DISSERTATION

SMART TRANSFERS: CHALLENGES AND OPPORTUNITIES IN BOOSTING LOW-RESOURCE LANGUAGE MODELS WITH HIGH-RESOURCE LANGUAGE POWER

Submitted by

Shadi Manafi

Department of Computer Science

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 2024

Doctoral Committee:

Advisor: Nikhil Krishnaswamy

Francisco R. Ortega

Nathaniel Blanchard

Edwin K. P. Chong

ABSTRACT

SMART TRANSFERS: CHALLENGES AND OPPORTUNITIES IN BOOSTING LOW-RESOURCE LANGUAGE MODELS WITH HIGH-RESOURCE LANGUAGE POWER

Large language models (LLMs) are predominantly built for high-resource languages (HRLs), leaving low-resource languages (LRLs) underrepresented. To bridge this gap, knowledge transfer from HRLs to LRLs is crucial, but it must be sensitive to low-resorce language (LRL)-specific traits and not biased toward an high-resource language (HRL) with larger training data. This dissertation addresses the opportunities and challenges of cross-lingual transfer in two main streams. The first stream explores cross-lingual zero-shot learning in Multilingual Language Models (MLLMs) like mBERT and XLM-R for tasks such as Named Entity Recognition (NER) and section-title prediction. The research introduces adversarial test sets by replacing named entities and modifying common words to evaluate transfer accuracy. Results show that word overlap between languages is essential for both tasks, highlighting the need to account for language-specific features and biases. The second stream develops sentence Transformers, which generate sentence embeddings by mean-pooling contextualized word embeddings. However, these embeddings often struggle to capture sentence similarities effectively. To address this, we fine-tuned an English sentence Transformer by leveraging a word-to-word translation approach and a triplet loss function. Despite using a pre-trained English BERT model and only word-by-word translations without accounting for sentence structure, the results were competitive. This suggests that mean-pooling may weaken attention mechanisms, causing the model to rely more on word embeddings than sentence structure, potentially limiting comprehension of sentence meaning. Together, these streams reveal the complexities of cross-lingual transfer, guiding more effective and equitable use of HRLs to support LRLs in NLP applications.

ACKNOWLEDGMENTS

The journey of completing a PhD dissertation is a profound and transformative experience, and it is one that I could not have undertaken alone. I am deeply indebted to the individuals whose support, guidance, and encouragement have been instrumental in helping me reach this milestone. First and foremost, I would like to express my sincere gratitude to my advisor, Prof. Nikhil Krishnaswamy, for his invaluable mentorship throughout the course of my doctoral studies. His expertise, insightful feedback, and unwavering commitment to my growth as a scholar have been a source of inspiration. I am profoundly grateful for his patience, constructive criticism, and the intellectual freedom he has provided, allowing me to explore new ideas while guiding me toward the completion of this dissertation. I am deeply thankful to my husband, Hadi, whose love, support, and understanding have been my greatest sources of strength throughout this journey. His belief in my abilities has given me the confidence to persevere through the most challenging moments. To my mother and father, Kobra and Mohsen, I owe a debt of gratitude that words cannot fully express. They have been my foundation, always encouraging me to pursue my dreams with determination and integrity. The values they instilled in me—hard work, perseverance, and the pursuit of knowledge—have shaped me into the person I am today. To my beloved sisters, Azadeh and Setareh, I extend my heartfelt thanks for their endless support, encouragement, and camaraderie throughout this journey. They have been my confidantes, cheerleaders, and greatest sources of comfort during both the highs and the lows. Finally, I would like to thank everyone who contributed, in ways both large and small, to the completion of this dissertation. Whether through academic collaboration, friendship, or emotional support, their contributions have been invaluable.

Dedication

I would like to dedicate this thesis to my FAMILY.

Their support has been invaluable throughout this journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

In this dissertation, we will explore two project streams focused on cross-lingual knowledge transfer, examining the reliability of leveraging LLMs trained on HRLs for LRLs.

Cross-lingual transfer for language models is an area of growing significance, particularly as models like mBERT and XLM-R [Conneau et al., 2020a] demonstrate impressive zero-shot transfer capabilities [Hu et al., 2020]. Zero-shot learning, where fine-tuning a model on a task in one language yields impressive results in other languages, further highlights the powerful transfer capabilities of these models. This phenomenon enables models trained on HRLs to perform downstream tasks in LRLs, offering significant promise for languages with limited online resources. LLMs leverage their multilingual capacity to perform tasks such as Named Entity Recognition (NER) and Part of Speech (POS) tagging across languages, as demonstrated by [Pires et al., 2019] and [Wu and Dredze, 2019].

Despite their strengths, challenges remain, particularly regarding vocabulary memorization [Patil et al., 2022], which may not reflect true language understanding. In this thesis, we aim to assess the robustness of cross-lingual transfer by evaluating how minor input changes affect performance and how language features, such as vocabulary overlap, impact this transfer. Specifically, we investigate the extent to which the performance of LLMs fine-tuned on HRLs transfers to LRLs, and we analyze how input perturbations influence model predictions for tasks like NER.

Hence, we explore the following questions in the first study:

- How does the accuracy of zero-shot learning change when introducing minor variations to the original test input?

- What impact do language features, such as vocabulary overlapping, have on zero-shot learning?

Our novel contributions are as follows:

- We conducted four perturbations to evaluate NER models' robustness in zero-shot learning across 21 languages. The two most crucial methods included replacing named entities shared between the HRL and the LRL with entities unique to the LRL, and modifying surrounding words to assess cross-lingual adaptability.
- We created a comprehensive section title dataset for 21 LRLs and performed two perturbations on section title prediction tasks: first, by substituting the common words between source and target languages with unique words in the target languages using the cosine similarity function, and second, by choosing substitutions randomly.
- We assess the relationship between vocabulary overlap, cross-lingual transfer robustness, and adversarial perturbations.

The second stream of work focuses on sentence Transformers and their ability to perform cross-lingual semantic transfer. Sentence embeddings map semantically similar sentences to close positions in vector space, supporting tasks such as semantic search, summarization, and question answering. However, monolingual models trained independently on different languages tend to generate sentence distributions that differ due to variations in initialization and training mechanisms. This can lead to misalignment in the vector space, reducing the effectiveness of cross-lingual transfer [Reimers and Gurevych, 2020].

[Reimers and Gurevych, 2020] uses knowledge distillation from a monolingual sentence Transformer to train a multilingual one. However, the model struggles with dissimilar sentences due to the design of the loss function. Since it is based on a translation dataset, it cannot accurately distinguish dissimilar sentences. To address this, we extend this work by developing a cross-lingual sentence embedding approach based on triplet loss. This method

ensures that knowledge learned from an HRL like English is effectively transferred to an LRL like Persian.

In addition, this powerful model, like other existing sentence Transformers, struggles with distinguishing texts of different lengths. It appears that texts from two different languages are not correctly clustered as either similar or dissimilar. To address this issue, we developed a model that is not pre-trained on non-English sentences but instead uses one-to-one translated sentences mapped into English. This model achieved performance comparable to the reference model, leading us to hypothesize that the proposed model may help identify the source of the reference model's difficulty in distinguishing texts with significantly different lengths.

In this stream, our contributions are as follows:

- We propose a triplet-based model for knowledge distillation, replacing the traditional binary loss function to better align dissimilar sentences and improve cross-lingual performance.

- We collect a new cross-lingual triplet dataset specifically for Persian-English sentence pairs, contributing valuable resources for this low-resource language.

- We replicate and analyze a model similar to the reference model to identify key challenges in handling long-short pair of texts similarities, providing insights into cross-lingual sentence embedding issues.

- Our experiments demonstrate that models trained on semantically similar texts, rather than direct translations, consistently outperform translation-based models, especially with on pairs whose lengths differ substiantially.

In both streams of work, we emphasize the importance of evaluating cross-lingual robustness, whether through perturbation experiments in NER or through fine-tuning competitive bilingual sentence Transformers. This work contributes to the growing understanding of how

knowledge can be effectively transferred between languages, improving performance on LRLs while addressing challenges such as vocabulary overlap and sentence alignment.

Chapter 2

Background and Related Works

In recent years, advancements in Transformer-based architectures have significantly impacted various natural language processing (NLP) tasks. This chapter explains two primary types of Transformer models: lexical-level Transformers and sentence-level Transformers, discussing their techniques, relevance, and applications. Additionally, it presents studies and visualizations we have conducted to investigate token representations in the BERT Transformer, particularly focusing on homographs and physical objects. Furthermore, the chapter examines Knowledge Transfer methods, including Fine-Tuning, Zero-Shot Learning, and Knowledge Distillation, which have been crucial for adapting models to different languages and tasks. Finally, we review the concept of Adversarial sets as a mechanism for testing model robustness during knowledge transfer.

2.1 Transformers

Transformer is a groundbreaking neural network architecture introduced by [Vaswani et al., 2017], which revolutionized sequence modeling by relying solely on self-attention mechanisms. Unlike traditional sequence models such as recurrent neural networks (RNNs) [Hochreiter and Schmidhuber, 1997, Bengio et al., 1994] and convolutional neural networks (CNNs) [LeCun and Bengio, 1995], which rely on sequential or localized operations, Transformer captures global dependencies within input sequences in a parallelizable manner. This allows it to overcome the limitations of RNNs, which suffer from vanishing gradients and struggle to model long-range dependencies [Hochreiter and Schmidhuber, 1997, Bengio et al., 1994], and CNNs, which require numerous layers or large kernel sizes to capture long-distance

relationships [Gehring et al., 2017].

A key innovation of the Transformer is its self-attention mechanism, which computes the importance of each word in the input sequence relative to all other words. This mechanism allows the model to focus on relevant parts of the sequence while ignoring irrelevant information. The ability to process sequences non-sequentially greatly enhances the computational efficiency of the Transformer, particularly for large datasets and long sequences. Consequently, the Transformer has become the foundation for many state-of-the-art models across various domains, including NLP tasks such as machine translation, language modeling, reading comprehension, and more [Radford et al., 2018, Devlin et al., 2018, Yang et al., 2019].

Self-attention, the core component of the Transformer, calculates attention scores by projecting the input sequence into query, key, and value vectors. The attention mechanism computes the relevance between queries and keys, and this relevance is used to weight the values. The result is an output that dynamically integrates information from different parts of the sequence, regardless of their distance in the original input. This mechanism has proven highly effective for learning contextual representations of words and sentences, which can be applied to various tasks, including sentence classification, named entity recognition, and syntactic parsing [Devlin et al., 2018].

One of the key strengths of the Transformer architecture is its flexibility, allowing it to model dependencies across words or sentences without being constrained by sequence length [Bahdanau et al., 2014, Kim et al., 2017]. This is particularly important in tasks such as machine translation, where the meaning of a word often depends on context that may be far removed from the word itself. In contrast, traditional RNNs are limited by their sequential nature, making it difficult to capture such long-range dependencies.

The original Transformer architecture follows an encoder-decoder structure, where the encoder processes the input sequence into a meaningful representation, and the decoder generates the output sequence step by step. These two components communicate via attention

mechanisms, enabling the decoder to selectively focus on relevant parts of the input during the generation process [Cho et al., 2014, Sutskever et al., 2014].

There are two main types of Transformers:

1. **Encoder-based Transformers**: These models, such as BERT, are designed to process and understand entire input sequences simultaneously, focusing on extracting context from the input data. They are widely used in tasks such as text classification, sentence embeddings, and named entity recognition. Encoder models operate in a non-sequential manner and are bi-directional, meaning they consider both the left and right contexts of each token, making them highly effective for various NLP tasks [Devlin et al., 2018].

2. **Generator-based Transformers (Auto-regressive models)**: These models, such as GPT, generate output one token at a time, with each new token conditioned on the previous ones. This auto-regressive approach makes them well-suited for tasks such as text generation and machine translation. These models are uni-directional, as each token is predicted based on previously generated tokens, making them ideal for language modeling and text completion tasks [Radford et al., 2019].

In this study, we will mostly focus on encoder-based Transformers, as our primary interest lies in understanding input representations and using Transformers for non-sequential tasks where generation is not required.

In summary, the Transformer architecture, with its self-attention mechanism and parallelizable nature, has become the backbone of modern NLP models. Its ability to capture global dependencies, combined with its efficiency, makes it suitable for both lexical-level and sentence-level tasks across various applications.

2.1.1 Lexical-Level Transformers

One of the most renowned Transformer models is BERT [Kenton and Toutanova, 2019], short for Bidirectional Encoder Representations from Transformers.

Masked Language Model (MLM) is one of the key components of BERT's pre-training, addressing the unidirectionality limitation of previous models. Traditional language models, which predict words in a left-to-right manner, can only use preceding tokens to make predictions, which limits the model's ability to capture full context.

BERT overcomes this by utilizing the masked language model (MLM) objective, inspired by the Cloze task [Taylor, 1953]. In the MLM task, a random subset of tokens in the input sequence is replaced with a special [**MASK**] token. The model's goal is to predict the original vocabulary ID of the masked word based solely on its context, which includes both preceding and following words.

This bidirectional approach enables the model to capture context from both sides of a token, unlike traditional left-to-right models. By pre-training in this way, BERT is able to create deep contextualized representations that significantly improve performance on downstream tasks.

In addition to the MLM task, BERT includes a next sentence prediction (NSP) task, which helps the model understand the relationship between pairs of sentences. In the NSP task, the model is presented with two sentences and must predict whether the second sentence follows the first one in the original context.

During pre-training, the model is provided with sentence pairs in two formats:

- **Positive pairs**, where the second sentence is the actual next sentence in the original text.
- **Negative pairs**, where the second sentence is randomly chosen from another part of the corpus, and thus does not logically follow the first sentence.

The task is a binary classification problem where the model predicts whether the given sentence pair is coherent.

Moving forward, mBERT(multilingual BERT) [Kenton and Toutanova, 2019] extends the Transformer model's pre-training to multiple languages. mBERT is trained on a multilin-

gual corpus composed of monolingual data from multiple languages. Specifically, mBERT is trained on the concatenation of Wikipedia pages from 104 different languages. However, this data is monolingual for each language—meaning there is no explicit alignment or translation between languages during pre-training. The model learns from each language independently, but the same architecture and parameters are shared across all languages.

While mBERT is remarkable for covering different languages, experiments reveal a trade-off when it comes to scaling the number of languages for a fixed model capacity. Initially, including more languages enhances cross-lingual performance for LRLs. However, beyond a certain point, the overall performance on both monolingual and cross-lingual benchmarks begins to degrade.

Therefore, XLM-R [Conneau et al., 2020a], another Transformer model, explores the trade-off between HRLs and LRLs, examining the impact of language sampling and vocabulary size. This trade-off is referred to as the "curse of multilinguality." XLM-R suggests that this limitation can be mitigated by increasing model capacity.

Both mBERT and XLM-R employ Transformer models trained with the multilingual MLM objective, relying solely on monolingual data. Streams of text from each language are sampled, and the model is trained to predict the masked tokens within the input.

Before presenting sentences to Transformer models, they must undergo tokenization, a crucial process that involves breaking down text into smaller units called tokens. These tokens can encompass words, phrases, subwords, or even individual characters. Tokenization plays a pivotal role in Transformer models, serving as a fundamental preparatory step that enhances the model's ability to process text data for various natural language understanding.

In essence, tokenization is a critical step in the text preprocessing pipeline for Transformer models, significantly influencing the model's overall performance. Notably, mBERT and XLM-R employ different tokenization methods: mBERT uses WordPiece encoding, while XLM-R employs Byte-Pair Encoding (BPE).

