Extracting and Modeling Relations with Graph Convolutional Networks

Ivan Titov

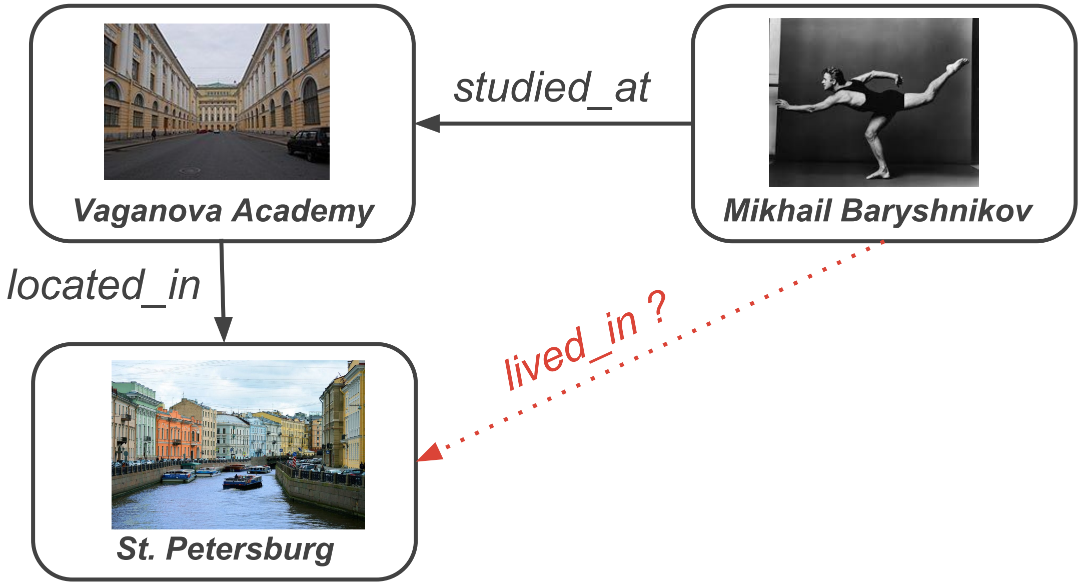
with Diego Marcheggiani, Michael Schlichtkrull, Thomas Kipf, Max Welling, Rianne van den Berg and Peter Bloem

THE UNIVERSITY OF EDINBURGH

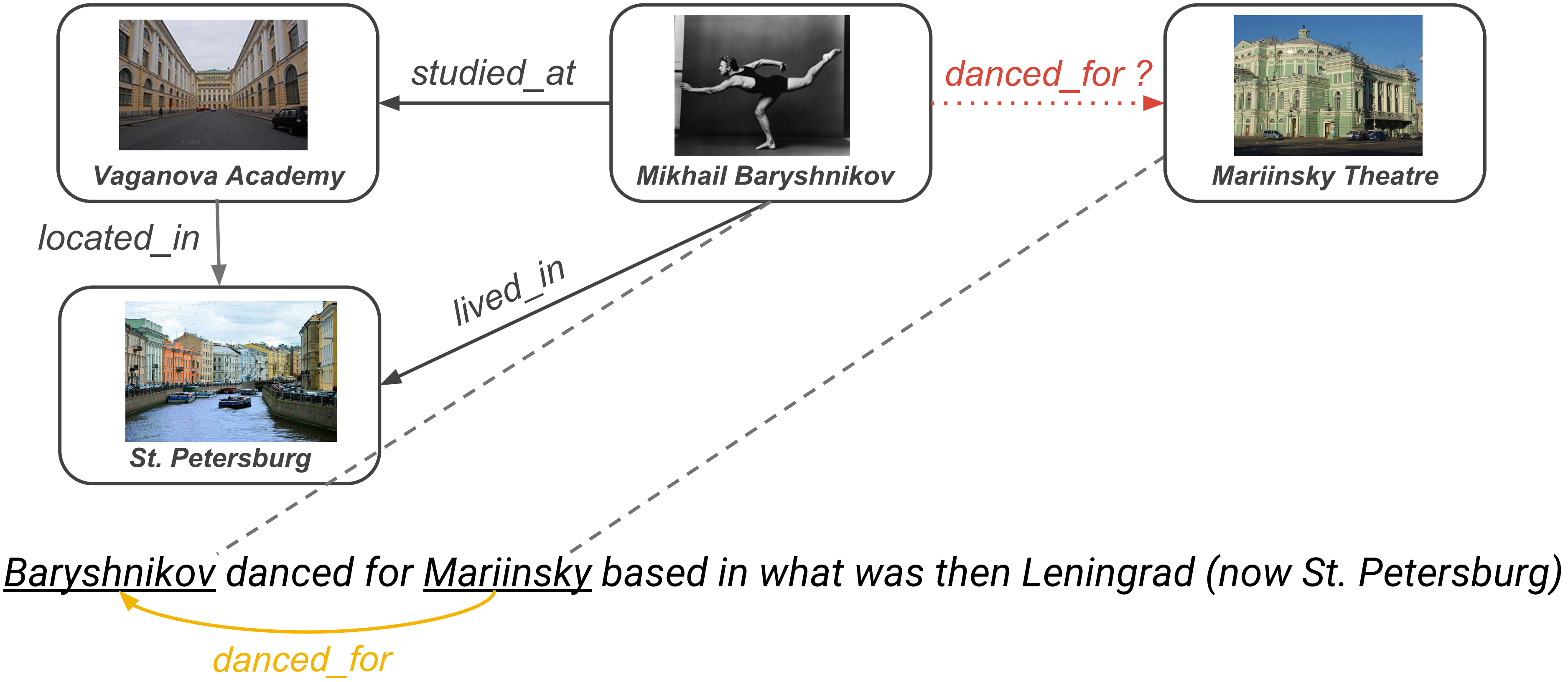
Edinburgh NLP
University of Edinburgh Natural Language Processing

UNIVERSITEIT VAN AMSTERDAM

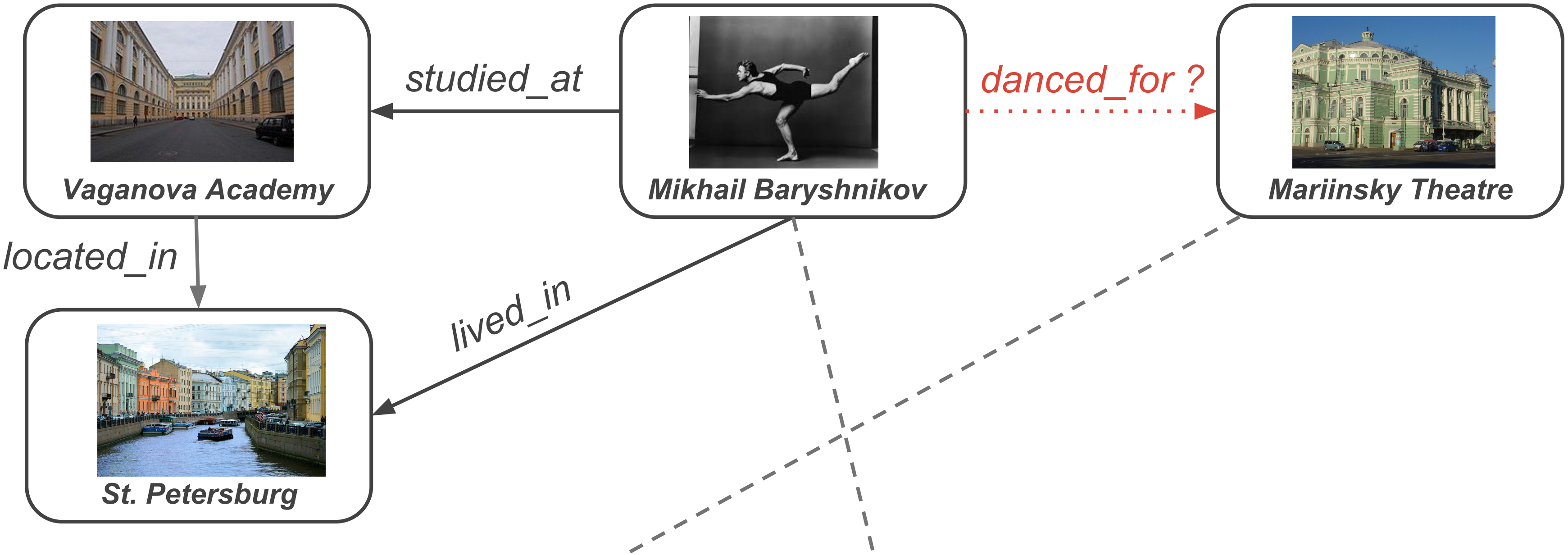
Inferring missing facts in knowledge bases: link prediction



Relation Extraction



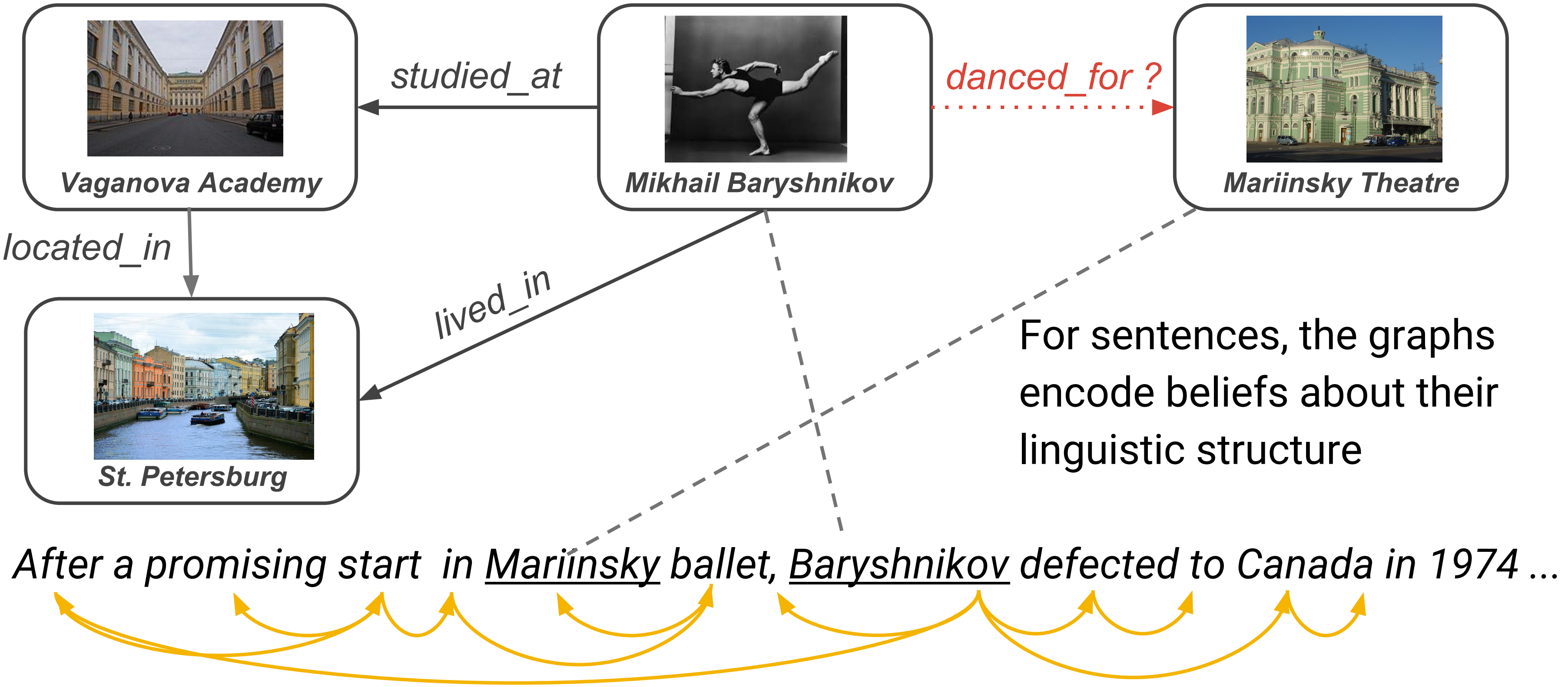
Generalization of link prediction and relation extraction



After a promising start in <u>Mariinsky</u> ballet, <u>Baryshnikov</u> defected to Canada in 1974 ...

E.g., Universal Schema (Reidel et al., 2013)

KBC: it is natural to represent both sentences and KB with graphs



For sentences, the graphs encode beliefs about their linguistic structure

How can we model (and exploit) these graphs with graph neural networks?

Outline

Graph Convolutional Networks (GCNs)

Link Prediction with Graph Neural Networks

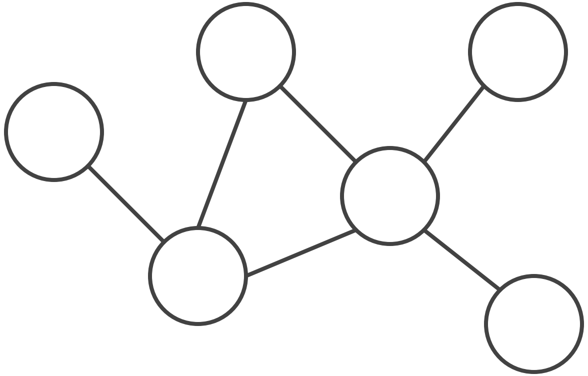
- Relational GCNs
- Denoising Graph Autoencoders for Link Prediction

Extracting Semantic Relations: Semantic Role Labeling

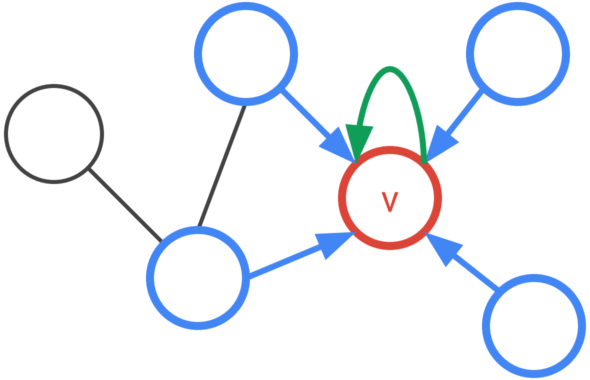
- Syntactic GCNs
- Semantic Role Labeling Model

Graph Convolutional Networks: Neural Message Passing

Graph Convolutional Networks: message passing



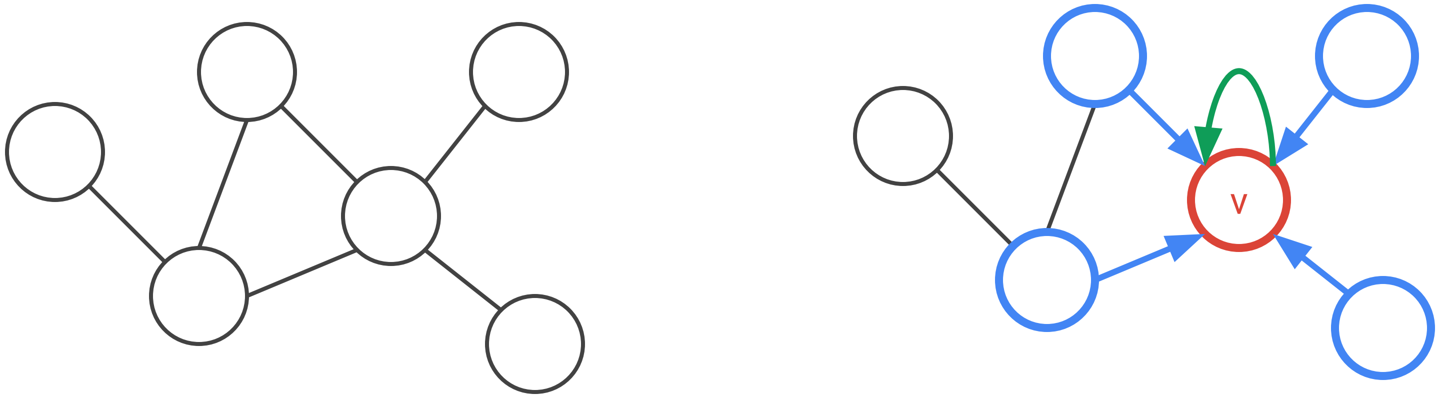
Undirected graph



Update for node v

Kipf & Welling (2017). Related ideas earlier, e.g., Scarselli et al. (2009).

