Appendices

A Super-Clusters

Table 4 (below) shows our eight super-clusters over TACRED sentence groups described in Section 3.2. Each group appears in exactly one cluster.

B Re-TACRED Corrected Prediction Errors

Table 5 (below) presents five handpicked sentences showing different types of errors that a SpanBERT model trained on TACRED makes on the Re-TACRED test split, and that the same model trained on Re-TACRED is able to correct on the same split.

C Effect of Non-Refined Labels

We also examine how models differ based on our *non-refined* label re-annotations. Non-refined relations are any for which we did not alter the TAC KBP relation definitions for (i.e. any label not discussed in Section 3.3). We conduct this analysis by comparing model performance over different combinations of train and test splits from TACRED and Re-TACRED. We denote train splits using $[\cdot]_{\texttt{train}}$ and test splits using $[\cdot]_{\texttt{test}}$, where $[\cdot]$ is either TACRED or Re-TACRED (e.g., $\text{TACRED}_{\texttt{train}}$). All models are then trained on $\text{TACRED}_{\texttt{train}}$ or $\text{Re-TACRED}_{\texttt{train}}$, and evaluated on $\text{TACRED}_{\texttt{test}}$ or $\text{Re-TACRED}_{\texttt{test}}$. Our results are shown in Table 6.

The results show several interesting differences between TACRED and Re-TACRED. First, all methods trained and evaluated on TACRED obtain significantly higher performance on the non-refined labels than over the full label set. We attribute this increase to the fact that these relations are less ambiguous compared than the refined ones. Second, methods trained on $\text{TACRED}_{\text{train}}$ achieve better performance on $\text{Re-TACRED}_{\text{test}}$ than on $\text{TACRED}_{\text{test}}$. This is consistent with the findings in Alt, Gabryszak, and Hennig (2020), and suggests that: (i) TACRED may be under-estimating model performance, and (ii) large improvements can be obtained simply by evaluating models on higher quality annotations.

Model	(Train Split, Test Split)	Metrics		
		F1	Precision	Recall
PA-LSTM	($\text{TACRED}_{\text{train}}$, $\text{TACRED}_{\text{test}}$)	72.3	71.3	73.3
	($\text{TACRED}_{\text{train}}$, $\text{Re-TACRED}_{\text{test}}$)	73.3	76.7	70.2
	($\text{Re-TACRED}_{\text{train}}$, $\text{TACRED}_{\text{test}}$)	68.3	65.9	70.9
	($\text{Re-TACRED}_{\text{train}}$, $\text{Re-TACRED}_{\text{test}}$)	75.9	75.8	76.1
C-GCN	($\text{TACRED}_{\text{train}}$, $\text{TACRED}_{\text{test}}$)	72.6	71.1	74.3
	($\text{TACRED}_{\text{train}}$, $\text{Re-TACRED}_{\text{test}}$)	73.2	76.0	70.6
	($\text{Re-TACRED}_{\text{train}}$, $\text{TACRED}_{\text{test}}$)	69.2	68.5	69.8
	($\text{Re-TACRED}_{\text{train}}$, $\text{Re-TACRED}_{\text{test}}$)	77.3	78.2	76.5
SpanBERT	($\text{TACRED}_{\text{train}}$, $\text{TACRED}_{\text{test}}$)	75.0	74.7	75.3
	($\text{TACRED}_{\text{train}}$, $\text{Re-TACRED}_{\text{test}}$)	76.8	81.2	72.8
	($\text{Re-TACRED}_{\text{train}}$, $\text{TACRED}_{\text{test}}$)	74.1	70.9	77.7
	($\text{Re-TACRED}_{\text{train}}$, $\text{Re-TACRED}_{\text{test}}$)	84.1	85.0	83.1

Table 6: Results for multiple RE models (leftmost column) on different train-and-evaluation combinations. Each combination is represented by a pair of the form ("train split", "test split"). For instance, ($\text{TACRED}_{\text{train}}$, $\text{Re-TACRED}_{\text{test}}$) indicates that a method is trained on the TACRED train partition and evaluated on the Re-TACRED test split. The remaining columns show metric results.

Super-Cluster	Subject Type	Object Types
org2miscmulti	ORGANIZATION	URL, DATE, NUMBER, RELIGION, IDEOLOGY, MISC
org2locmulti		CITY, COUNTRY, STATE_OR_PROVINCE, LOCATION
org2org		ORGANIZATION
org2per		PERSON
per2miscmulti	PERSON	TITLE, DATE, CRIMINAL_CHARGE, RELIGION, NUMBER, CAUSE_OF_DEATH, DURATION, MISC
per2locmulti		NATIONALITY, COUNTRY, STATE_OR_PROVINCE, CITY, LOCATION
per2org		ORGANIZATION
per2per		PERSON

Table 4: Mappings between super-clusters and sentence groups. Sentence groups are defined by the pair, (SUBJECT_TYPE, OBJECT_TYPE), which describes the subject and object type of all sentences in the group. The leftmost column denotes each super-cluster name. The middle column lists the two possible subject types (ORGANIZATION and PERSON), while the rightmost column shows the list of object types whose pairing with the corresponding subject type is an element of the respective super-cluster. For instance, (PERSON, TITLE) represents the sentence group where all sentence subject types are PERSON and all object types are TITLE. From the table, this group is an element of the per2miscmulti super-cluster.

