# A Experimental Details

## A.1 Hyperparameters

**TACRED** We set LSTM hidden size to 200 in all neural models. We also use hidden size 200 for the output feedforward layers in the GCN model. We use 2 GCN layers and 2 feedforward (FFNN) layers in our experiments. We employ the ReLU function for all nonlinearities in the GCN layers and the standard max pooling operations in all pooling layers. For the Tree-LSTM model, we find a 2-layer architecture works substantially better than the vanilla 1-layer model, and use it in all our experiments. For both the Tree-LSTM and our models, we apply path-centric pruning with $K = 1$, as we find that this generates best results for all models (also see Figure 3). We use the pre-trained 300-dimensional GloVe vectors (Pennington et al., 2014) to initialize word embeddings, and we use embedding size of 30 for all other embeddings (i.e., POS, NER). We use the dependency parse trees, POS and NER sequences as included in the original release of the dataset, which was generated with Stanford CoreNLP (Manning et al., 2014). For regularization we apply dropout with $p = 0.5$ to all LSTM layers and all but the last GCN layers.

**SemEval** We use LSTM hidden size of 100 and use 1 GCN layer for the SemEval dataset. We preprocess the dataset with Stanford CoreNLP to generate the dependency parse trees, POS and NER annotations. All other hyperparameters are set to be the same.

For both datasets, we work with the Universal Dependencies v1 formalism (Nivre et al., 2016).

## A.2 Training

For training we use Stochastic Gradient Descent with an initial learning rate of 1.0. We use a cutoff of 5 for gradient clipping. For GCN models, we train every model for 100 epochs on the TACRED dataset, and from epoch 5 we start to anneal the learning rate by a factor of 0.9 every time the $\text{F}_1$ score on the dev set does not increase after an epoch. For Tree-LSTM models we find 30 total epochs to be enough. Due to the small size of the SemEval dataset, we train all models for 150 epochs, and use an initial learning rate of 0.5 with a decay rate of 0.95.

In our experiments we found that the output vector $h_{\text{sent}}$ tends to have large magnitude, and therefore adding the following regularization term to the cross entropy loss of each example improves the results:

$$\ell_{\text{reg}} = \beta \cdot \|h_{\text{sent}}\|^2. \tag{6}$$

Here, $\ell_{\text{reg}}$ functions as an $l_2$ regularization on the learned sentence representations. $\beta$ controls the regularization strength and we set $\beta = 0.003$. We empirically found this to be more effective than applying $l_2$ regularization on the convolutional weights.

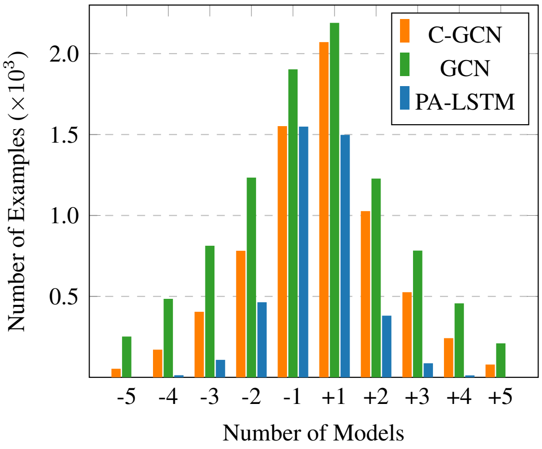
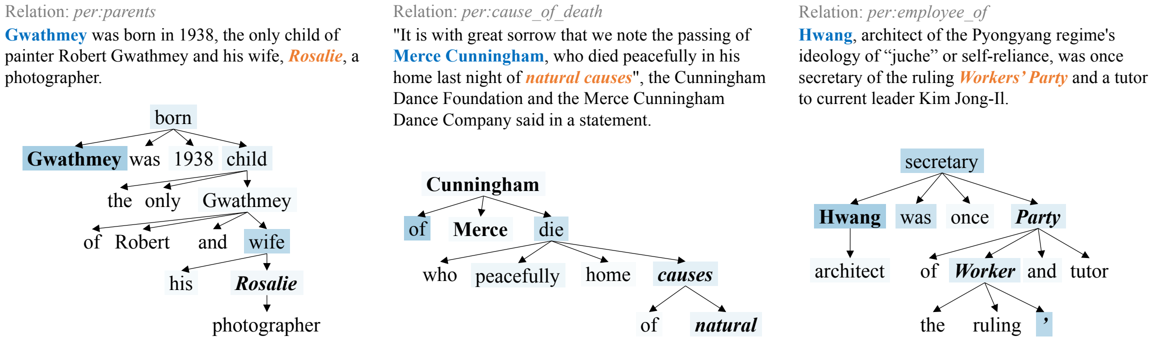


Figure 7: Aggregated 5-run difference compared to PA-LSTM on the TACRED dev set. For each example, if $X$ out of 5 GCN models predicted its label correctly and $Y$ PA-LSTM models did, it is aggregated in the bar labeled $X - Y$. "0" is omitted due to redundancy.