Graph Convolutional Networks: message passing



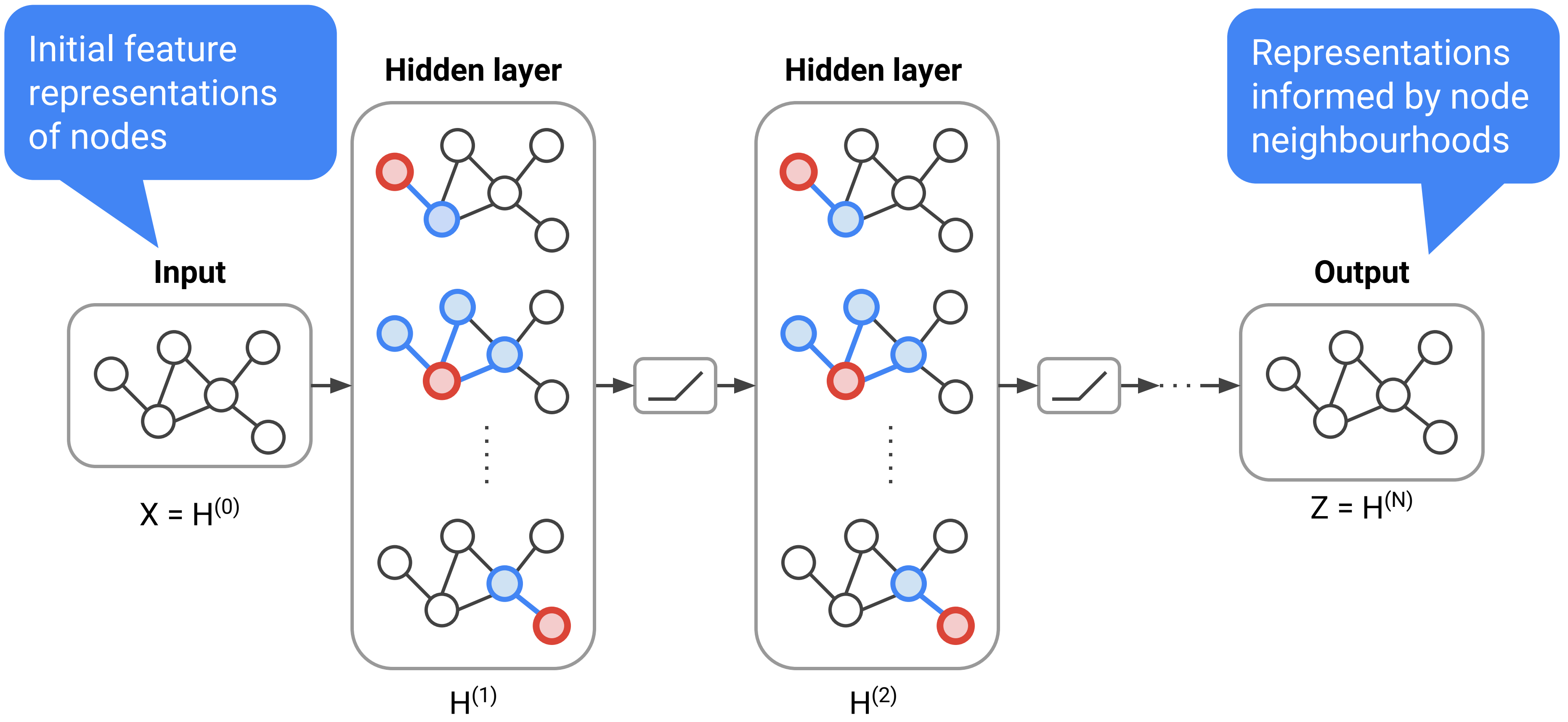
Undirected graph

Update for node v

$$\mathbf{h}_v = \mathrm{ReLU}\left(W_{\mathrm{loop}}\mathbf{h}_v + \sum_{u \in \mathcal{N}(v)} W\mathbf{h}_u\right)$$

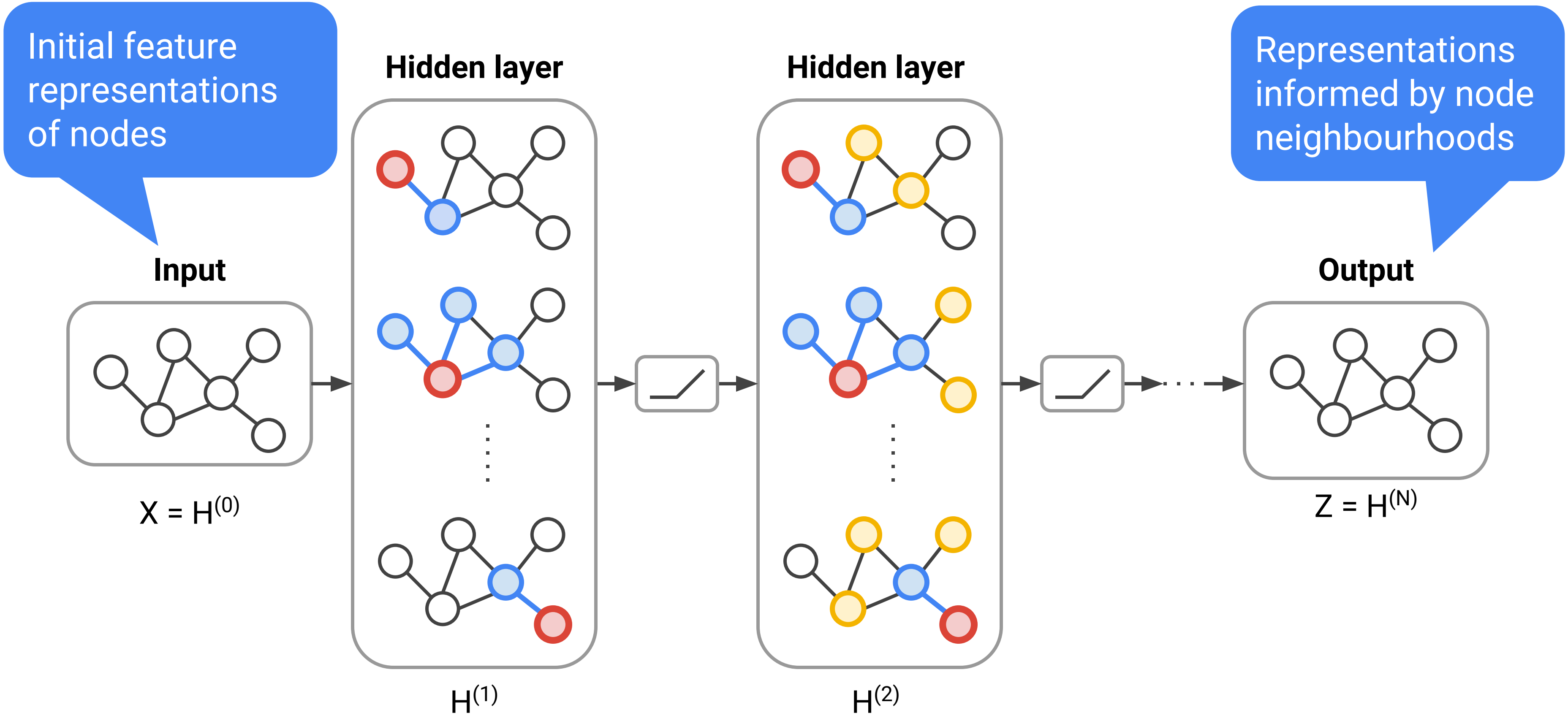
Kipf & Welling (2017). Related ideas earlier, e.g., Scarselli et al. (2009).

GCNs: multilayer convolution operation



Parallelizable computation, can be made quite efficient (e.g., Hamilton, Ying and Leskovec (2017)).

GCNs: multilayer convolution operation



Parallelizable computation, can be made quite efficient (e.g., Hamilton, Ying and Leskovec (2017)).

Graph Convolutional Networks: Previous work

Shown very effective on a range of problems - citations graphs, chemistry, ...

Mostly:

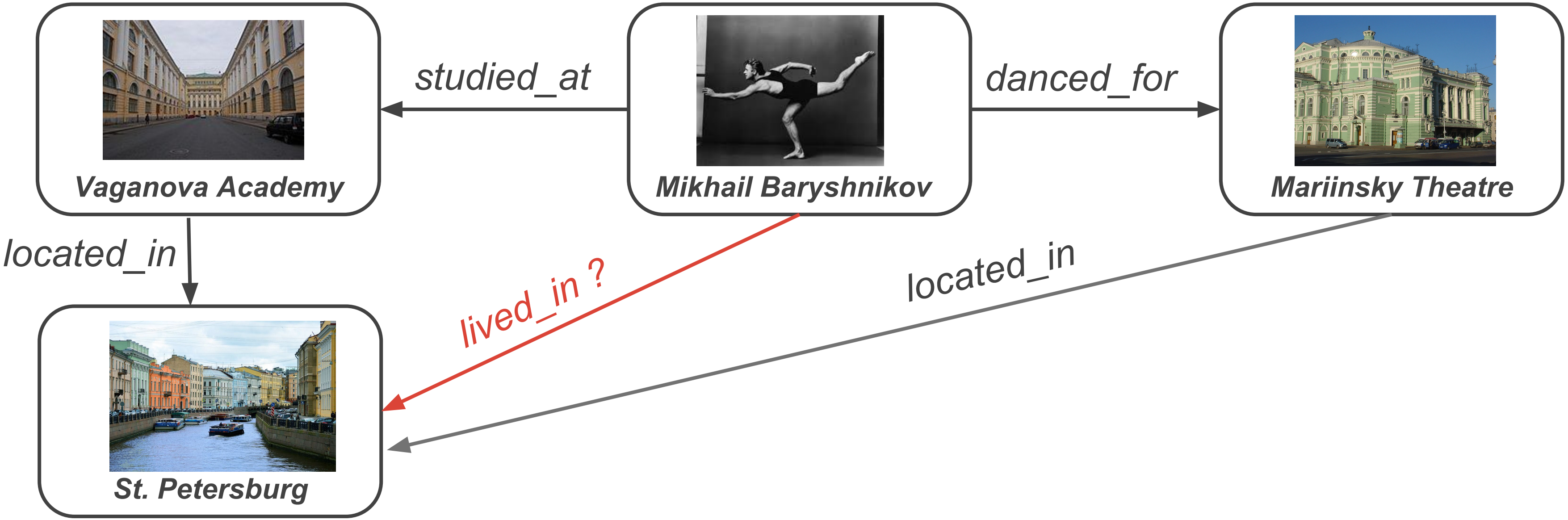
- Unlabeled and undirected graphs
- Node labeling in a single large graph (transductive setting)
- Classification of graphlets

How to apply GCNs to graphs we have in knowledge based completion / construction?

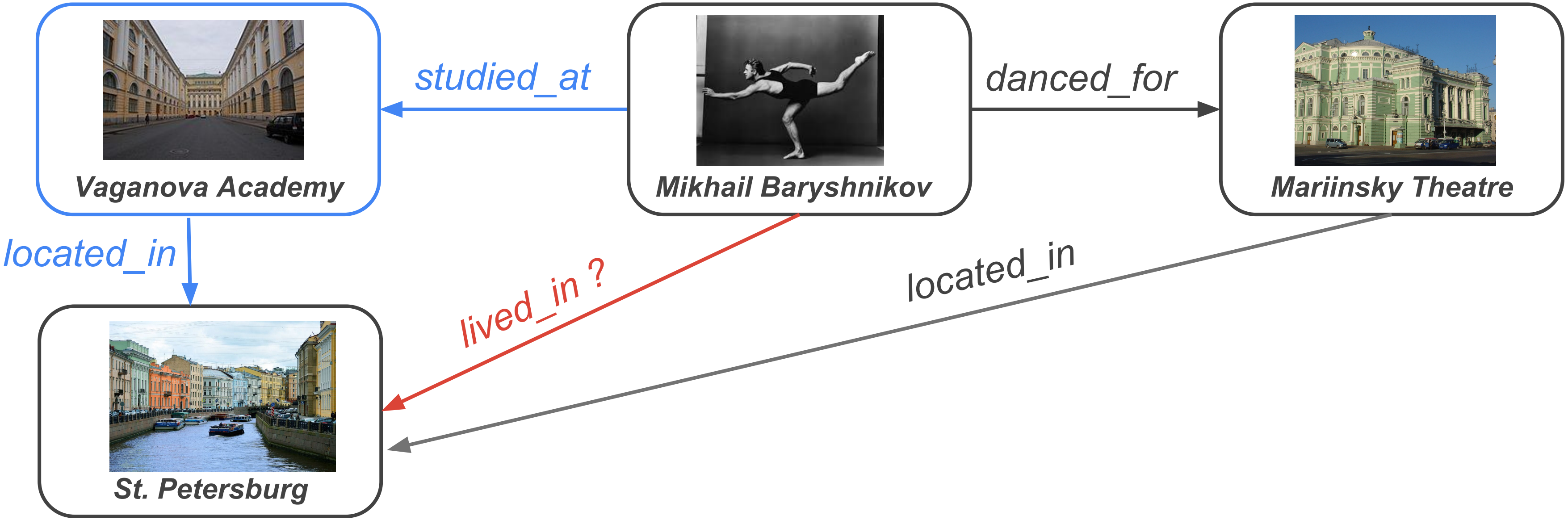
See Bronstein et al. (Signal Processing, 2017) for an overview