WordPiece is the subword tokenization algorithm used by mBERT. In WordPiece, the

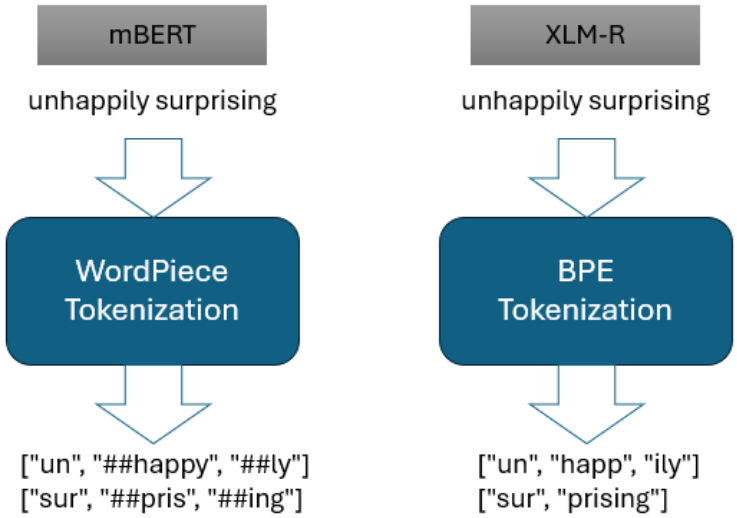

Figure 2.1: WordPiece tokenization in mBERT vs Byte-Pair Encoding in XLM-R.

vocabulary is initially initialized with every individual character found in the training data. The algorithm then iteratively learns a set number of merge rules. Unlike BPE, which selects the most frequent symbol pair, WordPiece selects the symbol pair that maximizes the likelihood of the training data when added to the vocabulary, leading to the creation of more subwords. This distinction in the selection process differentiates WordPiece from BPE. An example is provided in Figure 2.1.

MLLMs such as mBERT [Kenton and Toutanova, 2019] and XLM-R [Conneau et al., 2020a], demonstrate effective cross-lingual transfer capabilities for downstream tasks [Hu et al., 2020]. The objective of cross-lingual transfer is to leverage knowledge learned by a model in a source language and transfer it to a target language. Multilingual models like mBERT and XLM-R have shown surprising effectiveness in zero-shot transfer, where fine-tuning the model on a task in a source language often results in impressive performance on the same task in other languages. This is particularly beneficial for languages with limited web resources, such as LRLs, which often struggle with inadequate vocabulary for downstream tasks.

2.1.2 Sentence Transformers

Sentences are not considered similar to words to generate their embeddings as their distributions are not as many as words. How many times can we see a whole sentences are repeated through texts?

Researchers have employed BERT to generate fixed-size sentence embeddings by inputting individual sentences and deriving embeddings from either the average output layer of BERT or the [CLS] token. Unfortunately, this approach often results in suboptimal sentence embeddings, occasionally performing worse than traditional methods such as averaging GloVe embeddings [Pennington et al., 2014]. To address this issue, Sentence-BERT (SBERT) was proposed by [Reimers and Gurevych, 2019], significantly improving sentence embedding quality.

The distributional semantic hypothesis [Harris, 1954, Firth, 1957] states that words occurring in similar contexts carry similar information and meaning. This intuition has led to the development of robust, fixed-sized word embeddings where similar words cluster together in high dimensional space [Mikolov et al., 2013].

With Transformer models such as BERT [Devlin et al., 2019a], the intuition can be extended to semantically similar *sentences* through contrastive learning methods. A well-optimized model will cluster similar sentences, but optimization is dependent on having enough training data, which is often only the case for HRLs. For LRLs, there is often insufficient high-quality data available to train a well-optimized sentence Transformer, which limits accessibility to the benefits of such models for LRLs.

In addition to the limitations of sentence Transformers for LRLs, which restrict their application to these languages, there is another significant challenge: we cannot directly transfer sentence embeddings from HRLs to LRLs, or vice versa, as their embeddings reside in different vector spaces, making direct comparisons impossible.

However, this capability is crucial for tasks like cross-lingual information retrieval (CLIR),

where knowledge must be transferred across languages, despite structural and linguistic differences. As the saying goes, "Knowledge is universal, regardless of the language—same dish, different seasoning."

In multilingual and cross-lingual contexts, only a limited number of sentence embedding models have been developed.

[Reimers and Gurevych, 2020] proposed creating multilingual versions of monolingual models based on the principle that a translated sentence should occupy the same position in the vector space as the original sentence. They distill the knowledge learned from English in BERT, using it as a teacher model, and inject this knowledge into XLM-R, which serves as the student model, followed by mean pooling over the entire output like Figure 2.2.

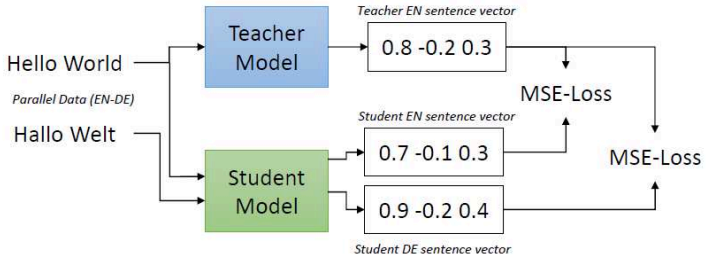

Figure 2.2: Given parallel data (e.g. English and German), train the student model such that the produced vectors for the English and German sentences are close to the teacher English sentence vector [Reimers and Gurevych, 2020].

SENT2VEC [Pagliardini et al., 2018] obtains representation of a sentence by averaging the word n-gram embeddings (including unigrams). However, due to the significant differences between non-contextual word embeddings and contextualized embeddings, the model may encounter performance issues.

The WACSE (Word Alignment Cross-lingual Sentence Embedding) model [Miao et al., 2024] enhances cross-lingual sentence embeddings by incorporating word alignment information between HRLs and LRLs. Its primary function is to leverage word-level alignments during training to ensure that semantically equivalent words in different languages are closely embedded in a shared vector space. However, relying on word alignment can lead to over-

fitting on lexical similarity, causing the model to focus more on individual word correspondences rather than the broader sentence meaning. This can result in suboptimal embeddings for tasks that require understanding sentence-level structures, such as sentiment analysis, paraphrase detection, or discourse-level tasks.

Before then, methods like Word Mover's Distance (WMD) [Kusner et al., 2015] use similar technique to assign similar sentences to each other. WMD focuses on word-level alignment and ignores the sentence-level structure and context, which are crucial for capturing cross-lingual sentence embeddings.

LASER (Language-Agnostic SEntence Representations) [Artetxe and Schwenk, 2019] trains a single encoder for multiple languages using a large multilingual parallel corpus. The model generates universal sentence embeddings that work across languages, enabling tasks like cross-lingual document retrieval and machine translation.

Furthermore, LaBSE [Feng et al., 2022] learns cross-lingual sentence embeddings by utilizing a dual-encoder architecture for translation ranking combined with additive margin softmax. The model is further enhanced by incorporating MLM and translation language modeling (TLM). LaBSE is trained on a vast dataset comprising 17 billion monolingual sentences and 6 billion translation pairs, allowing it to effectively align sentence representations across languages and perform well on a wide range of cross-lingual tasks.

However, these models depends heavily on the availability of large-scale parallel data, which limits its application to LRLs.

The paper [Sabet et al., 2019] introduces a method for generating cross-lingual sentence embeddings by leveraging parallel sentence pairs from bilingual corpora. It utilizes a dual-encoder architecture and a contrastive learning objective to align embeddings of semantically equivalent sentences across different languages in a shared vector space. By training on both monolingual and parallel data, the approach ensures that the embeddings capture both within-language and cross-language semantic relationships. This method is particularly robust in handling languages with less parallel data, making it effective for cross-lingual tasks

such as retrieval and machine translation. Since the method primarily focuses on parallel data for training, it may not generalize well to languages that were not part of the training set or to those with very different linguistic structures.

The aforementioned cross-lingual embedding models mostly rely on translation-based tasks to learn alignments between languages. Instead of relying solely on translation-based tasks, [Goswami et al., 2021] combine multiple objectives, including translation ranking, multilingual natural language inference (NLI), and paraphrase identification, to train the model. This multi-task setup allows the model to capture both semantic meaning and cross-lingual alignment, resulting in embeddings that perform better across a variety of languages and tasks, compared to single-task methods.

Its architecture uses two separate encoders for input sentences in different languages. This helps the model learn language-specific embeddings that can then be aligned in a shared vector space. One encoder handles the source language, and the other handles the target language. The model uses a dual-encoder architecture, but it does not train a separate model for each language pair. Instead, it uses shared encoders that are fine-tuned on tasks like translation ranking, multilingual NLI) and paraphrase identification across different languages.

Since the model is trained using multi-task learning, it may develop biases toward the tasks that have more data or are easier to learn. For example, tasks like translation ranking may dominate, potentially resulting in less optimal performance on tasks like NLI or paraphrase identification. Moreover, While the dual-encoder architecture effectively handles cross-lingual sentence embeddings, it requires two separate encoders for input languages. This increases computational and memory costs during inference compared to single-encoder models like BERT, especially when dealing with large datasets or deploying the model in real-time applications.

mSimCSE [Wang et al., 2022] shows that applying contrastive learning to English data alone can produce universal cross-lingual sentence embeddings, eliminating the need for

parallel data. However, the model may fail to capture cross-lingual semantic equivalence effectively when there are significant structural or vocabulary differences between HRLs and LRLs.

2.2 Token Representations in BERT

Our research journey embarked upon an in-depth analysis of token representations within the BERT word Transformer. Central to this exploration was our aim to unravel the intricate layer-wise representations that underlie the processing of tokens within BERT.

2.2.1 Insights of Homographs Representation

The journey of Transformer interpretation led us to an intriguing subset of words: homographs. These words, sharing the same spelling but possessing multiple meanings, offered a unique lens through which to understand how BERT captures and represents subtle shades of semantics.

Additionally, we conducted an analysis of the post-processing of embeddings to examine their impact on performance, as they are claimed to enhance the overall effectiveness of feature-based approaches in [Sajjad et al., 2022].

Diving deeper, we investigated the representation of homographs and their synonyms. We explored potential mis-tagged instances, indicating the possibility of utilizing activations for tag revision. This intersection of representation analysis and tag revision presents intriguing prospects for improving NLP systems.

So, in this section, we investigate the representations across three distinct but related branches:

- Homographs
- Normalizations

- Homographs and Their Synonyms

Dataset

We selected the "Brown" dataset to extract homographs [1]. This dataset comprises 500 English texts, each containing just over 2,000 words, sampled from 15 distinct text categories, with varying text counts in each category.

Furthermore, there is a "tagged" version of the Brown dataset in which each individual word (token) is assigned one of 81 grammatical tags. These tags fall into six categories: 1. Major form-classes (referred to as "parts of speech") such as noun (common and proper), verb, adjective, and adverb, encompassing the open lexical classes. 2. Function words, including determiners, prepositions, conjunctions, pronouns, and more, forming the closed lexical and grammatical classes. 3. Specific essential individual words, like "not," "existential there," "infinitival to," and various forms of the verbs "do," "be," and "have," whether used as auxiliaries or full verbs. 4. Punctuation marks with syntactic significance. 5. Inflectional morphemes, notably for noun plural and possessive, verb past, present, and past participle, 3rd singular concord markers, comparative and superlative adjective and adverb suffixes. 6. Two tags, FM and NC, are hyphenated to the regular tags to indicate that a word is a foreign word or a cited word, respectively.

Methodology and Evaluation

To narrow our focus on homographs, we applied specific criteria, isolating homographs with less than four parts of speech tags while also excluding common English stop words. Each representation of a layer includes 768 neurons but we cannot visualize the words in this number of dimensions. Therefore, we employed dimensionality reduction techniques to visualize the layer-wise representations of homographs. Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) served as our tools of choice.

[1] http://icame.uib.no/brown/bcm.html

For post-processing analysis, *z-score*, *min-max* normalization methods are chosen. Z-score normalization as a widely employed preprocessing step in numerous machine learning workflows, normalizes input vectors by centering and scaling to achieve a mean of zero and a unit variance. Specifically, for each feature (in the selected architecture, one of the 768 dimensions extracted from each layer's representation), we calculate the mean and variance across all words in the dataset, denoted as D. Subsequently, we subtract the mean and divide by the standard deviation for each feature's value within each embedding. Min-max normalization transforms each feature's range to span from 0 to 1.

We investigated the representation of homographs and their synonyms, applying t-SNE to capture their interplay in representation space. This revealed intriguing relationships between homographs and their synonyms, shedding light on how certain synonyms cluster together.

We used the `BERT-base-uncased` variant to get the words representations. Moreover, there appears to be a discrepancy in the layer representation — while BERT has been mentioned as a 12-layer structure [Kenton and Toutanova, 2019], the code exhibits 13 layers. This discrepancy arises because the initial layer is a static embedding layer, utilized before the application of attention-based encoding; thus, technically, the operational layers amount to 12.

Results and Discussion

The first occurrence and all occurrence representations of the most frequent homographs in the Brown dataset are shown in Figures 2.3 and 2.4, respectively. In these figures, the colors are applied at both word and layer levels, distinguishing various elements of the visualizations. Figure 2.3 exclusively displays representations for the first instance of each word, while Figure 2.4 encompasses representations from all word occurrences. Surprisingly, the representations of the last layer are located closer together while they must be much further for better classification.

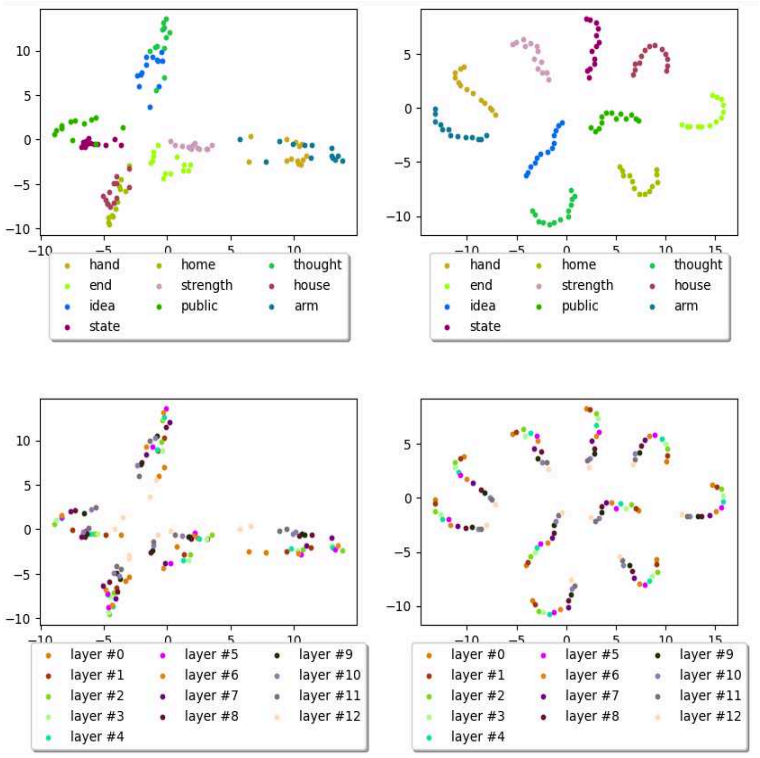

Figure 2.3: Representation of the first occurrence of most frequent homographs in the Brown dataset, (top: word-wise, bottom: layer-wise, left: PCA, right: TSNE)

These visualizations proved the pre-assumed patterns for tokens across different layers of BERT. Of particular interest was the PCA representation, which exhibited a division into three distinct branches. This suggested that the distribution of representations wasn't purely random but instead bore meaningful patterns. The inference would be that these visualizations suggest a relatively uniform distribution of each word across different segments of speech. Subsequently, we proceed to visualize one homograph, individually 2.5. However, we cannot conclude anything specific because these data points have not undergone reduction with coexisting data, which raises the possibility that they might not accurately represent the activations they aim to depict.

