Punctuation restoration improves structure understanding without supervision

We propose punctuation restoration (PR) as an unsupervised objective that improves transformer-based language models' syntactic, semantic structure understanding. We observe that PR improves structure understanding. **Introduction**. The modern natural language processing paradigm centered around pre-trained language models (PLMs; Peters et al., 2018, Radford et al., 2018; Devlin et al., 2019) trained on masked language modeling (MLM) and autoregressive language modeling is a powerful method to approach various problems in natural language processing. However, the following suggests there is room for improvement in current language models' structure understanding needed to perform structure-related tasks reliably, robustly.

1. The **reversal curse** where language models fail to infer "B is A" from "A is B" (Kitouni et al., 2024).
2. The **curse of performance instability**, where model checkpoint initialization and training dataset order strongly affects syntactic structure sensitivity (Zhou et al., 2020).
3. **Poor out-of-distribution generalization** in structural tasks. Systems report close-to-human-baseline performance on one dataset yet perform poorly on others representing the same task due to their picking up **spurious correlations** rather than learning the task (McCoy et al., 2020)
4. Structure hints improve model task performance. PLMs perform better when input is reinforced with structure information, such as syntactic parses (Zhang et al., 2019; He et al., 2020; Sachan et al., 2021) or semantic dependencies (Wu et al., 2021), in the form of additional neural network layers. This suggests structure encoding in LMs is either incomplete or not completely used for predictions.

Background. Efforts to mitigate such shortcoming include data-oriented approaches: syntactic augmentations (Min et al., 2020) and reversing input to mitigate the reversal curse (Golovneva et al. 2024). Architecture-oriented efforts include explicit graph network layers to encode structure, resulting in improvements in benchmark scores (Zhang et al. 2019; Sachan et al. 2021; Wu et al. 2021) and generalization abilities (He et al. 2020). **Our claims.** The methodologies are all human-in-the-loop methods with at least some human input in their design or an automatic tagging system previously trained on human-annotated data. We believe the pre-training stage of current PLMs may be further improved and propose PR as an unsupervised learning objective that improves structure understanding. As seen in Figure 1, punctuations, along with capitalization, often serve as boundary markers between different syntactic components of the sentence. We hypothesize that the model's ability to predict punctuation from plain text entails its ability to encode syntactic boundaries and thus structure, and the additional training in PR, which will involve the model predicting the *target* from the *source* in Figure 1, will translate to improvements in structure-related task performance. Similarly to popular pre-training objectives like MLM, autoregressive language modeling, and next sentence prediction, the objective requires no human input. The objective is also architecture-agnostic and requires no architecture- or language-specific engineering. **Experimental setup.** We gather a set of 16 datasets across 7 tasks: NER, OIE, SBD, chunking, POS tagging, SRL, and relation classification (RE). We evaluate on in-distribution (test set from the same source as training set) and out-of-distribution performance; in discriminative (classification) and generative settings, and in a multi-task setting. We additionally train the t5-base model (Raffel et al. 2019) on a PR dataset, before fine-tuning the vanilla and PR-trained models. Each performance measurement is an average over 5 initializations. **Results**. We observe higher and performance across the board, in all 7 tasks and 16 datasets and in all settings (in- and out-of-distribution; discriminative, generative, and multi-task), as well as smaller variance across initializations in a stability study. **Discussion.** Our results suggest that current methods for representation learning during the pre-training stage lacks sufficient structural signal, and effective distillation of implicit structural sensitivity via additional training should improve structure understanding. We claim that PR is an effective unsupervised training objective that improves structure understanding of pre-trained language models, directly and indirectly mitigating the four known issues discussed above. Analogous to how prosody helps disambiguate syntactic structure in human speech processing (Price et al., 1991, Kahn et al., 2005), punctuation can be a useful guide in forming a robust computational representation of text. A computational system fully proficient in PR is likely to be proficient in predicting syntactic and semantic structure within said text, even when there are no punctuation marks to be restored. Finally, we claim democratic in that we propose a non-intrusive unsupervised learning objective that is orthogonal to other engineering choices and can be implemented with little computational or human resources.

- Source: lee faker sang-hyeok (hangul: 이상혁) is a league of legends esports player currently mid laner and part owner at t1
- Target: **L**ee **"F**aker" **S**ang-hyeok (**H**angul: 이상혁) is a **L**eague of **L**egends esports player, currently mid laner and part owner at **T**1.

Figure 1