Link Prediction with Graph Neural Networks

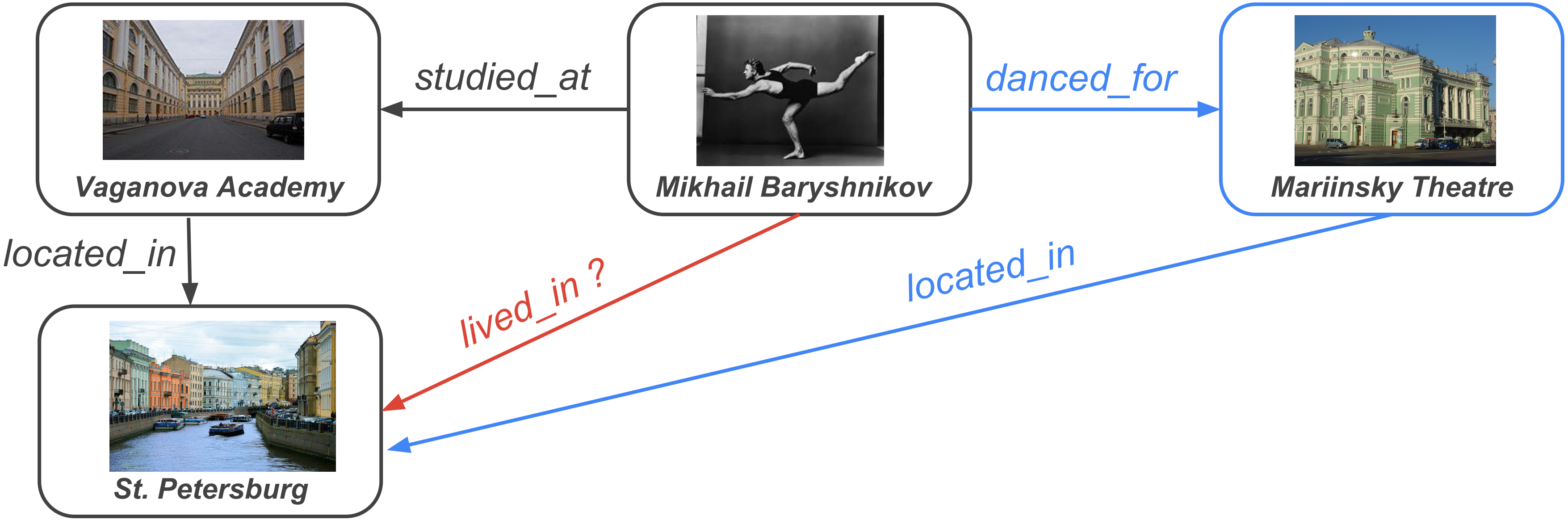
Link Prediction



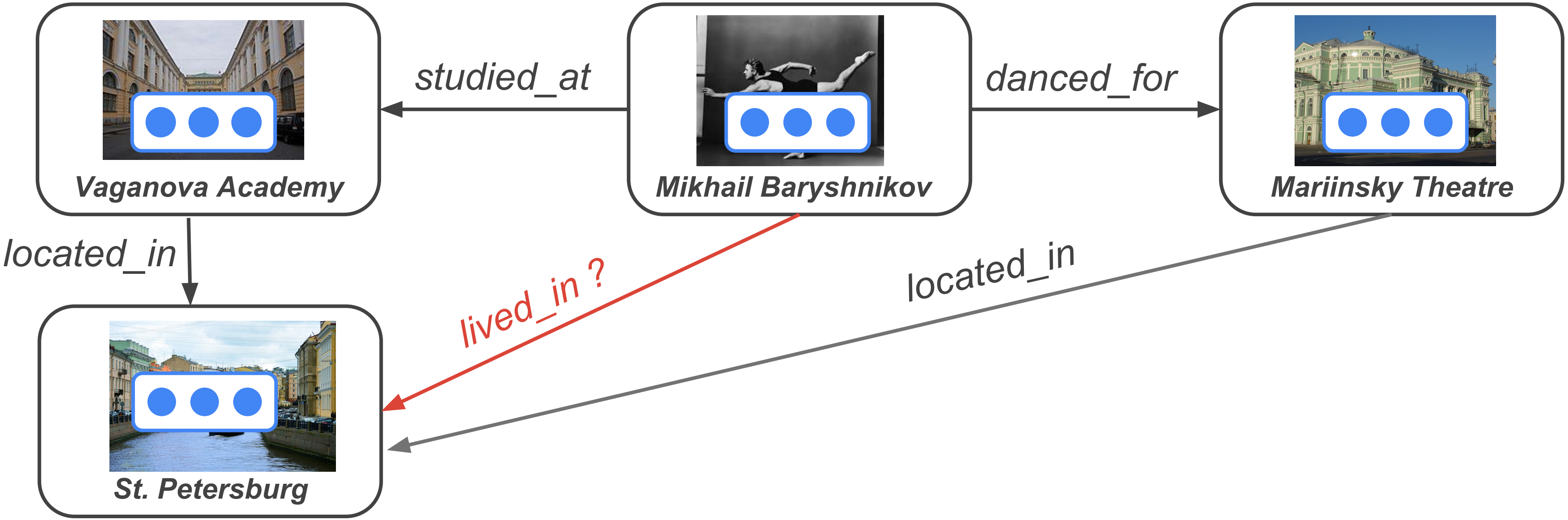
Link Prediction



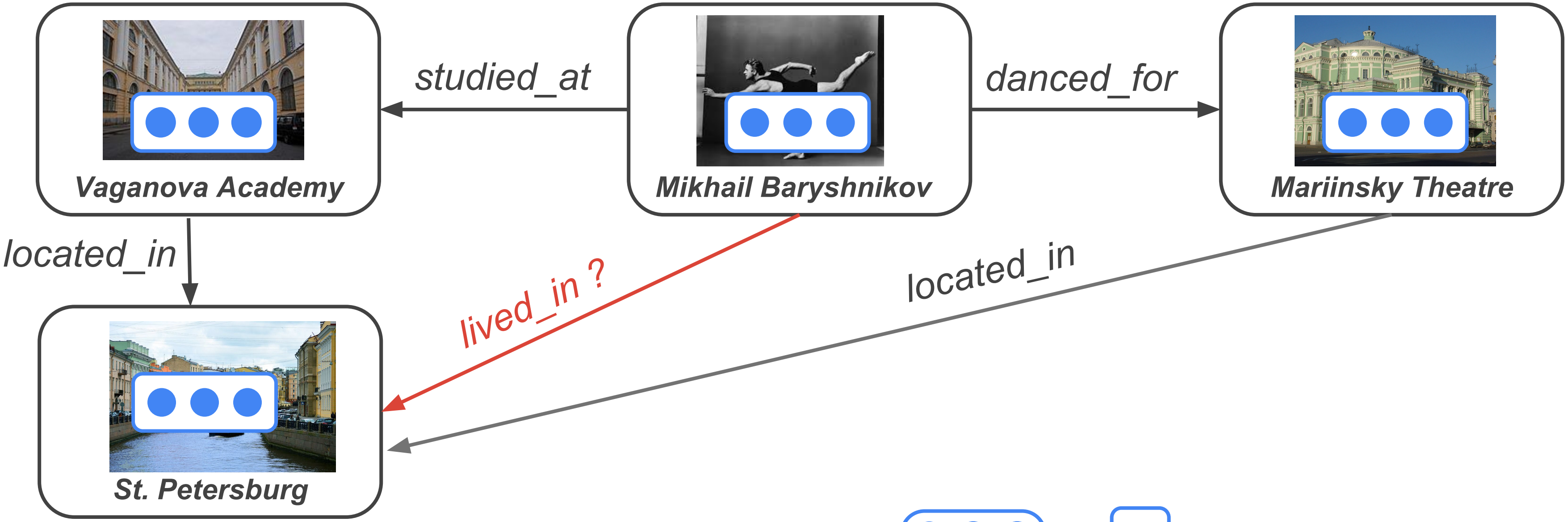
Link Prediction



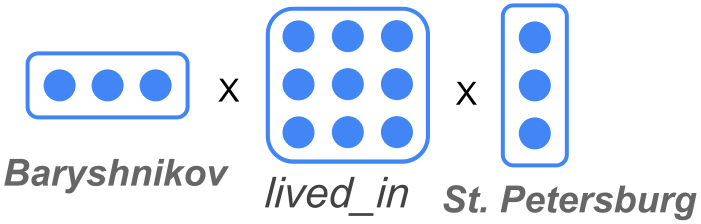
KB Factorization



KB Factorization

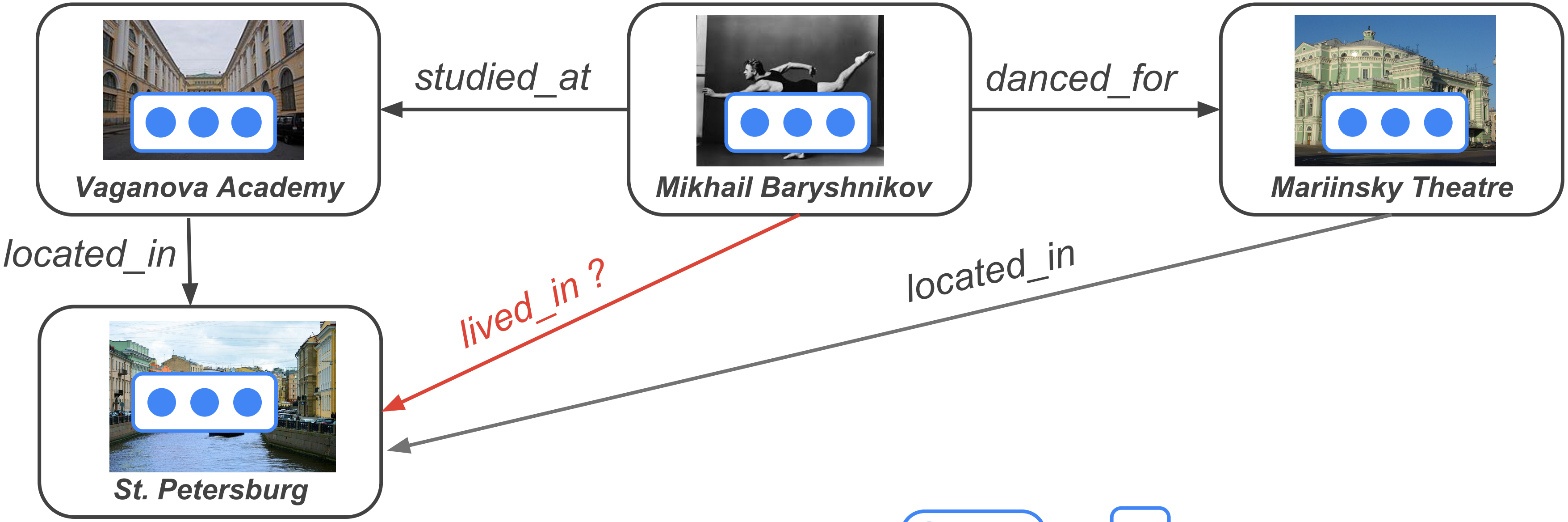


A scoring function is used to predict whether a relation holds:

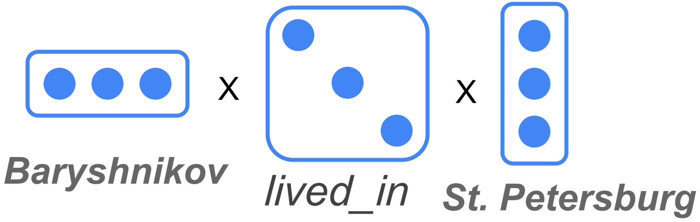


RESCAL
(Nickel et al., 2011)

KB Factorization

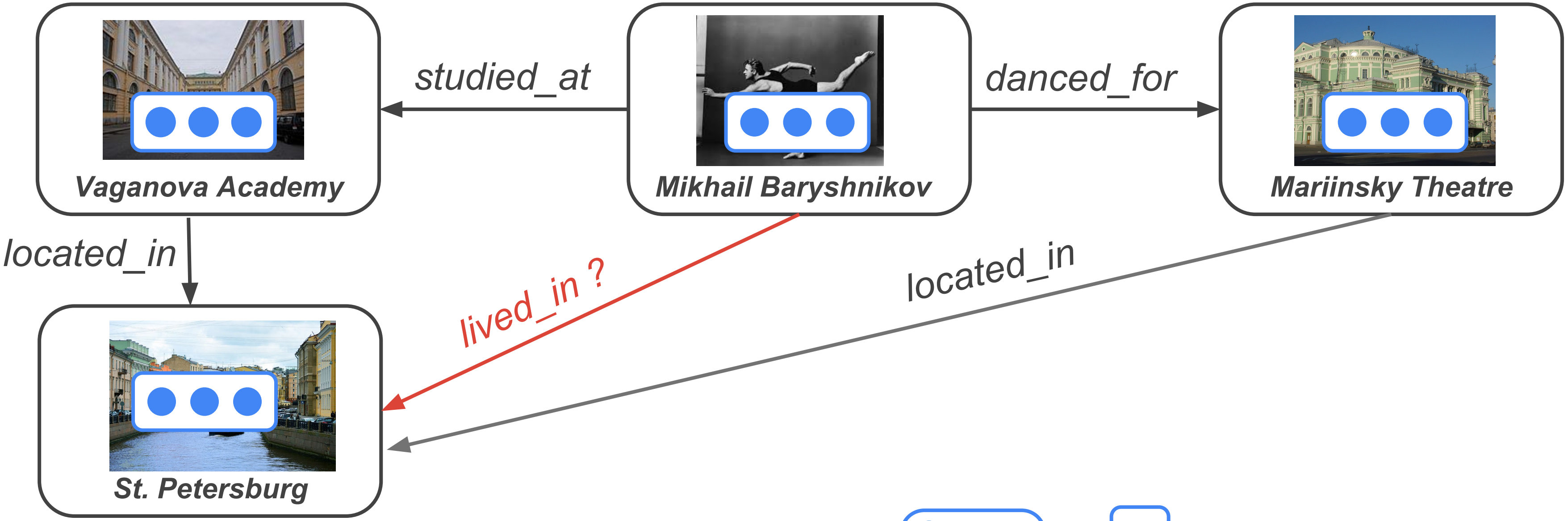


A scoring function is used to predict whether a relation holds:

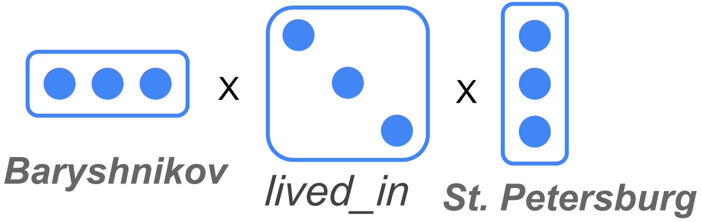


DistMult
(Yang et al., 2014)

KB Factorization



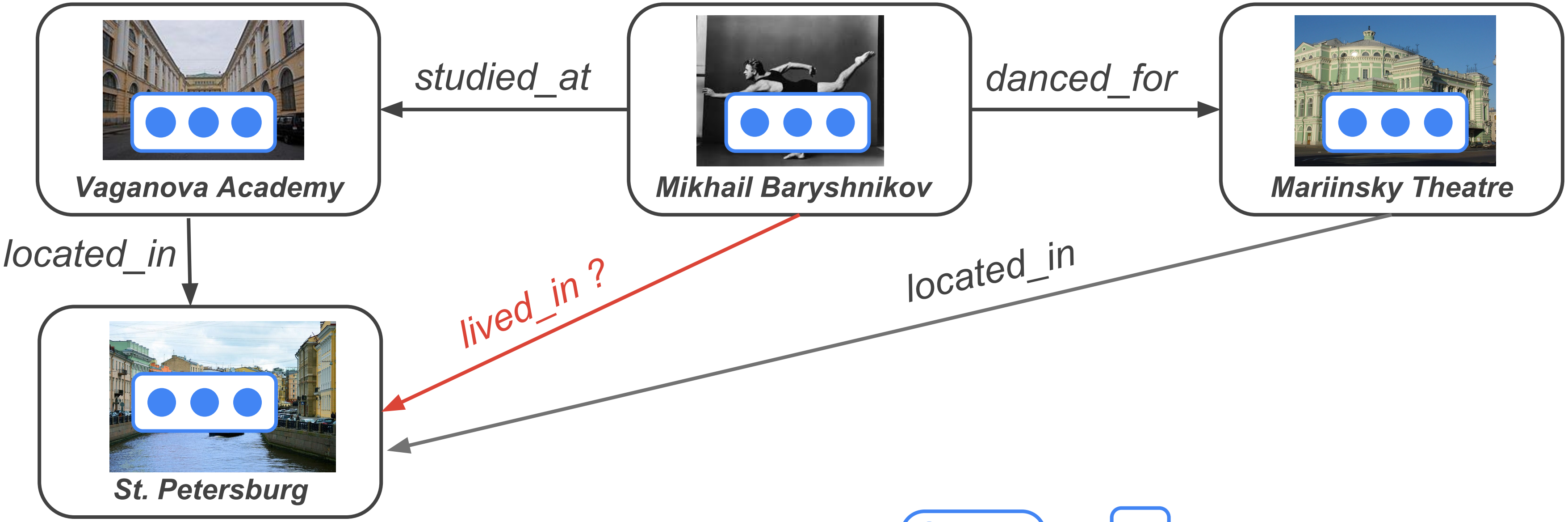
A scoring function is used to predict whether a relation holds:



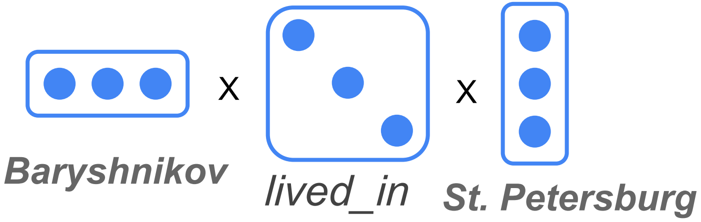
DistMult
(Yang et al., 2014)

Relies on SGD to propagate information across the graph

Relational GCNs



A scoring function is used to predict whether a relation holds:

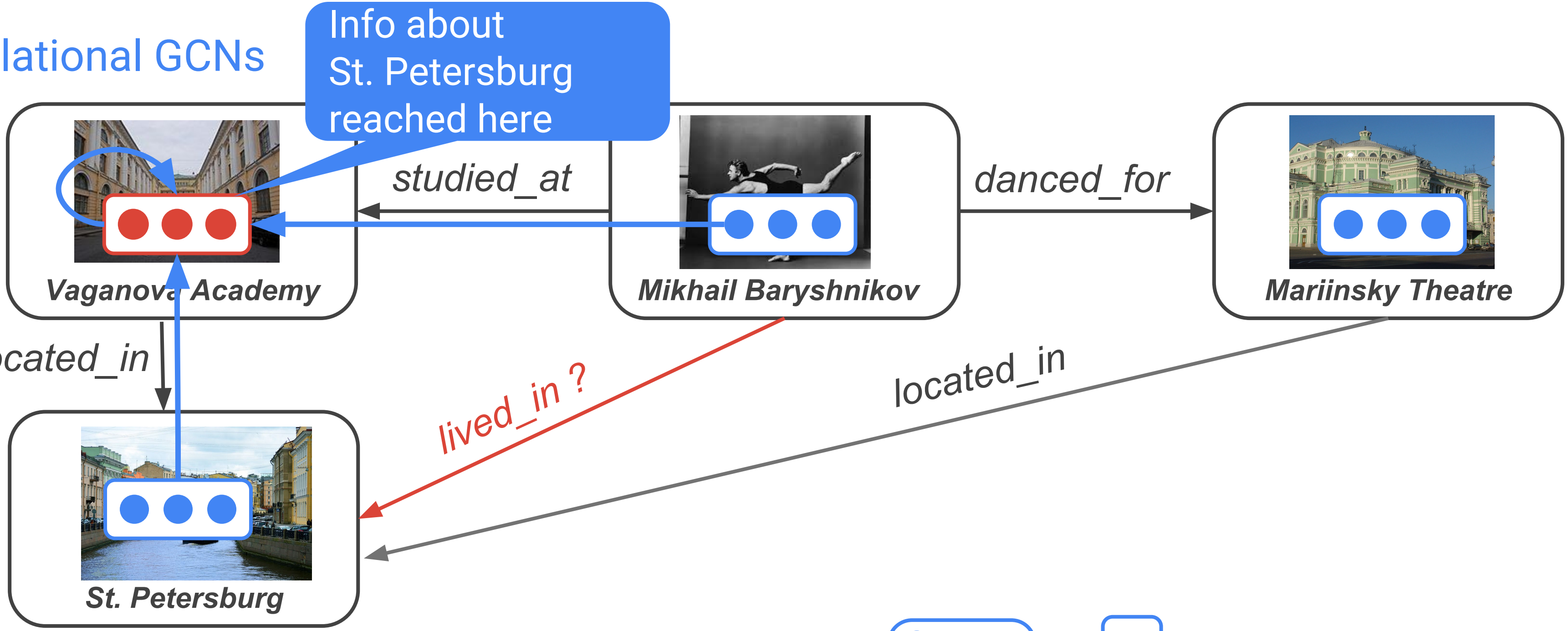


DistMult
(Yang et al., 2014)

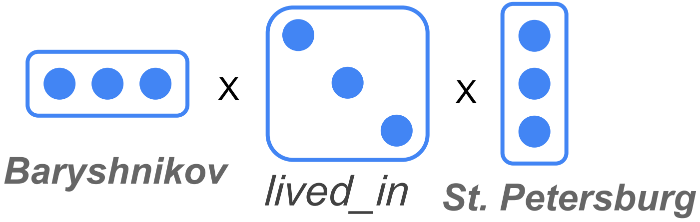
Use the same scoring function but with GCN node representations rather than parameter vectors

Schlichtkrull et al., 2017

Relational GCNs



A scoring function is used to predict whether a relation holds:

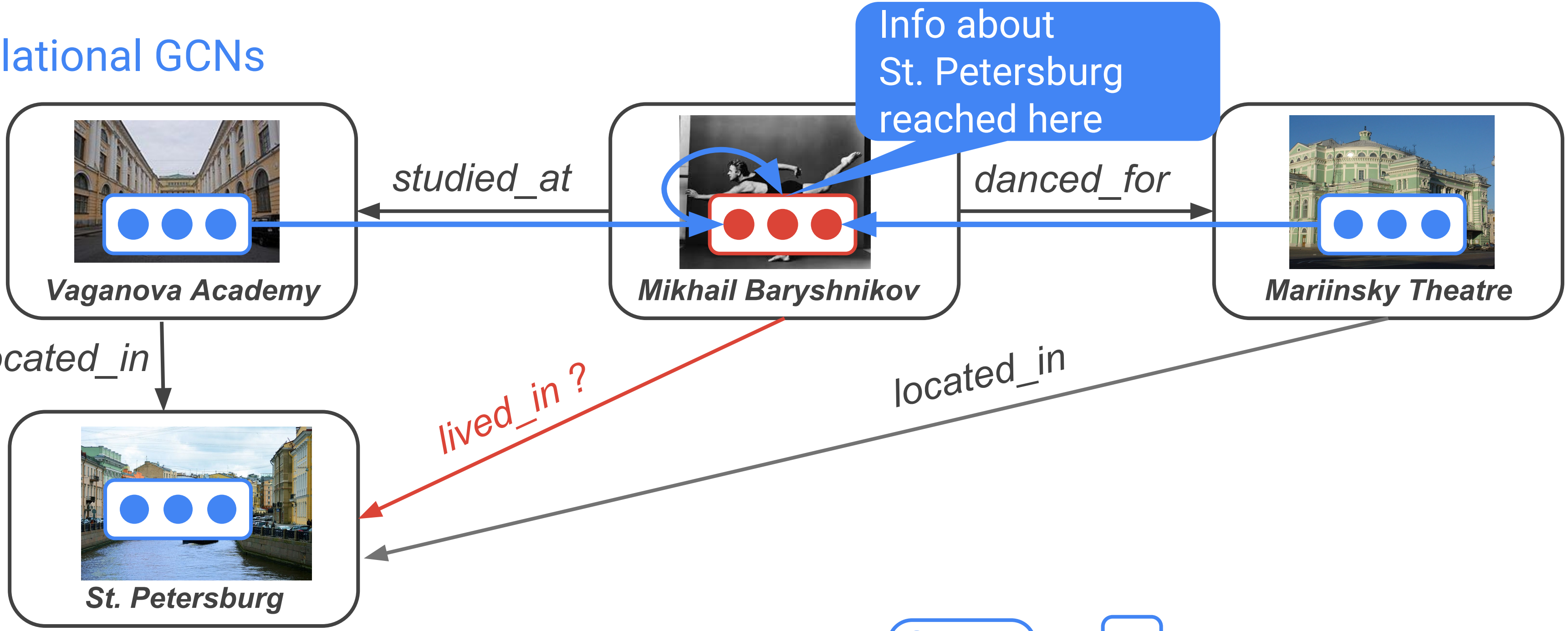


DistMult
(Yang et al., 2014)

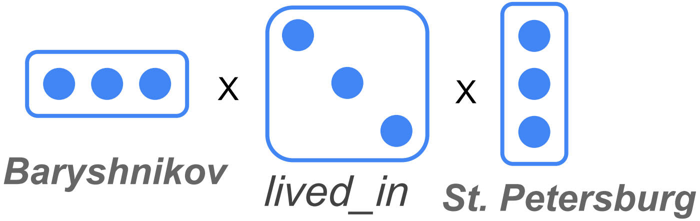
Use the same scoring function but with GCN node representations rather than parameter vectors

Schlichtkrull et al., 2017

Relational GCNs



A scoring function is used to predict whether a relation holds:

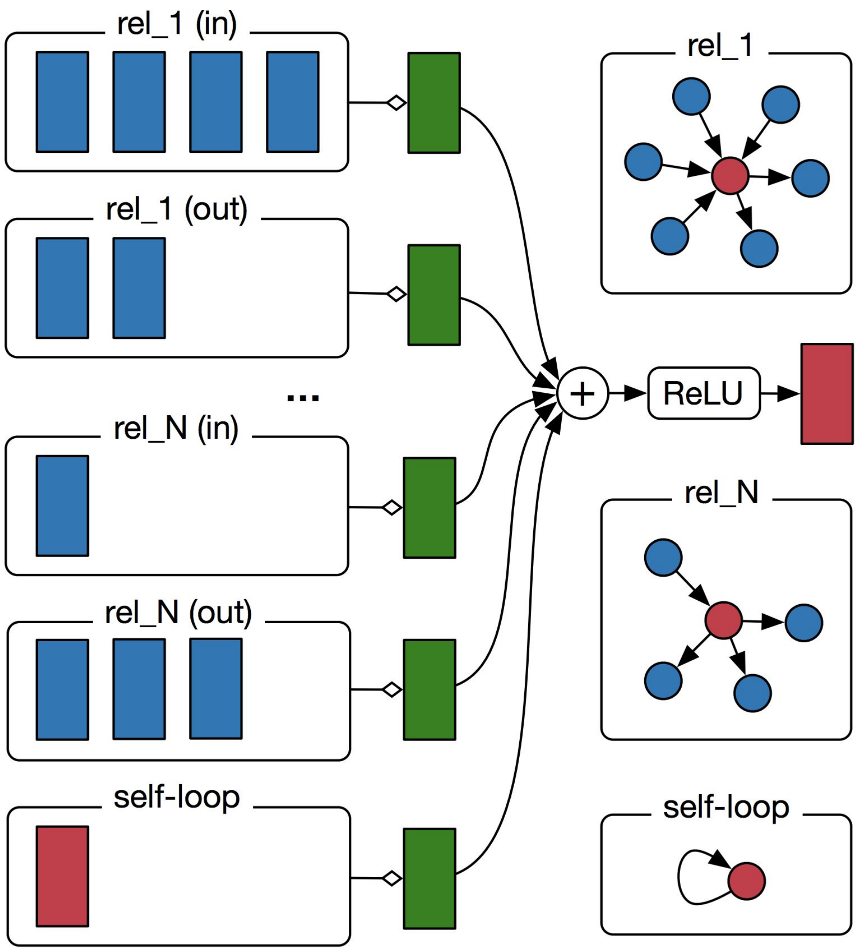


DistMult
(Yang et al., 2014)

Use the same scoring function but with GCN node representations rather than parameter vectors

Schlichtkrull et al., 2017

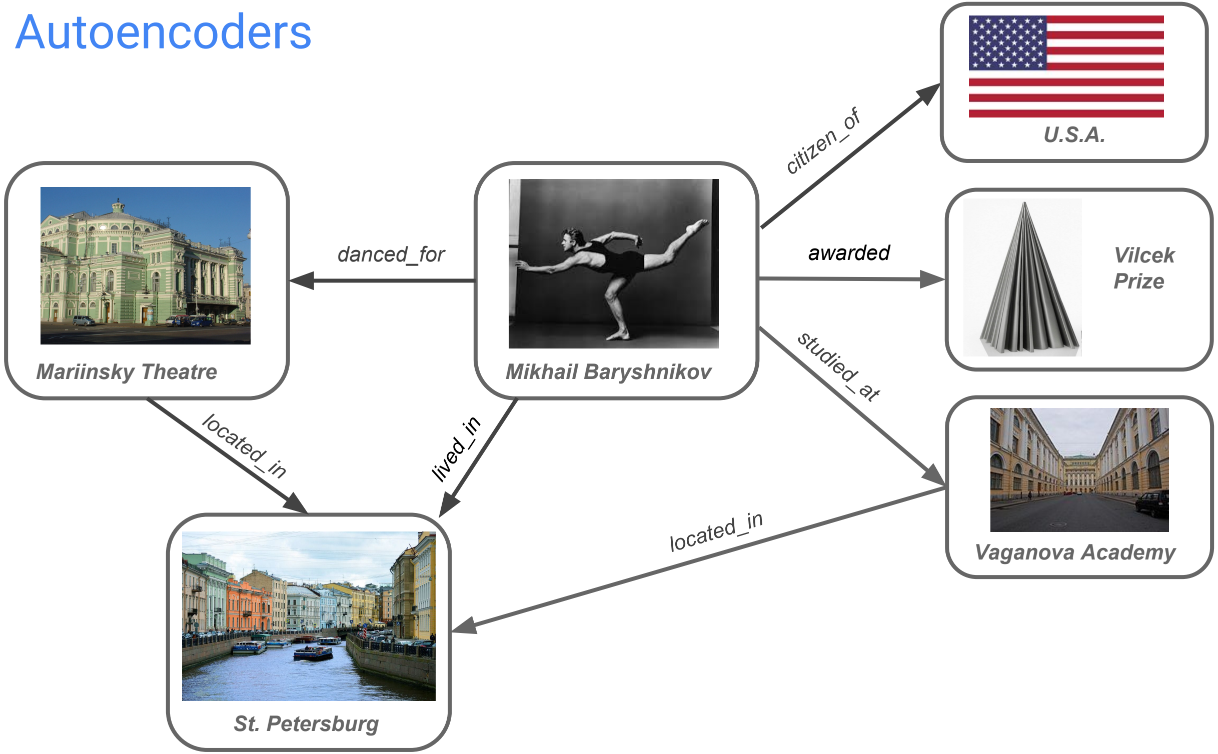
Relational GCNs



How do we train Relational GCNs?

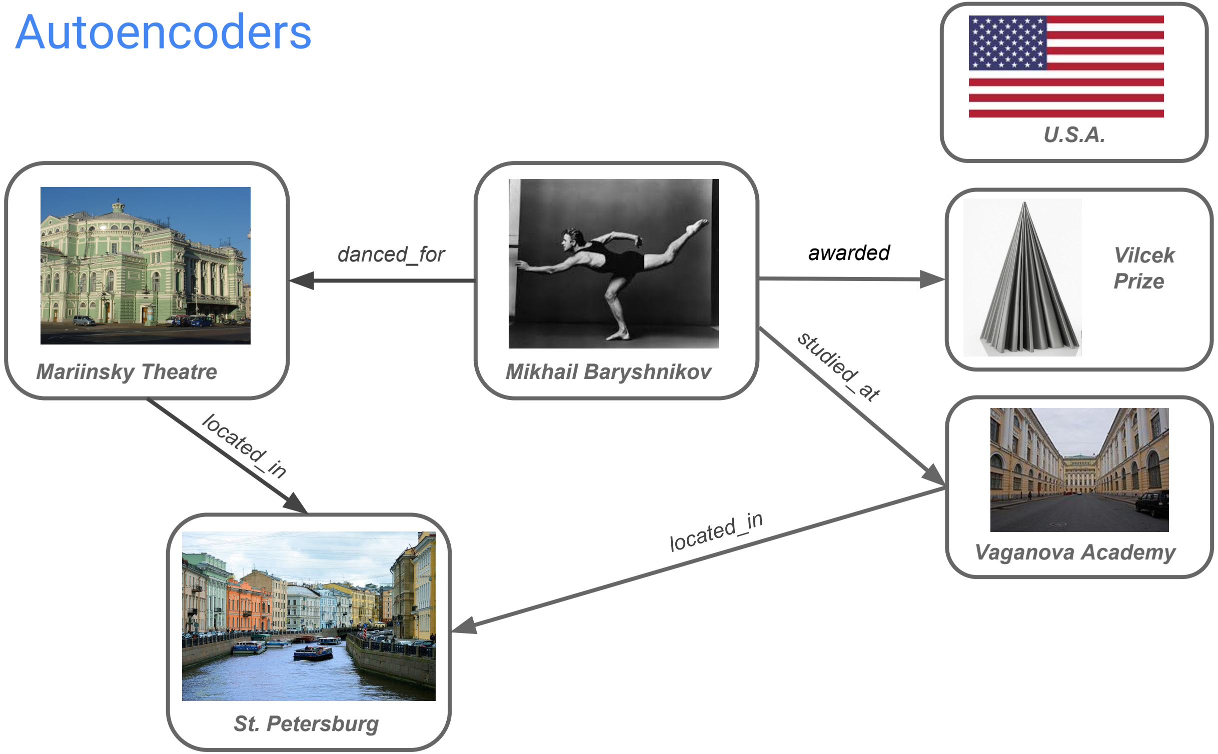
How do we compactly parameterize Relational GCNs?

GCN Denoising Autoencoders



Take the training graph

GCN Denoising Autoencoders

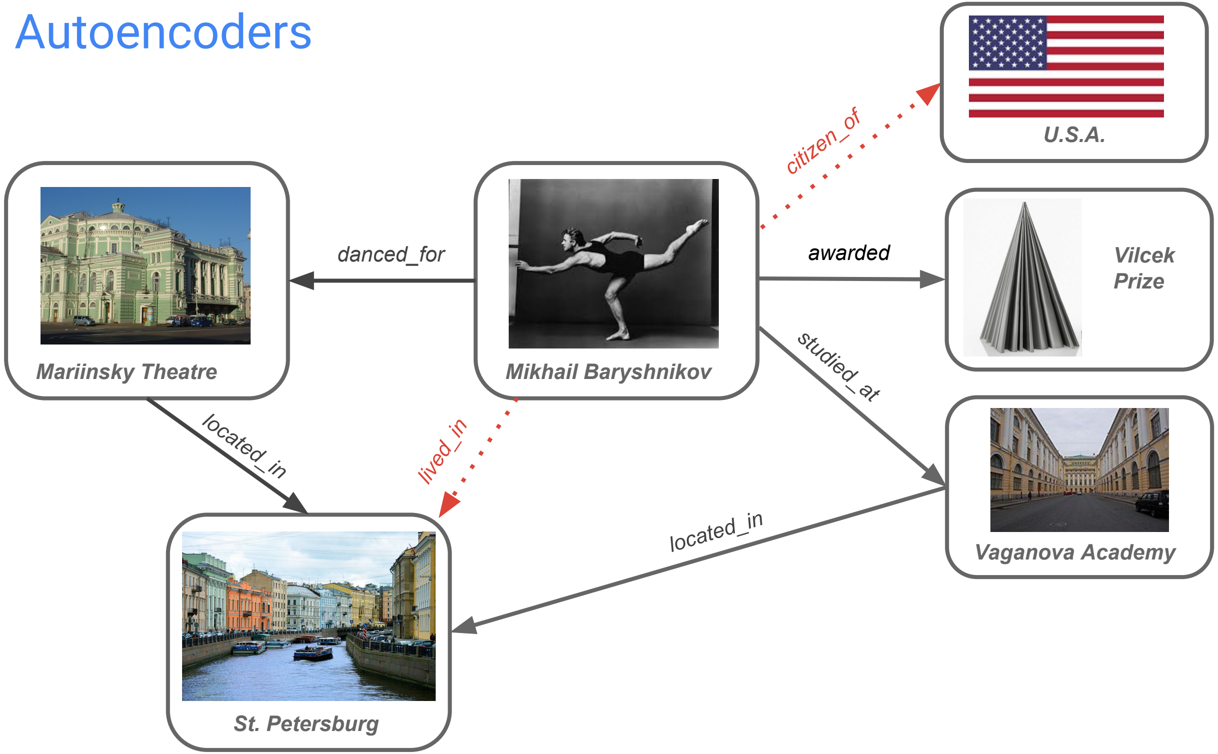


Produce a noisy version: drop some random edges

Use this graph for encoding nodes with GCNs

Schlichtkrull et al., 2017

GCN Denoising Autoencoders

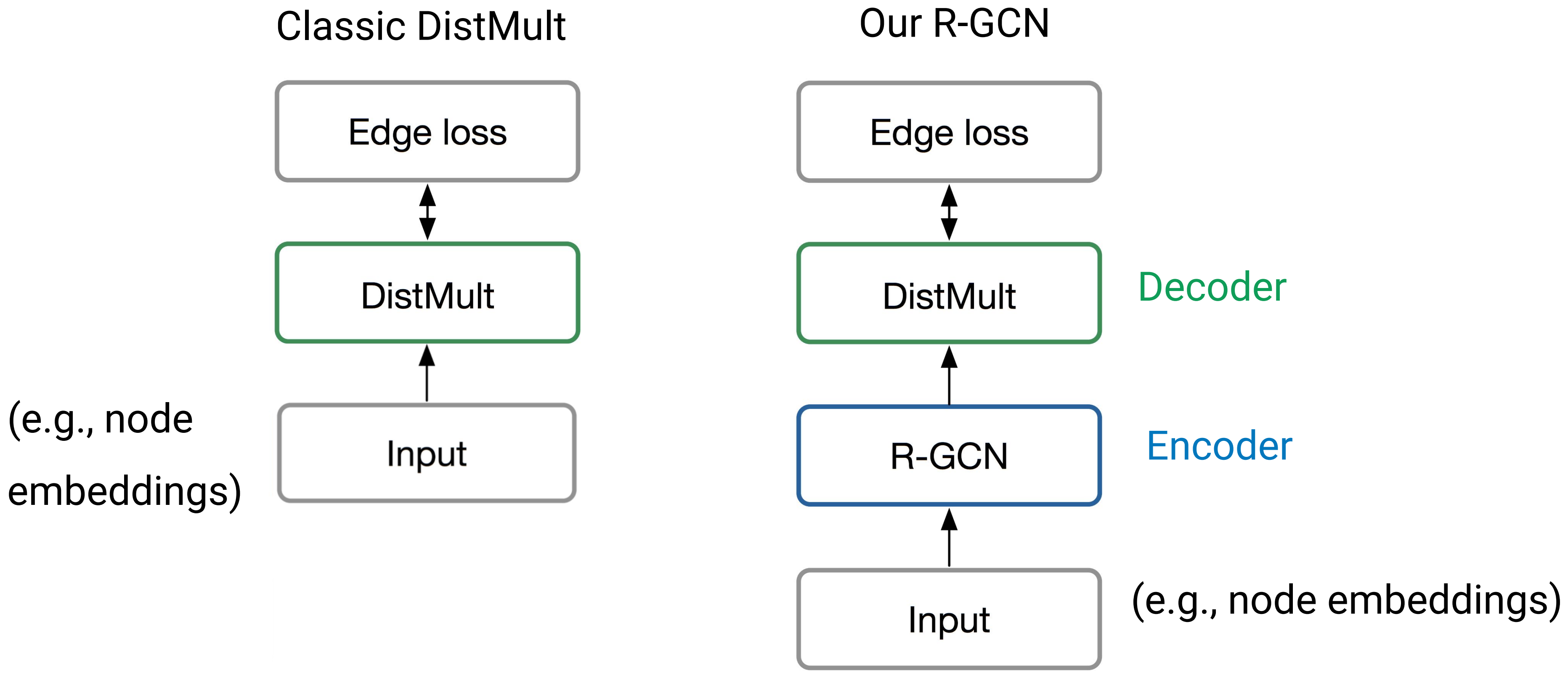


Force the model to reconstruct the original graph (including dropped edges)

