PoSTACRED: Stacked Encoders & Attention-GCN Relation Extraction

Jonathon Dilworth
Department of Computer Science
University of Manchester
jonathon.dilworth@postgrad.manchester.ac.uk

Emma O'Brien
Department of Computer Science
University of Manchester
emma.obrien@postgrad.manchester.ac.uk

Rojs Aktumanis
Department of Computer Science
University of Manchester
rojs.aktumanis@postgrad.manchester.ac.uk

Alexandros Michaelides
Department of Computer Science
University of Manchester
alexandros.michaelides@postgrad.manchester.ac.uk

1 Introduction

Relation extraction (RE) is a learning task in text mining, which involves two components: (1) determining whether a relation exists between a subject entity and an object entity within a document, and (2) what kind of relation is present. A common approach is to view this task as a multi-class classification problem, where the set of classes is the set of relations and the special class *no_relation*.

1.1 Problem Statement

Including typed entity markers during tokenisation for fine-tuning BERT improves downstream performance on RE tasks (Zhong and Chen, 2021). To the best of our knowledge, nobody has tried enriching embeddings with part-of-speech (PoS) tags (or dependency relations) during fine-tuning (of BERT) in the manner described by (Zhou and Chen, 2022a). Given the reported performance benefits, we hypothesise that adding syntactic markers at the token level (such as PoS tags) during fine-tuning will act as a heuristic mechanism for the learning algorithm, leading to faster convergence and possibly a performance benefit during downstream RE.

Attention has proven to be a powerful architectural component for RE as mentioned above (BERT). In addition, dependency tree-based features have been shown to increase the performance of RE models (Zhang et al., 2018). There have been attempts to incorporate both approaches (Guo et al., 2019). However, as far as we are aware, there have been no attempts at incorporating attention over the output embeddings of the GCN. We hypothesise that using attention in this way may enhance the contextualisation capabilities of the GCN model, leading to an increase in performance.

1.1.1 Hypotheses

H1. Including PoS tags or dependency relationships (in addition to named entity tags) during the masking procedure in BERT will yield some performance benefit in downstream sentence-level RE.
H2. Including syntactic features (described in H1) within sentence-level embeddings during fine-tuning will reduce the requisite time for convergence.
H3. Extending the GCN architecture to use attention will increase the performance of the model for longer sentences.

1.2 Contributions

(1) We demonstrate that BERT is efficient at encoding syntactic properties of language by nature of its existing implementation. (2) Adopting an attention mechanism in GCN marginally improves overall performance for the RE task.

2 Related Work

2.1 BERT

Historically, RE has relied on hand-crafted, rule-based systems to extract relations between subject-object pairs (Zhao et al., 2024). More recently, machine learning and deep learning-based approaches have significantly outperformed their historical counterparts. Specifically, fine-tuning transformer-based (Vaswani et al., 2017) models, such as BERT (Devlin et al., 2019), sets state-of-the-art (SOTA) level performance on downstream RE tasks (Zhong and Chen, 2021). Many variations of BERT-BASED models exist; for instance, (Joshi et al., 2020) describe SpanBERT, a BERT-BASED model

that applies the special *[MASK]* token (or cloze test) over spans of text rather than individual tokens. SpanBERT better captures context (syntactic and semantic qualities) over contiguous spans (which entities can take the form of), outperforming BERT on downstream RE. In addition, (Baldini Soares et al., 2019) demonstrated that applying entity markers (to better produce relation representations) to training examples further improves downstream RE. Further adaptations of entity markers: (1) typed entity markers and (2) their *'punct'* variants proposed through (Zhong and Chen, 2021) and (Zhou and Chen, 2022b) also contribute significantly to successfully identifying and extracting relations on datasets such as SemEval (Hendrickx et al., 2010) and TACRED (Zhang et al., 2017).

2.2 GCN

Recent studies have shown the efficacy of using the dependency tree features to enhance RE performance. Such information can be naturally parsed using a graph convolution neural network (GCN). For example, Zhang et al. (2018) applied a GCN over a pruned version of the dependency tree. This enables the model to embed syntactic information within the token embeddings, which benefits the RE task. They also proposed a contextual GCN (C-GCN), which uses a bidirectional LSTM (BiLSTM) layer to generate input embeddings to the GCN. Due to the computational overhead BiLSTMs introduce, we focus our investigation on the non-contextual variant of GCN. However, the GCN model performs worse than C-GCN due to its inability to capture semantic context, which LSTMs are able to do. Guo et al. (2019) introduced an attention-guided GCN mechanism for adding syntactic context to the embeddings. We take a similar approach to the problem by using attention (Vaswani et al., 2017) to increase the amount of context contained within each token embedding.