If we apply z-score and min-max normalizations to the activations, the result would change from Figure 2.4 to Figures 2.6 and 2.7, respectively. The results reveal a notable shift in the distributions of embeddings, a factor that has the potential to influence the outcomes

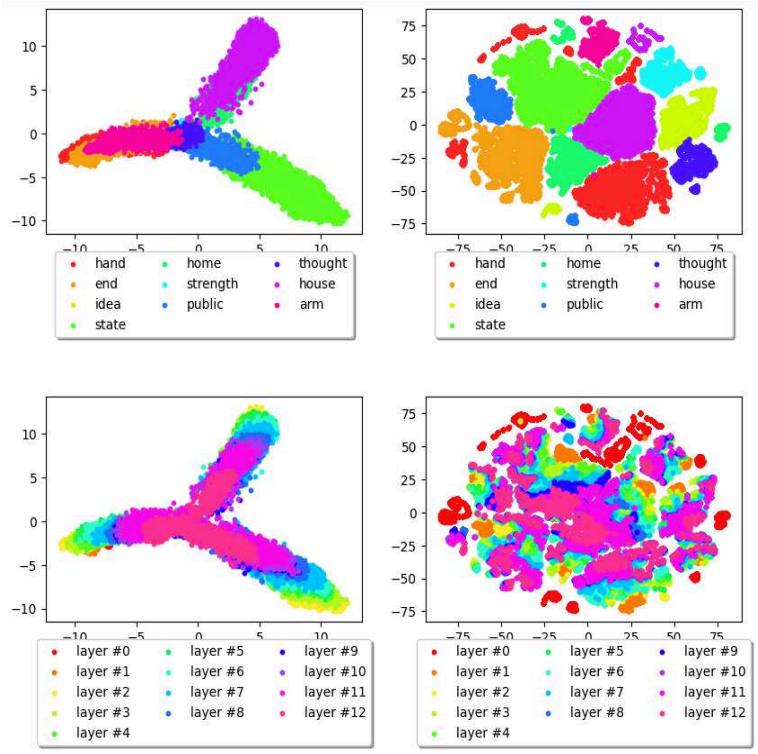

Figure 2.4: Representations of all occurrences of most frequent homographs in the Brown dataset, (top: word-wise, bottom: layer-wise, left: PCA, right: TSNE)

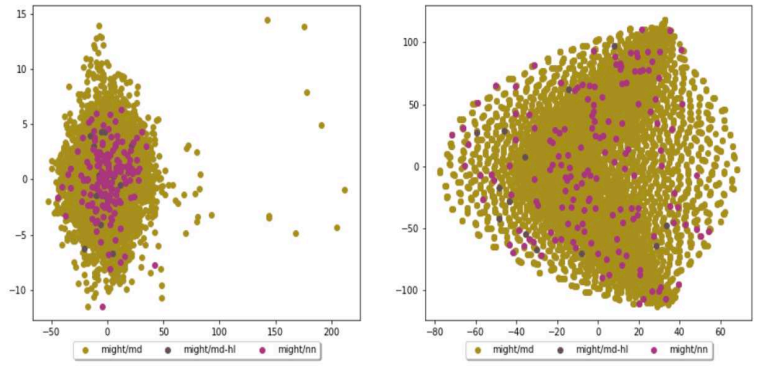

Figure 2.5: Representations of all occurrences of one homograph in the Brown dataset, colored word-wise (PCA: left, TSNE: right)

of feature-based approaches, such as classification results. Additionally, the representations in the final layer tend to overpower those in the intermediate layers, leading to a loss of fine-grained information within the model.

Now, we explore the results of homographs and synonyms. As illustrated in Figure

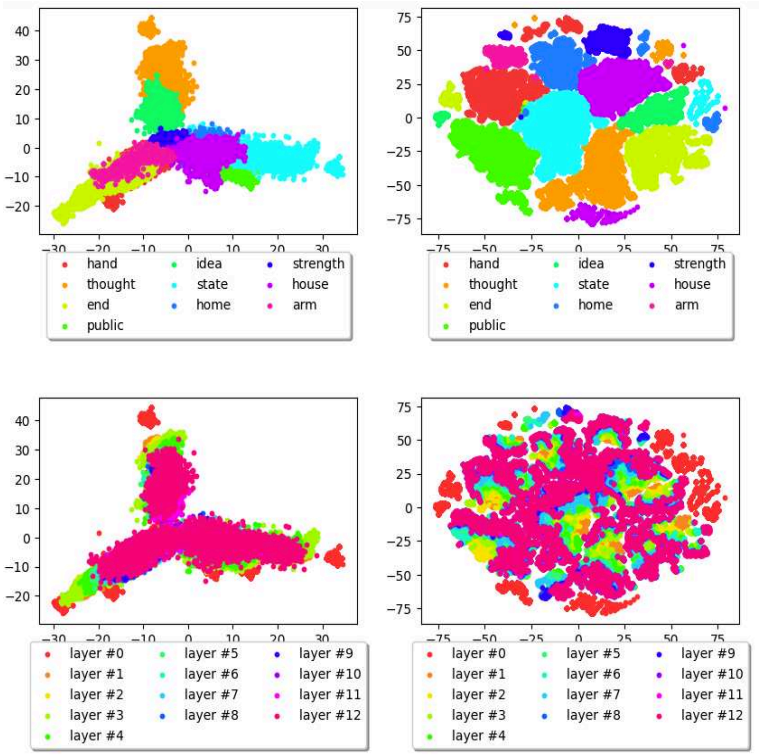

Figure 2.6: Z-score normalization of representations for all occurrences of most frequent homographs in the Brown dataset, (top: word-wise, bottom: layer-wise, left: PCA, right: TSNE)

2.8, certain representations of "might/nn" appear unexpectedly close to "could/md" and "might/md." Intrigued by this observation, we delved into the words that co-occur with "might/nn." The instances of "might/nn" are divided into two categories: "might/nn" with green points located in the black box representing occurrences where the representations are unconventional, and "might/nn" with green points located out of the box where the representations adhere more closely to expected patterns.

To elucidate, consider the sentence in Figure 2.9, which falls under the category of unconventional representations and contains the term "might/nn". Indeed, the model predicts "might/md" concurrently to other modal verbs like "would," "could," "may," and "should" which suggests a potential avenue for leveraging activations to rectify incorrect Part-of-Speech (POS) tags. The activations seem to contribute valuable insights that could aid in refining and correcting such misclassification. Upon observation, it became evident that

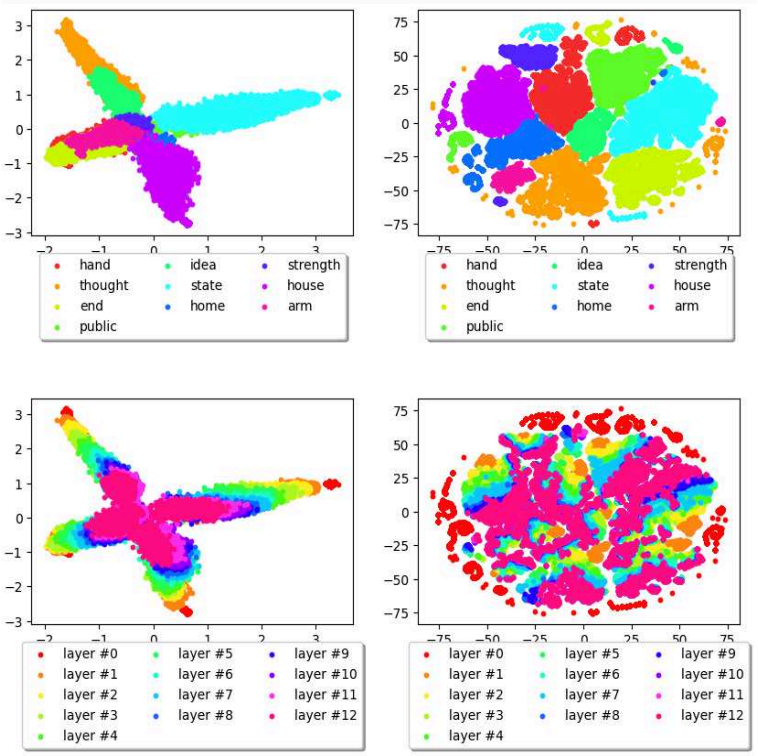

Figure 2.7: Min-max normalization of representations for all occurrences of most frequent homographs in the Brown dataset, (top: word-wise, bottom: layer-wise, left: PCA, right: TSNE)

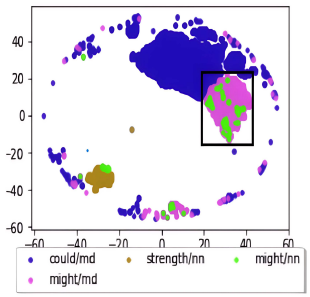

Figure 2.8: Representations of the words “might” and “could” in the Brown dataset.

certain sentences within this dataset are inaccurately tagged. This suggests a potential application: leveraging the reduced activations through t-SNE for the purpose of revising these tags and improving accuracy.

Consequently, we’ve chosen to analyze the representations of individual layers separately. Our rationale for this approach is rooted in the belief that not all layers are uniformly rep-

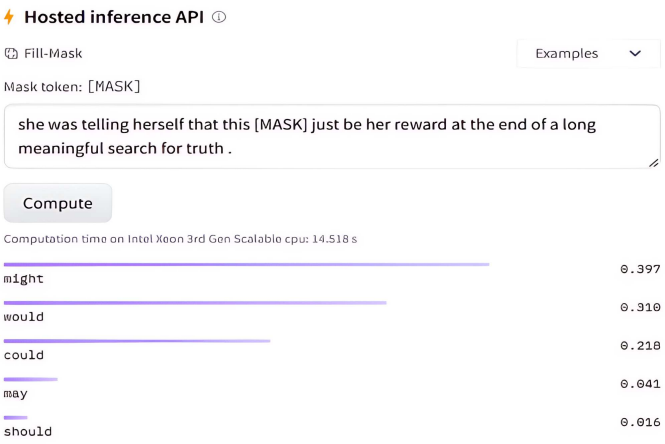

Figure 2.9: Predicting the masked word with BERT model.

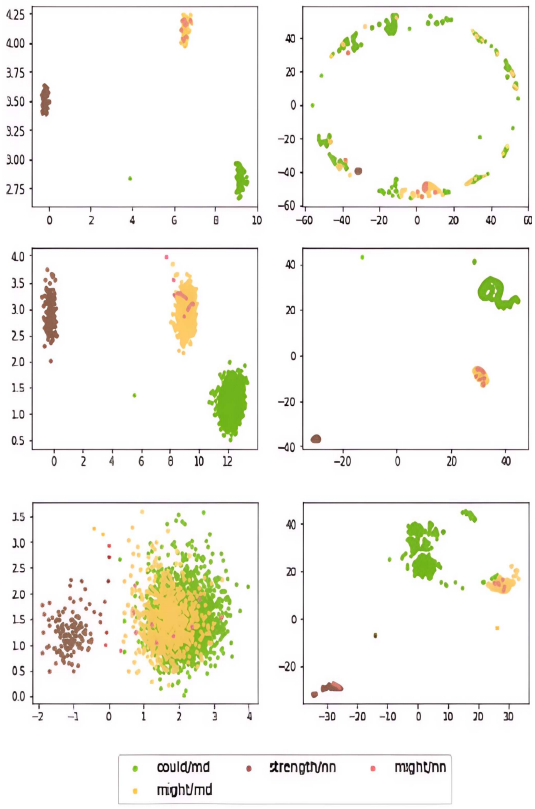

Figure 2.10: Top: First layer, center: second layer and bottom: thirteenth layer representations

resentative of Part-of-Speech (POS) tags, as they could potentially carry diverse types of information. Using Figure 2.10, you'll find a detailed presentation of these representations.

This separation aims to uncover layer-specific insights that contribute to a more comprehensive understanding of the data. Take the word "might" as an example. In the initial layer, there seems to be a lack of distinctive information related to this word. Moving on to the second layer, "might" instances are grouped together without regard for their tags, indicating a certain level of initial word similarity. However, as the layers progress, instances of "might" progressively distance themselves from each other. In the final layer succeeds in indicating the Part-of-Speech (POS) tags for six out of seven examples, signifying its effectiveness in capturing the relevant linguistic characteristics.

This granular analysis at the layer level enhances our understanding of how BERT's representations evolve, potentially uncovering brighter insights than a single, unified representation could offer.

2.2.2 Relationship of Physical Objects Representation

In this visualization, we dissect the representation of various physical objects alongside their inherent properties to elucidate their relational dynamics. The objective is to map out the proximity of these objects and properties to each other, based on their embeddings.

Dataset

We selected a specific dataset [Ghaffari and Krishnaswamy, 2023] that includes a vocabulary set of physical objects in different sentences. This study involves an agent operating within a simulated environment, constructed using the VoxWorld platform [Krshnaswamy et al., 2022]. The agent's task is to stack nine distinct types of theme objects onto a cube. The behavior of each object during stacking varies, contingent upon its geometric structure and the associated affordances, as described by Gibson [Gibson, 1977]. For instance, a cube, when correctly positioned on another cube, will maintain its stacked position, whereas a sphere in the same location will roll off and continue moving. Therefore, the Transformers provide different contextualized representations for the intended words.

Methodology and Evaluation

We used the `BERT-base-uncased` variant to get the words representations. As shown in Figure 2.12, for better differentiation and analysis, we have color-coded the points representing different sets of words in the visualization as follows: The first group, delineated in one color, includes: “cube,” “flat,” “stack,” “stable,” “stability,” “stand,” “blocks,” and “pyramid.” The second set, showcased in a divergent hue, consists of: “spheres,” “round,” “unstable,” “fall,” “instability,” “eggs,” “roll,” “capsule,” and “balls.” A third unique shade is reserved for the entities “cylinders” and “cones.” This categorization aims to more clearly reveal the intrinsic relationships between physical objects and their attributes.

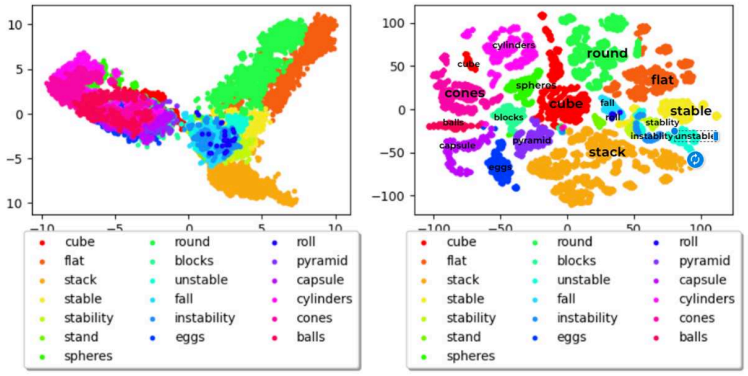

Figure 2.11: Activations of the physical objects

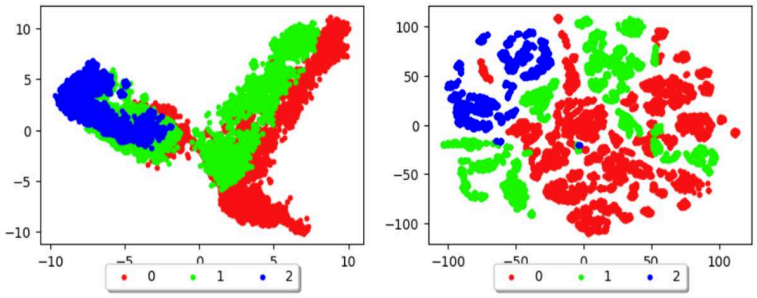

Figure 2.12: Groups of the physical objects activations

It is pertinent to note that the model lacks the specific vocabulary entry for “cubes,” consequently breaking it down into “cube” and “##s.” It then derives an average representation

from both fragments to achieve a coherent embedding.

Results and Discussion

The usual and color-coded representations for activation visualization are shown in Figure 2.11 and 2.12, respectively. It clearly shows how physical objects are connected to their characteristics.

However, there are exceptions in this delineation; for instance, "cylinder" and "roll" have not been grouped together. This is predicated on the observation that cylinders can exhibit dual behaviors — rolling or remaining stationary — contingent upon their orientation. A similar rationale has been applied to the categorization of "cones." In corpus, such behaviors of cylinders and cones in different contexts are evident, affirming the complexity of their behavioral properties. Upon analyzing the PCA plot, a noteworthy observation is the proximate placement of "flat" and "round," suggesting a potential relationship or similarity in the representation space.

We aspire that this visualization will offer a deeper understanding of the interrelationships and distinctive attributes of these physical objects and their properties, grounded in real-world behaviors and characteristics.

