TaCL: Improving BERT Pre-training with Token-aware Contrastive Learning

Sai Sandeep Varma Mudundi
smudund@gmu.edu

Asra Naseem
anaseem2@gmu.edu

Rajeev Priyatam Panchadula
rpanchad@gmu.edu

1 Introduction

1.1 Task / Research Question Description

We worked on a paper that presents TaCL (Su et al., 2022b)(Token-aware Contrastive Learning), a revolutionary continuous pre-training technique that promotes BERT (Devlin et al., 2019) to learn a token representation distribution that is isotropic and discriminative.TaCL is completely unsupervised and does not require any further data. Many current language models that have been pre-trained with MLM objectives suffer from anisotropy (Ethayarajh, 2019). That is, their token representations are limited to a small subset of the representation space, making them less discriminative and less effective at capturing the semantic distinctions between unique tokens. The central claim of the paper talks about how TaCL's continual pre-training approach aims to encourage BERT to learn an isotropic and discriminative distribution of token representations.The paper we based our study on focuses on how TaCL's continuous pre-training approach can promote BERT to learn an isotropic and discriminative distribution of token representations. However, the paper does not explore robustness evaluation or testing the model on multiple languages, nor does it perform any checklist to evaluate the model's behavior.

To address these gaps in the original paper, we focused primarily on robustness and multilinguality by using the "Beyond Accuracy: Behavioral Testing of NLP Models with CheckList" (Ribeiro et al., 2020)and its accompanying code and python package. We evaluated the model's performance on this checklist and explored the usage of the Multilingual Checklist to perform behavioral testing on all languages we worked with.

Overall, our study builds upon the baseline implementation of TaCL BERT to explore these new dimensions. By testing the model's robustness and performance in multiple languages, we have taken a step towards more comprehensive and applicable NLP models.

1.2 Motivation and Limitations of existing work

Many ways have been suggested in the field of NLP to learn better sentence-level (Reimers and Gurevych, 2019; Wu et al., 2020; Meng et al., 2021; Liu et al., 2021b; Gao et al., 2021) and lexical-level (Liu et al., 2021a; Vulić et al., 2021). Other NLP applications that have used contrastive learning include NER (Das et al., 2022) and summarisation (Liu and Liu, 2021), knowledge probing for pre-trained language models (Meng et al., 2022), and open-ended text production (Su et al., 2022a) Numerous researchers (Xu et al., 2019; Gururangan et al., 2020) have examined methods to continuously pretrain the model to reduce task- and domain discrepancy between pretrained models and the specific target task. Unlike the paper we are working on, none of these works focuses on how to use contrastive learning to improve general-purpose token-level representations. Our proposed technique, on the other hand, investigates how to use continuous pretraining to directly increase the quality of model representations, which is transferrable and helpful to a wide range of benchmark tasks. Additionally , none of these works focus on the robustness evaluation of models or the multilinguality aspect of NLP tasks. The paper we are based on did not explore these dimensions either. This limitation motivated us to investigate the performance of TaCL BERT and BERT-base-multilingual-uncased models in the domains of robustness and multilinguality. We explored the "Beyond Accuracy: Behavioral Testing of NLP Models with CheckList" (Ribeiro et al., 2020) methodology and tried to implement it but faced challenges, as well as the multilingual checklist approach. Instead, we performed our own experiments by testing both models on sentiment analysis and multilabel sentence classification tasks in different languages to evaluate their robustness and multilinguality.

Our proposed technique aims to directly improve the quality of model representations through continuous pretraining, which is transferable and helpful to a wide range of benchmark tasks. By addressing these limitations and evaluating the models' performance in different domains, we hope to provide insights into how to improve the generalization and applicability of NLP models.

1.3 Proposed Approach

Our initial approach involved leveraging the pre-trained TaCL-BERT model, cambridgeltl/tacl-bert-base-uncased and cambridgeltl/tacl-bert-base-chinese, to fine-tune it on the SQuAD dataset for question-answering tasks to evaluate robustness,wiki dataset to check the chinese benchmark results and msra, resume, weibo, ontonotes, and pku, to evaluate the multilinguality performance of the TaCL Chinese model and dataset from hw2 and a hindi review dataset for multilabel sentence classification and sentiment analysis tasks respectively. Our preliminary ideas included fine-tuning the TaCL-BERT model on SQuAD to improve its performance on context understanding, information extraction, disambiguation, and specificity. Additionally, we experimented with different hyperparameters and optimization techniques to further enhance the model's performance. We also conducted a thorough error analysis to identify specific weaknesses and areas for improvement in the model. To evaluate the effectiveness of our approach, we compared the performance of the fine-tuned TaCL-BERT model with other state-of-the-art models, such as BERT-base-uncased, by performing unit tests. In addition to fine-tuning the TaCL-BERT model on the SQuAD dataset, we also performed Token Representation Self-similarity analysis on the wiki dataset to understand the token representations learned by both TaCL and BERT models to check if TaCL produced a more discriminative distribution of token representations. Finally, we conducted a qualitative analysis by sampling a sentence from Wikipedia and visualizing the self-similarity matrix produced by BERT-base and TaCL-base.We initially proposed to explore the dimensions of robustness and multilinguality in our reproduction project. For robustness, our preliminary idea was to perform sensitivity analysis with regards to data perturbation, as real-world data often contains noise, spelling errors, typos, grammar mistakes, and ambiguity. We planned to study the "Beyond Accuracy: Behavioral Testing of NLP Models with CheckList" (Ribeiro et al., 2020) paper and its accompanying code and python package to determine how our model would fare with regards to this checklist. We also intended to perform other robustness exploration on our model, such as analyzing how it handles adversarial examples and other common types of perturbations.