3 Methodology

3.1 Data

TACRED is a large-scale, crowd-sourced relation extraction (RE) dataset built as an alternative to distant supervision in Knowledge Base Population (KBP) "slot-filling" tasks. While both datasets capture many of the same relation types, Re-TACRED re-annotates TACRED (see Table 1) by removing examples containing partial entity spans, multilingual sentences and culturally ambiguous references. As Re-TACRED consistently provides more accurate benchmarking scores for RE than TACRED, we opt to use it within the context of our experiments.

Table 1: Comparing TACRED & Re-TACRED

Dataset	Total	Train	Dev	Test
TACRED	106,264	68,124	22,631	15,509
Re-TACRED	92,303	68,124	10,761	13,418

3.2 BERT

Our experiments extend the entity marking method proposed by Zhou and Chen (2022a), who demonstrated that marking entities using existing special tokens improved downstream performance. Rather than masking entity tokens, this approach inserts special markers around entities, clearly defining their boundaries.

We further enhance these markers by integrating additional syntactic features:

- **PoS-based marking**: The first part-of-speech (PoS) tag of the first word of the entity is embedded within the markers.
- **Dependency-based marking**: We integrate the set of dependency relations *(deprel)* of the entity within the marker.

Both PoS tags and dependency relation sets were obtained directly from annotations provided in the original TACRED and Re-TACRED datasets. These enhancements are tested independently to evaluate their respective contributions clearly. The modified entity marking schemes are defined as follows:

- PoS-based marking:
  `@` ⋆ *NER-type* **+first PoS tag** ⋆ subject `@` ... # ^ *NER-type* **+first PoS tag** ^ object #
- Dependency-based marking:
  `@` ⋆ *NER-type* **+deprel set** ⋆ subject `@` ... # ^ *NER-type* **+deprel set** ^ object #

For instance, given the sentence: "Tim Cook is the CEO of Apple." The entities could be marked with enhanced entity markers as follows:

- PoS-based marking:
  `@` ⋆ *PERSON*+**NNP** ⋆ `Tim Cook@ is the CEO of #` ^ *ORG*+**NNP** ^ `Apple#.`

- Dependency-based marking:
  `@ *` *PERSON*+**nsubj** `* Tim Cook@ is the CEO of # ^` *ORG*+**pobj** `^ Apple#.`

3.3 GCN

We propose an extension to the GCN architecture utilising a Scaled Dot-Product Attention (Vaswani et al., 2017) layer over the outputs of GCN. We hypothesise that this should improve the model's ability to reason over lengthier sequences since GCN can only propagate information over fixed distance L, where L is the number of layers of the GCN (Zhang et al., 2018). Both, LSTMs and attention can play similar roles in language modeling - they enhance the existing embedding with the context of the sequence. Therefore, we hypothesise that our approach should improve the GCN model in a similar way C-GCN does.

The first part of our model remains unchanged from the original. We have $\mathbf{x}_1, \ldots, \mathbf{x}_n$ as our non-contextualised GloVe embeddings (Pennington et al., 2014) for tokens in the pruned dependency tree. The subject is defined as a span within the original sequence $[\mathbf{x}_{s_1}, \ldots, \mathbf{x}_{s_2}]$, where (s_1, s_2) denotes its start and end index. Similar notation is applied to the object. These embeddings are then updated with syntactic information using the GCN model, $[\mathbf{h}_1, \ldots, \mathbf{h}_n] = \mathbf{H} = \text{GCN}([\mathbf{x}_1, \ldots, \mathbf{x}_n])$. Feature-wise max-pooling is applied to get the subject, object and sentence embeddings, as shown in the following equations:

$$\mathbf{h}_{\text{subj}} = \text{pool}_{\max}(\left[\mathbf{h}_{s_1}, \ldots, \mathbf{h}_{s_2}\right]) \quad (1)$$

$$\mathbf{h}_{\text{obj}} = \text{pool}_{\max}(\left[\mathbf{h}_{o_1}, \ldots, \mathbf{h}_{o_2}\right]) \quad (2)$$

$$\mathbf{h}_{\text{sent}} = \text{pool}_{\max}(\left[\mathbf{h}_1, \ldots, \mathbf{h}_n\right]) \quad (3)$$

