm{H}(\mathrm{r}_2 \mid \mathrm{r}_1, \mathrm{r}_3)$$

$$\wr\wr \qquad\qquad \wr\wr \qquad\qquad \wr\wr$$

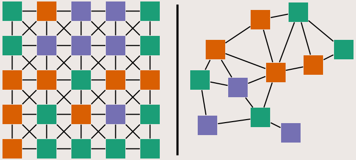
6.95 bits — 8.01 bits — 1.06 bits

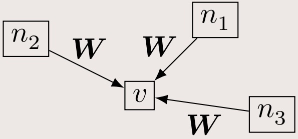
Modeling Hypothesis

$\mathscr{H}_{\text{1-NEIGHBORHOOD}}$: Two samples with the same neighborhood in the relation extraction graph convey the same relation.

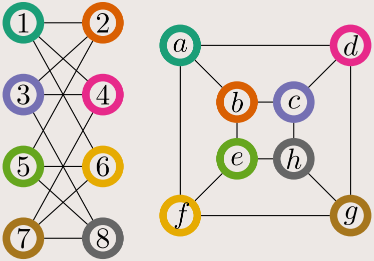
$\forall a, a' \in \mathcal{A}: \mathcal{N}(a) = \mathcal{N}(a') \implies \rho(a) = \rho(a')$

Graph Convolutional Network

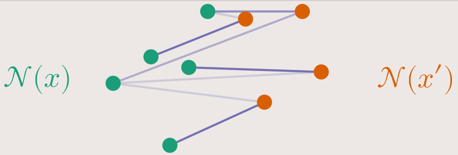




Graph Isomorphism



Earth Mover Distance



Compare Topological Features

Skip recoloring, directly compare neighborhoods in $\mathbb{R}^d$:

$S(x,k) =$ set of samples at distance k of x

$\mathfrak{S}(x,k) = \{\, \text{BERTcoder}(y) \in \mathbb{R}^d \mid y \in S(x,k) \,\}$

$$\boxed{W_1(\mathfrak{S}(x,1), \mathfrak{S}(x',1))}$$

algorithm WEISFEILER–LEMAN

Inputs: $G = (V,E)$ graph
k dimensionality

Output: χ_∞ coloring of k-tuples

$\chi_0(\boldsymbol{x}) \leftarrow \mathrm{iso}(\boldsymbol{x}) \quad \forall \boldsymbol{x} \in V^k$
for $\ell = 1, 2, \ldots$ **do**
  $\mathfrak{I}_\ell \leftarrow$ new color index
  for all $\boldsymbol{x} \in V^k$ **do**
    $c_\ell(\boldsymbol{x}) \hookleftarrow \{\!\{ \chi_{\ell-1}(\boldsymbol{y}) \mid \boldsymbol{y} \in N^k(\boldsymbol{x}) \}\!\}$
    $\chi_\ell(\boldsymbol{x}) \hookleftarrow (\chi_{\ell-1}(\boldsymbol{x}), c_\ell(\boldsymbol{x}))$ in $\mathfrak{I}_\ell$
until $\chi_\ell = \chi_{\ell-1}$
output χ_ℓ

Redefining similarity

We keep the linguistic similarity from MTB:

$$\text{sim}_{\text{ling}}(x, x') = \text{sigmoid}\left(\text{BERTcoder}(x)^{\mathsf{T}}\,\text{BERTcoder}(x')\right)$$

But also define a topological similarity:

Either using GCN:

$$\text{sim}_{\text{topo}}^{\text{GCN}}(x, x') = \text{sigmoid}\left(\text{GCN}(G)_x^{\mathsf{T}}\,\text{GCN}(G)_{x'}\right)$$

Or 1-Wasserstein:

$$\text{sim}_{\text{topo}}^{W_1}(x, x') = -W_1(\mathfrak{S}(x, 1), \mathfrak{S}(x', 1))$$

Define the topolinguistic similarity as:

$$\text{sim}_{\text{topoling}}(x, x') = \text{sim}_{\text{ling}}(x, x') + \lambda\,\text{sim}_{\text{topo}}(x, x')$$

Model	Accuracy
Pre-trained	
Linguistic (BERT)	69.46
Topological (W_1)	65.75
Topolinguistic	72.18
Fine-tuned	
MTB	78.83
MTB GCN–Chebyshev	76.10

Few-Shot Evaluation

1 query
5 candidates
Which candidate conveys the same relation as the query?
Random model score 20% accuracy.

Results: FewRel 5 way 1 shot Accuracies

Model	Accuracy
Pre-trained	
Linguistic (BERT)	69.46
Topological (W_1)	65.75
Topolinguistic	72.18
Fine-tuned	
MTB	78.83
MTB GCN–Chebyshev	76.10

Few-Shot Evaluation

1 query
5 candidates
Which candidate conveys the same relation as the query?
Random model score 20% accuracy.