2.3 Knowledge Transfer

Knowledge transfer is a key concept in machine learning that focuses on leveraging knowledge learned from one task or domain to improve performance in another. The main goal of knowledge transfer is to utilize existing models or data to reduce training time and enhance performance in new tasks or domains, especially when resources are limited.

There are various approaches to achieving knowledge transfer, each designed to address different scenarios:

- **Fine-tuning**

- **Zero-shot learning**
- **Knowledge Distillation**

In general, you can see the differences between the models' usage in Figure 2.13.

These techniques provide different ways to transfer knowledge effectively, enabling models to perform well even in low-resource environments. In the following sections, we will delve deeper into the goals, methods, and techniques of knowledge transfer.

2.3.1 Fine-Tuning

Fine-tuning is a widely adopted technique in NLP that builds upon a pre-trained model by adjusting its parameters on a specific downstream task or dataset. This allows models to adapt their generalized knowledge to particular use cases while retaining the information learned during pre-training.

One of the seminal works in fine-tuning is by [Devlin et al., 2019b], which introduced BERT. The authors demonstrated that fine-tuning a model, pre-trained on a large corpus of unlabeled text, on a smaller, task-specific dataset could yield state-of-the-art results across various tasks, such as named entity recognition. This two-stage approach —pre-training followed by fine-tuning— has since become the standard for modern NLP models, influencing subsequent models like GPT, RoBERTa, and XLM-R.

In cross-lingual learning, fine-tuning allows a model to adapt its embeddings in a HRL, such as English, to perform more effectively for LRLs. For instance, XLM-R, introduced by [Conneau et al., 2020a], extended BERT's architecture to multilingual settings, showing that fine-tuning on task-specific multilingual data significantly improved performance over zero-shot learning, especially for LRLs.

Fine-tuning offers several benefits:

- **Task-Specific Adaptation**: Fine-tuning adjusts the model's representations to focus on the specific features of the downstream task, enhancing performance on nuanced

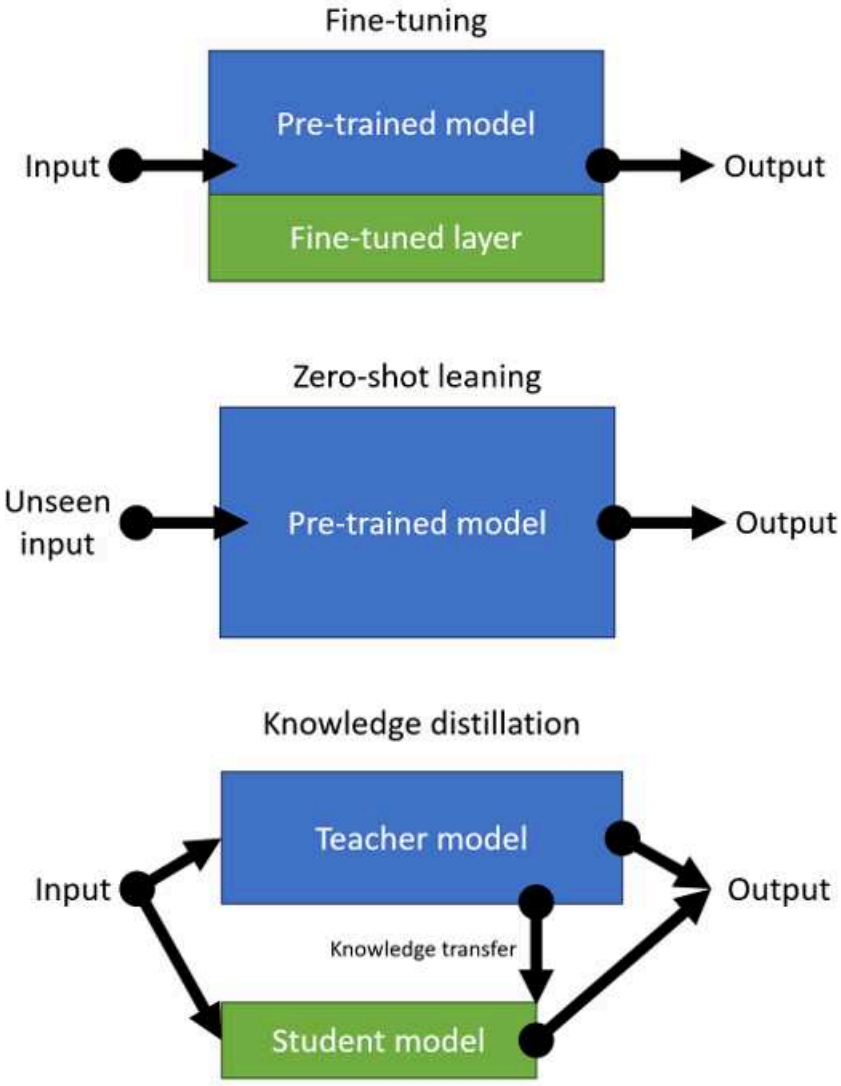

Figure 2.13: In cross-lingual applications, the blue model primarily focuses on the HRL, while the green model represents the LRL. In the top figure, the model is pre-trained on HRL, and the same model layers are fine-tuned on an LRL downstream task. In the middle figure, the entire model is first pre-trained on both HRL and LRL, then fine-tuned on HRL, with the test inputs provided from the LRL. It is assumed that the inputs of HRL and LRL are similar. In the bottom figure, the teacher model is trained on HRL, and the student model, pre-trained on LRL (and HRL), attempts to mimic the teacher's behavior for LRL.

tasks.

- **Preservation of Pre-Trained Knowledge**: Unlike training from scratch, fine-tuning retains the general knowledge acquired during pre-training, making it useful for related tasks.

- **Improved Performance Over Zero-Shot Learning**: Fine-tuning bridges the gap between source and target languages or tasks, offering superior results compared to zero-shot learning, as demonstrated in [Pires et al., 2019].

While knowledge distillation, as explored by [Hinton et al., 2015], offers the benefit of producing smaller and faster models, it can result in the loss of nuanced knowledge from the teacher model. Fine-tuning avoids this by preserving the full richness of the learned embeddings. This is especially important in cross-lingual sentence embeddings, where retaining the detailed meaning of sentence representations is crucial. For example, instead of using knowledge distillation, a sentence Transformer can be fine-tuned on new language to preserve the learned sentence distribution in the primary language while improving cross-lingual understanding.

In comparison to zero-shot learning, fine-tuning has shown superior performance across tasks. For instance, in [Conneau et al., 2020a], fine-tuning multilingual models outperformed zero-shot learning, particularly for LRLs, by adapting the model to the specific characteristics of the target language.

The paper [Liu et al., 2020] introduces a novel approach to enhance cross-lingual transfer by employing multitask fine-tuning across multiple languages. The authors argue that training language models on several tasks simultaneously encourages knowledge transfer from high-resource to LRLs, thereby improving performance on cross-lingual tasks. Their method enables models to generalize better to LRLs without requiring substantial amounts of labeled data in these languages. By integrating multiple tasks during fine-tuning, the model learns shared linguistic features that contribute to cross-lingual understanding.

The results of this approach demonstrate superior performance over standard monolingual and bilingual models on several multilingual benchmarks. The multitask fine-tuning framework outperforms strong baselines, particularly in cases where labeled data for certain languages is sparse, highlighting its effectiveness in improving cross-lingual generalization. This work underscores the potential of multitask learning for building robust multilingual models capable of generalizing across a wide range of languages.

The paper [Eisenschlos et al., 2019] introduces MultiFiT, a novel approach for efficient fine-tuning of multilingual language models. Unlike traditional models that require large datasets and extensive computation for fine-tuning, MultiFiT focuses on LRLs by leveraging transfer learning across languages. The model architecture is designed to be lightweight and efficient, allowing for faster training while maintaining competitive performance on multilingual tasks.

The key innovation of MultiFiT lies in its ability to fine-tune models in a fraction of the time needed for traditional multilingual models, while still achieving strong results in LRLs. The paper presents extensive experimental results, demonstrating that MultiFiT outperforms previous approaches in terms of both efficiency and accuracy on a wide range of multilingual benchmarks. By making fine-tuning more accessible for LRLs, MultiFiT opens the door for more inclusive NLP across diverse linguistic backgrounds.

The paper [Xu et al., 2021] explores various strategies for fine-tuning pre-trained language models to improve cross-lingual performance. The authors investigate methods that allow models, initially trained on HRLs, to effectively adapt to LRLs with minimal additional data. The paper focuses on techniques such as parameter-efficient fine-tuning, selective freezing of model layers, and utilizing task-specific adapters, all aimed at enhancing cross-lingual transfer while minimizing the computational costs associated with traditional fine-tuning approaches.

The study presents empirical results across multiple cross-lingual benchmarks, demonstrating that these fine-tuning methods significantly improve performance in LRLs. The find-

ings indicate that by strategically adapting fine-tuning approaches, it is possible to maintain high performance while reducing the need for large amounts of target-language data. This work contributes valuable insights into developing more efficient and effective cross-lingual NLP systems, particularly for languages with limited resources.

The paper [Wang et al., 2020] presents an innovative approach to improve cross-lingual lemmatization and morphology tagging using a two-stage fine-tuning process applied to Multilingual BERT (mBERT). The authors propose a method where mBERT is first fine-tuned on a large corpus of multilingual data and then further fine-tuned on language-specific data for lemmatization and morphological tasks. This two-stage fine-tuning technique enhances the model's ability to generalize across languages, especially in low-resource settings.

The results demonstrate that the proposed method significantly outperforms baseline approaches on several cross-lingual lemmatization and morphology tagging benchmarks. The paper highlights the effectiveness of leveraging a multilingual pre-trained model, such as mBERT, and refining it through targeted fine-tuning strategies, thus improving cross-lingual performance on tasks that require deep linguistic understanding.

2.3.2 Zero-Shot Learning

The advancement of NLP faced significant obstacles for an extended period due to the demanding nature of linguistic annotation tasks. These tasks required extensive and labor-intensive annotations, at times making progress seem nearly unattainable. The cornerstone of supervised learning in NLP relies on adequately labeled datasets, which play a vital role in driving advancements in the field.

The process of linguistic annotation is notorious for its intricacy, subjectivity, and the substantial costs associated with it, as evidenced by previous studies [Dandapat et al., 2009, Sabou et al., 2012, Fort, 2016]. Moreover, the wide range of structurally diverse NLP tasks adds to the complexities of data collection and annotation.

Unfortunately, the challenge of data scarcity is particularly pronounced beyond the realm

of the English language and a handful of resource-rich languages, as highlighted in previous research [Bender, 2011, Ponti et al., 2019, Joshi et al., 2020]. Transformers were game-changers for NLP tasks by offering unsupervised pre-trained models. Two existing strategies for utilizing pre-trained Transformers for zero-shot learning are the feature-based and fine-tuning approaches.

The feature-based approach, exemplified by ELMo [Peters et al., 2018], employs task-specific architectures that incorporate pre-trained representations as additional features. In contrast, the fine-tuning approach, as seen in the Generative Pre-trained Transformer (OpenAI GPT) [Alec et al., 2018], introduces minimal task-specific parameters and transfers all pre-trained parameters on the downstream tasks.

While extensive labeled datasets are readily available for the English language, the situation is significantly more challenging for numerous other languages. This scarcity of data raises a critical question: How can we adeptly leverage the abundant labeled data available for resource-rich languages to make valuable predictions in languages that lack such resources?

Zero-shot learning strategies act as a vital bridge connecting languages with ample labeled data resources to those facing a shortage of such resources. By tapping into the knowledge embedded in models trained on resource-rich languages, the aim is to enable the creation of NLP applications in languages with limited resources. We review the innovative methods and approaches to overcome the challenges posed by data scarcity, thus broadening the scope of NLP research and applications beyond English and a few privileged languages.

In the most challenging scenario, known as zero-shot cross-lingual transfer, not a single annotated example is available for the target language. Recent research has been particularly focused on the zero-shot scenario, as it theoretically offers the broadest applicability across the world's 7,000+ languages [Cao et al., 2020, Pires et al., 2019, Hu et al., 2020, Lin et al., 2019, Artetxe et al., 2017].

The predominant approach to cross-lingual transfer in NLP currently relies on massively

multilingual Transformer models [Conneau et al., 2020a, Lample and Conneau, 2019, Devlin et al., 2018]. The study conducted by [Pires et al., 2019] demonstrates the effectiveness of zero-shot cross-lingual transfer using mBERT for POS tagging and NER tasks. They also observe that this effectiveness is more pronounced between languages that are closely related. In a similar vein, [Wu and Dredze, 2019] expand the analysis to encompass a wider range of tasks and languages. They establish that transfer via mBERT is competitive even when compared to the best task-specific zero-shot transfer approaches for each task.

[Hu et al., 2020] contribute to this research landscape by introducing a benchmark for assessing multilingual encoders across nine tasks and a total of 40 languages. They primarily focus on zero-shot transfer evaluation but also experiment with target-language fine-tuning, which leads to substantial performance improvements over zero-shot transfer.

[Lauscher et al., 2020], introduce an innovative method for predicting the zero-shot performance of these models across various languages. This approach frames the problem as a regression task, taking into account factors such as pretraining data size and typological similarities between languages. They delve into the factors that hinder effective zero-shot transfer across diverse tasks. Their work provides a deeper understanding of the challenges associated with this type of transfer. They empirically demonstrate that the same principle applies to zero-shot transfer with massively multilingual Transformers (MMTs), showing that transfer performance significantly decreases as the focus shifts to more distant target languages with smaller pretraining corpora.

Furthermore, [Wu et al., 2019] and [Artetxe et al., 2019] analyze monolingual BERT models in various languages to elucidate the transfer effectiveness of MMTs. They conclude that similar to static word embeddings, it's the topological similarities between the subspaces of individual languages captured by MMTs that enable effective cross-lingual transfer. It's worth noting that this assumption of approximate isomorphism does not hold for distant languages, especially when dealing with limited-size monolingual corpora.

In some cases, pre-trained models face evaluation challenges when dealing with new

languages that were not included in the pre-training phase. These challenges can manifest as very low fine-tuning accuracy for these languages. In such scenarios, users often opt to undertake the task of pre-training a new model specifically tailored to the languages of interest. An example of this approach is the creation of IndicBERT, as demonstrated by [Kakwani et al., 2020].

2.3.3 Knowledge Distillation

Knowledge Distillation (KD) is a powerful technique that transfers knowledge from a large, high-capacity model (the teacher) to a smaller, more efficient model (the student), allowing for model compression without sacrificing much in terms of performance. First introduced by [Hinton et al., 2015], KD has been widely adopted across a variety of NLP tasks due to its ability to create compact models that are highly efficient in terms of both computational and memory resources, which makes it advantageous for real-world applications where resource constraints are critical.

One of the primary advantages of KD is that it compresses a complex, high-performing model into a smaller one while retaining much of the original model's performance. Fine-tuning often retains the full size and complexity of the pre-trained model, which can be prohibitive in deployment, especially in environments with limited computational resources [Jiao et al., 2020, Sanh et al., 2019]. In contrast, KD distills the key knowledge from the teacher into a student model that is significantly smaller, reducing the memory and computational costs. This makes KD ideal for tasks requiring real-time performance or deployment in resource-constrained environments, such as mobile devices [Sanh et al., 2019, Tang et al., 2019].

KD also facilitates the transfer of knowledge across tasks and languages, similar to zero-shot learning but with a more structured approach. Zero-shot learning can often fail to generalize effectively when applied to new tasks or languages, as it does not benefit from task-specific training. On the other hand, KD allows the student to inherit knowledge learned by

the teacher across multiple domains and tasks [Hu et al., 2020, Clark et al., 2020]. Moreover, multilingual tasks greatly benefit from KD, as the student model can learn cross-lingual representations without needing to be fine-tuned individually on each language, which can be resource-intensive [Liang et al., 2020, Sun et al., 2019].