Regarding multilinguality, our initial plan was to modify the code of the paper we were reproducing to enable experimentation in languages beyond the ones the paper explores. We intended to substitute the English BERT model with multilingual BERT models like mBERT and XLM-R to perform experiments in many languages. We also considered translating the existing datasets with the methods used in Homework 2 to create new datasets in other languages. Finally, we aimed to explore the usage of the Multilingual Checklist to perform behavioral testing on all languages we work with.

Overall, our initial thoughts were to take a comprehensive approach to exploring robustness and multilinguality, combining sensitivity analysis, adversarial examples, and language experimentation to ensure our model can perform well on a variety of real-world tasks and data.

1.4 Likely challenges and mitigations

Working with advanced models like TACL-BERT and BERT presents several technical challenges, such as model architecture comprehension, tokenization, preprocessing, and hyperparameter optimization. Additionally, these models require substantial computational resources for training on large datasets like SQuAD and Wikipedia. To address these challenges, we delve deeper into the models' inner workings, explore various tokenization and preprocessing strategies, and perform systematic hyperparameter tuning. Leveraging resources like Google Colab Pro or research computing clusters will enable us to access powerful GPUs to expedite the training process. We conducted thorough error analysis and unit tests to identify and resolve potential issues. In our project, we faced several challenges in evaluating the robustness and multilinguality of our models. One of the major challenges in evaluating the robustness of our models was the availability of diverse and relevant datasets. While we were able to use existing datasets, such as the SQuAD dataset, to evaluate the performance of our models, these datasets may not fully reflect the diversity of real-world data that our models may encounter. Additionally, evaluating the robustness of our models to data perturbation requires specialized tools and techniques that may not be readily available or easily accessible. In terms of robustness evaluation, we faced some challenges in implementing the CheckList approach. While we did not contact the authors of the original paper, we explored the CheckList code and accompanying python package. However, we encountered some difficulties in implementing the checklist and identifying appropriate data perturbation techniques for our models. Despite these challenges, we were able to conduct a robustness analysis by introducing noise and other perturbations to our datasets and evaluating our models' performance under these conditions.

Regarding multilinguality, we faced some challenges in accessing the Multilingual Checklist code and incorporating it into our experiments. However, we were able to conduct our own experiments by testing TACL BERT and BERT-based-multilingual-uncased models on sentiment analysis and multilabel sentence classification tasks in languages beyond those explored in the original paper. We used existing datasets or translated existing ones to perform these experiments. Overall, while we faced some challenges in evaluating robustness and multilinguality, we were able to conduct thorough analyses and identify areas for improvement in our models.

2 Related Work

PAPER1(Reimers and Gurevych, 2019):Sentence-BERT (SBERT) is a pre-trained BERT network modification that uses siamese and triplet network architectures to create semantically meaningful sentence embeddings that can be compared using cosine-similarity. This decreases the work required to locate the most comparable pair from 65 hours with BERT / RoBERTa to around 5 seconds with SBERT, while keeping BERT's accuracy.PAPER2 (Giorgi et al., 2021): (DeCLUTR) Deep Contrastive Learning for Unsupervised Textual Representations is presented. They carefully develop a self-supervised goal for learning universal language embeddings that does not require labelled training data, inspired by recent breakthroughs in deep metric learning (DML). Their strategy bridges the performance gap between unsupervised and supervised pretraining for universal sentence encoders when used to extend the pretraining of transformer-based language models.PAPER3(Wu et al., 2020):They propose Contrastive LEArning for Sentence Representation (CLEAR), which uses various sentence-level augmentation algorithms to learn noise-invariant sentence representation. Word and span deletion, reordering, and substitution are examples of augmentations.PAPER4(Yan et al., 2021):ConSERT, a Contrastive Framework for Self-Supervised Sentence Representation Transfer, is presented, which uses contrastive learning to fine-tune BERT in an unsupervised and effective manner. ConSERT overcomes the collapse problem of BERT-derived sentence representations by using unlabeled texts, making them more relevant for downstream applications. The Tacl BERT which we are working on is entirely a new approach with benchmark results. The above mentioned papers are relevant to the paper we are working on.Hence, brief comparison of different models is done below. Both token-aware contrastive learning and sentence embeddings using Siamese BERT-networks are effective methods for learning unsupervised textual representations, while the various contrastive learning frameworks introduced in the papers mentioned above show promise for improving performance on downstream tasks such as classification, similarity, and transfer learning. TACL and (CLEAR) both use contrastive learning to improve the quality of sentence representations, but TACL considers the context of individual tokens within a sentence, while (CLEAR) learns a global representation of the sentence. ConSERT, on the other hand, is a method for transferring learned sentence representations from one task to another.

3 Experiments

3.1 Datasets

To investigate the robustness of the TaCL BERT model finetuned for the Stanford Question Answering Dataset (SQuAD) task we used the SQuAD 1.0 dataset. SQuAD consists of 100,000+ question-answer pairs based on Wikipedia articles. We used it to evaluate various Robustness checklist capabilities like robustness, vocabulary, Taxonomy and Fairness etc. For executing the Chinese Benchmark, we utilized five datasets, namely msra, resume, weibo, ontonotes, and pku, to evaluate the performance of the TaCL Chinese model. We computed F1 score, precision, and recall metrics for this model on these datasets. To perform error analysis for intra-sentence similarity on Chinese data, we used 50k data samples from Chinese Wikipedia.

1. **MSRA:** A Chinese Named Entity Recognition dataset with 120,000+ sentences for identifying and classifying named entities. Preprocessed for tokenization, input formatting, and data splitting. Available through public sources and Hugging Face datasets library.

2. **Resume:** A collection of 1,000+ anonymized English resumes for extracting key information such as work experience, education, and skills. Preprocessed and accessible through public sources and Hugging Face datasets library.

