Punctuation restoration improves structure understanding without supervision

Junghyun Min[1], Minho Lee[2], Woochul Lee, Yeonsoo Lee[3]

[1]Department of Linguistics, Georgetown University, Washington, D.C., USA
[2]KT Gen AI Lab, Seocho, Seoul, Republic of Korea
[3]NC AI, Seongnam, Gyeonggi, Republic of Korea

Motivation

Recent gains in NLP tasks that require understanding of syntactic, semantic, and discourse structure is certainly impressive. However, the following suggest there is still room for improvement in current transformer-based language models' abilities to understand, process and linguistic structure.

1. **The reversal or factorization curse**: language models fail to infer "B is A" from "A is B" [1] or their representations are highly dependent on the order of the input [2].
2. **The curse of performance instability**: model checkpoint initialization and training dataset order strongly affect sensitivity to syntactic structure [3].
3. **Poor out-of-distribution generalization**: models report close-to-human performance on one dataset yet perform poorly on other datasets representing the same task, due to their picking up **spurious correlations** rather than learning the task [4].
4. **Insufficient or underutilized structure information**: PLMs are poor few-shot structure predictors [5] and perform better when input is reinforced with linguistic structure information [6].

Introduction

- Pre-trained language models excel in generative tasks but struggle with syntactic and semantic structure understanding.
- We hypothesize that this gap is due to insufficient linguistic structure learning in popular pre-training objectives like MLM.
- Goal: Investigate whether punctuation restoration (PR) as an additional learning objective enhances structure understanding in NLP tasks.
- Structure understanding is measured with test set performance, generalization to other datasets that represent the same task, and performance stability.
- We test on 18 experiments, representing 7 tasks, 12 datasets, and 3 settings (generative, multi-task, and discriminative) in English.

Methods

Objective Design

Restore punctuations . , ? ! " ' and capitalization. Normalize opening and closing quotation marks as one symbol (' and ' as ', " and " as ").

- Input: lee faker sang-hyeok (hangul:이상혁) is a league of legends esports player currently mid laner and part owner at t1
- Output: **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 **T1.**

Our selection of punctuation marks reflect frequency of occurrence and syntactic significance. Because of the lexical overlap between the input x and output y, the model effectively learns to predict capitalization and punctuation information $m = y - x$:

$$m_t = f(x, y_{<t}) = \begin{cases} \phi \\ \mathrm{addPunct}(x_t, \theta) \\ \mathrm{addCap}(x_t, \theta) \end{cases}$$

Experimental Setup

Our experiments involve two stages. In the first stage, we take the pre-trained weights of the T5-base model [7] and perform additional pre-training on the punctuation restoration objective to produce PR-T5. Then, in the second stage, we fine-tune PR-T5 on downstream tasks and datasets. We fine-tune the publicly available pre-trained T5 weights on the same downstream tasks and use their performance as comparison baseline.

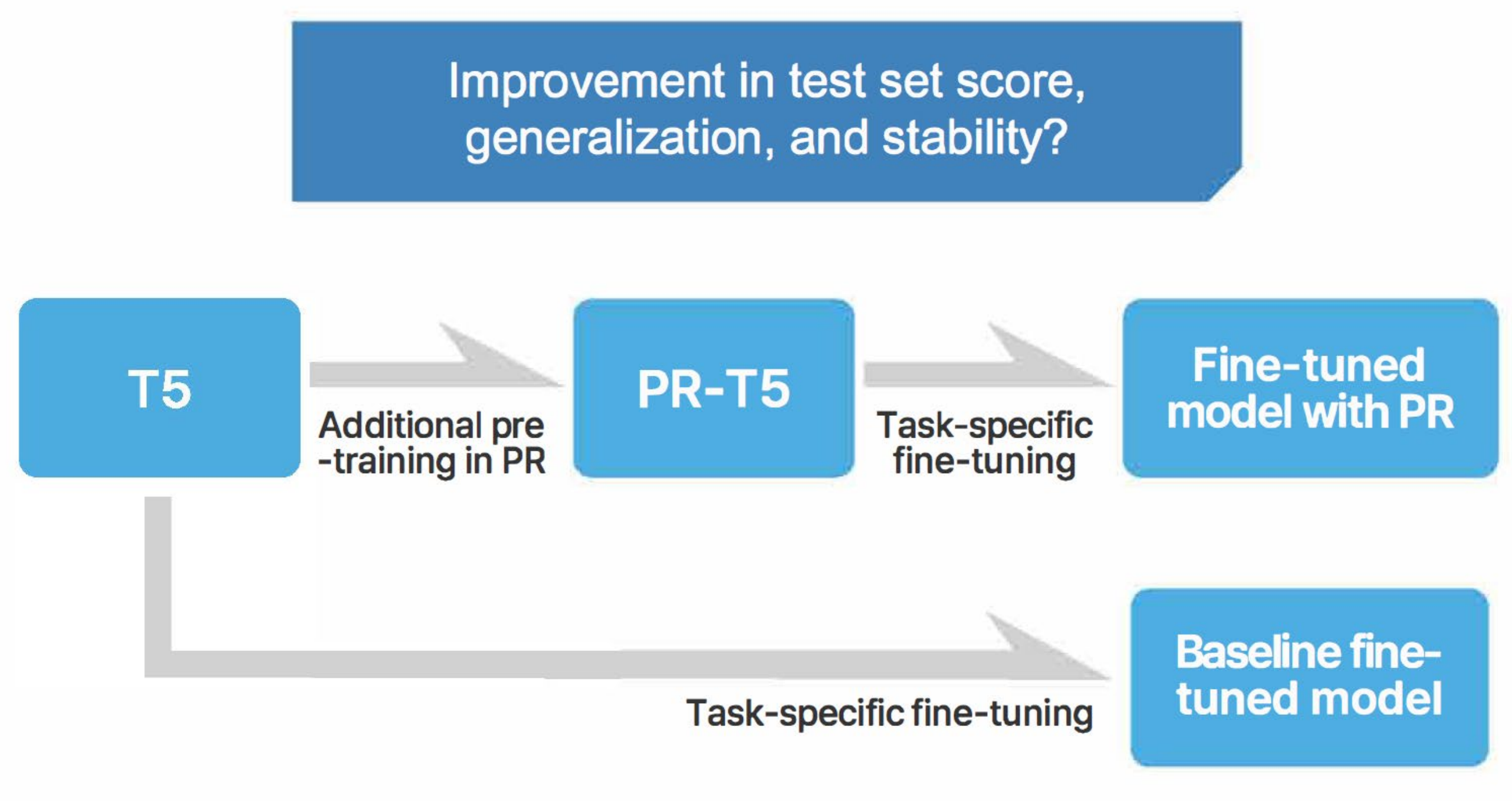
Improvement in test set score, generalization, and stability?
T5
Additional pre -training in PR
PR-T5
Task-specific fine-tuning
Fine-tuned model with PR
Baseline fine-tuned model
Task-specific fine-tuning