Fine-tuning, while effective at task-specific adaptation, requires training and retaining large models for each new task or domain. This is especially true for multilingual tasks, where fine-tuning a model on each target language separately may be inefficient and impractical [Conneau et al., 2020a]. KD provides a solution by compressing the teacher model while maintaining strong task performance. For example, works like [Sun et al., 2019] and [Jiao et al., 2020] show that student models can retain up to 96% of the original teacher's performance while being much smaller and faster. This allows KD to perform better than fine-tuning, where performance often suffers due to the lack of task-specific training.

Knowledge distillation has been successfully applied across various NLP tasks, from text classification to machine translation, and more recently in multilingual and cross-lingual tasks. [Sun et al., 2019] applied patient KD to gradually transfer knowledge from a teacher to a student in a step-by-step manner, improving model compression without performance loss. In the domain of multilingual NLP, [Hu et al., 2020] and [Conneau et al., 2020a] demonstrated that KD can be used to train smaller models that generalize across languages, providing substantial performance improvements over zero-shot learning.

Similarly, works like [Sanh et al., 2019] and [Tang et al., 2019] have shown the efficacy of distilling large Transformer models like BERT and GPT into smaller, more efficient versions, significantly reducing the model size while retaining competitive performance on a range of NLP tasks. These works underscore the benefits of KD in creating resource-efficient models suitable for real-world applications.

[Reimers and Gurevych, 2020] proposed creating multilingual versions of monolingual models based on the principle that a translated sentence should occupy the same position in vector space as the original sentence.

In one study [Nath et al., 2024], we extend traditional knowledge distillation by incorporating rationale generation into the distillation process. Event coreference is a challenging task that involves linking events mentioned in text based on their semantic relationships. This task requires deep contextual understanding and reasoning, making it a suitable application for enhanced knowledge distillation methods.

In our method, we use rationale-guided KD to improve the student's understanding of event coreference by not only transferring the predictions of the teacher model but also the explanations (or rationales) behind those predictions. This approach is inspired by previous work in explainable AI, where rationale generation helps models make more interpretable decisions [Jain and Wallace, 2019].

Our approach consists of the following steps:

- **Teacher Model Generation**: We train a large teacher model that generates rationales for event coreference decisions. These rationales provide insights into why certain events are linked, going beyond simple classification outputs.

- **Rationale-Guided KD**: During the distillation process, the student model is trained not only on the outputs of the teacher model but also on its rationales. This additional information allows the student model to learn both the final prediction and the underlying reasoning process, making the student model more robust and accurate in handling event coreference tasks.

- **Efficient and Deployable Model**: The result is a student model that is significantly smaller than the teacher model, making it more efficient for real-time deployment while still maintaining high performance on the event coreference task.

This approach builds on earlier work in knowledge distillation by extending it to more complex, reasoning-based tasks. Previous works such as [Clark et al., 2020] and [Jiao et al., 2020] have focused on improving task-specific performance through distillation, while our

method emphasizes both task performance and interpretability by incorporating rationales into the distillation process.

In conclusion, knowledge distillation provides several benefits over zero-shot learning and fine-tuning, particularly in terms of model efficiency, cross-task adaptability, and performance retention. Our work on rationale-guided KD for event coreference further enhances the distillation process by incorporating explainability into the student model, making it both efficient and interpretable.

2.4 Adversarial Sets

NLP systems traditionally undergo evaluation and comparison against an often unchanging gold standard. Recent NLP tasks, such as NER models, have exhibited remarkable performance on well-established benchmarks like CoNLL2003 [Sang and De Meulder, 2003] and OntoNotes 5.0 [Weischedel et al., 2011]. These state-of-the-art NER models frequently achieve F-scores surpassing 90% on standard English datasets. However, the reliability of these models in real-world applications, where entities or contextual words may deviate from the training data distribution, remains uncertain. We should seek comprehensive insights about these NER models beyond the conventional single metric, typically the micro-F score.

The majority of NER research predominantly revolves around training and evaluating models using standard train/dev/test splits, which can lead to problems such as overfitting to the test set. Despite recent calls for alternative evaluation approaches like random splits [Gorman and Bedrick, 2019], multiple test sets [Søgaard et al., 2020], and the introduction of a "tune-set" [van der Goot, 2021], these suggestions have not significantly influenced NER research practices. While some papers conduct experiments on multiple datasets [Bernier-Colborne and Langlais, 2020, Ushio and Camacho-Collados, 2022], they still primarily present results based on the standard splits of each dataset. Although a few papers acknowledge

variance across multiple training runs [Strubell et al., 2017], there is limited analysis of how model performance changes when subjected to non-standard splits or when using slightly modified test sets for NER evaluation.

Therefore, it holds significance to assess the robustness of NER systems through natural adversarial challenges. Adversarial datasets are constructed by leveraging human knowledge to target aspects of the model that are presumed to be more susceptible. Recent studies have highlighted that even NLP systems that excel on conventional test sets experience a notable decline in performance when subjected to minor alterations in input test data, spanning various NLP tasks [Gardner et al., 2020].

The paper [Rogers et al., 2020] provides a comprehensive overview of research that seeks to understand the inner workings of BERT. The authors examine various studies focused on understanding how BERT captures linguistic knowledge and semantic relationships, investigating aspects such as syntax, semantics, and the influence of pre-training on BERT's representations. The paper systematically covers both probing methods, which analyze the internal layers of BERT, and studies that test its performance on specific linguistic tasks.

The work highlights the complexities of interpreting BERT's behavior, given that it functions as a black-box model. However, it emphasizes that BERT does encode a significant amount of linguistic knowledge that emerges through unsupervised pre-training. By summarizing key findings from "BERTology," the paper serves as a foundational resource for understanding both the strengths and limitations of BERT. However, it does not focus on the investigation of cross-lingual models.

In the domain of NLP, the creation of adversarial data primarily revolves around introducing surface-level modifications to input text. These modifications encompass a spectrum of techniques, including the insertion, deletion, or swapping of words, characters, or sentences [Gao et al., 2018, Ribeiro et al., 2018, Jia and Liang, 2017]. Moreover, researchers have delved into alternative strategies, such as paraphrasing [Iyyer et al., 2018] and generating text with semantically analogous content using distinct deep learning models [Zhao

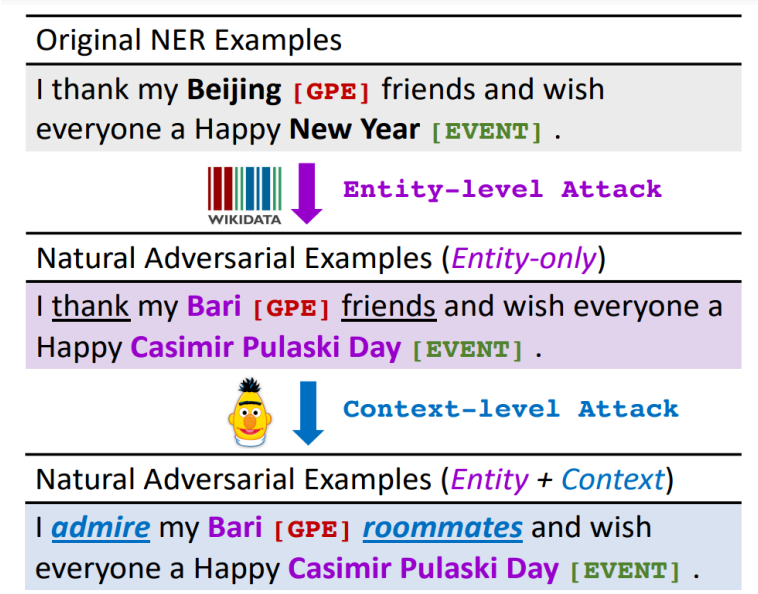

Figure 2.14: Illustration of Rock-NER attacking pipeline.

et al., 2017, Michel et al., 2019]. Some approaches involve human-in-the-loop interventions [Wallace et al., 2019a].

It's important to distinguish between black-box and white-box methodologies. Black-box approaches operate with minimal knowledge about the internal workings of NLP models, while white-box approaches delve deeper into the model's inner mechanisms [Wallace et al., 2019b, Liang et al., 2017, Blohm et al., 2018]. Certain models even combine elements of both [Li et al., 2018]. As aforementioned, within the scope of this thesis, we will explore both techniques; however, our emphasis will be on black-box methods when it comes to the NER task.

Many of the existing methods for generating adversarial attacks in NLP primarily target tasks such as sentence classification [Minervini and Riedel, 2018, Jin et al., 2020, Li et al., 2020], and question answering [Ribeiro et al., 2018, Jia and Liang, 2017, Gan and Ng, 2019]. However, these methods often lack specific designs that account for the inherent structures present in NER examples, including entity compositions and their contextual words.

In contrast, [Lin et al., 2021a] place a specific emphasis on crafting natural adversarial

examples that closely resemble real-world entities and possess human-readable context, offering a valuable approach for evaluating the robustness of NER models. As depicted in Figure 2.14, the ROCK-NER approach begins with a given NER example and initiates the adversarial attack process. Initially, it focuses on generating entity-level attacks by substituting the original entities with entities obtained from Wikidata. Subsequently, a pre-trained masked language model, such as BERT [Devlin et al., 2018], is employed to generate context-level attacks. To demonstrate the effectiveness of ROCK-NER, the authors select the OntoNotes dataset [Weischedel et al., 2011], chosen for its high annotation quality and extensive coverage of entity types. This choice leads to the creation of a novel benchmark, designed specifically to assess the robustness of a wide array of contemporary NER models. Their extensive experiments and subsequent analysis unveil a notable observation: even the most advanced NER models experience a significant decline in performance. Intriguingly, it becomes apparent that these models tend to rely on memorizing in-domain entity patterns rather than engaging in context-based reasoning.

In a further expansion of this research direction, [Vajjala and Balasubramaniam, 2022] take a notable step by introducing six new challenging adversarial test sets for evaluating NER, focusing on the English language. These sets are created using the Faker library [2], and they pose distinct challenges as follows:

1. All person names are replaced with the word "Dodo." This straightforward perturbation aims to assess how much models rely on context inference as opposed to mere memorization of specific lexical tokens. 2. Person names are replaced using the en-US locale in Faker. 3. Person names are replaced using the en-IN locale (India) in Faker. 4. All person names are replaced with female names, employing the en-TH locale (Thailand) in Faker. 5. All person names are substituted with female names, using the en-IN locale in Faker. 6. Geopolitical entity (GPE) names are replaced with GPE names from the en IE (Ireland) locale.

[2]https://faker.readthedocs.io/en/master/

Original: It was the second costly blunder by **Syria_LOC** in four minutes .
Altered: It was the second costly blunder by **Hyderabad_ORG** in four minutes .
Altered: It was the second costly blunder by **Hyderabad_LOC** in four hours .

Figure 2.15: Illustration of an NER model's predictions with minor changes to an original test set sentence

The study evaluated three state-of-the-art (SOTA) models: Spacy[3], Stanza[4], and SparkNLP[5]. These models showed significant differences in how well they performed for different named entity (NE) categories. They also noticed that their performance varied considerably depending on the specific types of entities in the dataset. Additionally, performance consistency varied across different text genres. Notably, all models displayed lower performance levels in telephone conversation (tc) and web blog (wb) genres compared to other genres. Furthermore, this study broke new ground by examining the potential presence of racial and gender biases in NER models, underscoring the importance of fairness in NER applications.

However, it's worth noting that these benchmarks and evaluations were limited to the English language, which can be considered a limitation of the study.

Additionally, the paper by [Srinivasan and Vajjala, 2023] explores the impact of minor perturbations, such as entity replacements and contextual alterations, across multiple languages, including English, German, and Hindi. Figure 2.15 illustrates a perturbation introduced in this paper, highlighting how predictions can vary with slight input modifications. This underscores the inconsistency in the robustness of NER models to such variations. Notably, for both German and Hindi NER, the study observes the most significant performance

[3]https://spacy.io/
[4]https://stanfordnlp.github.io/stanza/
[5]https://nlp.johnsnowlabs.com/

drop in masking+random sampling datasets in both languages. When comparing the two German datasets, it becomes apparent that the drop in F1 scores for adversarial datasets is notably larger for mconer21 [Malmasi et al., 2022] than for conll03 [Sang and De Meulder, 2003]. The plausible explanation could be the comparatively poorer performance of the original mconer21 model itself.

Looking forward, our research aims to dive deeper into multilingual model fine-tuning for different NLP tasks like NER and section title prediction results and scrutinize potential biases stemming from data or model factors. In conclusion, while substantial progress has been made in understanding the cross-lingual transfer capabilities of Transformer models, challenges such as biases and the intricate interplay of linguistic and structural factors continue to warrant further exploration. Our research endeavors to contribute to this evolving field by generating more challenging adversarial datasets and shedding light on delicate aspects of cross-lingual transfer.

Chapter 3

Multilingual Transformers Investigation

3.1 Cross-Lingual Transfer Robustness to Lower-Resource Languages on Adversarial Datasets

Multilingual Language Models (MLLMs) exhibit robust cross-lingual transfer capabilities, or the ability to leverage information acquired in a source language and applying it to a target language. These capabilities find practical applications in well-established NLP tasks such as Named Entity Recognition (NER). This study aims to investigate the effectiveness of a source language when applied to a target language, particularly in the context of perturbing the input test set. We evaluate on 13 pairs of languages, each including one HRL and one LRL with a geographic, genetic, or borrowing relationship. We evaluate two well-known MLLMs—mBERT and XLM-R—on these pairs, in native LRL and cross-lingual transfer settings, in two tasks, under a set of different perturbations. Our findings indicate that NER cross-lingual transfer depends largely on the overlap of entity chunks. If a source and target language have more entities in common, the transfer ability is stronger. Models using cross-lingual transfer also appear to be somewhat more robust to certain perturbations of the input, perhaps indicating an ability to leverage stronger representations derived from the HRL. Our research provides valuable insights into cross-lingual transfer and its implications for NLP applications, and underscores the need to consider linguistic nuances and potential limitations when employing MLLMs across distinct languages.

3.1.1 Datasets

Our exploration focuses on 13 language pairs from a pool of 21 languages: Arabic/Farsi, Arabic/Hindi, Czech/Slovak, Dutch/Afrikaans, English/Scots, English/Welsh, French/Breton, French/Occitan, Indonesian/Malay, Italian/Sicilian, Spanish/Aragonese, Spanish/Asturian, and Spanish/Catalan. These languages were chosen following the rationale established by Nath et al. [2022] for collecting loanword data: languages with a sufficiently high density for evaluation, with varying levels of expected vocabulary overlap. While Nath et al. [2022]'s data source is Wiktionary, we examined the WikiANN dataset [Pan et al., 2017], a common multilingual NER dataset, and selected language pairs consisting of one language with greater resources in the data and one with fewer resources, where a substantial level of overlap in the vocabulary should be expected due to the languages having some areal (e.g., French/Breton) or close genetic (same sub-family) relationship (e.g., Czech/Slovak), or known history of borrowing at large scale (e.g., Arabic/Farsi). See Table 3.1. One of these pairs—Arabic/Hindi—serves as a kind of "control" group; although there is a substantial amount of vocabulary shared due to borrowing, the two languages use different native scripts, and since we perform experiments over the raw text without transliteration, we expected there to be little token- or word-level overlap between these languages in terms of the raw text. In this research, we will follow the pairwise notation L1/L2, where L1 refers to the HRL in a pair and L2, the LRL. In the selected pairs, the HRL is often a major world or national language while the LRL is often a regional or minority language, providing an opportunity to examine where biases toward major world languages creep into NLP for minority or underrepresented languages.

WikiANN contains NER data for every language in our set, in standard BIO notation for person (`PER`), location (`LOC`), organization (`ORG`), and miscellaneous (`MISC`) categories, and so serves as the NER dataset for our experiments.

