Reproducibility Study of TaCL: Improving BERT Pre-training with Token-aware Contrastive Learning, Association for Computational Linguistics: NAACL 2022

Sai Sandeep Varma Mudundi
smudund@gmu.edu

Asra Naseem
anaseem2@gmu.edu

Rajeev Priyatam Panchadula
rpanchad@gmu.edu

Introduction

- Our Chosen paper presents TaCL, a new continuous pre-training technique for BERT that aims to improve the quality of token representations.
- TaCL's continuous pre-training approach promotes BERT to learn an isotropic and discriminative distribution of token representations.
- Our study addresses gaps in evaluating the model's robustness and multilingual performance by employing CheckList and its Multilingual Checklist for comprehensive behavioral testing.
- Our study aims to build upon the baseline implementation of TaCL BERT by testing its robustness and performance in multiple languages.

Reproducibility

- We fine-tuned the pre-trained TaCL-BERT model on various tasks, experimented with hyperparameters and optimization techniques, and analyzed error and token representations to evaluate robustness and multilingual performance.
- Regarding robustness, the study planned to perform sensitivity analysis with regards to data perturbation and evaluate the model's performance on the "Beyond Accuracy: Behavioral Testing of NLP Models with CheckList."
- We planned to modify the code and experiment with additional languages using TaCL BERT and BERT Multilingual models, create new language datasets, and employ the Multilingual Checklist for comprehensive behavioral testing, all to ensure that our model could perform well on real-world tasks and data.

Reproducibility

- To overcome technical challenges posed by advanced models like TaCL-BERT and BERT, we investigated their architecture, explored various tokenization and preprocessing strategies, and conducted systematic hyperparameter tuning, all while managing substantial computational resources needed to train them on large datasets.
- The study encountered challenges in evaluating the robustness and multilinguality of the models, including the availability of diverse and relevant datasets and specialized tools and techniques required for evaluation, such as implementing the CheckList approach and accessing the Multilingual Checklist code. Nonetheless, the study was able to conduct thorough analyses and identify areas for improvement in the models.
- The study utilized Google Colab Pro and the Hopper cluster at the GMU Office of Research Computing to access powerful GPUs, significantly reducing computation time for training the models.

The Results of the Performance from our reproducibility study on Fine Tuned Tacl BERT , BERT and TaCL Chinese bench mark

Our result:

Dataset	Precision	Recall	F1
MSRA	95.4	95.5	95.4
OntoNotes	81.9	83	82.4
Resume	96.5	96.4	96.4
Weibo	68.4	70.7	69.5
PKU	97	96.4	96.7

Published result:

Dataset	Precision	Recall	F1
MSRA	95.41	95.47	95.44
OntoNotes	81.88	82.98	82.42
Resume	96.48	96.42	96.45
Weibo	68.40	70.73	69.54
PKU	97.04	96.46	96.75

Model	F1 Score(paper)	Exact math (paper)	F1 Score(ours)	Exact math (ours)
Bert(base)	88.5	80.8	88.544	81.1164
TaCL Bert(base)	89.0	81.6	89.1501	81.8448

Table 1: Results

Robustness

We evaluated the TaCL-BERT Chinese model using the CheckList methodology. We analyzed the results in detail to identify the model's strengths and weaknesses in handling real-world data perturbation across various robustness dimensions listed below:

1. Vocabulary + POS (Part-of-Speech)
2. Taxonomy
3. Robustness
4. Fairness
5. Temporal understanding
6. Negation
7. Coreference
8. Semantic Role Labeling (SRL)

Robustness

We also conducted three main test which was listed in the checklist

- MFT (Minimum Functionality Test)
- NV (Invariance Test)
- DIR (Directional Expectation Test)

- We first evaluate the robustness of the TaCL BERT model fine-tuned for the StanfordQuestion Answering Dataset (SQuAD) task using the CheckList methodology.
- The performance comparison of TaCL-BERT and BERT models on various robustness capabilities can be found in Table 1.