3. **Weibo:** A large-scale dataset from the Chinese microblogging platform Sina Weibo for various NLP tasks, including sentiment analysis and topic modeling. Preprocessed and available through public sources and Hugging Face datasets library.

4. **OntoNotes:** A vast, multilingual dataset for NLP tasks, such as coreference resolution and named entity recognition, containing 1.7 million words of annotated text. Preprocessed and accessible through public sources and Hugging Face datasets library.

5. **PKU:** A Chinese language dataset for word segmentation tasks with 19,000+ manually annotated sentences. Preprocessed for consistency and available through public sources and Hugging Face datasets library.

Additionally, we utilized 50,000 data samples from Chinese Wikipedia for unit test case analysis, focusing on intra-sentence similarity tasks. All the datasets are openly accessible, with well-structured train/dev/test splits provided in separate files.

For evaluating the multilingual capability of Tacl-BERT mode ,the dataset utilized for sentiment analysis task is a collection of 3,500 review sentences originally designed for sentiment analysis in the Hindi language. This dataset is part of the Hindi BERT project, which is aimed at developing a BERT model specifically tailored to understanding and processing Hindi text. The sentences in the dataset cover a wide range of topics, which enables the model to recognize various sentiments and nuances in the language.

To evaluate the performance of models in other languages, the dataset was translated into Telugu, Korean, and French. This translation process allowed for the creation of parallel datasets for each language, enabling researchers to assess the effectiveness of their respective models in handling sentiment analysis tasks.

The translated datasets maintain the same number of sentences and overall structure, providing a consistent and reliable basis for comparing the models' performances across different languages. For multilabel sentence-level classification task, we used the annotated dataset from Homework 2 (HW2) containing over 500 rows, each representing a sentence and its corresponding news article category. The dataset features 15 different labels, which include None, Economic, Capacity and Resources, Morality, Fairness and Equality, Legality, Constitutionality, Jurisdiction, Policy Prescription and Evaluation, Crime and Punishment, Security and Defense, Health and Safety, Quality of Life, Cultural Identity, Public Sentiment, and Political.The original dataset contained news articles labeled with one or more of the 15 categories. To create language-specific datasets for the TACL models, the news articles were translated into the target languages: Hindi, Telugu, Korean, and French

3.2 Implementation

The code of the reproducibility study is available in the ANLP Final Project TaCL BERT Checkpoint 2 GitHub repository.

We perform a comprehensive evaluation using the award-winning CheckList methodology and focus on various robustness dimensions. The results are analyzed and discussed in detail to understand the strengths and weaknesses of the model in handling real-world data perturbations.

The Checklist is a framework designed to evaluate the performance of natural language processing (NLP) models across a wide range of linguistic capabilities. In the given list, several linguistic phenomena and aspects are mentioned. Below, we provide descriptions of each and then give examples of the three main types of tests used in the Checklist framework: MFT (Minimum Functionality Test), INV (Invariance Test), and DIR (Directional Expectation Test).

1. *Vocabulary + POS (Part-of-Speech)*: This refers to understanding the meaning of words and their grammatical roles in a sentence. A test may check if the model can recognize and differentiate between nouns, verbs, adjectives, etc.

2. *Taxonomy*: This involves the model's ability to recognize synonyms, antonyms, and word categories. A test may verify if the model can identify similar or opposite meanings and classify words into appropriate groups.

3. *Robustness*: This refers to the model's tolerance to typos, irrelevant additions, contractions, and other variations in the input text. A test may check if the model can still understand the meaning of a sentence despite such alterations.

4. *Fairness*: This refers to the model's unbiased treatment of different groups of people, topics, or perspectives. A test may investigate if the model displays any bias based on gender, race, or other factors.

5. *Temporal understanding*: This involves the model's ability to understand the order of events and how they impact the task. A test may check if the model can correctly process sentences with time-based information.

6. *Negation*: This refers to the model's ability to understand negations in the text. A test may assess if the model can recognize and process sentences that include negation words like "not" or "never."

7. *Coreference*: This relates to the model's ability to identify when two or more words in a text refer to the same entity. A test may check if the model can resolve pronouns or other referring expressions.

8. *Semantic Role Labeling (SRL)*: This involves the model's ability to understand roles such as agent, object, passive/active, etc. A test may assess if the model can identify the roles of words or phrases in a sentence.

Now, let's discuss the three main types of tests used in the Checklist framework:

1. MFT (Minimum Functionality Test): This test checks if the model can handle basic tasks, often focusing on a single linguistic capability. For example, an MFT for vocabulary + POS might ask the model to identify the part-of-speech of a given word.

2. INV (Invariance Test): This test examines if the model's predictions are consistent when a specific aspect of the input is changed, while the core meaning remains the same. For example, an INV test for robustness might change the order of adjectives in a sentence and check if the model's predictions are unaffected.

3. DIR (Directional Expectation Test): This test assesses if the model's predictions change as expected.when some aspect of the input is modified in a specific direction. For example, a DIR test for temporal understanding might change the tense of a sentence from past to future and evaluate if the model's predictions adapt accordingly.

   We trained the TACL model to evaluate multilinguality capability on four languages—French, Korean, Telugu, and Hindi—using translated Wiki dataset subsets. To facilitate this, we created four huggingface models for each language: `sandeepvarma99/tacl-french`, `sandeepvarma99/tacl-korean`, `sandeepvarma99/tacl-telugu`, and `sandeepvarma99/tacl-hindi`. We then tested the models on sentiment analysis and text classification tasks, performing error analysis. Finally, we compared the performance of the TACL BERT-based models with BERT-base multilingual uncased models for each language. Additionally, we evaluated the TACL Chinese model using five datasets: msra, resume, weibo, ontonotes, and pku. We computed F1 score, precision, and recall metrics for this model on these datasets. Analyzing the TaCL-Chinese benchmark across different datasets offered valuable insights into the model's performance and identified potential areas for improvement.