We chose NER as an experimental task as it is one of the most common NLP tasks in both research and industrial applications.[1] Another of the most common tasks is document

[1] https://gradientflow.com/2021nlpsurvey/

HRL	Size	LRL	Size	Relationship
Arabic (ar)	100K	Farsi (fa)	100K	Borrowing
Arabic (ar)	100K	Hindi (hi)	42.6K	Borrowing
Czech (cs)	100K	Slovak (sk)	61.1K	Areal, Genetic
Dutch (nl)	100K	Afrikaans (af)	29.7K	Genetic
English (en)	100K	Scots (sco)	5.1K	Areal, Genetic
English (en)	100K	Welsh (cy)	15.2K	Areal, Borrowing
French (fr)	100K	Breton (br)	8.1K	Areal, Borrowing
French (fr)	100K	Occitan (oc)	13.7K	Areal, Genetic
Indonesian (id)	100K	Malay (ms)	60.3K	Areal, Genetic
Italian (it)	100K	Sicilian (scn)	1.4K	Areal, Genetic
Spanish (es)	100K	Aragonese (an)	5.1K	Areal, Genetic
Spanish (es)	100K	Asturian (ast)	85.5K	Areal, Genetic
Spanish (es)	100K	Catalan (ca)	100K	Areal, Genetic

Table 3.1: Size of languages for section title prediction dataset, and relationship between languages in studied pair.

classification, but a document classification corpus in all of the languages we evaluate is not available.[2] We therefore approximate this task by creating a Wikipedia Section Title Prediction dataset for our languages of interest following the methodology of Kakwani et al. [2020]. Section title prediction is an appropriate approximation for document classification because Wikipedia articles are usually sectioned into distinct topics regarding the subject, such as "Early life" and "Career" for biographical articles, or "Government" and "Economy" for articles about states or localities, which parallel categories in document classification corpora like the Reuters Corpus dataset.

We built the Section Title corpus by crawling the Wikipedia pages corresponding to each specific language. We extracted the pages with at least 4 sections. Following this, we used the WikiExtractor tool [Attardi, 2015] to systematically extract sections along with their associated second and third-level titles from the Wikipedia pages. The dataset contains subsection text paired with four candidate titles, of which one is correct and the others are titles of other sections of the same article. We collected as many samples as possible for each language up to a limit of 100,000. Data sizes are given in Table 3.1.

All datasets were then divided into an 80:20 train/test split.

[2]Schwenk and Li [2018] come closest, with 8 languages.

3.1.2 Methodology

To assess the effect of zero-shot transfer between languages with overlapping vocabulary, we compare the performance of the MBERT and XLM-R models. These are two of the most well-known multilingual, publicly-available encoder-style models in use, notable for their abilities to align semantically similar representations across languages and for their multilingual task performance [Pires et al., 2019, Conneau et al., 2020a]. We evaluate both models in a native setting when they are fully fine-tuned on the two tasks in an LRL; in a transfer setting, where they are trained on an HRL and evaluated on the paired LRL; and under different perturbations of the data. Details are given in Sec. 3.1.3.

The two tasks chosen represent two fundamental NLP inference challenges: information extraction (IE) from unstructured texts, encompassing the identification of individuals' names, organizations, geographical locations, etc.; and the selection of the appropriate classification of a text from multiple options, requiring the selection of the most appropriate title for a section text among the four presented choices. NER is a valuable IE task to assess the effects of vocabulary overlap on cross-lingual transfer, because named entities in LRLs are often borrowed from HRLs with minimal modifications. Title selection is a valuable classification task for similar reasons: section texts represent "documents" where key words evidencing the section title options may be shared or similar between languages. For instance in the Welsh example "*mae logo'r ddarpar fanc*," *logo* overlaps with English and predicts the section title "logo."

Perturbation methods

Two additional small datasets were gathered for the perturbation process: 1) A dataset of given names for each target language scraped from the `[Language]_given_names` category of Wiktionary. 2) A dataset of places for each target language scraped from its `Places` category in Wiktionary.

We implemented four methods to generate adversarial sets:

1. Change given names (first element) of all PER entities to randomly-chosen elements of the given names dataset in the same language.

2. Change names of all LOC entities to randomly-chosen elements of the placenames dataset in the same language.

3. Replace named entities in the L2 test file that also occur in the L1 training file with a named entity with the same tag that occurs in L2 test but not in L1.

   For example, *Tour Eiffel* (Eiffel Tower) is the same in Breton and French, so it may be replaced with *Bolz-enor Pariz* (Arc de Triomphe), which is the same NER type, but non-overlapping.

4. Leave the the named entity unchanged, but instead take *surrounding* words in the L2 test file that occur in the L1 training file and replace them with words unique to L2 test that are not punctuation or stop words.[3] The substitute word is the word with the highest cosine similarity with the original word.[4]

   For example, in *An Tour Eiffel zo un tour metalek savet e Pariz gant Gustave Eiffel* ("The Eiffel Tower is a metal tower built in Paris by Gustave Eiffel"), the word *tour* is the same in Breton and French, so it is replaced with a semantically-similar Breton word that doesn't also occur in the French wordlist. Table 3.2 shows a sample of original words and their substitutes determined by cosine similarity according to MBERT and XLM-R. This sample shows how bias can creep into the substitutions, *viz.* "males" for "hijackers".

5. Combine the perturbations 3 and 4 to change the entities and surrounding words at the same time.

[3] https://github.com/stopwords-iso/stopwords-iso/ provided the stop word lists.

[4] Computing the most similar word follows Nath et al. [2022] and constructs a dummy "sentence" consisting of `[CLS]<word>[SEP]` (MBERT) or `<bos><word><eos>` (XLM-R) for each word, and computing the cosine distance between the contextualized first token representation.

Content word	MBERT option	XLM-R option
channels	shots	broadcasts
bred	lived	assistant
population	parted	people
serve	carried	arrangement
place	event	there
journalist	lawyer	activist
female	woman	woman
hijackers	triumphs	males
defeated	won	defeating

Table 3.2: Sample of highest cosine-similarity alternatives existing in the test split of the English dataset. Since we focus on non-English languages, these are note actual examples from the data but rather illustrative of the phenomenon and extracted using the same methdology in use.

For section title prediction, we use only perturbation 4, which does not require the training data to be tagged with NE labels.

Adversarial set generation was conducted automatically. For semantic-level perturbations like perturbation 4, a manually-created semantic resource like WordNet is appealing, but infeasible due to some of the languages we examine, for which there do not exist sufficient WordNet or WordNet-like resources. For instance, we considered Global WordNet[5] but even this resource does not exist for all of our languages of interest, for instance Aragonese. This is to be expected given that the domain in question is low-resource languages. Secondly, WordNets that do exist for the languages of interest are incomplete or very small, lacking a substantial number of words. Thirdly, WordNet synsets do not necessarily cover alternative potential substitutions that would be sensical, without traversing the synset tree quite far from the word of origin. For instance, instead of saying "I see my sister that day," a reasonable perturbation would be "I see my mother that day," but despite being semantically similary, "sister" and "mother" are not in a synset. Similarly, a WordNet-based approach for synonym replacement becomes challenging when dealing with homographs, such as (in English) *might/could* vs. *might/strength*. Therefore, for consistency across all language, we used the automated method as described.

Perturbations 3 and 4 are comprehensively explained below:

[5]https://omwn.org/

Entity replacement:

- Read examples in both languages, L1 and L2, to retrieve words and their corresponding labels.
- Filter out words with labels labeled as 'O' in both languages.
- Identify complete entities within the text. Each entity typically begins with a word labeled as "B-" (beginning part of the entity) and may contain other words labeled as "I-" (intermediate parts). We iterate through the words to extract complete entities, stopping when we encounter another "B-" labeled word or when we've parsed all the words.
- After extracting complete entities from L1 and L2, we find the intersection of these entities, referring to them as common entities.
- Additionally, we identify entities in L2 that do not exist in L1 and group them based on their labels, referring to these as unique entities.
- Finally, for each common entity, we randomly select an entity from the unique entities pool, ensuring that their labels are similar to the common entity.

Context Replacement:

- Extract individual words and their named entity (NE) tags from both the L1 train file and the L2 test file.
- Eliminate stop words, single-character words, words with punctuation, or those tagged as anything other than "O." Ensure that all words are in lowercase from this point forward.
- Find the intersection of these word sets, resulting in the "common words."
- Identify all unique words in L2, forming the "unique words" set.

- Retrieve embedding vectors for each word in both models. Construct a dummy sentence format such as [CLS]<word>[SEP] (for MBERT) or <bos><word><eos> (for XLMR) to obtain the vector for the [CLS]/<bos> token.
- Calculate the cosine similarity for each common word with all unique words.
- Select the unique word with the highest cosine similarity for each common word.
- Replace each common word with its corresponding selected word. If the common word is capitalized, ensure the selected word is also capitalized.

Computing vocabulary overlap

To investigate the correlation between vocabulary overlap and zero-shot knowledge transfer across languages, we started by extracting all labeled NER chunks within the datasets of the paired languages, and computing the percentage of identical words with identical labels—excluding those tagged `O`. For a pair, the percentage of overlap between L1 and L2 is considered to be number of words shared between the L1 training set and L2 test set divided by the total number of words in the L2 test set. See Table 3.3.

For section title prediction, we identified common and unique words for perturbation, and performed overlap computation, using the first 128 tokens from each section. Due to variances in tokenization between MBERT and XLM-R, there may be different values for overlap between the two models (see Table 3.4).

3.1.3 Evaluation

We used the `bert-base-multilingual-cased` [Devlin et al., 2019a] (MBERT) and `xlm-roberta-base` [Conneau et al., 2020a] (XLM-R) variants. We fine-tuned each of these models in the two tasks in multiple conditions:

1) *Native L2*: A standard fine-tuning on the L2 training data and testing on the L2 test data; 2) *Cross-lingual transfer*: Fine-tuning on the L1 training data and testing on the L2 test

L1	L2	% overlap
ar	hi	4.88
ar	fa	19.94
cs	sk	39.55
nl	af	31.57
en	sco	25.19
en	cy	22.07
fr	br	23.33
fr	oc	23.61
it	scn	43.17
id	ms	41.87
es	an	46.26
es	ast	47.66
es	ca	36.77

Table 3.3: Named entity overlap in L1-train/L2-test for NER.

data. Since the language pairs were selected for vocabulary overlap, this condition allowed us to assess the level to which performance on a LRL can be achieved by exposure to data only from an HRL that may contain similar task-relevant vocabulary. 3) *Perturbation*: In each of the two above conditions, the task-relevant perturbations are applied, to further assess the extent to which cross-lingual transfer or native performance is robust to adversarial changes to the input.

To account for randomness in training and testing sample selection, which could lead to disparate values, we averaged the results across three runs. Our primary evaluation metric is F_1 score relative to token overlap in the chosen sentences rather than the entire token pool. This methodology aligns with our earlier strategy, where we concentrated on examining the overlap between the L1 training set and the L2 testing set, rather than the entire datasets of L1 and L2, offering a more precise insight into the multilingual capacities of the models relative to specific training and testing data.

3.1.4 Results

Table 3.5 shows the performance on NER and title selection of MBERT and XLM-R, for all evaluated language pairs, with baseline (unperturbed) scores, and scores after all appli-

L1	L2	Model	% overlap
ar	hi	MBERT	2.12
ar	hi	XLM-R	1.98
ar	fa	MBERT	14.65
ar	fa	XLM-R	15.01
cs	sk	MBERT	24.26
cs	sk	XLM-R	24.18
nl	af	MBERT	22.63
nl	af	XLM-R	22.57
en	sco	MBERT	29.22
en	sco	XLM-R	29.19
en	cy	MBERT	17.31
en	cy	XLM-R	17.08
fr	br	MBERT	9.50
fr	br	XLM-R	9.44
fr	oc	MBERT	23.09
fr	oc	XLM-R	23.04
id	ms	MBERT	36.34
id	ms	XLM-R	36.34
it	scn	MBERT	25.99
it	scn	XLM-R	25.86
es	an	MBERT	24.80
es	an	XLM-R	24.77
es	ast	MBERT	29.59
es	ast	XLM-R	29.65
es	ca	MBERT	17.12
es	ca	XLM-R	17.20

Table 3.4: Word overlap in L1-train/L2-test for title prediction task.

		MBERT								XLM-R							
		NER						WikiTitle		NER						WikiTitle	
Train	**Test**	**Base**	**P1**	**P2**	**P3**	**P4**	**P5**	**Base**	**P4**	**Base**	**P1**	**P2**	**P3**	**P4**	**P5**	**Base**	**P4**
ar	hi	67.2	64.2	68.9	67.2	67.2	67.2	63.6	63.0	67.3	67.4	70.7	67.3	67.3	67.3	75.8	75.0
hi	hi	86.7	86.5	87.2	71.3	79.0	66.7	73.8	72.5	87.5	87.2	88.1	76.6	80.7	68.3	77.8	77.1
ar	fa	45.0	43.0	44.7	45.0	45.0	44.9	79.3	77.1	43.6	42.8	40.1	43.6	43.5	43.4	78.0	73.9
fa	fa	90.3	88.0	89.1	86.5	60.8	56.7	81.6	79.1	89.4	88.2	87.4	85.5	78.2	74.1	81.0	76.5
cs	sk	82.9	82.4	87.0	78.4	82.5	77.9	80.3	75.6	78.0	77.2	86.1	73.4	78.1	73.5	80.3	73.3
sk	sk	92.6	91.7	91.0	86.4	92.1	85.0	83.5	78.5	91.5	91.1	89.8	81.5	88.6	77.5	82.3	75.1
nl	af	81.2	81.0	83.8	78.4	81.2	78.6	78.5	71.6	79.9	80.0	81.5	77.8	79.3	76.9	75.4	71.6
af	af	92.2	91.6	92.1	81.1	89.5	78.5	81.3	74.3	89.8	90.0	90.8	77.9	86.2	76.0	76.8	66.9
en	sco	78.3	77.9	72.0	71.0	78.2	71.7	85.7	76.2	62.4	62.0	60.6	60.6	63.2	61.3	75.5	62.5
sco	sco	93.4	93.0	83.2	81.0	91.4	79.2	88.6	80.8	90.2	89.6	82.5	79.6	87.5	75.0	71.5	60.2
en	cy	62.5	61.8	65.3	61.3	62.4	61.6	67.5	63.6	61.5	61.2	64.9	60.4	61.4	60.4	61.7	58.8
cy	cy	92.6	91.9	87.1	77.0	89.5	75.0	76.6	73.5	90.9	90.4	85.1	76.1	83.1	67.8	72.1	67.3
fr	br	74.3	71.8	73.5	73.3	74.2	72.8	66.6	63.1	66.3	64.2	66.6	64.7	66.3	64.5	59.3	54.0
br	br	92.8	88.4	88.2	84.5	88.8	79.9	71.1	66.1	89.1	85.8	87.1	81.3	82.8	74.1	59.3	55.2
fr	oc	83.9	83.7	89.1	83.5	83.7	83.4	76.6	71.9	72.5	72.3	78.8	71.8	72.3	71.9	66.5	59.1
oc	oc	95.3	94.9	95.8	92.3	87.8	83.9	79.1	75.2	93.8	93.0	94.6	91.5	92.6	89.8	67.0	61.3
id	ms	68.7	67.7	76.7	64.8	68.5	64.8	79.9	68.4	69.7	69.5	79.9	66.2	69.5	65.8	78.3	58.4
ms	ms	92.4	92.6	83.5	81.7	81.8	70.5	82.7	71.8	92.4	91.9	89.1	71.7	79.7	59.5	80.3	62.4
it	scn	63.7	63.3	80.2	58.4	49.5	45.4	71.0	66.2	60.8	60.7	74.0	55.3	50.4	45.5	60.7	46.8
scn	scn	92.9	91.1	88.1	79.8	74.4	64.9	64.3	57.1	90.5	88.2	82.8	79.7	72.4	62.5	40.0	39.0
es	an	**88.0**	87.9	84.8	85.4	80.7	77.5	86.1	76.3	**86.1**	86.2	86.4	83.3	75.3	72.9	77.0	55.0
an	an	95.8	95.8	88.4	85.6	90.9	**79.1**	83.4	76.8	94.2	93.6	92.5	79.8	80.4	**66.1**	72.6	59.4
es	ast	**90.4**	90.2	86.0	85.1	89.6	84.6	84.1	77.5	**84.3**	84.2	86.0	77.0	84.1	76.3	76.7	59.6
ast	ast	93.6	92.8	90.1	82.7	93.3	**79.7**	85.2	78.4	89.6	89.2	90.1	77.7	90.0	**76.4**	80.3	68.0
es	ca	**85.1**	84.3	87.2	84.0	85.1	84.0	79.3	75.9	**82.6**	82.8	83.9	80.8	82.3	79.8	72.8	66.2
ca	ca	92.3	91.5	91.6	87.3	91.6	**86.5**	85.9	83.0	89.4	89.6	88.0	83.3	88.6	**82.1**	83.9	78.0

Table 3.5: $F_1 \times 100$ (NER) and accuracy (title prediction) scores for MBERT and XLM-R without perturbation (**Base**) and with all applicable perturbations on all evaluated language pairs. P1-5 references the different perturbations described in the list in Sec. 3.1.2. Bold numbers refer to native and cross-lingual NER accuracy values when the source language is Spanish, which are discussed further as noteworthy cases in Sec. 3.1.5.