(a ranking loss on edges)

Schlichtkrull et al., 2017

Training



Schlichtkrull et al., 2017

GCN Autoencoders: Denoising vs Variational

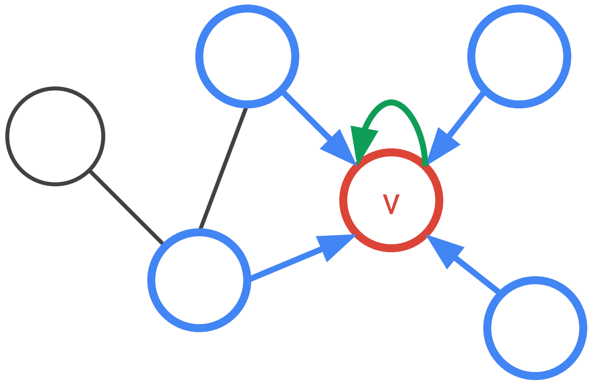
Instead of denoising AEs, we can use variational AEs to train R-GCNs

VAE R-GCN can be regarded as **an inference network** performing amortized variational inference

Intuition:

R-GCN AEs are amortized versions of factorization models

Relational GCN



$$\mathbf{h}_v = ReLU\left(\frac{1}{|\mathcal{N}(v)|}\sum_{u\in\mathcal{N}(v)} W_{r(u,v)}\mathbf{h_u}\right)$$

There are too many relations in realistic KBs, we cannot use full rank matrices W_r

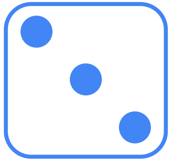
Schlichtkrull et al., 2017

Relational GCN

Naive logic:

We score with a diagonal matrix (DistMul), let’s use a diagonal one in GCN

$$W_r =$$

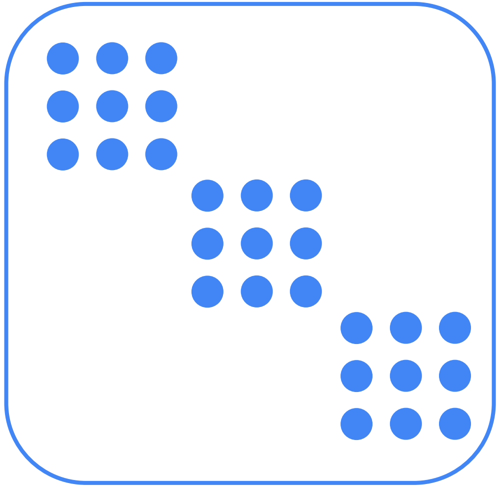


Relational GCN

Block diagonal assumption:

Latent features can be grouped into sets of tightly inter-related features, modeling dependencies across the sets is less important

$$W_r = $$



Relational GCN

Basis / Dictionary learning:

Represent every KB relation as a linear combination of basis transformations

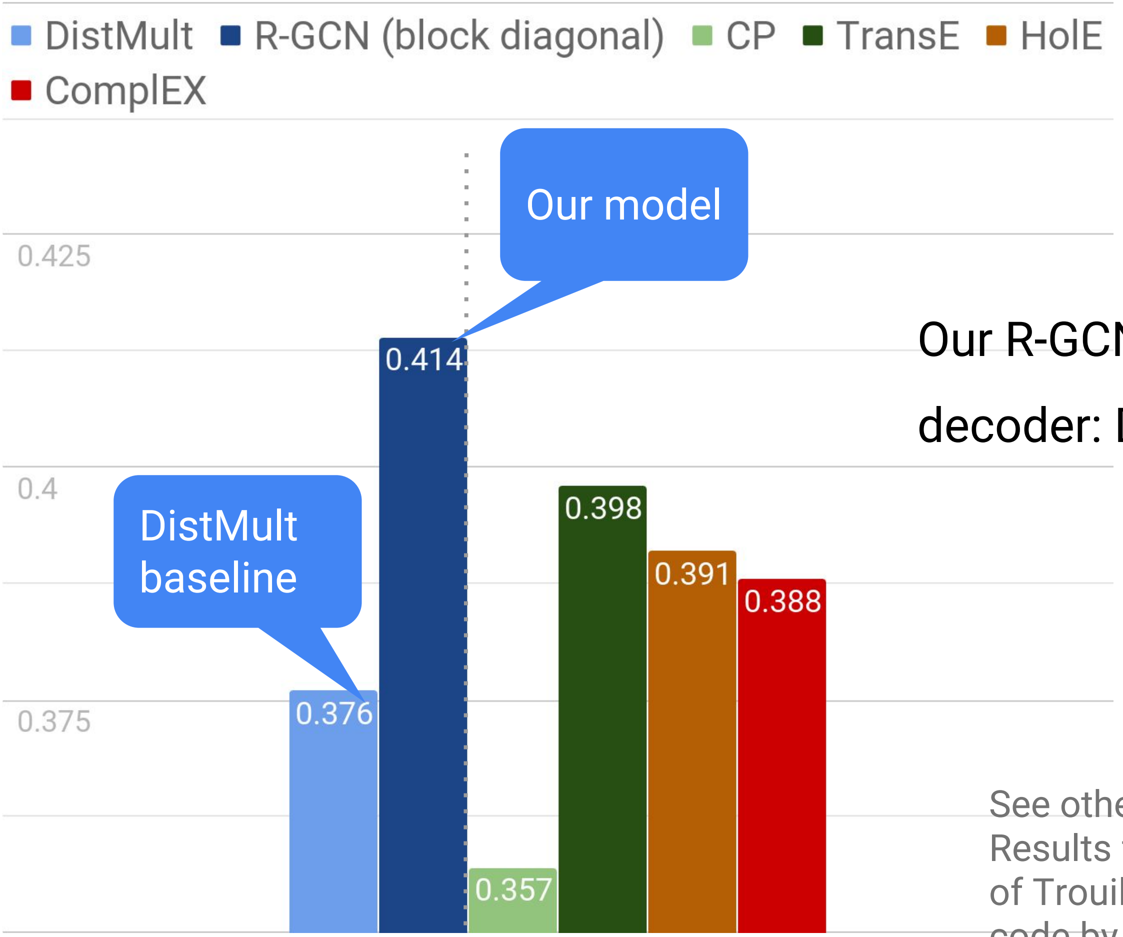
coefficients

$$W_r = \sum_{b=1}^{B} a_{rb} V_b$$

basis

transformations

Results on FB15k-237 (hits@10)



Our R-GCN relies on DistMult in the decoder: DistMult is its natural baseline

See other results and metrics in the paper. Results for ComplEX, TransE and HolE from code of Trouillon et al. (2016). Results for HolE using code by Nickel et al. (2015)

Relational GCNs

Fast and simple approach to Link Prediction

Captures multiple paths without the need to explicitly marginalize over them

Unlike factorizations, can be applied to **subgraphs unseen in training**

FUTURE WORK:

R-GCNs can be used in combination with **more powerful factorizations / decoders**

Objectives favouring **recovery of paths** rather than edges

Gates and memory may be effective

Extracting Semantic Relations

Semantic Role Labeling

Closely related to the relation extraction task

Discovering the predicate-argument structure of a sentence

Sequa makes and repairs jet engines

Semantic Role Labeling

Closely related to the relation extraction task

Discovering the predicate-argument structure of a sentence

- Discover predicates

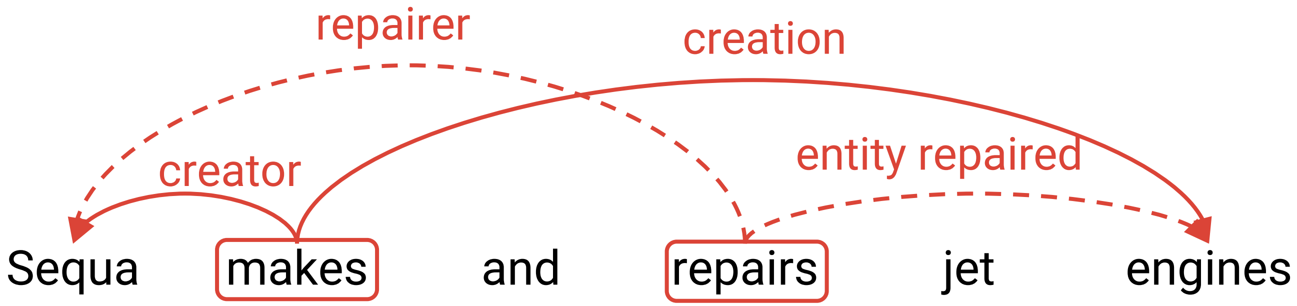
Sequa $\boxed{\text{makes}}$ and $\boxed{\text{repairs}}$ jet engines

Semantic Role Labeling

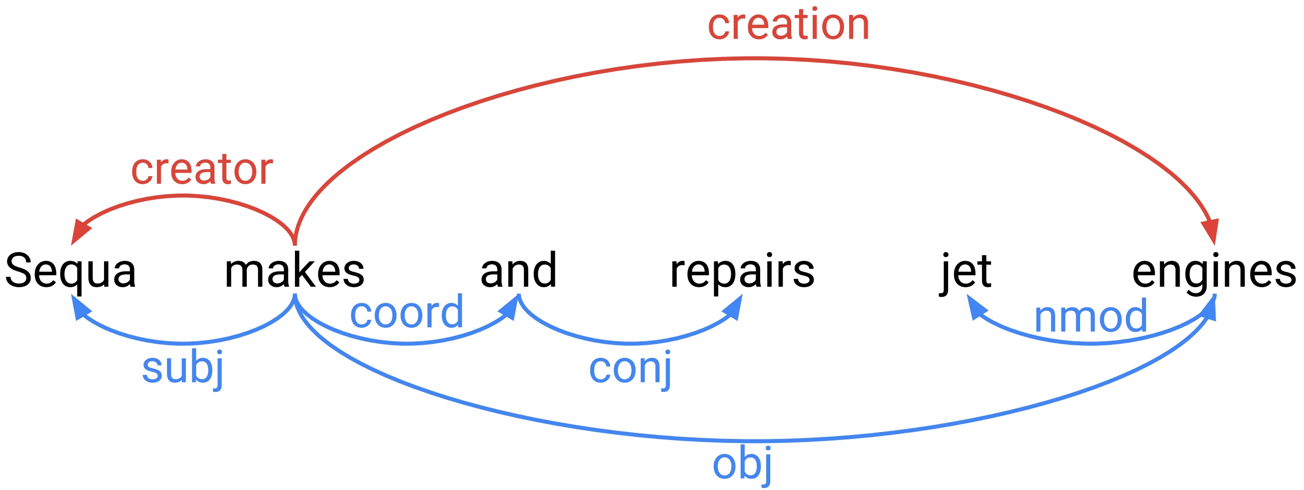
Closely related to the relation extraction task

Discovering the predicate-argument structure of a sentence

- Discover predicates
- Identify arguments and label them with their semantic roles

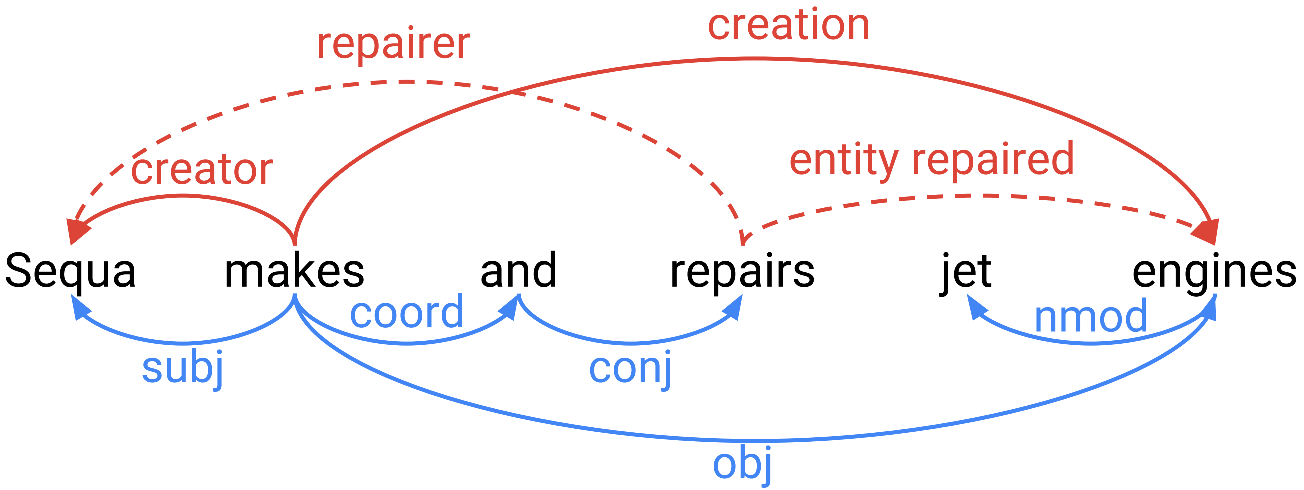


Syntax/semantics interaction



Some syntactic dependencies are **mirrored** in the semantic graph

Syntax/semantics interaction

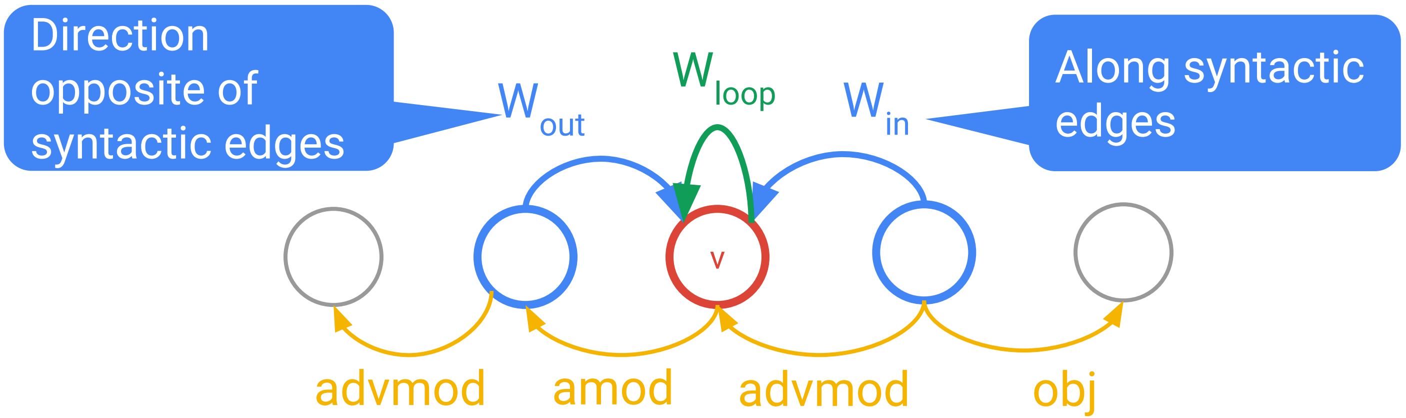


Some syntactic dependencies are **mirrored** in the semantic graph

… **but not all of them** – the syntax-semantics interface is far from trivial

GCNs provide a flexible framework for capturing interactions between the graphs

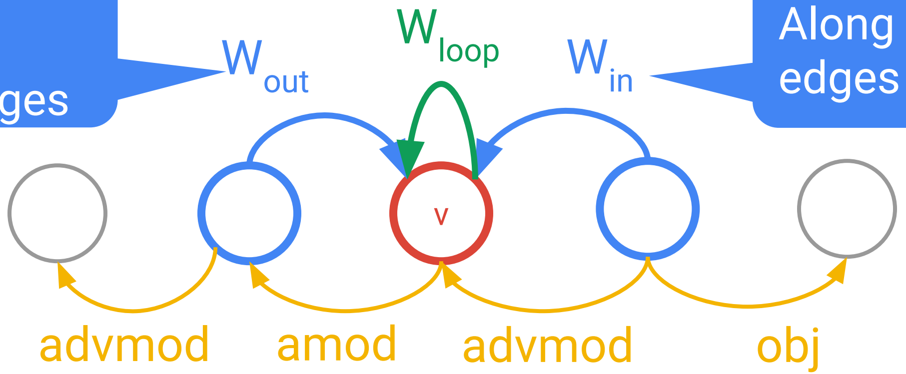
Syntactic GCNs: directionality and labels



Syntactic GCNs: directionality and labels

Direction opposite of syntactic edges

Along syntactic edges



Weight matrix for each direction: W_{out}, W_{in}, W_{loop}

Bias for each label + direction, e.g. $\mathbf{b}_{\text{in-subj}}$

$$\mathbf{h}_v = \mathrm{ReLU}\Big(\sum_{u \in \mathcal{N}(v)} W_{\mathrm{dir}(u,v)}\, \mathbf{h}_u + \mathbf{b}_{\mathrm{lab}(u,v)} \Big)$$