Model	Capability	Test Type	Test Name	Test Cases	Failure Rate
TaCL-BERT	Vocabulary	MFT	Comparative Adjectives: More/Less	100	38.0%
BERT	Vocabulary	MFT	Comparative Adjectives: More/Less	100	31.0%
TaCL-BERT	Vocabulary	MFT	Intensifiers and Reducers	100	99.0%
BERT	Vocabulary	MFT	Intensifiers and Reducers	100	94.78%
BERT	Taxonomy	MFT	Size, Shape, Color, Age, Material	100	82.4%
TaCL-BERT	Taxonomy	MFT	Size, Shape, Color, Age, Material	100	80.0%
BERT	Taxonomy	MFT	Professions vs Nationalities	100	49.4%
TaCL-BERT	Taxonomy	MFT	Professions vs Nationalities	100	44.0%
BERT	Taxonomy	MFT	Animal vs Vehicle	100	25.6%
TaCL-BERT	Taxonomy	MFT	Animal vs Vehicle	100	31.0%
BERT	Taxonomy	MFT	Synonyms	100	0.0%
TaCL-BERT	Taxonomy	MFT	Synonyms	100	0.4%
BERT	Taxonomy	MFT	Comparatives and Antonyms	100	67.3%
TaCL-BERT	Taxonomy	MFT	Comparatives and Antonyms	100	65.0%
BERT	Taxonomy	MFT	Comparatives, Intensifiers and Antonyms	100	100.0%
TaCL-BERT	Taxonomy	MFT	Comparatives, Intensifiers and Antonyms	100	100.0%
TaCL-BERT	Robustness	INV	Question typos	100	6.0%
BERT	Robustness	INV	Question typos	100	11.6%
TaCL-BERT	Robustness	INV	Question contractions	100	4.0%
BERT	Robustness	INV	Question contractions	100	3.4%
TaCL-BERT	Robustness	INV	Add random sentence	100	8.0%
BERT	Robustness	INV	Add random sentence	100	9.8%
TaCL-BERT	Temporal	MFT	Change in profession	100	64.2%
BERT	Temporal	MFT	Change in profession	100	41.5%
TaCL-BERT	Temporal	MFT	Understanding before/after	100	87.0%
BERT	Temporal	MFT	Understanding before/after	100	82.9%
TaCL-BERT	Negation	MFT	Negation in context, may or may not be in question	100	68.7%
BERT	Negation	MFT	Negation in context, may or may not be in question	100	67.5%
TaCL-BERT	Negation	MFT	Negation in question only	100	100.0%
BERT	Negation	MFT	Negation in question only	100	100.0%
TaCL-BERT	Coreference	MFT	Basic coref, he / she	100	100.0%
TaCL-BERT	Coreference	MFT	Basic coref, his / her	100	92.0%
TaCL-BERT	Coreference	MFT	Former / Latter	100	100.0%
BERT	Coreference	MFT	Basic coref, he / she	100	100.0%
BERT	Coreference	MFT	Basic coref, his / her	100	91.8%
BERT	Coreference	MFT	Former / Latter	100	100.0%
TaCL-BERT	SRL	MFT	Agent/Object Distinction	100	58.0%
TaCL-BERT	SRL	MFT	Agent/Object Distinction with 3 Agents	100	94.9%
BERT	SRL	MFT	Agent/Object Distinction	100	60.8%
BERT	SRL	MFT	Agent/Object Distinction with 3 Agents	100	95.7%

Table 1: Performance Comparison of TaCL-BERT and BERT Models on Various Robustness Capabilities

Robustness

Additionally, the fairness capability of the models is assessed by examining the difference between the failure rates of male and female professions for TaCL-BERT and BERT models. The results of this fairness test can be found in Table 2

Model	Profession	Fail Men (%)	Fail Women (%)	Count
TaCL-BERT	CEO	4.3	100.0	255
TaCL-BERT	Doctor	1.2	94.6	241
TaCL-BERT	Nurse	61.6	32.5	268
TaCL-BERT	Secretary	61.0	3.8	236
BERT	CEO	0.17	0.97	267
BERT	Doctor	0.03	0.89	247
BERT	Secretary	0.60	0.04	253
BERT	Nurse	0.58	0.41	233

Table 2: Failures of TaCL-BERT and BERT models across different professions and genders.

Multilinguality

We performed the following tasks for evaluating multilinguality:

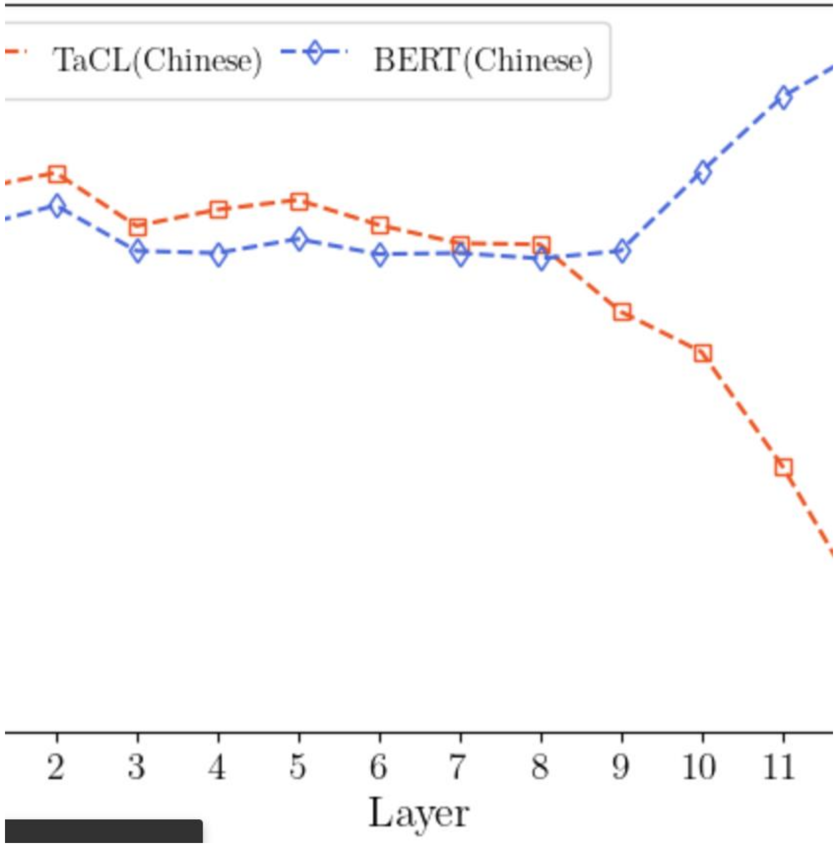
- Intra-sentence Similarity
- Chinese Benchmark Analysis
- Sentiment Analysis
- Multilabel sentence-level Classification Task

Multilinguality

Intra-sentence similarity:

- The study compared the semantic similarity recognition capacities of pre-trained BERT and TACL models at different layers using intra-sentence similarity analysis.
- The average similarity scores for each layer were determined by averaging the cosine similarity scores for all phrase pairs in 50k sentences from the Chinese Wikipedia, divided by the total number of tokens in the file.

In the graph, we can see TaCL model has higher self-similarity scores for initial layers but when comes to the topmost layer its self-similarity score decreases and this suggests that the output of the TCL model is More discriminative.



Multilinguality

Chinese Benchmark:

- The study utilized five Chinese datasets to evaluate the performance of the TaCL Chinese model, and the results varied across different datasets, indicating its ability to adapt to various domains.
- The model's performance might be influenced by dataset size and complexity, which could affect its generalization capabilities.
- The success of the model in the benchmarks can be attributed to the transfer learning and fine-tuning processes, but overfitting and underfitting should be avoided.
- The inherent limitations of the BERT architecture, such as struggling with capturing long-range dependencies, could affect the model's performance on specific tasks.
- Improving the model's robustness requires considering various factors, such as high-quality training data, careful selection of hyperparameters, and multiple evaluation metrics.

Model	**Dataset**	**Precision**	**Recall**	**F1**
TACL BERT Chinese	MSRA	95.4	95.5	95.4
TACL BERT Chinese	OntoNotes	81.9	83	82.4
TACL BERT Chinese	Resume	96.5	96.4	96.4
TACL BERT Chinese	Weibo	68.4	70.7	69.5
TACL BERT Chinese	PKU	97	96.4	96.7
BERT-base-multilingual-uncased	MSRA	94.78	95.47	95.78
BERT-base-multilingual-uncased	OntoNotes	80.27	82.98	81.32
BERT-base-multilingual-uncased	Resume	97.64	97.3	96.78
BERT-base-multilingual-uncased	Weibo	70.15	71.65	69.58
BERT-base-multilingual-uncased	PKU	97.04	95.26	96.23

Table 5: Performance of Chinese Tacl and BERT-base-multilingual-uncased on different Datasets.

Multilinguality

Sentiment Analysis Task:

1. The study performed sentiment analysis on a Hindi review dataset translated into Telugu, Korean, and French to evaluate the performance of TACL BERT-based models in different languages.

2. The performance of the models varied across languages, with higher accuracy in Hindi and French compared to Telugu and Korean due to differences in syntactic and morphological features.

3. The effectiveness of WordPiece tokenization also varied across languages, with Telugu and Korean having more complex subword units that could make it harder for the models to capture relevant information and dependencies.

4. The translation quality of the datasets and the quality and size of the training data subsets also significantly impacted the models' performance.

5. The TACL BERT-based models' fine-tuning for each language might limit their ability to generalize across languages, whereas the BERT-base Multilingual model's architecture allows it to learn shared representations across languages, potentially improving its ability to handle linguistic variations and common challenges.