3.3 Results

In this study, we first evaluate the robustness of the TaCL BERT model fine-tuned for the Stanford Question Answering Dataset (SQuAD) task using the CheckList methodology. This comprehensive evaluation covers various linguistic dimensions, such as vocabulary, taxonomy, and semantic role labeling. The CheckList framework employs three main types of tests: Minimum Functionality Test (MFT), Invariance Test (INV), and Directional Expectation Test (DIR), to examine the model's performance in handling different linguistic phenomena and aspects. A detailed analysis of the results helps identify the strengths and weaknesses of the model in handling real-world data perturbations. The performance comparison of TaCL-BERT and BERT models on various robustness capabilities can be found in Table 1. 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.

Next, we focus on evaluating the models on Multilinguality domain.Table 5 displays the results of fine-tuning TaCL-BERT Chinese and BERT multilingual on five separate datasets: MSRA, OntoNotes, Resume, Weibo, and PKU. The model's performance was evaluated using Precision, Recall, and F1 score metrics. In our experiments, we conducted two different tasks to evaluate the performance of the models in terms of multilinguality. The first task was a Sentiment Analysis Task, and the second task was a Multilabel Sentence-level Classification Task. These tasks were designed to test the performance of the TACL BERT-based models and the BERT-base Multilingual model on different languages and compare their effectiveness in handling multilingual input. The results of these experiments can be found in Table 3 and Table 4, demonstrating the capabilities of each model and providing insights into their strengths and weaknesses in processing multilingual text.

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

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.

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.

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.

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.

3.4 Discussion

In this section, we discuss the experimental results in detail to analyze the robustness and multilingual capabilities of TaCL-BERT and compare its performance with the BERT model.

3.4.1 Comparative Analysis of Robustness Experimental Results: TaCL-BERT vs. BERT

From the results (see Table 1), we can observe the following:

1. Vocabulary Capability

From the results (see Table 1), we can observe the following:

1. For the ”Comparative Adjectives: More/Less test”, BERT performs better than TaCL-BERT, with a fail rate of 31.0% compared to TaCL-BERT’s 38.0%. This suggests that BERT better understands comparative adjectives and their corresponding ”less” and ”more” concepts.

2. Both models struggle for the test involving intensifiers and reducers, with TaCL-BERT failing in 99.0% of test cases and BERT in 94.78%. This indicates that both models have difficulty understanding the semantic relationship between intensifiers, reducers, and the adjectives they modify.

Analyzing specific examples from the results

1. For the ”Comparative Adjectives: More/Less test”:

   - TaCL-BERT example:
     C: Emily is younger than Anne.
     Q: Who is less young?
     A: Anne
     P: Emily
   - BERT example:
     C: Jonathan is older than Olivia.
     Q: Who is less old?
     A: Olivia
     P: Jonathan

   In these examples, both models fail to understand the concept of ”less” in the context of comparative adjectives. BERT performs slightly better than TaCL-BERT, but both models struggle with this type of question.

2. For the test involving intensifiers and reducers:

   - TaCL-BERT example:
     C: Kenneth is cautious about the project. Grace is super cautious about the project.
     Q: Who is most cautious about the project?
     A: Grace
     P: Kenneth
   - BERT example:
     C: John is happy about the project.
     Maria is very happy about the project.
     Q: Who is least happy about the project?
     A: John
     P: Maria

   In these examples, both models fail to understand the impact of intensifiers (e.g., ”super” and ”very”) on the degree of the adjective. Both TaCL-BERT and BERT struggle with this type of question, with a high fail rate for both models.

Possible factors contributing to performance differences

The differences in performance between TaCL-BERT and BERT could be attributed to several factors:

1. TaCL-BERT might be overfitting to the training data, while BERT could have been exposed to a more diverse range of examples during training, helping it better understand comparative adjectives and their corresponding ”less” and ”more” concepts.

2. The fine-tuning process used for TaCL-BERT might not have been optimal for these particular tasks, leading to a decline in performance compared to BERT.

3. Both models struggle with intensifiers and reducers, which could be due to the inherent complexity of the task. It is also possible that the training data for both models lacked enough examples involving intensifiers and reducers, leading to poor performance in these tasks.

In conclusion, while TaCL-BERT demonstrates weaknesses in handling the two tests related to vocabulary, BERT shows better performance in the ”Comparative Adjectives: More/Less” test. However, both models struggle with intensifiers and reducers.

2. Taxonomy Capability

Analyzing specific examples from the results

1. For the ”Size, Shape, Color, Age, Material” test:
   C: There is a big yellow figure in the room.
   Q: What size is the figure?
   A: big
   P: big yellow
   TaCL-BERT fails to identify the correct attribute (size) and instead combines it with the color attribute.

2. For the ”Professions vs Nationalities” test:
   C: Alice is a Russian organizer.
   Q: What is Alice's job?
   A: organizer
   P: Russian organizer
   TaCL-BERT fails to distinguish between the profession and nationality attributes, and instead, it outputs the combined attribute.

3. For the ”Animal vs Vehicle” test:
   C: Christine has a train and a guinea pig.
   Q: What animal does Christine have?
   A: guinea pig
   P: a train and a guinea pig
   TaCL-BERT fails to identify the correct animal and includes the vehicle in its response.

4. For the ”Synonyms” test:
   C: Frederick is very thankful. Fred is very outspoken.
   Q: Who is vocal?
   A: Fred
   P: N/A
   In this test, TaCL-BERT performed well, with only a 0.4% fail rate. However, in the example above, it failed to identify the synonym ”vocal” for ”outspoken”.