Figure 1
A flowchart illustrating our experimental setup.

Results

We observe that additional pre-training in punctuation restoration can enhance representations of linguistic structure, as shown by improved test set scores, out-of-distribution generalization, and stability across initialization.

Out of 18 experiments, we observe ▲≥2%p increase in 16 experiments (6 of 7 tasks), and ▲≥5%p in 10 experiments (5 of 7 tasks). Improvements persist across setting as well, supporting the utility of the objective across architecture and decoding method.

Task	Training set	Evaluation set	t5-base			+ PR			Δ
			P	R	F1	P	R	F1	F1
NER	Econ-mNER	ID	.69	.65	.67	.90	.89	.89	▲.22
		Econ-sNER	.67	.76	.71	.74	.81	.77	▲.06
	GENIA	ID	.57	.73	.64	.64	.76	.69	▲.05
	CoNLL03	ID	.89	.90	.89	.92	.92	.92	▲.03
	ontonotes	ID	.87	.88	.88	.91	.91	.91	▲.03
OpenIE	EconIE-PRO	ID	.47	.43	.45	.60	.63	.62	▲.17
		CaRB	.22	.16	.19	.62	.42	.50	▲.31
	OIE2016	ID	.16	.19	.18	.19	.19	.19	●.01
		CaRB	.10	.15	.12	.26	.27	.27	▲.15
Chunking	CoNLL00	ID	.94	.94	.94	.96	.96	.96	▲.02
		CoNLL03	.41	.41	.41	.41	.42	.42	●.01
SRL	CoNLL12	ID	.75	.79	.77	.84	.86	.85	▲.08
SBD	PTB	ID	.97	.72	.81	.98	.98	.98	▲.17
POS	CoNLL00	ID	.96	.96	.96	.98	.98	.98	▲.02
		CoNLL03	.74	.87	.79	.84	.88	.86	▲.07
RE	TACRED	ID			.67			.83	▲.16

Figure 2
In the generative setting, we observe ▲≥2%p improvements in 14 out of 16 experiments. ID stands for in-distribution and denotes evaluation on test set from the same distribution as training set.

	t5-base (joint)			+ PR			Δ
	P	R	F1	P	R	F1	F1
NER	.86	.84	.85	.87	.86	.87	▲.02
OIE	.57	.60	.58	.60	.62	.61	▲.03

	t5-base (EO)			+ PR			Δ
	P	R	F1	P	R	F1	F1
min	.67	.91	.78	.74	.90	.82	▲.04
max	.88	. 94	.91	.90	.94	.91	●.00
avg	.78	.93	.85	.83	.92	.88	▲.03
sdev	.061	.009	.035	.048	.010	.027	▼.008

Figure 3
In the multitask (join) and discriminative (encoder-only) settings, we observe ▲improvement in in-distribution test score and ▼decrease in min-max range and standard deviation.

Limitations or next steps

Our analysis is limited to base-sized encoder-only and encoder-decoder models for English. Its effects in larger models, implications to generative or conversational systems, and generalization to other languages warrant additional investigation.

It is also likely that punctuation restoration is far from the best or only unsupervised learning objective that improves representation learning. Other forms of unsupervised structure learning, possibly simpler and more effective methods than punctuation restoration, as well as optimizations on objective combination (e.g. with word prediction methods) should be studied in future work.

Conclusion

Our results support our hypothesis that complementing MLM with a more structure-related objective improves structure understanding. Because of the wide range of experiments in which improvement is observed, we interpret this to be a general improvement of structure understanding rather than fortunate task-specific artifacts from the additional training. We conclude that our methods yield a more reliable and robust representation that can be easily implemented and do not interfere with architectural additions.

References

1. Berglund et al., 2023
2. Kitouni et al., 2024
3. Zhou et al., 2020; McCoy et al., 2020; Du and Nguyen, 2023
4. Gururangan et al., 2018; McCoy et al., 2019; Serrano et al., 2023
5. Zhao et al., 2023; Bai et al., 2023
6. Strubell et al., 2018; He et al., 2020; Sachan et al., 2021; Wu et al., 2021; Fei et al., 2021; Xie et al., 2023, Huang et al., 2024
7. Raffel et al., 2019
8. Brown et al., 2020

Acknowledgements

We used OpenAI's GPT-3.5 Turbo [8] as a debugging assistant during the project's technical implementation. A paper under the same title is currently under review. We thank Tatsuya Aoyama and Ethan Wilcox (Georgetown) and Yerang Kim (Samsung SDS) for their helpful comments.