Syntactic GCNs: edge-wise gating

We use parsers to predict syntax

Not all edges are equally informative for the downstream task or reliable

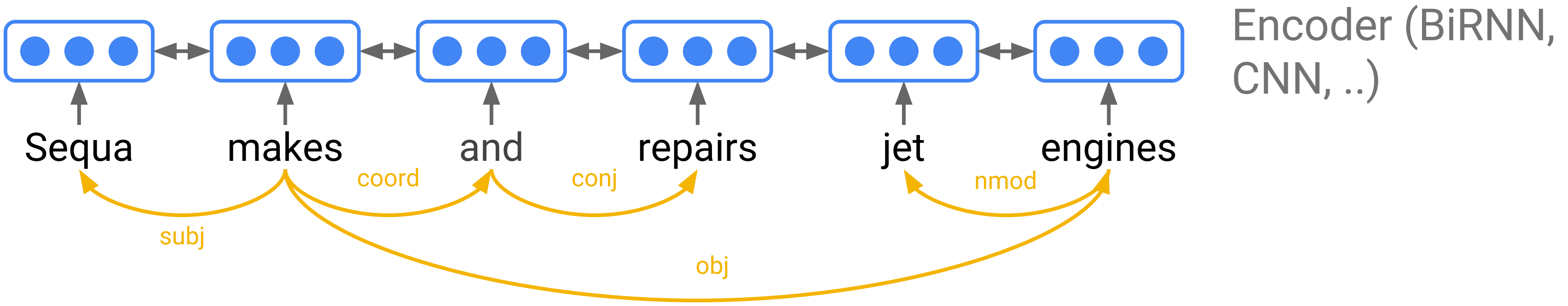
$$g_{u,v} = \sigma\left(\mathbf{h}_u \cdot \hat{\mathbf{w}}_{\mathrm{dir}(u,v)} + \hat{b}_{\mathrm{lab}(u,v)}\right)$$

$$\mathbf{h}_v = \mathrm{ReLU}\left(\sum_{u \in \mathcal{N}(v)} g_{u,v}\left(W_{\mathrm{dir}(u,v)}\,\mathbf{h}_u + \mathbf{b}_{\mathrm{lab}(u,v)}\right)\right)$$