	MBERT				XLM-R			
	NER: L2	avg. Δ F_1	NER: L1→L2	avg. Δ F_1	NER: L2	avg. Δ F_1	NER: L1→L2	avg. Δ F_1
P1	$p = 0.0118$	-1.00	$p = 0.0046$	-0.92	$p = 0.0116$	-0.80	$p = 0.0655$	-0.34
P2	$p = 0.0033$	-3.65	$p = 0.2096$	2.15	$p = 0.0165$	-2.33	$p = 0.0246$	3.42
P3	$p < 0.0001$	-9.66	$p = 0.0013$	-2.72	$p < 0.0001$	-10.46	$p = 0.0013$	-2.52
P4	$p = 0.0105$	-7.07	$p = 0.1500$	-1.80	$p = 0.0004$	-6.73	$p = 0.1499$	-1.69
P5	$p < 0.0001$	-16.71	$p = 0.0106$	-4.36	$p < 0.0001$	-17.62	$p = 0.0090$	-4.26
	Titles: L2	avg. Δ acc.	Titles: L1→L2	avg. Δ acc.	Titles: L2	avg. Δ acc.	Titles: L1→L2	avg. Δ acc.
P4	$p < 0.0001$	-5.38	$p < 0.0001$	-5.54	$p = 0.0002$	-7.57	$p = 0.0003$	-9.52

Table 3.6: Effects of different perturbations, per model type by a paired, two-tailed t-test, and average change in F1/accuracy. The average change in F1/accuracy metrics after perturbation appears significantly less during cross-lingual transfer than in the native setting. While XLM-R demonstrates nearly equivalent robustness to perturbation in both settings in NER when compared to MBERT, its robustness diminishes in the sentence-level task—section title prediction—where word memorization might be more applicable.

cable perturbations. Table 3.6 shows the statistical significance of the performance changes associated with each perturbation, given the model and cross-lingual transfer setting.

We can see that using an HRL→LRL transfer setting never reaches the performance of the native LRL fine-tuning, falling below by ∼1-30% F_1/accuracy. Where cross-lingual transfer comes closest is in language pairs that are geographically close and genetically close (e.g., Spanish/Asturian), because core vocabulary is likely to be similar already, and the document sets in the training data likely share named entities like names of people and locations that are commonly discussed in the two languages. Interestingly, though, we see that the cross-lingual transfer models appear to be more robust to certain perturbations, such as P4 (perturbing context words), which by itself did not significantly change the NER results for MBERT or XLM-R using cross-lingual transfer. MBERT cross-lingual models are also more robust to P2 than XLM-R cross-lingual models, as the perturbation of LOC tags also did not affect results to a statistically significant extent. XLM-R was more robust to random replacement of B-PER tags. On average, MBERT appears more robust to the perturbations we applied, where even the performance changes that were statistically significant were less so than those of XLM-R. However, we should note that even the simple perturbation of changing context words in the title selection task degraded performance to a very significant level nearly across the board.

3.1.5 Discussion

NER

Under perturbation of `B-PER` tokens (P1), macro F_1 score changes between 0–4% and F1 of `PER` classes changes from 1–13%, depending on the `B-PER` distribution and the number of available alternatives. Under perturbation of `LOC` tokens (P2), macro F_1 score changes from 1–13% and `LOC` F_1 (averaged across `B-LOC` and `I-LOC`) changes from 1–27%. Although `LOC` tokens form a greater proportion of the test data than `B-PER` tokens alone, perturbing the `LOC` tokens causes far less drop in performance. One reason may be that many *LOC* entities in the test sets include 3–6 individual tokens, while the alternate candidates scraped from Wiktionary mostly include 2 tokens, making it easier to segment shorter NE chunks. Notably, `LOC` perturbation frequently causes performance on a language pair to *rise*, perhaps significantly (see Italian/Sicilian), signaling cases where cross-lingual transfer provides increased robustness to adversarial data, relative to baseline performance. Distributions of `B-PER` and `LOC` tokens in the LRL test sets are shown in Fig. 3.1.

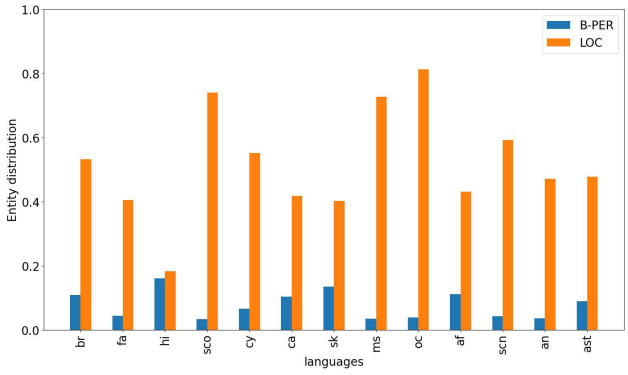

Figure 3.1: WikiANN distribution of B-PER and LOC for different LRLs.

Fig. 3.2 (first row) shows macro F_1 change under perturbations 3, 4, and 5 as a function of the degree of vocabulary overlap between L1 and L2 for all pairs. We observe a clear correlation between the proportion of shared vocabulary items between the train and test sets and the performance degradation when test entities are perturbed to substitute those

that are shared between the sets with unique entities. In the cross-lingual transfer models, this removes words that are common between L1 and L2 and replaces them with words unique to L2. This suggests that multilingual models' NER performance for LRLs depends to some extent on word memorization, and the extent to which this is true is a function of vocabulary overlap with other, more well-resourced languages; the model may not be recognizing that a term is a named entity in Occitan or Catalan, but rather one from a French or Spanish corpus and is "riding" its ability to perform in those languages.

Degradation of performance under P4 (changing surrounding context words—see Fig. 3.2, top center) is pronounced in the native L2 models, but for most cross-lingual transfer models, degradation is small. Two notable exceptions are Italian/Sicilian and Spanish/Aragonese, where perturbing context words causes a drop of ∼8–20 F_1 points.

Under perturbation 5 (the combination of perturbations 3 and 4), NER performance suffers a fairly precipitous drop. The three pairs involving Spanish (Spanish/Aragonese, Spanish/Asturian, and Spanish/Catalan; bolded in Table 3.5) are notable here, in that P5 brings the native model down to the performance level of the unperturbed cross-lingual transfer model (in the case of Aragonese and Asturian, perturbed native model performance drops *below* the performance of the model trained for cross-lingual transfer from Spanish). This also suggests that on these LRLs, MLLMs may be leveraging their capabilities in Spanish to achieve their initial performances.

Section Title Prediction

In this task as with NER, in the case of cross-lingual transfer, the degree of overlap has a significant impact on the F_1 score, with a noticeable drop in F_1 score when key words in the target language are removed.

In the lower center of Fig. 3.2, the plot shows the effect of perturbation 4 (perturbing context words) on the title selection, when the substitutes are chosen randomly instead by choosing the most similar candidate according to the cosine function. We can see that

MBERT suffers more from the random perturbation than from the cosine perturbation in the cross-lingual transfer condition, but both models suffer more from this perturbation when compared to the other one in the native condition.

One point of note is that in the case of Arabic/Hindi, the one pair where the two languages use different native scripts, none of the perturbations appeared to have much effect in the cross-lingual setting. This is expected, due to the low default token overlap: a model fine-tuned on Arabic will have difficulty handling Hindi words written in Devanagari, regardless of what they are. What is interesting is that in the case of Arabic/*Persian*, which do share the same script, the same is true, and Arabic/Persian cross-lingual transfer performance on NER is substantially lower than on Arabic/Hindi, despite the differences in script and therefore tokenization.

I, as a native Persian speaker conducted a qualitative analysis of words judged similar for use in P4 and P5, according to their cosine similarity. In both MBERT and XLM, similar words were often found to rhyme or share subwords: e.g., *mârk* ("brand") vs. *mârd* ("evil"), or *sard* ("story") vs. *sardard* ("headache"). This implies that subword tokens are being overvalued when computing vectors from `[CLS]/<bos>` tokens in Persian, and perhaps other LRLs.

In the native condition, overlap is computed using non-stop words found in both the training and test files. Consequently, when the value is low, as is the case with Breton, we would expect performance under perturbation to remain relatively unchanged (compare Hindi), but Breton still suffers a performance loss of ∼4–5 F_1 points. This suggests that this task relies heavily on word memorization of the training data, as a similar drop in performance is observed when words are substituted randomly. The semantic similarities of the substitute words under P4 seem to not matter. Sicilian performance in MBERT substantially exceeds that of XLM-R, but also suffers more under perturbation. Sicilian training data is included in the pretraining data for MBERT but not for XLM-R, which partially explains this trend, but the much lower performance of the native Sicilian XLM-R

model on title selection compared to NER suggests that NER fine-tuning can leverage other representations (e.g., common named entities between Italian and Sicilian) in a way that a task that requires inference over more common words, like title selection, cannot.

3.2 Challenges of Cross-Lingual Transfer in Sentence Transformers from HRLs to LRLs

Unlike words, sentences have more complex distributions and do not repeat frequently across texts, making it harder to generate reliable embeddings for them. Word-based embeddings, while effective for lexical-level tasks such as word similarity or syntactic parsing, struggle to capture the intricate meaning conveyed at the sentence level. This difficulty arises from the fact that sentence embeddings need to encapsulate not just the meaning of individual words, but also the syntactic structure, semantic roles, and relationships between words within a sentence. Given this, it is rational to employ word Transformer models, which have proven effective in understanding word contexts, and use them as a foundation for training on sentence-based tasks.

There could be different methods to transfer knowledge from HRL sentence Transformers to LRLs. Knowledge transfer is a crucial aspect of enabling the development of sentence Transformers for LRLs, especially in multilingual and cross-lingual contexts. [Reimers and Gurevych, 2020] proposed creating multilingual versions of monolingual models by training them on the principle that a translated sentence in one language should occupy the same position in the vector space as the original sentence in another language. This principle aligns well with cross-lingual sentence embedding tasks, where semantic similarity across languages is essential. In this approach, the model distills the learned knowledge from English in SBERT as a teacher model and injects it into XLM-R, a multilingual model, followed by a mean-pooling process to generate sentence embeddings in the student model.

However, despite its promise, this method may not always be effective for distinguish-

ing between semantically dissimilar sentences in different languages, as it relies on translation datasets rather than the triplet network approach used for training the teacher model, SBERT [Reimers and Gurevych, 2019].

We conducted an experiment to illustrate this issue using the model sentence-Transformers/all-MiniLM-L12-v2 [6]. The model produced a cosine similarity score of 0.73 for the pair "I am going to go home." and "Everyone is going to go home, not me.", while it gave a score of 0.70 for "I am going to go home." and "Going home is the plan." These results demonstrate the limitations of sentence Transformers in capturing subtle nuances in sentence meaning, particularly when dealing with cross-lingual contexts.

Following this work [Reimers and Gurevych, 2020], we modified the model to train on a triplet dataset, which has been widely used in training for sentence similarity tasks. The knowledge distillation method proposed in their work did not fully take advantage of the benefits of training sentence Transformers with a triplet network and dataset [Reimers and Gurevych, 2019]. Triplet loss allows the model to learn better sentence embeddings by considering not just positive sentence pairs, but also negative examples, thereby improving the model's ability to distinguish between semantically similar and dissimilar sentences.

We addressed the shortcomings of prior approaches by gathering a large cross-lingual triplet dataset and incorporating triplet loss in the knowledge distillation process. We generated the cross-lingual triplet dataset as each triplet consists of one Persian sentence (the anchor) and two English sentences (the positive and negative examples). By using a triplet dataset, the model learns to map semantically similar sentences close together in the vector space, while pushing dissimilar sentences apart, even if they exist in different languages. This approach is particularly effective in multilingual scenarios, as it does not require strict binary labels.

However, although our model outperforms the previous one in identifying dissimilar sentences, its performance still faces limitations, particularly with longer sentences. These mod-

[6] https://huggingface.co/sentence-Transformers/all-MiniLM-L12-v2

els struggle to fully capture the complexity and structure of extended phrases, which affects their ability to generate accurate embeddings for such inputs.

To address this issue, we developed a more targeted approach. First, we designed a specific model that incorporates a word-level translation mapping between two languages, ensuring that training remains effective across linguistic boundaries. This mapping is critical, as without it, words in Latin-based languages risk being broken into short, uninformative character sequences, while words from non-Latin alphabets may incorrectly map to the UNK token. By establishing this mapping, we improve the model's ability to handle the structural and linguistic diversity present in multilingual datasets.

Following this step, we fine-tuned an English sentence Transformer, specifically SBERT, to create a bilingual model capable of mapping sentences from a non-English language (Persian in this case) into the same vector space as English.

This methodology highlights an important insight: even though a mapped sentence might not seem intuitively meaningful to an attention-based Transformer due to differences in sentence structure, it significantly aids the model in finding similar sentences through word embeddings. To gain a clearer understanding of what happens during the mapping process and how the model produces its results, refer to Table 3.7.

3.2.1 Datasets

Data collection for this project involved systematically crawling Wikipedia pages in both Persian and English. Specifically, we retrieved all Persian Wikipedia pages that contained links to their corresponding English counterparts, along with their English equivalents. This alignment of linked pages allowed us to create a dataset grounded in real-world, multilingual content. The Persian Wikipedia currently consists of 1,013,254 pages, providing a vast pool of data for our cross-lingual study.

To extract marked-down texts in both HTML and non-HTML formats, we employed the WikiExtractor tool, which efficiently processes large volumes of Wikipedia data. This

	Correct	Incorrect
Original	*Az daheh 1960, zamani ke keshvarhaye san'ati kârkhaneh-hayi ra mo'arrefi kardand, gostarish-e ziad-i yaft.*	*Yeki digar az vijegi-haye mavadd-e ceramiki, arzani va faravani-ye nesbi-ye in mavadd ast.*
Translation	It has significantly expanded since the 1960s, when industrial countries introduced factories.	Another characteristic of ceramic materials is their low cost and relative abundance.
Mapped	from Decade 1960 when that Countries industrial factory those particle for direct object introduction they did expansion a lot found	one other from features Material ceramic cheapness and Abundance relative this Material is
Predicted target	It expanded greatly from the 1960s when industrialized countries introduced factories. (Score: 0.9143)	Almost all types of materials, including metals, ceramics, polymers, and composites can be coated on similar or dissimilar materials. (Score: 0.6869)

Table 3.7: Example preliminary results showing Persian sentences (transliterated) and their predicted English equivalents in a bilingual sentence Transformer. "Score" denotes the cosine similarity between the source and target sentences' embedding vectors.

tool enabled us to focus on relevant sections of text while discarding irrelevant metadata, ensuring clean and usable data for further analysis.

After extracting the contents of each page, we focused on extracting all section titles from both the Persian and English versions. One of the key challenges we encountered when working with Persian Wikipedia compared to the English version was the inconsistency in formatting section titles. Persian Wikipedia uses a variety of formats for designating section headers, necessitating a robust pattern-matching approach. To handle this, we used a regular expression pattern, `r'(={1,6})(.*?)(={1,6})`(<h[1-6]>)(.*?)(</h[1-6]>)'|, which captures different patterns for titles in both HTML and markdown formats.