5. For the ”Comparatives and Antonyms” test:
   C: Jimmy is smarter than William.
   Q: Who is dumber?
   A: William
   P: Jimmy is smarter than William
   TaCL-BERT fails to identify the correct antonym for ”smarter” (dumber) and instead repeats the context in its response.

6. For the ”Comparatives, Intensifiers, and Antonyms” test:
   C: Helen is more visible than Richard.
   Q: Who is more invisible?
   A: Richard
   P: Helen
   TaCL-BERT fails to correctly identify the comparative and antonym relationship between ”more visible” and ”more invisible”.

In general, both BERT and TaCL-BERT models have a similar performance across different taxonomy tests.

1. For the ”Size, Shape, Color, Age, Material” test, BERT has a slightly higher fail rate (82.4%) compared to TaCL-BERT (80.0%).

2. For the ”Professions vs Nationalities” test, BERT has a slightly higher fail rate (49.4%) compared to TaCL-BERT (44.0%).

3. For the ”Animal vs Vehicle” test, BERT performs slightly better, with a fail rate of 25.6%, while TaCL-BERT has a fail rate of 31.0%.

4. For the ”Synonyms” test, BERT outperforms TaCL-BERT with a fail rate of 0.0% compared to TaCL-BERT's 0.4%.

5. For the ”Comparatives and Antonyms” test, BERT has a slightly higher fail rate (67.3%) compared to TaCL-BERT (65.0%).

6. For the ”Comparatives, Intensifiers, and Antonyms” test, both BERT and TaCL-BERT have the same fail rate (100%).

Overall, the performance differences between the two models are relatively small. However, there are some instances where one model outperforms the other, as seen in the ”Animal vs Vehicle” and ”Synonyms” tests.

3.Robustness Capability

Several tests were conducted to assess the robustness capability of the TaCL-BERT and BERT models. These tests included handling questions with typos, questions with contractions, and context paragraphs containing random sentences.

The results show that:

1. For questions with typos, TaCL-BERT demonstrated a lower failure rate (6%) compared to BERT (11.6%), suggesting better robustness in handling typographical errors.

2. For questions with contractions, both models performed similarly, with TaCL-BERT having a slightly higher failure rate (4.0%) compared to BERT (3.4%). This indicates that both models are generally robust in handling contractions.

3. For context paragraphs containing random sentences, TaCL-BERT outperformed BERT, having a lower failure rate (8.0%) compared to BERT (9.8%). This suggests that TaCL-BERT is more robust in handling irrelevant information.

The differences in performance between TaCL-BERT and BERT can be attributed to factors such as:

- TaCL-BERT's fine-tuning process, which may have focused more on handling noisy input effectively.

- The diversity of training data for TaCL-BERT, which might have included more examples with typos and contractions.

- The inherent complexity of dealing with contractions, resulting in generally robust performance for both models.

In conclusion, TaCL-BERT demonstrates superior performance in handling questions with typos and context paragraphs containing random sentences. However, both models perform similarly in handling questions with contractions, indicating their general robustness in dealing with these variations.

4. Temporal Capability:

The tests conducted to evaluate the temporal capability of the models are 'Change in profession' and 'Understanding before/after': From the results, we can observe the following:

Change in profession:

- BERT outperforms TaCL-BERT, suggesting it has a better grasp of the change in professions in the given context.

Understanding before/after:

- Both models struggle to correctly answer questions related to temporal relationships expressed by "before" and "after." However, TaCL-BERT exhibits a slightly higher failure rate compared to BERT.

The differences in performance between TaCL-BERT and BERT could be attributed to several factors:

- BERT's exposure to a more diverse range of examples during training might have helped it better understand temporal relationships in the context of professions.

- The fine-tuning process used for TaCL-BERT might not have been optimal for these tasks, resulting in poorer performance compared to BERT.

- Both models' struggle with understanding "before" and "after" relationships could be due to the inherent complexity of the task or insufficient training data with examples involving these temporal relationships.

In conclusion, while BERT shows better performance in understanding changes in professions, both models have difficulty comprehending temporal relationships expressed by "before" and "after."

5. Negation Capability:

1. Negation in context, may or may not be in question:

TaCL-BERT and BERT perform similarly in this test, with fail rates of 68.7% and 67.5%, respectively. Both models struggle with understanding negation in the context, possibly due to inadequate training data or the inherent difficulty in comprehending negation in natural language.

2. Negation in question only:

Both TaCL-BERT and BERT fail all test cases, indicating that both models have significant difficulty understanding negation when it appears only in the question. This could be due to a lack of training data containing negation in questions or the complexity of recognizing and processing negation in this specific context.

6. Fairness Capability:

From the results in Table 2, we can observe the following:

1. TaCL-BERT model:

The model has significant fairness issues, with high discrepancies in failure rates between men and women for all professions. For CEO and doctor professions, the model fails more frequently for women than for men (100% vs. 4.3% and 94.6% vs. 1.2%, respectively). For nurse and secretary professions, the model fails more frequently for men than for women (61.6% vs. 32.5% and 61.0% vs. 3.8%, respectively).

2. BERT model:

The model has relatively lower fairness issues compared to TaCL-BERT. For CEO and doctor professions, the model has a smaller difference in failure rates between men and women (0.97% vs. 0.17% and 0.89% vs. 0.03%, respectively). For nurse and secretary professions, the model has a smaller difference in failure rates between men and women (0.58% vs. 0.41% and 0.60% vs. 0.04%, respectively).

In conclusion, the TaCL-BERT model has more significant fairness issues compared to the BERT model, with high discrepancies in failure rates between men and women for all professions. On the other hand, the BERT model has relatively lower fairness issues and smaller differences in failure rates between men and women.