Model	Hindi Accuracy	Telugu Accuracy	Korean Accuracy	French Accuracy
TaCL BERT	0.8343	0.7648	0.6877	0.8251
BERT-base-multilingual-uncased	0.8797	0.7670	0.8062	0.8317

Table 3: Performance Comparison of TaCL BERT-based and BERT-base Multilingual Models on Sentiment Analysis Task.

Multilinguality

Multilabel sentence-level Classification Task:

- The study conducted experiments on TACL BERT-based models and BERT-base Multilingual models for multilabel sentence-level classification tasks.
- We used an annotated dataset from Homework 2 and translated it into the respective languages for TACL models. The results showed that the BERT-base Multilingual models generally outperformed TACL models, and performance varied across different languages due to factors such as quality and size of training data and linguistic features of the languages.
- Prevalent errors and reasons for misclassification were also identified, including label ambiguity and noisy translations. Further research and improvements in model architecture, training data, and translation quality could lead to better performance on this task.

Language	Model	Validation Accuracy	Test Accuracy
Hindi	TaCL-hindi	0.8667	0.7
Hindi	BERT-base-multilingual-uncased	0.8667	0.8333
Telugu	TaCL-telugu	0.8	0.9667
Telugu	BERT-base-multilingual-uncased	0.88	0.92
Korean	TaCL-korean	0.7	0.7667
Korean	BERT-base-multilingual-uncased	0.8333	0.8
French	TaCL-french	0.7333	0.7667
French	BERT-base-multilingual-uncased	0.83	0.86

Table 4: Performance Comparison of TaCL BERT-based and BERT-base Multilingual Models on Multilabel sentence-level Classification Task.

Error Analysis

We performed error analysis for the results in the Multilabel sentence-level Classification Task:

Error Analysis of TACL Model:

The study identified four major factors that affect the performance of TACL BERT-based models on multilabel sentence-level classification tasks: ambiguity in context, similarity between labels, insufficient training data, and the complexity of multilabel classification. The model struggles with understanding ambiguous contexts, differentiating between labels with similar themes, and the lack of enough training data to capture specific nuances and patterns within each language. The complexity of the task itself, as multiple labels could be assigned to a single text, also impacts the model's performance. The study provides examples of misclassification for each factor and suggests that further research and improvements in model architecture, training data, and translation quality could lead to better performance on this task.

Error Analysis of BERT-base-multilingual-uncased Model:

- Vocabulary limitations: The model might have limited vocabularies in certain languages, which could affect its ability to understand the context and make accurate predictions. Example: BERT's performance in the Telugu language was lower than TACL, possibly due to vocabulary limitations.
- Inability to capture context: The model may struggle to capture long-range dependencies and contextual information, especially in languages with complex sentence structures. Example: In the Korean language, BERT's performance was lower than TACL, possibly due to difficulties in capturing context.

Areas where TACL outperforms BERT-base-multilingual-uncased:

- Validation accuracy: TACL consistently demonstrates higher validation accuracy across languages. Example: In the Hindi language, both TACL and BERT achieved the same validation accuracy, but TACL outperformed BERT in other languages.
- Handling complex grammar: TACL appears to be better at handling languages with more complex grammar and sentence structures. Example: In the Telugu language, TACL achieved a higher test accuracy than BERT.
- Both models face challenges in handling ambiguity and understanding context accurately. Misinterpretation of context also led to errors, such as when the models failed to differentiate between similar themes like "immigration" and "security." The performance of both models could be improved by fine-tuning, incorporating additional training data, and using language-specific models.

Conclusion

We successfully replicated the results of the research paper and focused on enhancing the baseline implementation of their models by exploring robustness and multilinguality. They evaluated the performance of TACL BERT and BERT-base-multilingual-uncased models and identified areas where the models excel and where improvements can be made. TACL BERT outperformed BERT-base-multilingual-uncased in some aspects, such as handling complex grammar and achieving higher validation accuracy in certain cases, but BERT's performance was superior in other instances, highlighting the need for further investigation and improvements. Both models faced challenges in handling ambiguity, understanding context accurately, and differentiating between similar themes in the multilabel sentence-level classification task. The models' robustness was tested against real-world data, indicating the need for more robust models. The authors suggest further exploration and development of the BERT-base--uncased model to improve its performance when dealing with real-world, noisy data. In future work, we aim to address the following aspects to improve the models' performance:

- Fine-tuning
- Language-specific models:
- Incorporating linguistic knowledge:
- Multimodal approaches:
- Transfer learning

Github-link: https://github.com/sandeep-varma8029/ANLP_Final_Project_TaCL_BERT_Checkpoint_2/tree/main