# B Comparing GCN models and PA-LSTM on TACRED

We compared the performance of both GCN models with the PA-LSTM on the TACRED dev set. To minimize randomness that is not inherent to these models, we accumulate statistics over 5 independent runs of each model, and report them in Figure 7. As is shown in the figure, both GCN models capture very different examples from the PA-LSTM model. In the entire dev set of 22,631 examples, 1,450 had at least 3 more GCN models predicting the label correctly compared to the PA-LSTM, and 1,550 saw an improvement from using the PA-LSTM. The C-GCN, on the other hand, outperformed the PA-LSTM by at least 3 models on a total of 847 examples, and lost by a margin of at least 3 on another 629 examples, as reported in the main text. This smaller difference is also reflected in the diminished gain from ensembling with the PA-LSTM shown in Table 1. We hypoth-



Relation: *per:parents*
**Gwathmey** was born in 1938, the only child of painter Robert Gwathmey and his wife, ***Rosalie***, a photographer.

Relation: *per:cause_of_death*
"It is with great sorrow that we note the passing of **Merce Cunningham**, who died peacefully in his home last night of ***natural causes***", the Cunningham Dance Foundation and the Merce Cunningham Dance Company said in a statement.

Relation: *per:employee_of*
**Hwang**, architect of the Pyongyang regime's ideology of "juche" or self-reliance, was once secretary of the ruling ***Workers' Party*** and a tutor to current leader Kim Jong-Il.

Figure 8: More examples and the pruned dependency trees the C-GCN predicted correctly. Words are shaded by the number of dimensions they contributed to $h_{\text{sent}}$ in the pooling operation, with punctuation omitted.

| Relation | Dependency Tree Edges | | |
|---|---|---|---|
| *per:children* | S-PER ← son | son → O-PER | S-PER ← survived |
| *per:parents* | S-PER ← born | O-PER ← son | S-PER ← mother |
| *per:siblings* | S-PER ← sister | sister → O-PER | brother → O-PER |
| *per:other_family* | S-PER ← stepson | niece → O-PER | O-PER ← stepdaughter |
| *per:spouse* | wife → O-PER | S-PER ← wife | his ← wife |
| *per:city_of_death* | S-PER ← died | died → O-CITY | ROOT → died |
| *per:city_of_birth* | S-PER ← born | was ← born | born → O-CITY |
| *per:cities_of_residence* | in ← O-CITY | O-CITY ← S-PER | S-PER ← lived |
| *per:employee_of* | a ← member | S-PER ← worked | S-PER ← played |
| *per:schools_attended* | S-PER ← graduated | S-PER ← earned | S-PER ← attended |
| *per:title* | O-TITLE ← S-PER | as ← O-TITLE | former ← S-PER |
| *per:charges* | S-PER ← charged | O-CHARGE ← charges | S-PER ← faces |
| *per:cause_of_death* | died → O-CAUSE | S-PER ← died | from ← O-CAUSE |
| *per:age* | S-PER → O-NUMBER | S-PER ← died | age → O-NUMBER |
| *org:alternate_names* | S-ORG → O-ORG | O-ORG → ) | ( ← O-ORG |
| *org:founded* | founded → O-DATE | established → O-DATE | was ← founded |
| *org:founded_by* | O-PER → founder | S-ORG ← O-PER | founder → S-ORG |
| *org:top_members* | S-ORG ← O-PER | director → S-ORG | O-PER ← said |
| *org:subsidiaries* | S-ORG ← O-ORG | S-ORG → 's | O-ORG → division |
| *org:num_of_employees* | S-ORG ← has | S-ORG → employs | O-NUMBER ← employees |
| *org:shareholders* | buffett ← O-PER | shareholder → S-ORG | largest ← shareholder |
| *org:website* | S-ORG → O-URL | ROOT → S-ORG | S-ORG → : |
| *org:dissolved* | S-ORG ← forced | forced → file | file → insolvency |
| *org:political/religious_affiliation* | S-ORG → group | O-IDEOLOGY ← group | group → established |

Table 5: The three dependency edges that contribute the most to the classification of different relations in the dev set of TACRED. For clarity, we removed edges which 1) connect to common punctuation (i.e., commas, periods, and quotation marks), 2) connect to common preposition (i.e., of, to, by), and 3) connect tokens within the same entities. We use PER, ORG, CHARGE, CAUSE for entity types of PERSON, ORGANIZATION, CRIMINAL_CHARGE and CAUSE_OF_DEATH, respectively. We use S- and O- to denote subject and object entities, respectively. ROOT denotes the root node of the tree.

esize that the diminishing difference results from the LSTM contextualization layer, which incorporates more information readily available at the surface form, rendering the model's behavior more similar to a sequence model.

For reference, we also include in Figure 7 the comparison of another 5 different runs (with different seeds) of the PA-LSTM to the original 5 runs of the PA-LSTM. This is to confirm that the difference shown in the figure between the model classes is indeed due a to model difference, rather than an effect of different random seeds. More specifically, the two groups of PA-LSTM only see 99 and 121 examples exceeding the 3-model margin on either side over the 5 runs, much lower than the numbers reported above for the GCN models.

## C Understanding Model Behavior

We present visualization of more TACRED dev set examples in Figure 8. We also show the dependency edges that contribute the most to more relation types in Table 5.