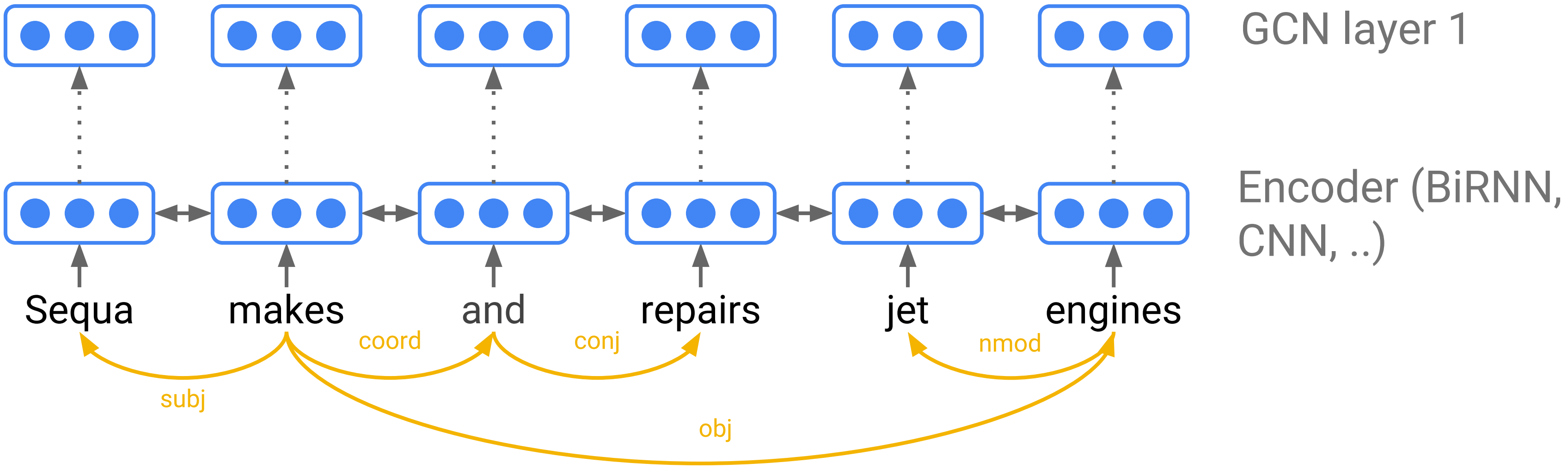
The gate weights the message

Marcheggiani et al., EMNLP 2017

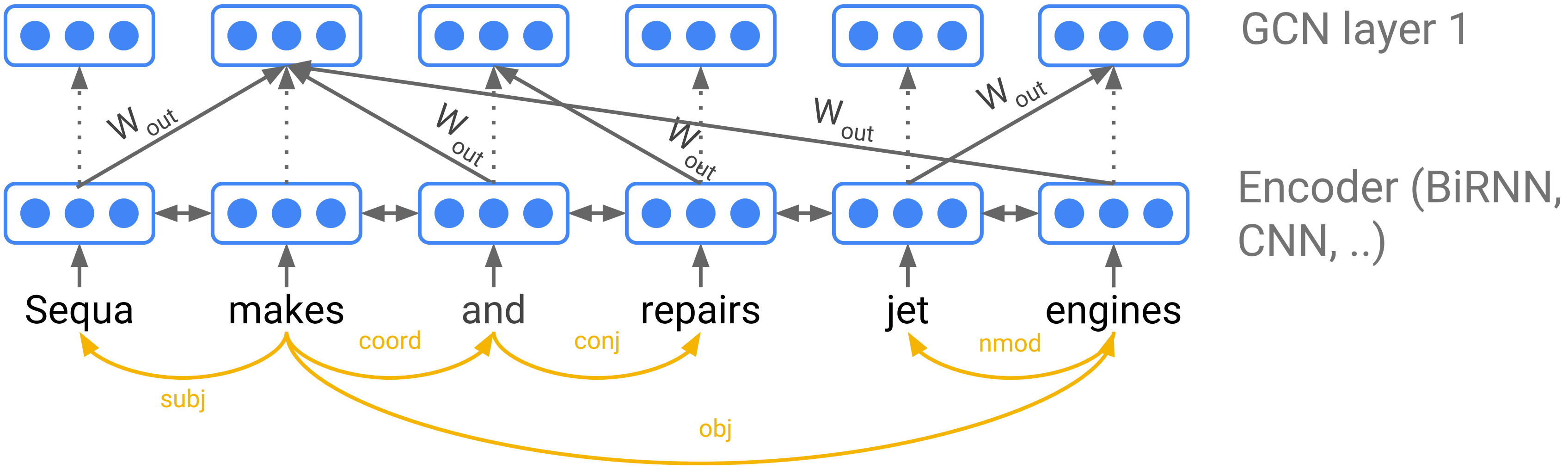
Graph Convolutional Encoders



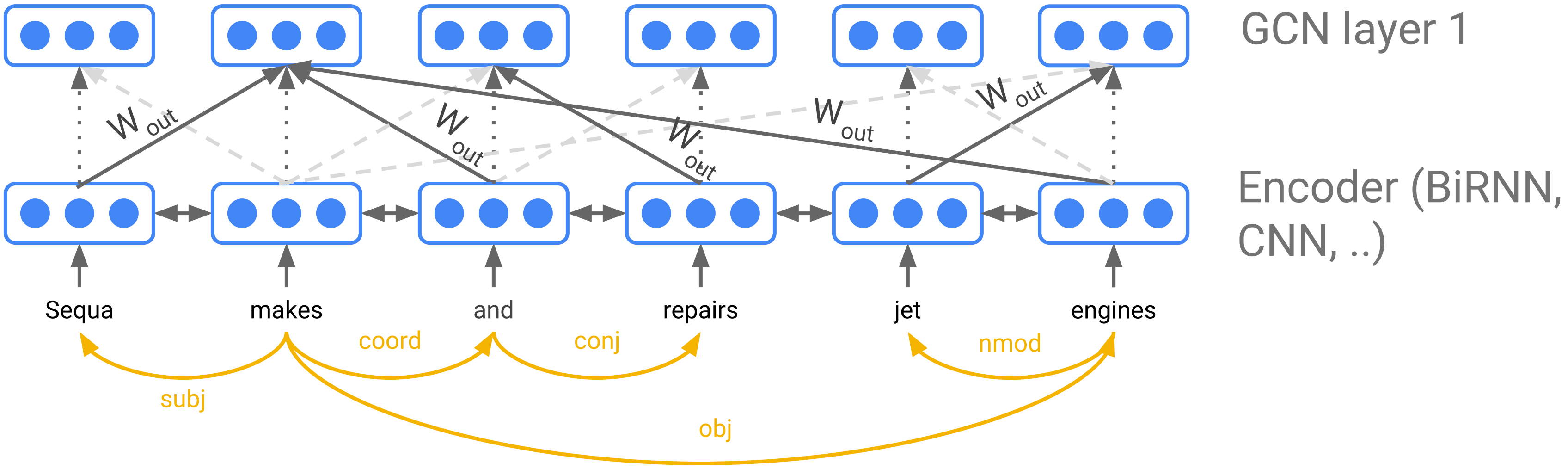
Graph Convolutional Encoders



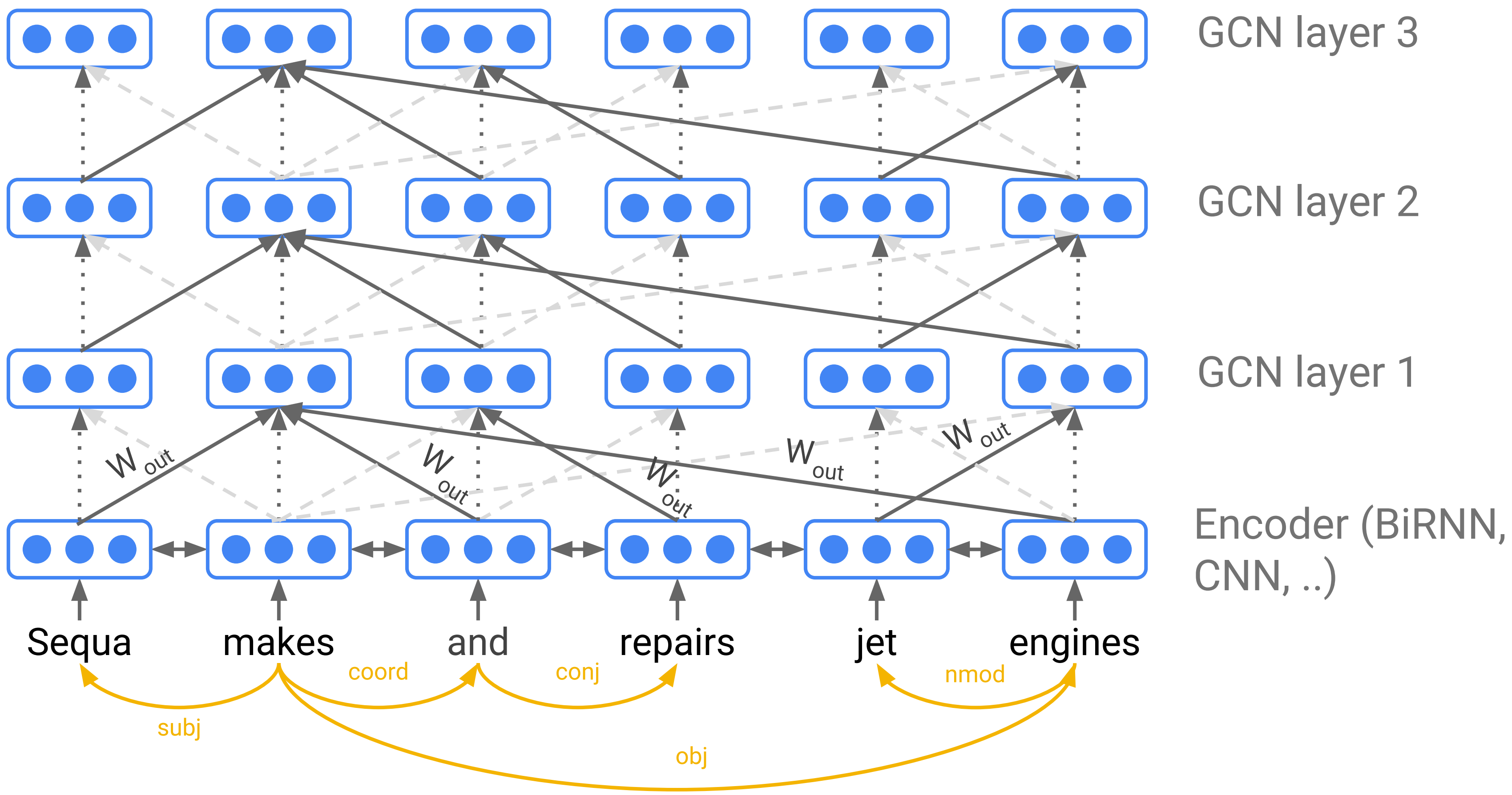
Graph Convolutional Encoders



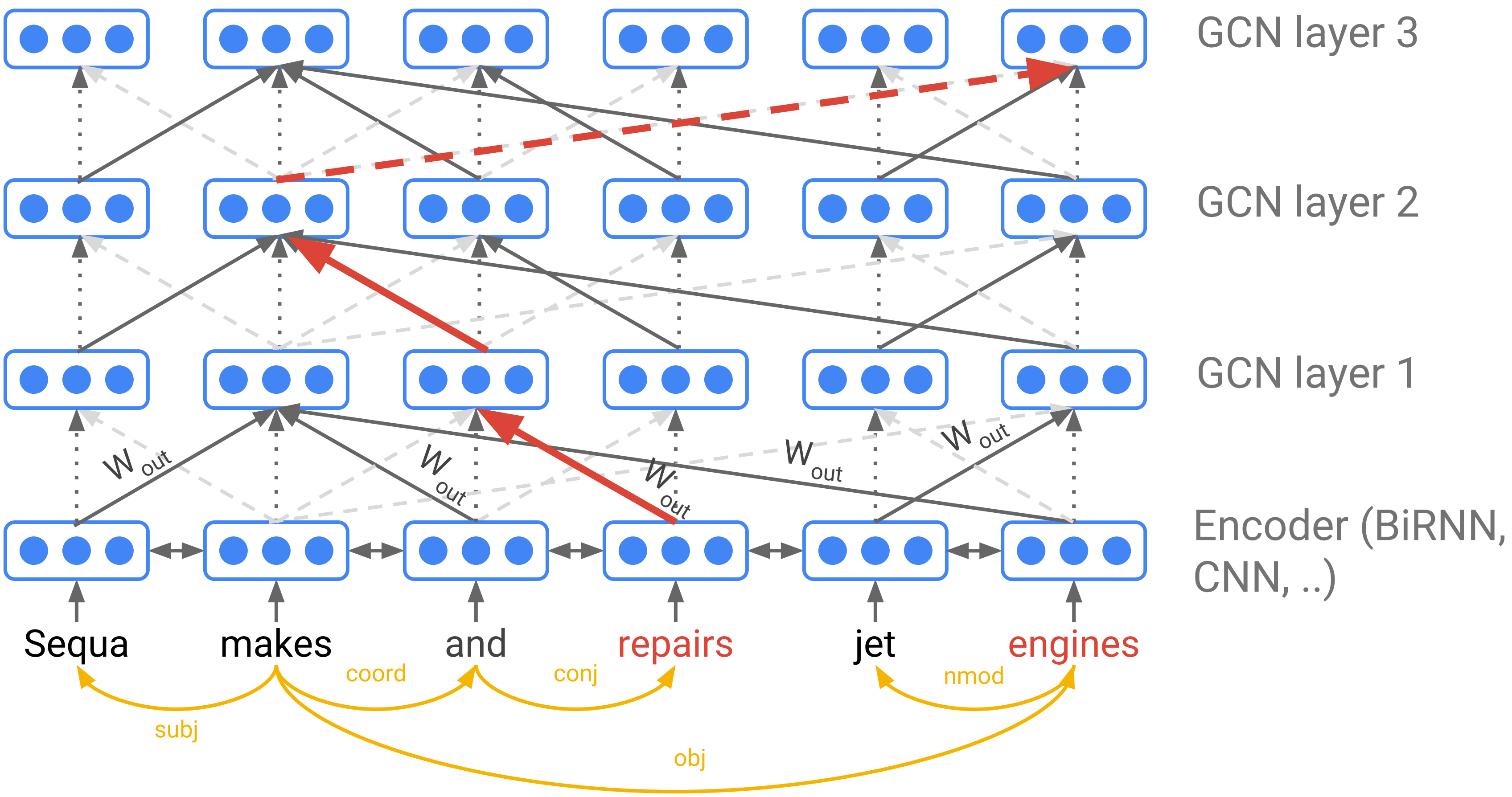
Graph Convolutional Encoders



Graph Convolutional Encoders

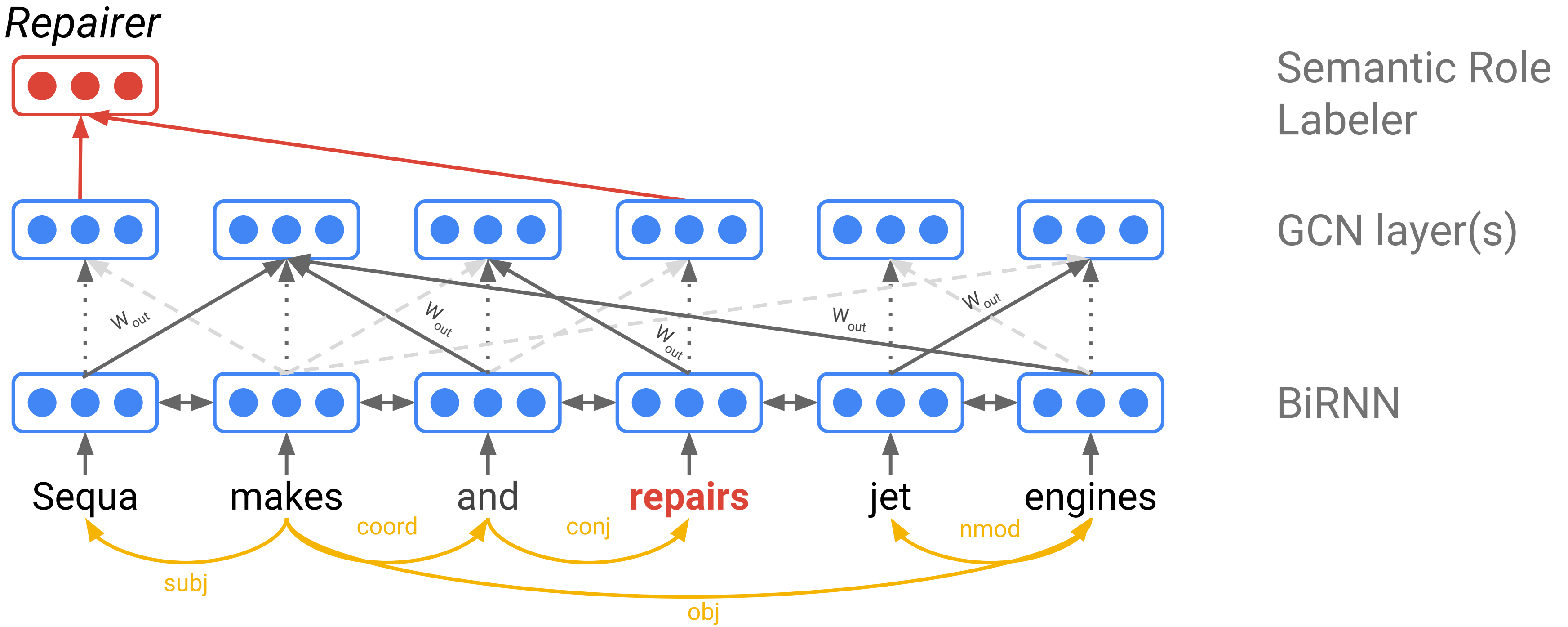


Graph Convolutional Encoders



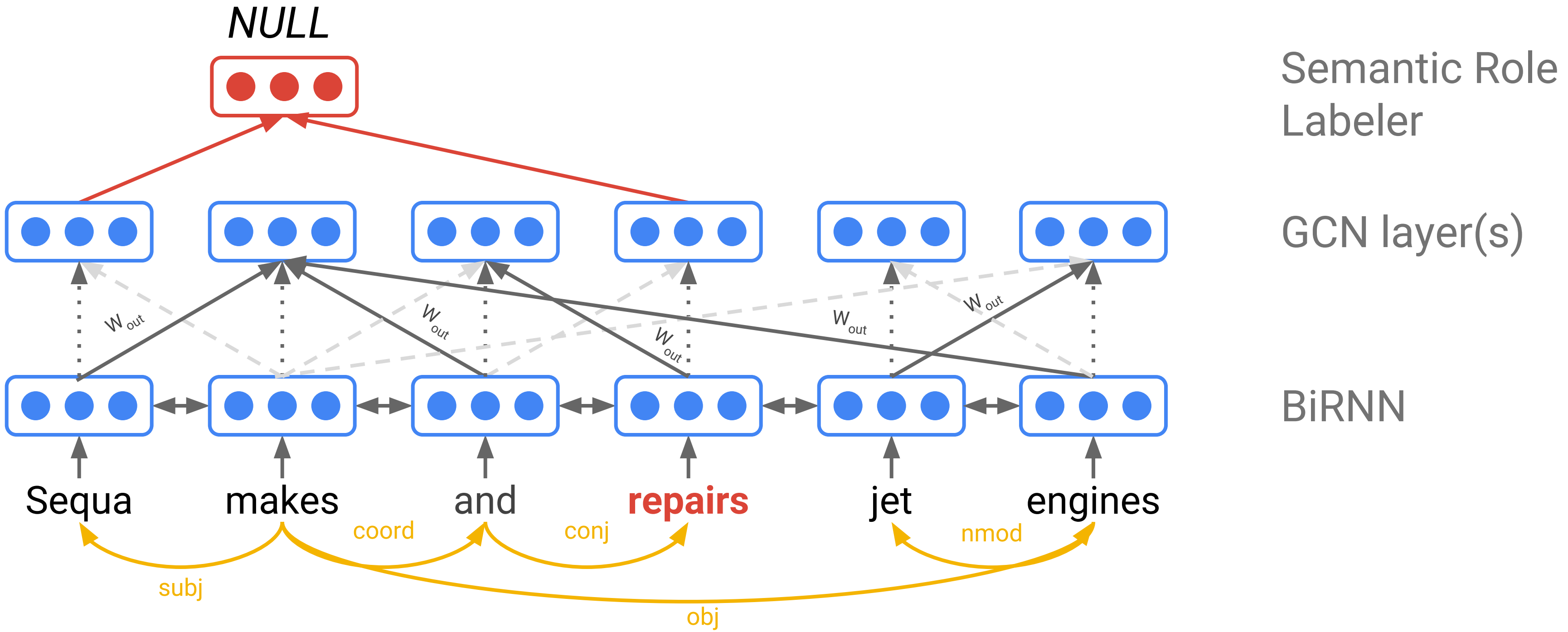
How do we construct a GCN-based semantic role labeler?