Here, we introduce our adaptations to the approach. Instead of using $\mathbf{h}_{\text{subj}}, \mathbf{h}_{\text{obj}}$, and $\mathbf{h}_{\text{sent}}$ as input to a multi-layer perceptron (MLP), we propose an additional processing step using attention. To allow the model to reason about the ordering of the tokens, we apply RoPE embeddings (Su et al., 2023) to the sequence. Since the word embeddings $\mathbf{H}$ contain syntactic information, applying attention at this stage enables the model to utilise both syntactic and semantic context. This allows the dependency tree information to be propagated further than L edges, and lead to increase in performance for longer sentences. We use the subject, object and sentence embeddings as queries, and the sequence of vectors $\mathbf{H}$ as keys and values in an attention mechanism.

$$\tilde{\mathbf{h}}_{\text{i}} = \text{Attention}(\mathbf{h}_{\text{i}}, H, H) : \forall i \in \{\text{subj}, \text{obj}, \text{sent}\}$$

Further steps remain unmodified from the original implementation of GCN. Namely, we concatenate these embeddings and apply an MLP classification layer to get our relation estimate, $\hat{r} = \text{Softmax}(\text{MLP}(\left[\tilde{\mathbf{h}}_{\text{subj}}; \tilde{\mathbf{h}}_{\text{obj}}; \tilde{\mathbf{h}}_{\text{sent}}\right]))$.

3.4 Evaluation

To measure the performance of the models, we used the micro-averaged F1 score[1] as is common in related work (Stoica et al. (2021), Guo et al. (2019)). We use this metric to fine-tune our models on the development set, and compare it to other models from the literature.

4 Results & Analysis

4.1 BERT

F_1 scores and their averages for BERT_{BASE} and BERT_{LARGE} on RE-TACRED and TACRED are shown in Table 2. Measures are provided for methods: *typed entity marker (punct)*, in addition to variations: (1) part-of-speech (*-pos*) tag, and (2) dependency relation (*-deprel*) inclusion.

Effects of Syntactic Markers. Rather than improving the performance as hypothesised (**H1**), we observe a *minor* performance drop when adding part-of-speech tags and a further reduction when concatenating dependency relations. While these deviations are borderline negligible, the pattern is consistent across RE-TACRED and TACRED, suggesting that these variations introduce noise or that the original model is learning sentence-level syntactic representations adequately. Furthermore, introducing multiple tokens (in *-deprel*) versus singular tokens (in *-pos*) results in a more significant decrease in performance, further disproving our original hypothesis (**H1**). This could be attributed to the highly imbalanced distribution of syntactic markers in the dataset, where NNP (Proper Noun, Singular) appears far more often than other PoS tags and compound relations are unusually frequent in dependency structures.

Impact on Time to Convergence. Figure 1 shows the performance (dev F_1 score) as a function of the timestep during training. It is clearly

[1]We evaluated model performance using the micro-averaged F1 score by excluding the *no_relation* class to ensure a fair comparison, as is done in related work.

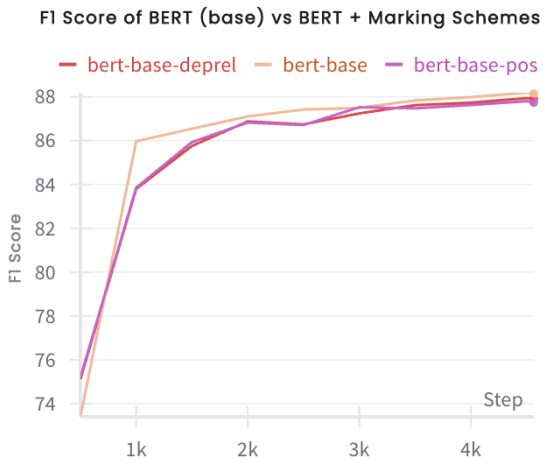

Figure 1: F1 vs timestep for BERT_{BASE}

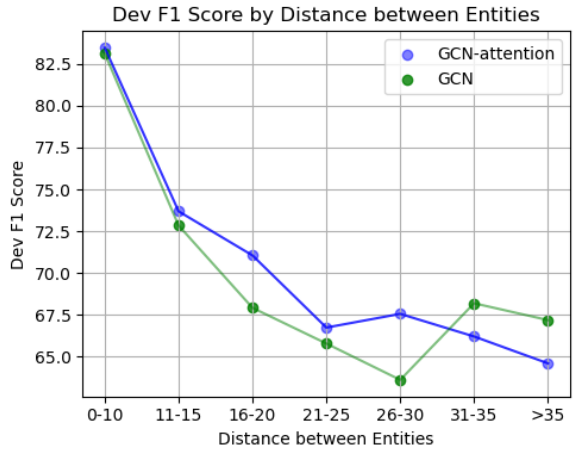

Figure 2: Dev set performance against the distance between the entities in the sentence for C-GCN and the GCN-attention models

shown that $\text{BERT}_{\text{BASE}}$ more steadily converges and improves over a shorter timestep. Again, further reinforcing the problems discussed above.