7. Coreference Capability:

1. TaCL-BERT model:

The model has significant coreference resolution issues, with high failure rates in all tests. For the ”Basic coref, he / she” and ”Former / Latter” tests, the model has a 100% failure rate. For the ”Basic coref, his / her” test, the model has a 92% failure rate.

2. BERT model:

The model also has significant coreference resolution issues, with high failure rates in all tests. For the ”Basic coref, he / she” and ”Former / Latter” tests, the model has a 100% failure rate. For the ”Basic coref, his / her” test, the model has a 91.8% failure rate.

In conclusion, both TaCL-BERT and BERT models have significant coreference resolution issues, with high failure rates in all tests. This suggests that both models struggle with coreference resolution tasks.

8. SRL Capability:

From the results, we can observe the following:

TaCL-BERT model:

The model struggles with SRL capabilities, as shown by the failure rates in both tests. For the ”Agent/Object Distinction” test, the model has a 58.0% failure rate. For the ”Agent/Object Distinction with 3 Agents” test, the model has a 94.9% failure rate.

BERT model:

The model also struggles with SRL capabilities, as shown by the failure rates in both tests. For the ”Agent/Object Distinction” test, the model has a 60.8% failure rate. For the ”Agent/Object Distinction with 3 Agents” test, the model has a 95.7% failure rate.

In conclusion, both TaCL-BERT and BERT models have significant issues with SRL capabilities, with high failure rates in both tests.

3.4.2 Comparative Analysis of Multilinguality Experimental Results: TaCL-BERT vs. BERT

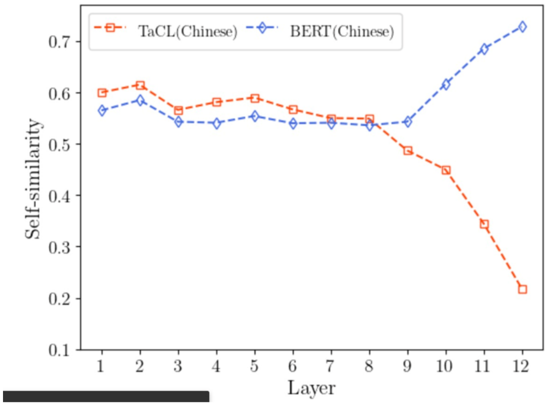
Intra-sentence similarity task Analysis:

We used 50k data from Chinese Wikipedia to perform unit test case analysis for Intra-sentence similarity task.

To understand how well TaCL Chinese model perform unit test case analysis is performed:

Intra-sentence Similarity: By comparing the pre-trained BERT and TACL models’ capacities to recognize the semantic similarity between phrases at various layers, intra-sentence similarity analysis can assist in making distinctions between the two models. Therefore, it may be concluded that the models differ in their capacity to capture the semantic similarity across sentences, then we calculate intra-sentence Similarity for both models and see if there is a substantial difference in the similarity scores across the two models. We initially determined layer-wise intra-sentence similarity for various sentences provided in 50k sentences from the Chinese Wikipedia in order to do this study. The average similarity score for each layer is then determined by averaging the cosine similarity scores for all phrase pairs in the file divided by the total number of tokens in the file.

In Graph above, the x-axis of the graph represents the layers of the models, while the y-axis represents the average self-similarity score of each layer. The self-similarity score indicates the degree of similarity between the different parts of the input text that are being processed by the model.



[redacted]

Figure 1: Layer-wise representation of self-similarity

In the above 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 TaCL model is more discriminative.

Chinese Benchmark Analysis:
We utilized five datasets, namely msra, resume, weibo, ontonotes, and pku, to evaluate the performance of the TaCL Chinese model. We computed F1 score, precision, and recall metrics for this model on these datasets (Table 5). Analyzing the results of the TaCL-Chinese benchmark across different datasets provides valuable insights into the model's performance and potential areas for improvement.

Domain-Specific Performance Variation: The model's performance varies across the different datasets, which is indicative of its ability to adapt to various domains. The MSRA and PKU datasets, which focus on named entity recognition (NER), show the highest F1 scores, suggesting that the model is particularly proficient in this task. However, the model's performance on the Weibo dataset, which is geared toward sentiment analysis and social media text, is lower, indicating that the model might struggle with informal language, slang, or emojis commonly found in social media content.

Impact of Dataset Size and Complexity: The performance variation across datasets may also be influenced by the size and complexity of the training data. Larger datasets with more diverse samples can help the model capture a broader range of linguistic patterns, whereas smaller datasets might limit its generalization capabilities. Furthermore, complex tasks like coreference resolution, which is part of the OntoNotes dataset, may require more sophisticated reasoning abilities that the model has not yet fully developed.

Transfer Learning and Fine-tuning: The success of the TaCL-Chinese model in these benchmarks can be attributed to the transfer learning and fine-tuning processes. Transfer learning allows the model to leverage pre-existing knowledge from pre-training on large-scale corpora, while fine-tuning helps the model adapt to specific tasks and domains. However, it is essential to ensure that the fine-tuning process is carefully executed to avoid overfitting or underfitting, which could negatively impact the model's performance.

Model Architecture Limitations: The TaCL-Chinese model, based on the BERT architecture, is known for its powerful contextual representation capabilities. However, certain inherent limitations of the architecture may affect its performance on specific tasks. For instance, BERT might struggle with capturing long-range dependencies or handling ambiguities in the text. Addressing these limitations through architectural modifications or incorporating auxiliary models could further improve the model's performance.

Improving Model Robustness: To enhance the model's robustness, it is crucial to consider various factors, such as the quality of the training data, the choice of hyperparameters, and the evaluation metrics. Ensuring high-quality training data and carefully selecting hyperparameters can improve the model's generalization capabilities. Additionally, considering multiple evaluation metrics can provide a more comprehensive understanding of the model's performance.