Error Type	Sentence	TACRED Prediction	Correct Label
Neg → Pos	"…leave of absence from [his]$_{\text{SUB}}$ posts as [Cephalon]$_{\text{OBJ}}$'s chairman and chief executive."	NO_RELATION	PERSON:EMPLOYEE_OF
	"…Pakistani [journalist]$_{\text{OBJ}}$ and Taliban expert [Ahmed Rashid]$_{\text{SUB}}$, from Madrid."	NO_RELATION	PERSON:TITLE
	"…[National Taiwan Symphony Orchestra]$_{\text{SUB}}$ (NTSO) …an [NTSO]$_{\text{OBJ}}$ spokesman…"	NO_RELATION	ORGANIZATION:ALTERNATE_NAMES
Pos → Neg	"[His]$_{\text{SUB}}$ [therapist]$_{\text{OBJ}}$ told him to politely decline, 'which helped."	PERSON:TITLE	NO_RELATION
Pos → Pos	"…[her]$_{\text{SUB}}$ stepchildren, Susan, …, Stephen and [Maggie]$_{\text{OBJ}}$ Mailer; …."	PERSON:SIBLINGS	PERSON:CHILDREN

Table 5: Five handpicked sentences from the Re-TACRED test split that a TACRED-trained SpanBERT model misclassifies but a Re-TACRED-trained SpanBERT method correctly classifies. Sentence subjects and objects are defined as in Section 1, and the complete TACRED-trained SpanBERT predictions and gold labels are provided. Additionally, each sentence is marked by a specific "error type" (the leftmost column) decribing whether the error is due to predicting a negative sentence label when the correct relation is positive (Neg → Pos), predicting a positive relation when the correct label is negative (Pos → Neg), or inferring the incorrect positive label (Pos → Pos).

Third, methods trained on Re-TACRED$_{\text{train}}$ and evaluated on TACRED$_{\text{test}}$ perform worse than those evaluated on Re-TACRED$_{\text{test}}$. A deeper inspection of the data reveals that such models exhibit significantly fewer correct positively labeled predictions in TACRED$_{\text{test}}$ than in Re-TACRED$_{\text{test}}$, resulting in substantially lower scores. For instance, SpanBERT trained on Re-TACRED$_{\text{train}}$ exhibits 16.5% fewer correct positively labeled instances in TACRED$_{\text{test}}$ compared to Re-TACRED$_{\text{test}}$. This highlights the effects of our label changes described in Section 4.1; many positively labeled sentences in Re-TACRED are either negatively labeled or assigned another positive relation in TACRED. Fourth, models trained and evaluated on Re-TACRED perform significantly better than any other combination. Thus, while methods trained on TACRED$_{\text{train}}$ achieve performance boosts when testing on Re-TACRED$_{\text{test}}$ (compared to evaluating on TACRED$_{\text{test}}$), training on Re-TACRED$_{\text{train}}$ is critical to achieving the strongest performance on Re-TACRED$_{\text{test}}$.

D Hyperparameters

We train all our TACRED-based models using the reported hyperparameters by their respective contributors. All hyperparameter details for our Re-TACRED-based methods can be found below. Additionally, all code required to reproduce our results and our new dataset can be found in our repository at https://github.com/gstoica27/Re-TACRED. We train our PA-LSTM and C-GCN models on a single Nvidia Titan X GPU, and utilized a single Nvidia Tesla V100 GPU to train SpanBERT.

Re-TACRED PA-LSTM. We perform an extensive grid-search over LSTM hidden dimension sizes from {100, 150, 200, 250, 300}, LSTM depth of {1, 2, 3}, word dropout from {0.0, 0.01, 0.04, 0.1, 0.25, .5}, and position-encoding dimension size among {15, 20, 25, 30, 50, 75, 100}. However, we observe the best performance with the hyperparameters reported by Zhang et al. (2017). In addition, we employ the equivalent training strategy as is reported in Zhang et al. (2017) (detailed under Appendix B of their publication).

Re-TACRED C-GCN. Similar to our PA-LSTM experiments, we find that keeping the majority of hyperparemeters equivalent to those reported by Zhang, Qi, and Manning (2018) yield the best results. The sole parameter we alter is increasing the residual neural network hidden dimension from 200 to 300. In addition, we use the same training procedure as Zhang, Qi, and Manning (2018) (described in Appendix A of their publication).

Re-TACRED SpanBERT. For SpanBERT, we perform a grid-search over learning rate sizes in {1e-6, 2e-6, 2e-5} and warm-up proportions in {.1, .2}. However, we observe the best performance using the reported parameters by Joshi et al. (2019). We refer readers to Joshi et al. (2019) (detailed in Section 4.2 and Appendix B in their publication) for further details on training strategy.

E Amazon Mechanical Turk

Our annotation process can be broken down by super-cluster, where each cluster represents as distinct annotation task. Overall, we had 8 such tasks (number of super-clusters), and each consisted of 2-4 labeling rounds. The first round gathered 2 distinct annotations for every sentence in the respective super-cluster. Any disagreements were then given to a third annotator in the second round. Then, if necessary, the third and fourth label rounds asked an additional worker to annotate remaining disagreements. Sentences to be annotated in label rounds were grouped into Human Intelligence Tasks (HIT). Each HIT consisted of 5 sentences, of which one was gold (i.e., its correct label was known). We priced HITs competitively at $.15 per HIT. We utilized annotations from 243 total workers, and the total time taken to sequentially annotate all our label rounds across all annotation tasks was $\approx$784 hours. While this number is large, it is important to note that many labeling rounds were completed in parallel, significantly decreasing the overall time.