4.2 GCN

As shown in Figure 2, our GCN-attention model outperformed the original GCN for sentences, where the entities are of medium distance apart. However, for entities far apart, our model performed worse than GCN. This disproves our hypothesis that utilising attention after GCN should allow the model to parse longer dependencies better. We are currently unsure as to what could have caused this performance degradation; further research is required.

Table 2 illustrates the performance benefits the attention mechanism can bring to a GCN model. However, our proposed attention-based adaptation to the GCN architecture does not manage to outperform the BiLSTM-based C-GCN model. Although our attention-based GCN performs slightly worse than the C-GCN model (Table 2), the computational cost to train the model is significantly lower. This is due to the marginal computational overhead introduced by attention relative to BiLSTM.

Dataset	Model	F1(1)	F1(2)	F1(3)	Mean F1
Re-TACRED	bert-large	89.77	89.67	89.51	89.65
	bert-large-pos	89.96	89.46	89.41	89.61
	bert-base	87.93	88.29	88.00	88.08
	bert-base-pos	88.12	87.75	88.16	88.01
	bert-base-deprel	88.09	87.92	87.80	87.94
Re-TACRED	GCN (baseline) *	76.30	-	-	76.30
	GCN-attention *	77.70	-	-	77.70
	C-GCN +	80.30	-	-	80.30
TACRED	bert-large	72.19	73.21	72.34	72.58
	bert-large-pos	73.09	72.65	72.01	72.58
	bert-base	70.93	71.36	70.86	71.05
	bert-base-pos	70.36	71.02	71.38	70.92
	bert-base-deprel	70.14	71.19	70.40	70.57

Table 2: Model F1 scores (repeated runs) and their respective means. * marks the GCN models that were trained only once due to time constraints. + marks the C-GCN model that was evaluated by the authors of Re-TACRED (Stoica et al., 2021).

5 Conclusion & Future Work

As expected, the transformer-based model outperforms the GCN-based approach in all cases, as shown in Table 2 and supported by contemporary literature (discussed in Section 2).

We adapted the GCN architecture for RE by including an attention mechanism. We showed marginal performance benefits over the original GCN architecture. Although the performance of our proposed method does not surpass that of C-GCN, the computational requirements to train our model are significantly lower. Further investigation is required to achieve SOTA performance for GCN-based models.

We incorporated PoS tags and dependency relations into BERT-based entity markers for relation extraction. While adding PoS tags led to a small performance drop, and appending dependency relations reduced performance further, the overall impact was minor. These enhancements may still be beneficial for specific domains—such as biomedical, legal, or scientific text—where explicit syntactic cues could be more beneficial.

References

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhijiang Guo, Yan Zhang, and Wei Lu. 2019. Attention guided graph convolutional networks for relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 241–251, Florence, Italy. Association for Computational Linguistics.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden. Association for Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

George Stoica, Emmanouil Antonios Platanios, and Barnabás Póczos. 2021. Re-tacred: Addressing shortcomings of the tacred dataset.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2023. Roformer: Enhanced transformer with rotary position embedding.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Yuhao Zhang, Peng Qi, and Christopher D. Manning. 2018. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2205–2215, Brussels, Belgium. Association for Computational Linguistics.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark. Association for Computational Linguistics.

Xiaoyan Zhao, Yang Deng, Min Yang, Lingzhi Wang, Rui Zhang, Hong Cheng, Wai Lam, Ying Shen, and Ruifeng Xu. 2024. A comprehensive survey on relation extraction: Recent advances and new frontiers.

Zexuan Zhong and Danqi Chen. 2021. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61, Online. Association for Computational Linguistics.

Wenxuan Zhou and Muhao Chen. 2022a. An improved baseline for sentence-level relation extraction. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 161–168, Online only. Association for Computational Linguistics.

Wenxuan Zhou and Muhao Chen. 2022b. An Improved Baseline for Sentence-level Relation Extraction. ArXiv:2102.01373 [cs] version: 4.

A BERT Fine-tuning Configuration

We use standard fine-tuning settings for BERT-based models. Specifically, we fine-tune both baseline and enhanced entity marking models using a learning rate of 5e-5, batch size of 64, and Adam optimizer over 5 epochs. Each experiment is repeated three times, with results averaged to ensure statistical reliability. Models are evaluated using the F1-score metric on both TACRED and Re-TACRED datasets.