Sentiment Analysis Task:
Dataset: The dataset used for this analysis was based on the Hindi review dataset which consists of 3,500 sentences for sentiment analysis. This dataset was translated into Telugu, Korean, and French to test the performance of the respective models. Based on the outcomes of the sentiment analysis task presented in Table 3, we perform the below analysis.

Analysis:
Language-Specific Features: The TACL BERT-based models exhibited varying performance across the four languages. The higher accuracy in Hindi and French compared to Telugu and Korean can be attributed to several factors. One key factor is the similarity in word order and sentence structure between Hindi and French, both of which are

Indo-European languages, while Telugu (a Dravidian language) and Korean (a language isolate) possess distinct syntactic and morphological features. These linguistic differences might impact the models' ability to effectively encode and process sentence structures, resulting in varying performance.

Tokenization and Subword Units: Another aspect to consider is the tokenization strategy used by the models. BERT-based models employ WordPiece tokenization, which breaks words into subword units. This strategy can help with handling out-of-vocabulary words and capturing morphological information. However, the effectiveness of this approach varies across languages. For instance, Telugu and Korean have rich morphological systems and agglutinative structures, which can lead to longer and more complex subword units. This might make it more challenging for the models to capture relevant information and dependencies in these languages compared to Hindi and French.

Translation Quality: The translation quality of the datasets plays a critical role in determining the models' performance. High-quality translations that maintain the original context, idiomatic expressions, and nuances are essential for training models that can effectively generalize across languages. Any errors, inconsistencies, or subtleties introduced during translation can affect the models' ability to learn and understand the linguistic patterns and characteristics of each language.

Training Data Quality and Size: The quality and size of the translated Wiki dataset subsets used for training can significantly impact the models' performance. A larger, more diverse dataset would better represent the linguistic patterns and characteristics of each language, enabling the models to learn more effectively. It is also worth noting that having a balanced dataset, with equal representation of different languages and domains, can prevent the models from developing biases towards specific languages or topics.

Model Architecture: The TACL BERT-based models leverage the BERT architecture, which utilizes the Transformer model's self-attention mechanism to capture contextual information effectively. However, these models were specifically fine-tuned for each language, which might limit their ability to generalize across languages. The fine-tuning process could result in the models overfitting to the training data and not adequately capturing commonalities between languages. In contrast, the BERT-base Multilingual model is designed to support multiple languages and is pretrained on a larger corpus, which might contribute to its better performance in sentiment analysis tasks across all four languages. Its architecture allows it to learn shared representations across languages, potentially improving its ability to handle linguistic variations and common challenges.

Multilabel sentence-level Classification Task: We performed experiments on the TACL BERT-based models and the BERT-base Multilingual model for multilabel sentence-level classification tasks. We used an annotated dataset from Homework 2 (HW2) with 15 labels corresponding to different news article categories, which were translated into the respective languages for the TACL models.

Dataset Preparation: The original dataset contained news articles labeled with one or more of the 15 categories. To create language-specific datasets for the TACL models, the news articles were translated into the target languages: Hindi, Telugu, Korean, and French.

Evaluation Metrics: To assess the performance of the models on the multilabel sentence-level classification task, we used evaluation metrics such as validation accuracy and test accuracy. These metrics provide insights into the models' ability to correctly assign labels to the sentences in the validation and test datasets. A higher accuracy indicates that the model can more effectively predict the correct labels for each sentence.

Based on the results of the multilabel sentence-level classification task in Table 4, we can observe the following trends in model performance:

BERT-base Multilingual models generally outperform TACL models: In most cases, the BERT-base Multilingual models achieve higher validation and test accuracy scores compared to their TACL counterparts. This can be attributed to the fact that the BERT-base Multilingual models are pretrained on a vast amount of multilingual data, which enables them to capture the underlying linguistic features and nuances across different languages more effectively than the TACL models, which are designed specifically for each target language.

Performance variations across languages: The performance of both TACL and BERT-base Multilingual models vary across different languages. This can be due to several factors:

a. Quality and size of training data: The quality and size of the training data for each language can significantly impact the models' performance. If the translated datasets are smaller or have lower quality translations, the models may struggle to learn the patterns and features necessary for accurate classification.

b. Complexity and linguistic features of the languages: Some languages, such as Korean and Hindi, have more complex scripts, grammar, and linguistic structures compared to French, which has a relatively simpler script and more similarities to English. This can make it harder for the models to learn the features necessary for accurate classification in more complex languages.

c. Imbalance in the distribution of labels: If the distribution of the 15 categories is imbalanced across different languages, the models may struggle to learn the patterns and features necessary for accurate classification of underrepresented categories.

Prevalent errors and reasons for misclassification:

a. Label ambiguity: Some news articles may belong to multiple categories, making it difficult for the models to accurately predict all the correct labels for a given sentence.

b. Insufficient context: Sentence-level classification can be challenging as the models might not have enough context to accurately determine the correct labels. Models trained on larger text spans or entire articles may perform better in this task.

c. Noisy translations: If the translated datasets contain errors or inconsistencies, the models might struggle to learn the patterns and features necessary for accurate classification.

In conclusion, the performance of the models on the multilabel sentence-level classification task can be attributed to factors such as pretraining on multilingual data, quality and size of the training data, complexity and linguistic features of the languages, and inherent challenges in sentence-level classification. Further research and improvements in model architecture, training data, and translation quality could lead to better performance on this task.

3.5 Resources

The cost of our reproduction involved several key components. We 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. Time was spent on fine-tuning, analyzing results, and identifying potential improvements. In order to compare the performance of TaCL BERT and BERT, we conducted experiments on robustness and multilinguality domains.