Soares et al. “Matching the Blanks: Distributional Similarity for Relation Learning” ACL 2019

Results: FewRel 5 way 1 shot Accuracies

Model	Accuracy
Pre-trained	
Linguistic (BERT)	69.46
Topological (W_1)	65.75
Topolinguistic	72.18
Fine-tuned	
MTB	78.83
MTB GCN–Chebyshev	76.10

Few-Shot Evaluation

1 query
5 candidates
Which candidate conveys the same relation as the query?
Random model score 20% accuracy.

Results: FewRel 5 way 1 shot Accuracies

Model	Accuracy
Pre-trained	
Linguistic (BERT)	69.46
Topological (W_1)	65.75
Topolinguistic	72.18
Fine-tuned	
MTB	78.83
MTB GCN–Chebyshev	76.10

Few-Shot Evaluation

1 query
5 candidates
Which candidate conveys the same relation as the query?
Random model score 20% accuracy.

Model	Accuracy
Pre-trained	
Linguistic (BERT)	69.46
Topological (W_1)	65.75
Topolinguistic	72.18
Fine-tuned	
MTB	78.83
MTB GCN–Chebyshev	76.10

Few-Shot Evaluation

1 query
5 candidates
Which candidate conveys the same relation as the query?
Random model score 20% accuracy.

Soares et al. "Matching the Blanks: Distributional Similarity for Relation Learning" ACL 2019

Model	Accuracy
Pre-trained	
Linguistic (BERT)	69.46
Topological (W_1)	65.75
Topolinguistic	72.18
Fine-tuned	
MTB	78.83
MTB GCN–Chebyshev	76.10

Few-Shot Evaluation

1 query
5 candidates
Which candidate conveys the same relation as the query?
Random model score 20% accuracy.

Take-home Message

Topological information can be leverage for unsupervised relation extraction.

Contributions

- Explicitly modeled the aggregate setup for the unsupervised problem.
- Provided proof on the quality of topological information.
- Proposed an approach to exploit the mutual information between topological and linguistic features.

Several directions still need to be explored.

Use the topological features to identify the relational information in the linguistic features.

$$\mathcal{L}_{\text{LT}}(x_1, x_2, x_3) = \max\begin{pmatrix} 0,\ \zeta + 2\left(\text{sim}_{\text{ling}}(x_1, x_2) - \text{sim}_{\text{topo}}(x_1, x_2)\right)^2 \\ \qquad - \left(\text{sim}_{\text{ling}}(x_1, x_2) - \text{sim}_{\text{topo}}(x_1, x_3)\right)^2 \\ \qquad - \left(\text{sim}_{\text{ling}}(x_1, x_3) - \text{sim}_{\text{topo}}(x_1, x_2)\right)^2 \end{pmatrix}$$

Use the topological features to identify the relational information in the linguistic features.

$$\mathcal{L}_{\text{LT}}(x_1, x_2, x_3) = \max\left(\begin{aligned} 0,\ \zeta + 2&\left(\text{sim}_{\text{ling}}(x_1, x_2) - \text{sim}_{\text{topo}}(x_1, x_2)\right)^2 \\ &- \left(\text{sim}_{\text{ling}}(x_1, x_2) - \text{sim}_{\text{topo}}(x_1, x_3)\right)^2 \\ &- \left(\text{sim}_{\text{ling}}(x_1, x_3) - \text{sim}_{\text{topo}}(x_1, x_2)\right)^2 \end{aligned}\right)$$

- Idealy we want to align the two similarities.

Use the topological features to identify the relational information in the linguistic features.

$$\mathcal{L}_{\text{LT}}(x_1, x_2, x_3) = \max\begin{pmatrix} 0, \zeta + 2\left(\text{sim}_{\text{ling}}(x_1, x_2) - \text{sim}_{\text{topo}}(x_1, x_2)\right)^2 \\ - \left(\text{sim}_{\text{ling}}(x_1, x_2) - \text{sim}_{\text{topo}}(x_1, x_3)\right)^2 \\ - \left(\text{sim}_{\text{ling}}(x_1, x_3) - \text{sim}_{\text{topo}}(x_1, x_2)\right)^2 \end{pmatrix}$$

- Idealy we want to align the two similarities.
- However to stabilize the loss we need to use negative samples.