GCNs for Semantic Role Labeling



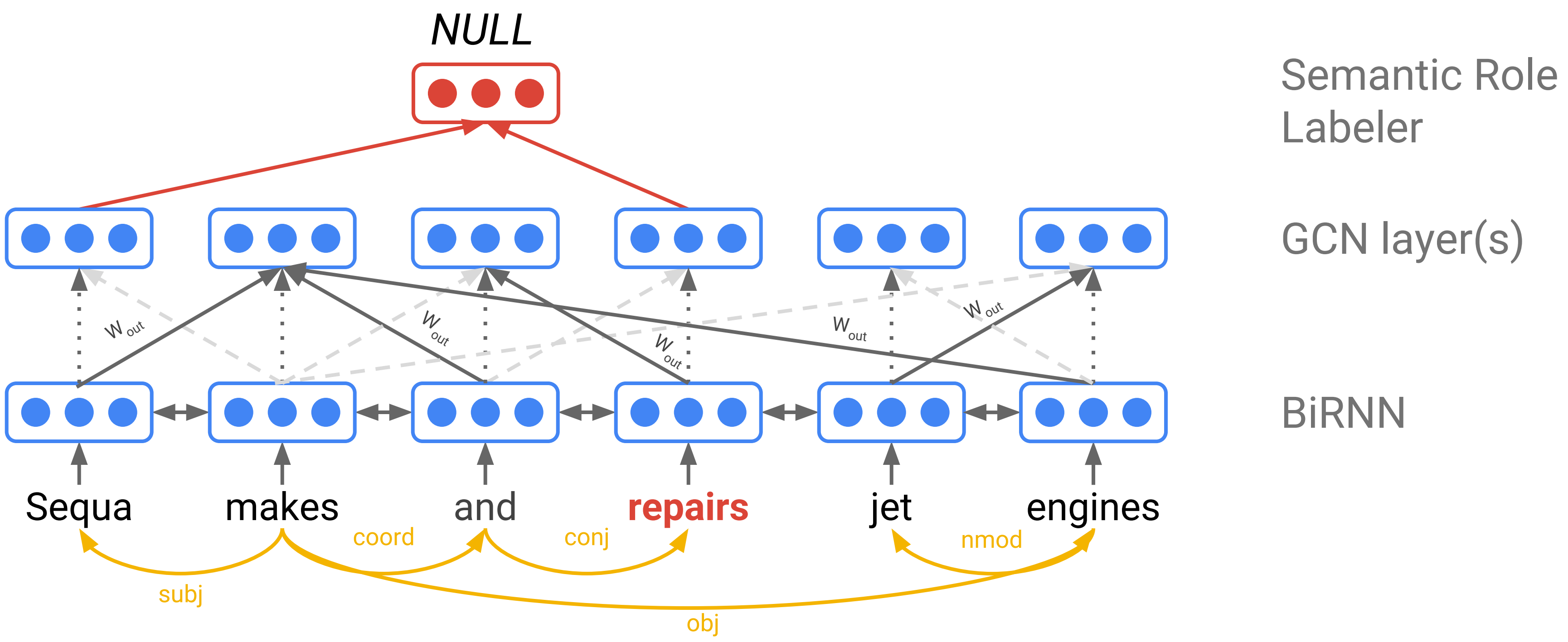
Marcheggiani et al., EMNLP 2017

GCNs for Semantic Role Labeling



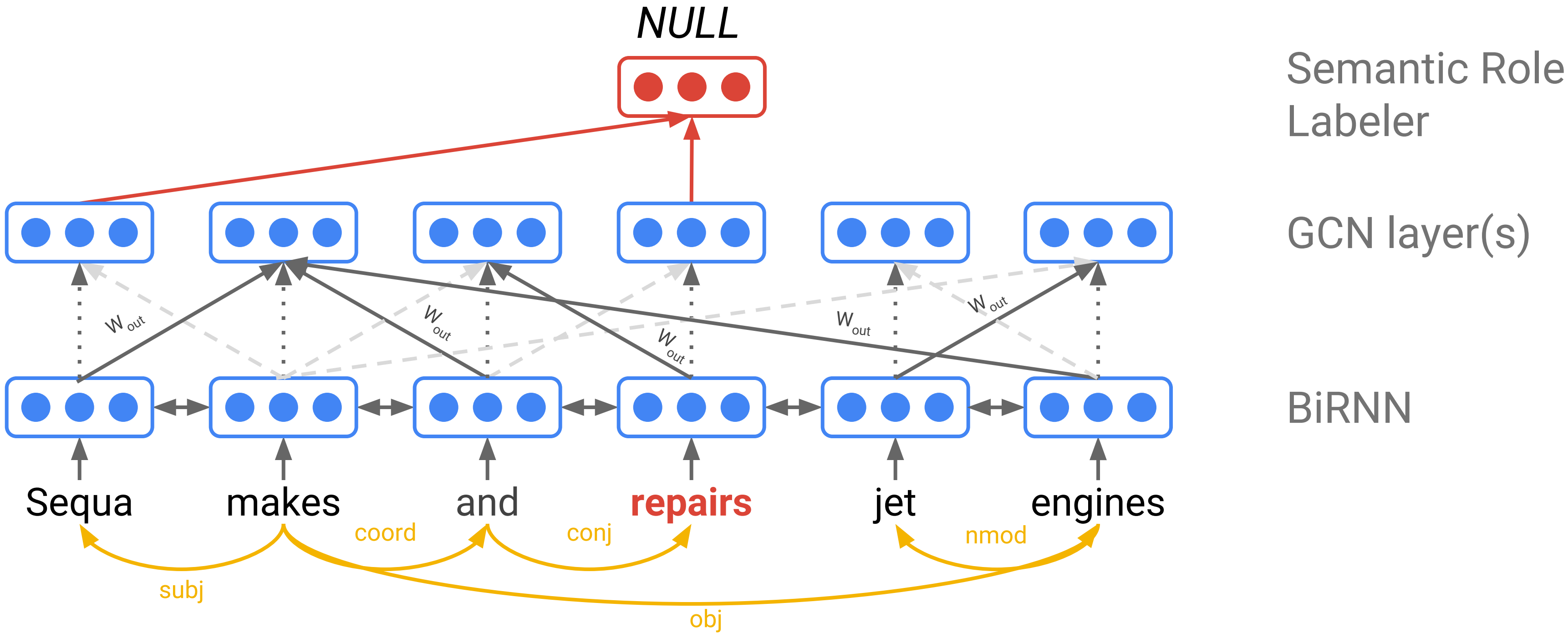
Marcheggiani et al., EMNLP 2017

GCNs for Semantic Role Labeling



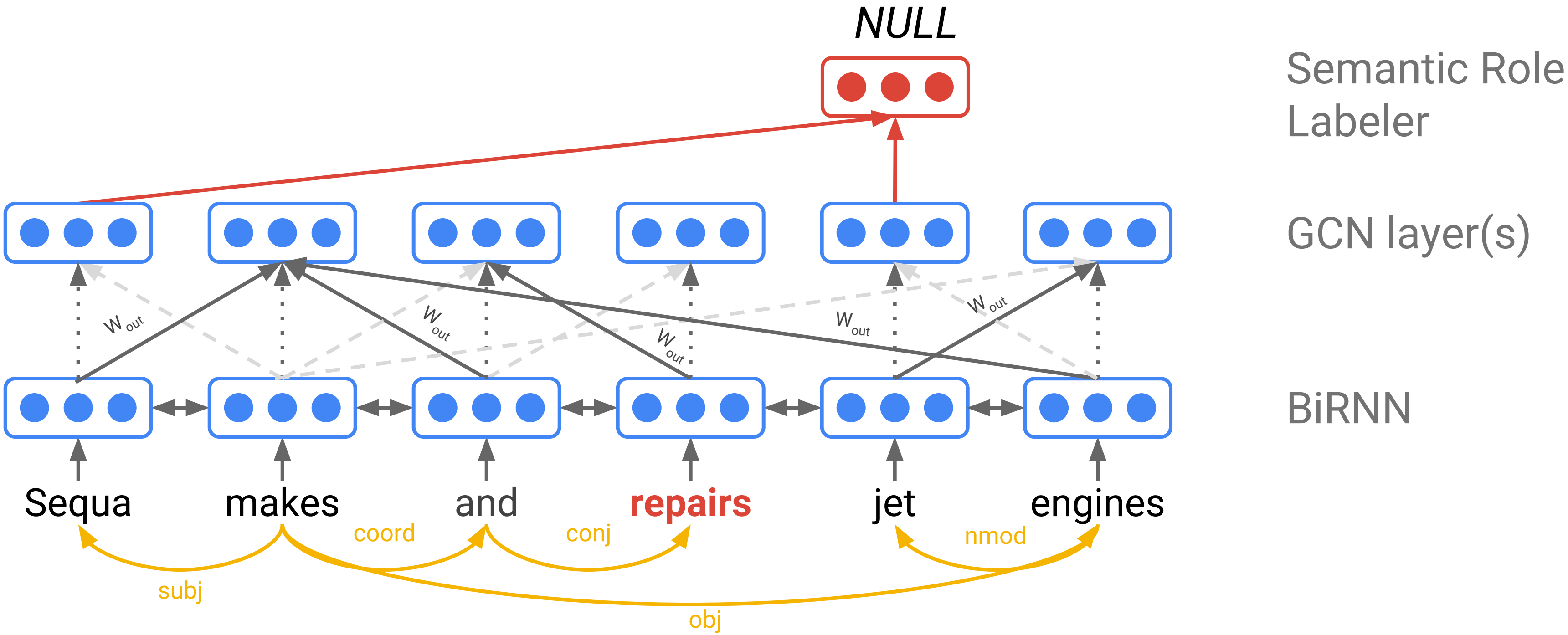
Marcheggiani et al., EMNLP 2017

GCNs for Semantic Role Labeling



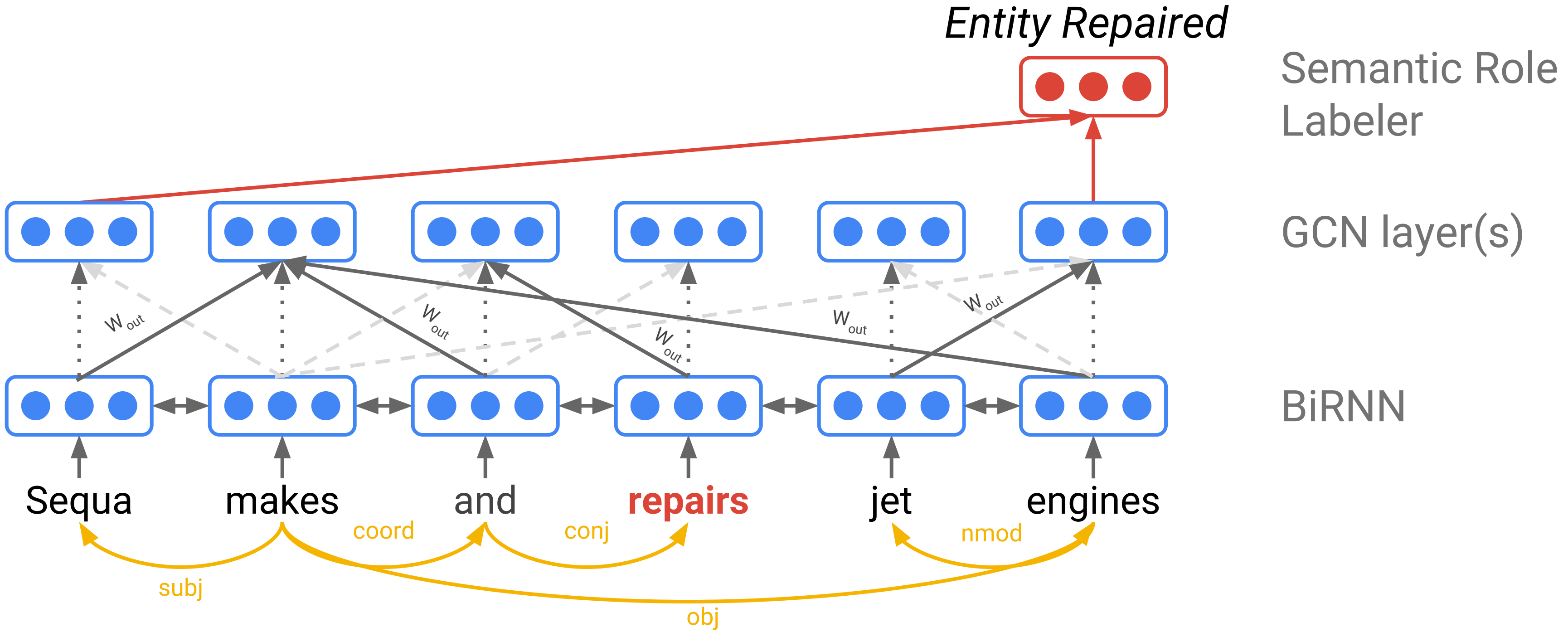
Marcheggiani et al., EMNLP 2017

GCNs for Semantic Role Labeling



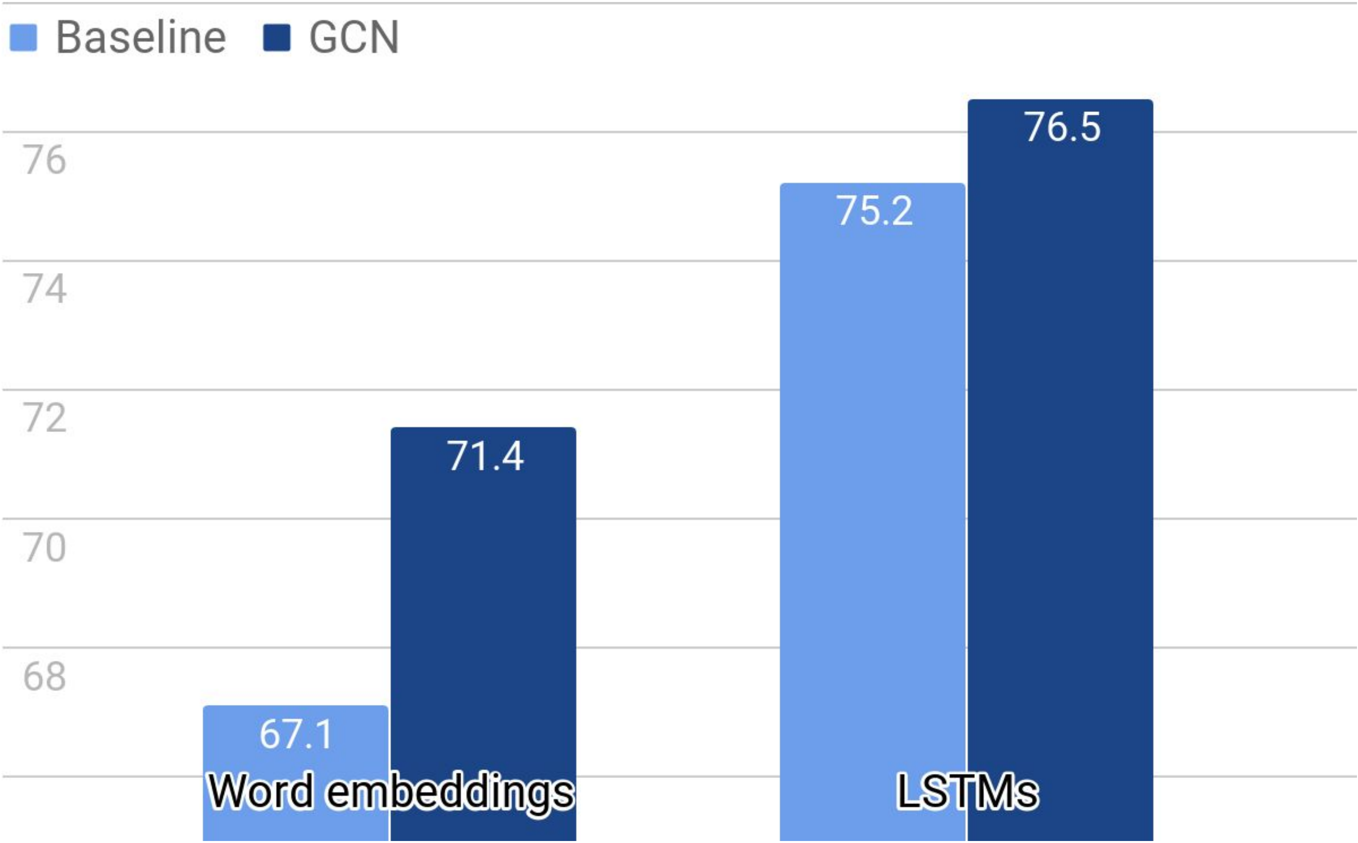
Marcheggiani et al., EMNLP 2017

GCNs for Semantic Role Labeling



Marcheggiani et al., EMNLP 2017

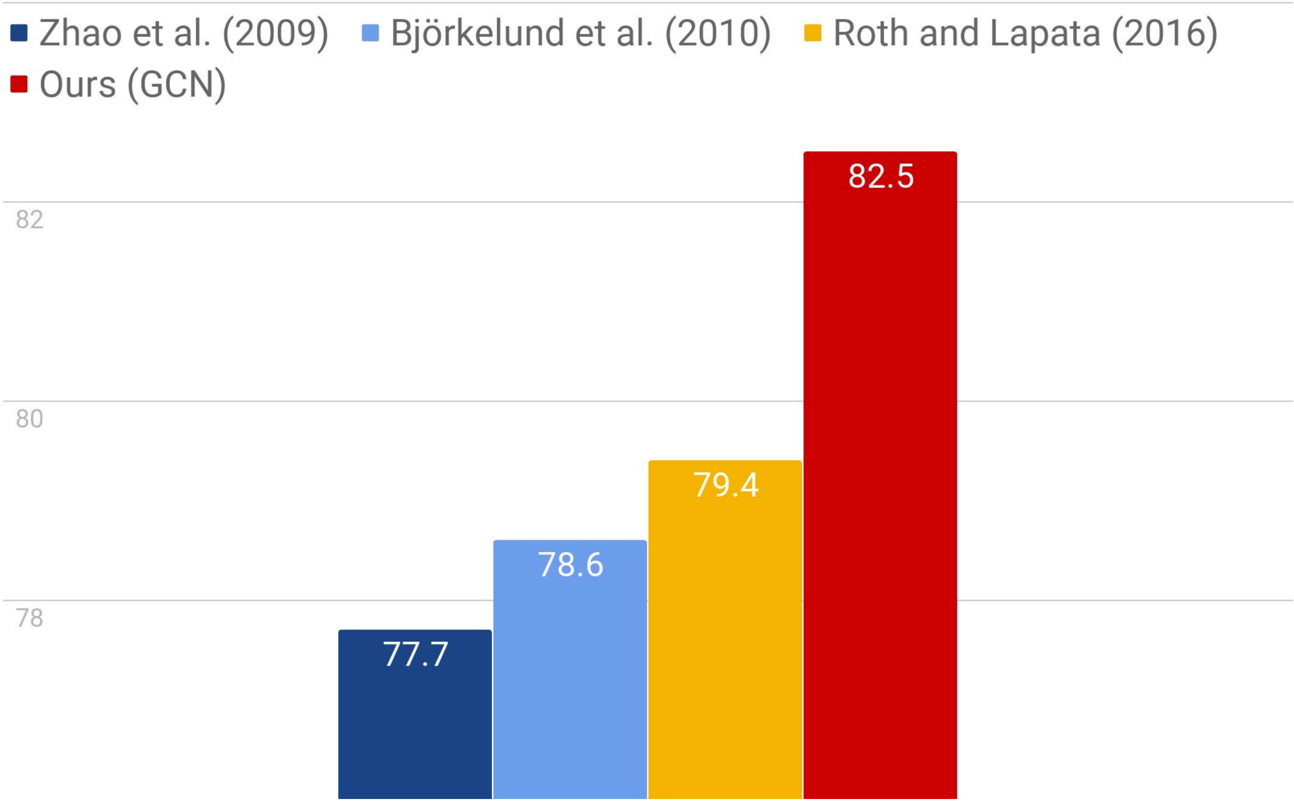
Results (F1) on Chinese (CoNLL-2009, dev set)



Marcheggiani & Titov (EMNLP, 2017)

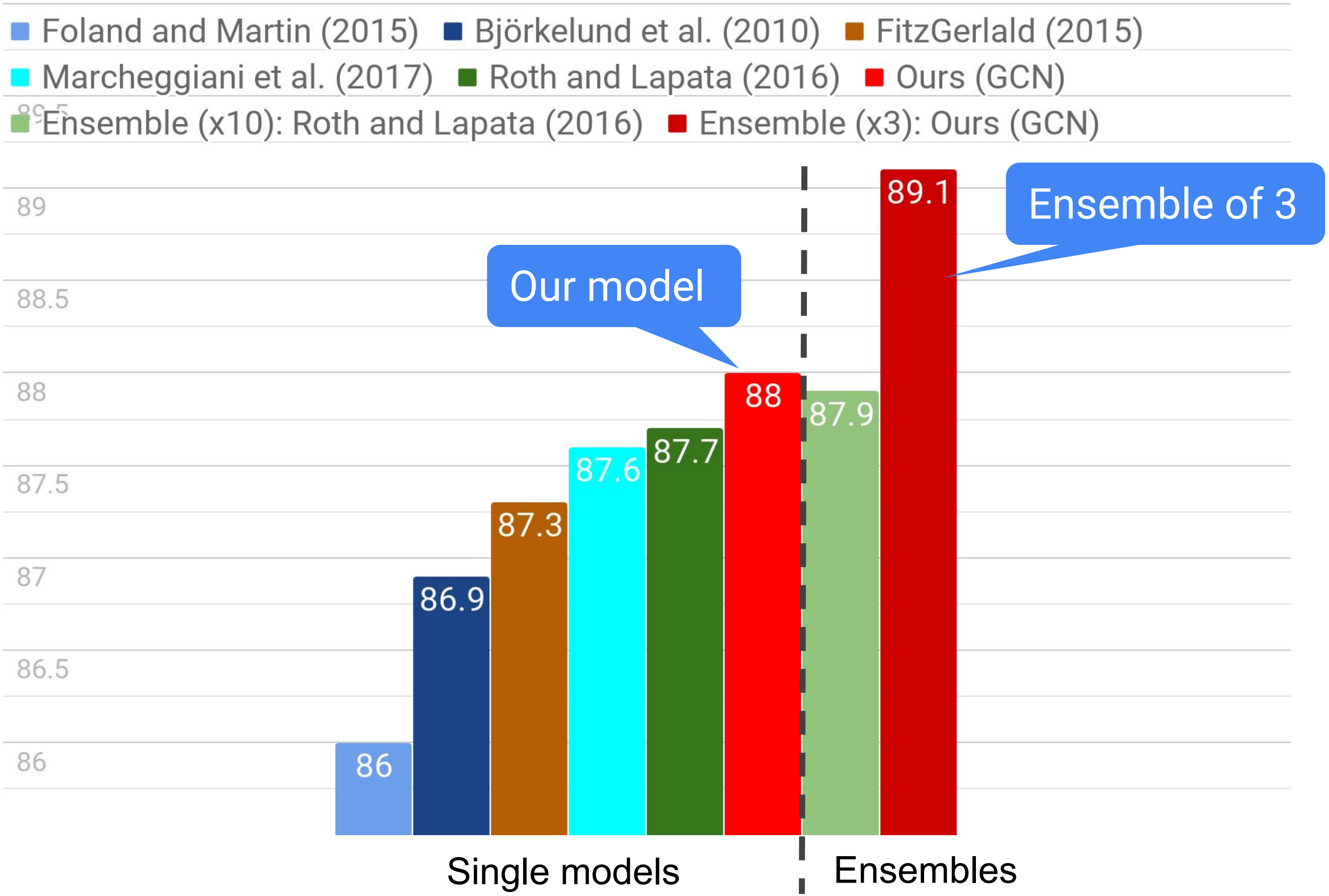
Predicate disambiguation is excluded from the F1 metric

Results (F1) on Chinese (CoNLL-2009, test set)



Marcheggiani & Titov (EMNLP, 2017)

Results (F1) on English (CoNLL-2009)



Marcheggiani & Titov (EMNLP, 2017)

Flexibility of GCN encoders

Simple and fast approach to integrating linguistic structure into encoders

In principle we can exploit almost **any kind** of linguistic structure:

Semantic role labeling structure

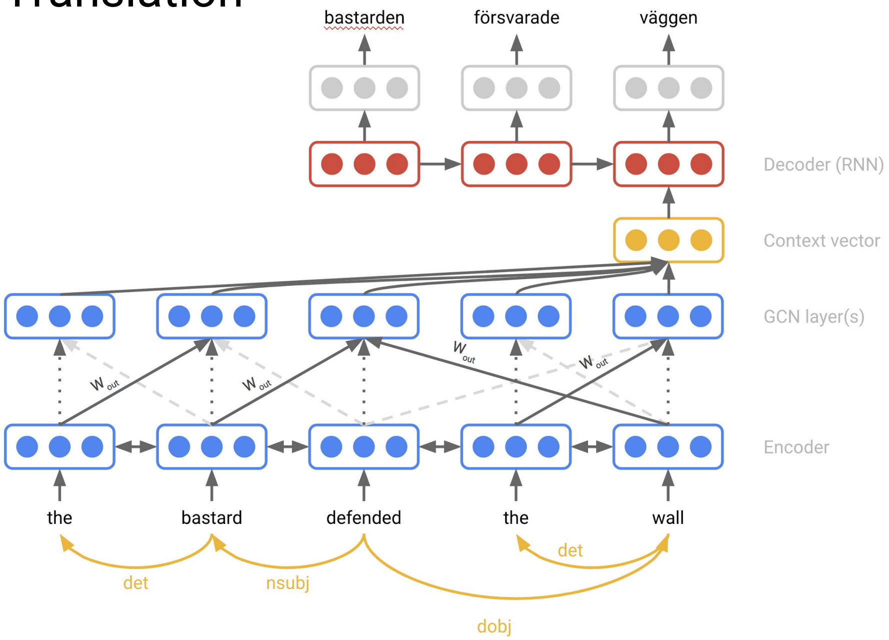
Co-reference chains

AMR semantic graphs

Their combination

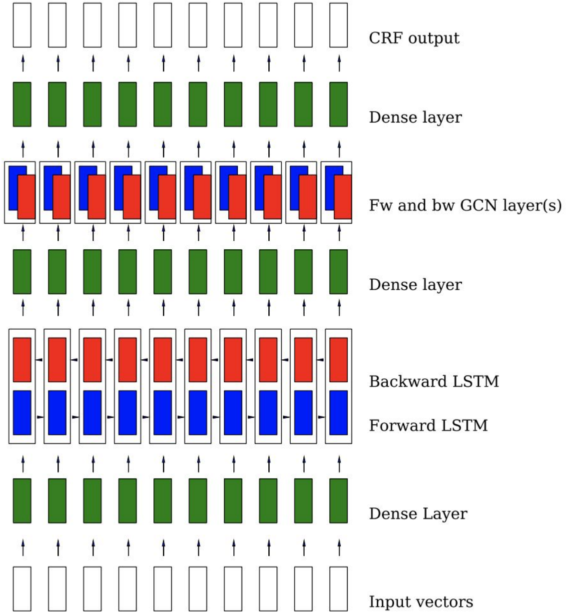
Other applications of syntactic GCN encoders

We also showed them effective as encoders in Neural Machine Translation



Bastings et al. (EMNLP, 2017)

Others recently applied them to NER



Cetoli et al. (arXiv:1709.10053)

Conclusions

GCNs are in subtasks of KBC (and in NLP beyond KBC):

- Semantic Roles: we proposed GCNs for encoding linguistic knowledge
- Link prediction: GCNs for link prediction (and entity classification) in multi-relational knowledge bases

Code available

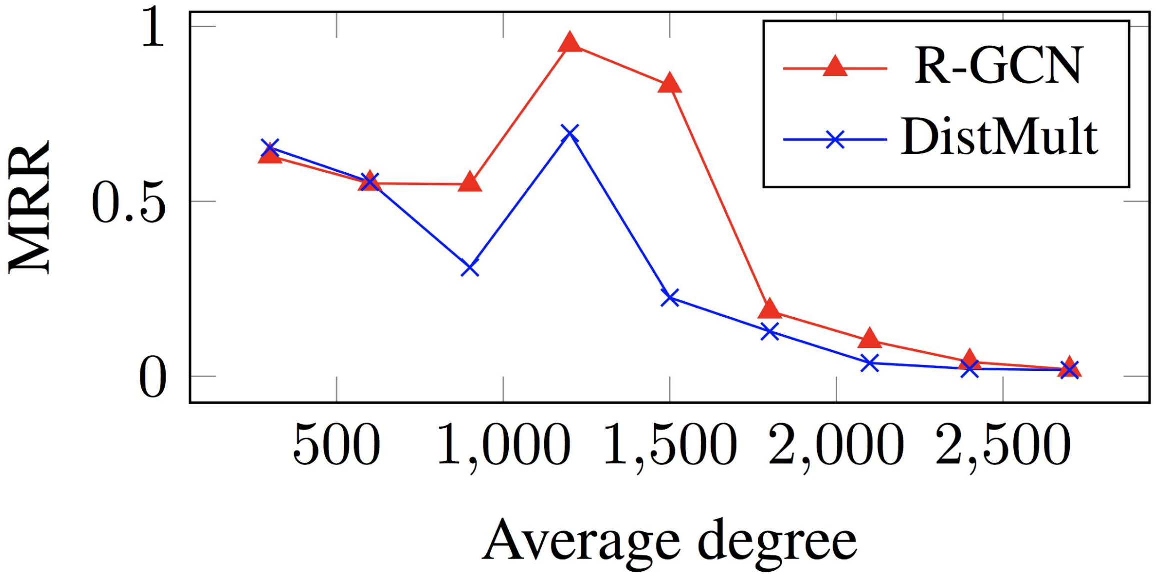
We are hiring! (PhD students / postdocs)

erc European Research Council

NWO Nederlandse Organisatie voor Wetenschappelijk Onderzoek

SAP

Analysis / Discussion



- Improvement across the board, especially in the middle of the range

Effect of Distance between Argument and Predicate (English)