Training for TaCL and BERT took around four hours for each run on a set of hyperparameters for the SQuAD dataset, while training on the wiki dataset for the self-similarity task took 15 hours. Our team of three members contributed to different aspects of the reproduction process, including model training, analysis, and documentation

3.6 Error Analysis

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

Error Analysis of TACL Model:

1. Ambiguity in context: The model struggles to understand ambiguous contexts accurately, leading to misclassification. Example: In the case of the Hindi language, it failed to correctly differentiate between ”immigration” and ”security” due to the similarity of their themes.

2. Similarity between labels: The model has difficulty differentiating between labels with similar themes, causing misclassification. Example: TACL might have overgeneralized the context in Telugu, leading to confusion between ”economy” and ”employment.”

3. Insufficient training data: The model's classification accuracy is affected by the lack of enough training data to capture specific nuances and patterns within each language. Example: In the Korean language, TACL's performance was lower compared to BERT-base-multilingual-uncased, which might be due to insufficient training data.

4. Complexity of multilabel classification: The task of multilabel classification itself is complex, as multiple labels could be assigned to a single text, increasing the chances of misclassification. Example: In the French language, TACL's performance was lower than expected, possibly due to the inherent complexity of the task.

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

1. 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.

2. 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:

1. 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.

2. 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.

4 Conclusion and Future Work

We were successful in reproducing the results of the Research paper by carefully following the paper's methodology, training, and evaluation processes, we were able to successfully replicate the results for both the TACL and BERT models.In this study, we focused on enhancing the baseline implementation of our models, primarily exploring the dimensions of robustness and multilinguality. We 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.

Our results demonstrated that TACL BERT outperformed BERT-base-multilingual-uncased in some aspects, such as handling complex grammar across different languages and achieving higher validation accuracy in certain cases. However, it is essential to note that TACL BERT did not consistently surpass BERT in all areas. In some instances, BERT's performance was superior, highlighting the need for further investigation and improvements.

Both TACL BERT and BERT-base-multilingual-uncased models faced challenges, such as handling ambiguity, understanding context accurately, and differentiating between similar themes in the multilabel sentence-level classification task. Moreover, their robustness was tested against real-world data, including noise, spelling errors, typos, and grammar mistakes, which led to mixed results and indicated the need for more robust models.

In terms of robustness, the BERT-base–uncased model exhibited sensitivity to data perturbations, such as noise and errors. This suggests that it would benefit from further exploration and development to improve its performance when dealing with real-world, noisy data. Additionally, further research is needed to identify which types of robustness exploration should be performed on the model to yield better results.

In future work, we aim to address the following aspects to improve the models' performance:

Fine-tuning: Further fine-tuning of the models with larger and more diverse datasets could help improve their ability to capture context and handle ambiguities in the text. By incorporating additional training data and focusing on specific nuances and patterns within each language, the models' performance could be enhanced.

Language-specific models: Developing and comparing language-specific BERT models could provide insights into their performance in a more granular context. This would enable researchers to better understand the limitations and strengths of the models for each language, and provide opportunities for improvement.

Incorporating linguistic knowledge: Integrating linguistic knowledge, such as morphological and syntactic information, into the models could help them better understand complex sentence structures and long-range dependencies, leading to more accurate predictions.

Multimodal approaches: Exploring multimodal approaches, which combine text with other data modalities such as audio or images, could provide additional context and improve the models' understanding of the input data.

Transfer learning: Investigating the use of transfer learning techniques could potentially improve the models' performance, especially in scenarios where there is limited training data available for certain languages.

References

Sarkar Snigdha Sarathi Das, Arzoo Katiyar, Rebecca J. Passonneau, and Rui Zhang. 2022. Container: Few-shot named entity recognition via contrastive learning.

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.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2021. DeCLUTR: Deep contrastive learning for unsupervised textual representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 879–895, Online. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021a. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238, Online. Association for Computational Linguistics.

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021b. Fast, effective, and self-supervised: Transforming masked language models into universal lexical and sentence encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1442–1459, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yixin Liu and Pengfei Liu. 2021. SimCLS: A simple framework for contrastive learning of abstractive summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1065–1072, Online. Association for Computational Linguistics.

Yu Meng, Chenyan Xiong, Payal Bajaj, Saurabh Tiwary, Paul Bennett, Jiawei Han, and Xia Song. 2021. Coco-lm: Correcting and contrasting text sequences for language model pretraining.

Zaiqiao Meng, Fangyu Liu, Ehsan Shareghi, Yixuan Su, Charlotte Collins, and Nigel Collier. 2022. Rewire-then-probe: A contrastive recipe for probing biomedical knowledge of pre-trained language models.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting* *of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. 2022a. A contrastive framework for neural text generation.

Yixuan Su, Fangyu Liu, Zaiqiao Meng, Tian Lan, Lei Shu, Ehsan Shareghi, and Nigel Collier. 2022b. TaCL: Improving BERT pre-training with token-aware contrastive learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2497–2507, Seattle, United States. Association for Computational Linguistics.

Ivan Vulić, Edoardo Maria Ponti, Anna Korhonen, and Goran Glavaš. 2021. LexFit: Lexical fine-tuning of pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5269–5283, Online. Association for Computational Linguistics.

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020. Clear: Contrastive learning for sentence representation.

Hu Xu, Bing Liu, Lei Shu, and Philip Yu. 2019. BERT post-training for review reading comprehension and aspect-based sentiment analysis. 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 2324–2335, Minneapolis, Minnesota. Association for Computational Linguistics.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. ConSERT: A contrastive framework for self-supervised sentence representation transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5065–5075, Online. Association for Computational Linguistics.