Use the topological features to identify the relational information in the linguistic features.

$$\mathcal{L}_{\mathsf{LT}}(x_1, x_2, x_3) = \max\begin{pmatrix} 0, \zeta + 2\left(\mathrm{sim}_{\mathsf{ling}}(x_1, x_2) - \mathrm{sim}_{\mathsf{topo}}(x_1, x_2)\right)^2 \\ \quad - \left(\mathrm{sim}_{\mathsf{ling}}(x_1, x_2) - \mathrm{sim}_{\mathsf{topo}}(x_1, x_3)\right)^2 \\ \quad - \left(\mathrm{sim}_{\mathsf{ling}}(x_1, x_3) - \mathrm{sim}_{\mathsf{topo}}(x_1, x_2)\right)^2 \end{pmatrix}$$

- Idealy we want to align the two similarities.
- However to stabilize the loss we need to use negative samples.
- Up to a margin ζ.

Conclusion

Contributions

Regularizing Discriminative Methods

- Trained a deep (PCNN) classifier.
- Introduced two regularizing losses:
  - A skewness loss to ensure confidence.
  - A distribution distance loss to ensure diversity.
- Improved experimental setup:
  - 2 metrics (V-measure, ARI).
  - 2 datasets (T-REXes).

Graph-based Aggregate Methods

- Explicitly modeled the aggregate setup for the unsupervised problem.
- Provided proof on the quality of topological information.
- Proposed an approach to exploit the mutual information between topological and linguistic features.

Short-term

- Replace uniform assumption with Zipf-like distribution.
- Masking neighbors to enforce an information bottleneck.
- Make soft-positives stronger in triplet loss.
- Data distribution problem of graph-based models.

Long-term

- Using language modeling for relation extraction.
- Dataset-level modeling hypotheses.
- Complex relations:
  - n-ary relations,
  - fact qualifiers.

Questions?

Supplementary Material

$\mathscr{H}_{\text{DISTANT}}$

A sentence conveys all the possible relations between all the entities it contains.

$\mathcal{D}_{\mathcal{R}} = \mathcal{D} \bowtie \mathcal{D}_{\text{KB}}$

where $\bowtie$ denotes the natural join operator:

$$\mathcal{D} \bowtie \mathcal{D}_{\text{KB}} = \{\, (s, e_1, e_2, r) \mid (s, e_1, e_2) \in \mathcal{D} \wedge (e_1, e_2, r) \in \mathcal{D}_{\text{KB}} \,\}\,.$$

Features used by Marcheggiani and rel-LDA

1. the bag of words of the infix;
2. the surface form of the entities;
3. the lemma words on the dependency path;
4. the POS of the infix words;
5. the type of the entity pair (e.g. person–location);
6. the type of the head entity (e.g. person);
7. the type of the tail entity (e.g. location);
8. the words on the dependency path between the two entities.

$$\mathrm{B}^3\,\text{precision}(g,c) = \mathop{\mathbb{E}}_{\mathrm{X},\mathrm{Y}\sim\mathcal{U}(\mathcal{D}_{\mathcal{R}})} P(g(\mathrm{X}) = g(\mathrm{Y}) \mid c(\mathrm{X}) = c(\mathrm{Y}))$$

$$\mathrm{B}^3\,\text{recall}(g,c) = \mathop{\mathbb{E}}_{\mathrm{X},\mathrm{Y}\sim\mathcal{U}(\mathcal{D}_{\mathcal{R}})} P(c(\mathrm{X}) = c(\mathrm{Y}) \mid g(\mathrm{X}) = g(\mathrm{Y}))$$

$$\mathrm{B}^3 F_1(g,c) = \frac{2}{\mathrm{B}^3\,\text{precision}(g,c)^{-1} + \mathrm{B}^3\,\text{recall}(g,c)^{-1}}$$

$$\text{homogeneity}(g, c) = 1 - \frac{\mathrm{H}\left(c(\mathrm{X}) \mid g(\mathrm{X})\right)}{\mathrm{H}\left(c(\mathrm{X})\right)}$$

$$\text{completeness}(g, c) = 1 - \frac{\mathrm{H}\left(g(\mathrm{X}) \mid c(\mathrm{X})\right)}{\mathrm{H}\left(g(\mathrm{X})\right)}$$

$$\text{V-measure}(g, c) = \frac{2}{\text{homogeneity}(g, c)^{-1} + \text{completeness}(g, c)^{-1}}$$

$$\mathsf{RI}(g, c) = \mathop{\mathbb{E}}_{\mathrm{X},\mathrm{Y}} \left[P(c(\mathrm{X}) = c(\mathrm{Y}) \Leftrightarrow g(\mathrm{X}) = g(\mathrm{Y}))\right]$$

$$\mathsf{ARI}(g, c) = \frac{\mathsf{RI}(g, c) - \mathop{\mathbb{E}}_{c \sim \mathcal{U}(\mathcal{R}^{\mathcal{D}})} [\mathsf{RI}(g, c)]}{\max_{c \in \mathcal{R}^{\mathcal{D}}} \mathsf{RI}(g, c) - \mathop{\mathbb{E}}_{c \sim \mathcal{U}(\mathcal{R}^{\mathcal{D}})} [\mathsf{RI}(g, c)]}$$

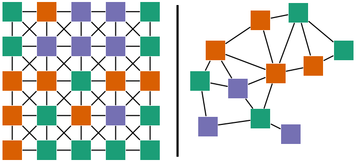
$$\pi_r = \frac{(\exp(y_r) + \mathrm{G}_r) \;/\; \tau}{\sum_{r' \in \mathcal{R}} (\exp(y_{r'}) + \mathrm{G}_{r'}) \;/\; \tau}$$

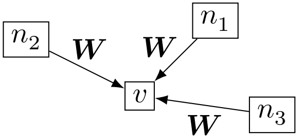
Confidence	B^3			V-measure			ARI
	F_1	Prec.	Rec.	F_1	Hom.	Comp.	
$\mathcal{L}_S$ regularization	39.4	32.2	50.7	38.3	32.2	47.2	33.8
Gumbel–Softmax	35.0	29.9	42.2	33.2	28.3	40.2	25.1

$$P(\mathrm{r} = r \mid s, \boldsymbol{e}; \boldsymbol{\theta}, \boldsymbol{\phi}) = P(\mathrm{r}_s = r \mid s; \boldsymbol{\phi}) P(\mathrm{r}_e = r \mid \boldsymbol{e}; \boldsymbol{\theta})$$

$$\mathcal{L}_{\mathsf{ALIGN}}(\boldsymbol{\theta}, \boldsymbol{\phi}) = -\log \sum_{r \in \mathcal{R}} P(r \mid s, \boldsymbol{e}; \boldsymbol{\theta}, \boldsymbol{\phi}) + \mathcal{L}_{\mathsf{D}}(\boldsymbol{\theta}) + \mathcal{L}_{\mathsf{D}}(\boldsymbol{\phi}).$$

Model	B^3			V-measure			ARI
	F_1	Prec.	Rec.	F_1	Hom.	Comp.	
$\mathcal{L}_{\mathsf{EP}} + \mathcal{L}_{\mathsf{S}} + \mathcal{L}_{\mathsf{D}}$	39.4	32.2	50.7	38.3	32.2	47.2	33.8
$\mathcal{L}_{\mathsf{ALIGN}}$ average	37.6	30.3	49.7	39.4	33.1	48.8	20.3
$\mathcal{L}_{\mathsf{ALIGN}}$ maximum	41.2	33.6	53.4	43.5	36.9	53.1	29.5
$\mathcal{L}_{\mathsf{ALIGN}}$ minimum	34.5	26.5	49.3	35.9	29.6	45.7	15.3





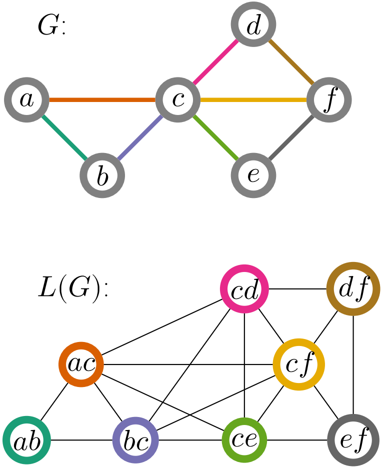
Spectral (convolution is multiplication in Fourier space)

	Graph	**Euclidean**
Laplacian	$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{M}$	∇^2
↪ Eigenfunctions	$\boldsymbol{U}$ s.t. $\boldsymbol{L} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{-1}$	$\xi \mapsto e^{2\pi i \xi x}$
Fourier transform	$\boldsymbol{U}^\top \boldsymbol{f}$	$\mathscr{F}(f) = \int_{-\infty}^{\infty} f(x) e^{2\pi i \xi x} \, \mathrm{d}x$
Convolution	$\boldsymbol{U}(\boldsymbol{U}^\top \boldsymbol{w} \boldsymbol{U}^\top \boldsymbol{f})$	$\mathscr{F}^{-1}(\mathscr{F}(w)\,\mathscr{F}(f))$

Spatial

$$\mathrm{GCN}(\boldsymbol{X};\boldsymbol{W})_v = \mathrm{ReLU}\left(\frac{1}{|N(v)|}\sum_{n_i \in N(v)} \boldsymbol{W}\boldsymbol{X}_{n_i}\right)$$

Line Graph



Emergency Otter