The "\u200c" is a Unicode character known as the "Zero Width Non-Joiner" (ZWNJ). It is used in writing systems like Persian, Arabic, and others to prevent the automatic joining of characters that are otherwise connected in cursive scripts. When placed between two characters, it ensures that they do not form a ligature, and instead, appear as separate characters. When parsing a Wikipedia page text, the Zero Width Non-Joiner ("\u200c") appears in its Unicode form. As part of the text-cleaning process, we removed it throughout the titles and texts.

Given that sentence Transformers perform more effectively with shorter sentences, we automated the identification of sections within each pair of pages that shared equivalent or closely related titles. Titles were considered similar if the similarity threshold between the embeddings from sentence Transformer exceeded 0.7. A critical constraint we applied was that each title could only be matched with one equivalent title from the other language. In cases where multiple titles exhibited similarity scores greater than 0.7, the pairing was determined based on the highest similarity score, rather than the order of appearance on the page. This process ensured that the thematic and semantic similarity between sections was prioritized, leading to coherent and meaningful cross-lingual content alignment [Ein Dor et al., 2018][7].

To maintain the quality of the aligned sections, we applied additional filtering criteria. Only sections with equivalent counterparts that exceeded 200 characters in length were considered. Furthermore, we filtered out sentences that contained 10 or fewer words, as these were deemed insufficiently informative for training a sentence Transformer model.

Crawling this vast number of pages sequentially would have been highly inefficient, taking up to a month to complete. To optimize the process, we implemented concurrency, parallelizing the crawling operation. This approach significantly reduced the processing time by allowing multiple pages to be crawled simultaneously. Given the need to use concurrency for increased speed and the requirement to load the sentence model on CUDA, we opted for multithreading rather than multiprocessing to avoid re-initializing the model every time a new task was executed.

By applying these stringent filtering criteria and optimizing the crawling process, we ensured that the final dataset consisted of high-quality sentence pairs. These pairs were highly suitable for building and evaluating cross-lingual similarity datasets. The combination of multithreading, strict filtering criteria, and automated parsing guarantees the creation of a dataset that is not only large in scale but also rich in content, tailored for training and

[7]These pairings were verified by a fluent bilingual speaker to ensure accuracy and relevance.

assessing multilingual models.

Each sample in the dataset is structured as a triplet, consisting of:

1. **Anchor:** A Persian sentence serving as the anchor or reference point in the triplet. These sentences were drawn from a wide range of topics and genres in Persian Wikipedia, which provides broad coverage of various linguistic constructs and nuances in the Persian language.

2. **Positive Sample:** An English sentence that is semantically aligned with the anchor sentence. The positive samples were chosen based on the sentence position within a corresponding section. For instance, for the n^{th} sentence in the identified Persian section, the corresponding n^{th} English sentence from the paired English section was selected. This pairing ensured that positive samples were closely aligned in both content and context.

3. **Negative Sample:** An English sentence randomly selected from a different section of the same Wikipedia page. By ensuring the negative sample came from a different section, we aimed to introduce a degree of semantic dissimilarity that would challenge the model to differentiate between relevant and irrelevant cross-lingual sentence pairs.

Overall, we generated a total of 545,233 triplet samples. To evaluate the performance of the model, we randomly held out 50K samples as test data, resulting in an approximately 90:10:10 split between the training/validation/test sets. This division was strategically made to ensure that the model would have sufficient data for training while preserving a substantial portion for testing and validation purposes.

There were two other datasets on which we could evaluate the models:

The first dataset we used comes from the OPUS website [8], a large-scale open-source collection of parallel corpora for multiple languages [Tiedemann, 2012]. OPUS offers a vast repository of aligned sentences for hundreds of language pairs, making it a valuable resource

[8] http://opus.nlpl.eu/

for cross-lingual NLP tasks such as machine translation, semantic textual similarity (STS), and more. For this study, we specifically selected the Persian-English parallel corpus from OPUS, which provides sentence pairs aligned by meaning in both languages. This rich dataset allows models to learn and improve their understanding of linguistic patterns and semantic relationships between Persian and English sentences. By leveraging this parallel data, our models can be trained to recognize similarities and differences in sentence structures and meanings across languages, making it crucial for cross-lingual STS tasks.

However, we applied certain preprocessing steps to ensure the quality of the dataset. One significant decision was to exclude sentences containing very few words, particularly sentences with only 1, 2, or 3 words. These extremely short sentences often lack sufficient context, making it difficult for machine learning models to capture meaningful semantic relationships between them. Including such short sentences in training can decrease model performance, as the brevity of these examples introduces noise and leads to poorer generalization in real-world tasks. By filtering out these sentences, we aimed to improve the model's ability to handle more complex and contextually rich sentence pairs, ultimately boosting its performance in evaluating semantic textual similarity.

Moreover, we incorporated a scored dataset specifically for Persian-English sentence pairs. This dataset provides sentence similarity scores ranging from 1 to 5, offering a more nuanced evaluation of semantic similarity. This dataset is PESTS: Persian-English Cross-Lingual Corpus for Semantic Textual Similarity [Abdous et al., 2024], which aims to bridge the gap in cross-lingual semantic textual similarity (STS) research between Persian and English. Given the limited availability of high-quality STS resources for Persian, PESTS plays a crucial role in advancing NLP models for the Persian language.

While PESTS is useful for evaluation purposes, it consists of only about 4,300 pairs, with more than half representing either irrelevant or low-similarity pairs. This imbalance presents a challenge, as the irrelevant pairs often involve completely unrelated topics, which can make it difficult for models to effectively measure dissimilarity. Additionally, the relatively

small size of this dataset makes it unsuitable for training large models, limiting its use to evaluation and fine-tuning tasks. Despite these limitations, PESTS offers valuable insights for cross-lingual STS, though larger and more diverse datasets are needed for robust model training.

It's important to note that while building the dataset, we initially considered using GPT models to generate synthetic anchor, positive, and negative sentence pairs. However, relying on these models for data generation wouldn't have been ideal, as they are not 100% accurate. In fact, the approach proposed here, as well as similar techniques, are specifically designed to interpret these models and assess their accuracy. This creates a paradox: if we use the same models to generate training data, we risk an infinite loop where the model's limitations are reflected in the dataset itself, perpetuating the same inaccuracies we aim to evaluate and improve.

For this reason, we opted to use Wikipedia pages as a source for constructing the dataset. While this choice offers the advantage of working with real-world data, it also introduces significant challenges. Extracting meaningful sentence triplets (anchor, positive, and negative examples) from Wikipedia is not straightforward. The process requires extensive filtering, alignment, and translation to ensure high-quality, accurate data, which is both time-consuming and labor-intensive. Despite these challenges, this approach avoids the pitfalls of synthetic data generation and allows for a more robust evaluation of the model's performance in a real-world context.

3.2.2 Methodology

We develop our model using English as the HRL and Persian as the LRL as a use case. Though distantly related, the two languages have significant typological differences. For instance, English word order is usually SVO while standard Persian word order is SOV. Persian is a pro-drop language while English is not. Persian lacks a definite article "the." So, by default, language technologies developed for one would not be assumed to work

for the other, as might be the case with languages that share stronger areal, genetic, or borrowing relationships [Nath et al., 2022]. Nonetheless, many prior works (e.g., Conneau et al. [2020b]) indicate that representations of similar concepts across language also develop similarly in the latent space during pretraining, and so despite the structural differences of English and Persian, similar semantics may be represented in a multilingual model with appropriate training.

SBERT [Reimers and Gurevych, 2019] serves as the foundation of our approach. Unlike BERT, which is optimized for token-level tasks, SBERT is fine-tuned to generate fixed-size sentence-level embeddings using mean pooling, making it particularly useful for tasks like semantic textual similarity, clustering, and sentence classification. Our primary comparison concerns the approach taken in Reimers and Gurevych [2020], which utilizes the multilingual model XLM-R [Conneau et al., 2019] to improve cross-lingual transfer learning through knowledge distillation and equalizing English SBERT sentence embeddings with corresponding non-English embeddings from XLM-R.

Distill-XLMR-Triplet model

The core idea behind our proposed model is that the student model should mimic the behavior of the teacher model as closely as possible. Instead of merely approximating this behavior, we aim to distill the knowledge embedded in the teacher model and train the student model in a way that mirrors how the teacher was originally trained.

To achieve this, we use the triplet dataset described in Sec. 3.2.1, where a Persian sentence serves as the anchor, and two English sentences represent the positive and negative examples. The goal is to distill knowledge from a SBERT teacher model into an XLM-R student model, ensuring the student model accurately learns the semantic relationships between the anchor, positive, and negative sentences. This is accomplished by combining *triplet loss*, which reinforces relative distance learning, with *knowledge distillation loss*, allowing the student model to capture the teacher's output distribution effectively. The architecture and

approach to triplet knowledge distillation is given in Fig. 3.5.

To begin, we will examine the architecture of SBERT [Reimers and Gurevych, 2019] as a foundation for comparing it with more recent advancements in multilingual models. SBERT is a modification of the original BERT model, specifically designed to produce high-quality sentence embeddings. Unlike BERT, which is optimized for token-level tasks, SBERT is fine-tuned to generate sentence-level embeddings, making it particularly useful for tasks like semantic textual similarity, clustering, and sentence classification.

The overall architecture of SBERT is illustrated in Figure 3.3 . It utilizes the transformer-based BERT as its backbone but introduces a crucial change: instead of feeding individual tokens into a classifier, SBERT pools the output from BERT to generate a fixed-size vector for the entire sentence. This vector represents the sentence embedding, which can then be used in downstream tasks. Additionally, SBERT can be fine-tuned using Siamese and triplet network structures, enabling the model to learn embeddings that directly optimize for sentence similarity. This architecture has set a strong precedent for efficient and accurate sentence representation, and it serves as a baseline for evaluating the performance of newer models, especially in multilingual settings.

The primary work for comparison, as outlined in [Reimers and Gurevych, 2020], utilizes the architecture shown in Figure 3.4. This approach focuses on cross-lingual sentence embedding alignment through a knowledge distillation process. Specifically, it aims to equalize the sentence embeddings produced by SBERT for English sentences with those generated by XLM-R (a multilingual transformer model) for corresponding non-English sentences. This alignment enables XLM-R to improve cross-lingual transfer learning, allowing it to handle multilingual tasks more effectively.

The core idea behind our proposed model is that the student model should mimic the behavior of the teacher model as closely as possible. Instead of merely approximating this behavior, we aim to distill the knowledge embedded in the teacher model and train the student model in a way that mirrors how the teacher was originally trained.

To achieve this, we use a triplet dataset, where a Persian sentence serves as the anchor, and two English sentences represent the positive and negative examples. The goal is to distill knowledge from a SBERT teacher model into an XLM-R student model, ensuring the student model accurately learns the semantic relationships between the anchor, positive, and negative examples. This is accomplished by combining triplet loss, which reinforces relative distance learning, with knowledge distillation loss, allowing the student model to capture the teacher's output distribution effectively. You can see the main architecture and approach to triplet knowledge distillation in Figure 3.5.

The triplet loss, denoted as $\mathcal{L}_{\text{triplet}}$, encourages the model to ensure that the distance between the anchor ($\mathbf{a}$) and the positive example ($\mathbf{p}$) is smaller than the distance between the anchor and the negative example ($\mathbf{n}$). Triplet loss with a soft margin can be defined as follows:

$$\mathcal{L}_{\text{triplet}} = \frac{1}{N} \sum_{i=1}^{N} \log \left(1 + \exp \left(d(\mathbf{a}_i, \mathbf{p}_i) - d(\mathbf{a}_i, \mathbf{n}_i) + \alpha\right)\right) \tag{3.1}$$

where $d(\mathbf{a}, \mathbf{p}) = \|\mathbf{a} - \mathbf{p}\|_2$ represents the Euclidean distance between the anchor and positive embeddings, and similarly for $d(\mathbf{a}, \mathbf{n})$. This loss ensures that the anchor is closer to the positive than to the negative, driving the model to learn effective representations.

Here, the margin α is a hyperparameter that enforces a smaller distance between the anchor ($\mathbf{a}$) and the positive sentence ($\mathbf{p}$) than the distance between the anchor and the negative sentence ($\mathbf{n}$), encouraging the model to learn finer distinctions between positive translations and semantically unrelated negative pairs. The inclusion of α reflects the expectation that translations (positive pairs) are closer in meaning and should be positioned nearer to each other in the embedding space compared to unrelated sentences. To account for the fact that the positive sentences in our triplet setup are often translations of the anchor, we increase α, to further reduce the distance between the anchor and the positive examples.

To facilitate the transfer of knowledge from the teacher model (SBERT) to the student model (XLM-R), we incorporate a knowledge distillation loss, $\mathcal{L}_{\text{KD}}$. This loss aims to min-

imize the difference between the embeddings produced by the student and teacher models for both positive and negative examples. Specifically, we use the mean squared error (MSE) to measure the distance between the embeddings:

$$\mathcal{L}_{\text{KD}} = \frac{1}{N} \sum_{i=1}^{N} \left(\|\mathbf{p}_i^{\text{s}} - \mathbf{p}_i^{\text{t}}\|_2^2 + \|\mathbf{n}_i^{\text{s}} - \mathbf{n}_i^{\text{t}}\|_2^2 \right) \tag{3.2}$$

where $\mathbf{p}_i^{\text{s}}$ and $\mathbf{n}_i^{\text{s}}$ are the embeddings for the positive and negative sentences produced by the student model, and $\mathbf{p}_i^{\text{t}}$ and $\mathbf{n}_i^{\text{t}}$ are the corresponding embeddings produced by the teacher model.

The total loss function for training the student model is a weighted combination of the triplet loss and the knowledge distillation loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{triplet}} + \lambda_{\text{KD}} \cdot \mathcal{L}_{\text{KD}} \tag{3.3}$$

where λ_{KD} is a hyperparameter that controls the contribution of the knowledge distillation loss to the overall objective. Using the grid search technique, we found that a value of 0.9 is appropriate for λ_{KD}. This combined loss encourages the student model to not only learn the relative distances between anchor, positive, and negative embeddings but also to closely mimic the teacher model's outputs, facilitating more effective knowledge transfer.

In this setup, the triplet dataset is processed as follows: for each triplet, the student model produces embeddings for the anchor sentence $\mathbf{a}$, the positive sentence $\mathbf{p}$, and the negative sentence $\mathbf{n}$. These embeddings are then used in two ways:

- The triplet loss uses the student embeddings to measure the relative distances between the anchor, positive, and negative sentences, ensuring that the student model learns to rank the positive sentence closer to the anchor than the negative one.
- The knowledge distillation loss directly compares the student model's positive and negative embeddings with the teacher model's corresponding embeddings, enforcing the student model to approximate the teacher's behavior.

Fine-tuned SBERT model

As a second comparison, we designed a model intended to mirror the internal mechanisms of the previously mentioned multilingual sentence transformer, allowing us to gain a deeper understanding of its operation. This sister model is pre-trained solely on monolingual English data, for comparison to the model trained on a multilingual corpus. The goal is to investigate how the model's sentence embeddings behave when it lacks any direct knowledge or exposure to relationship and sequence of the non-English words, shedding light on the impact of not being aware of non-English grammar and syntax.

The architecture consists of the following key components:

1. **Word Mapping Layer:** To address the discrepancy between the word representations of Persian and English, we apply a word-to-word gloss from source sentence to the target sentence. By aligning word translations, the model can better recognize and process words from the non-English language, thereby mitigating suboptimal performance that arises due to the differences in language structure and vocabulary.

2. **Sentence Transformer:** The mapped word embeddings are subsequently passed through a sentence transformer. This sentence transformer is initialized with pretrained SBERT weights [Reimers and Gurevych, 2019] and fine-tuned on the cross-lingual triplet dataset.

To fine-tune the SBERT component of the model, we applied the triplet loss function (Eq. 3.1) to account for the distances between the embedding vector of the mapped sentence (translated from Persian) and the embedding vectors of the positive and negative sentences. During this process, the weights of the SBERT model are adjusted to improve its performance on the cross-lingual task.

Henceforth, we will refer to the reference model from Reimers and Gurevych [2020] as `Distill-XLMR`, our knowledge distillation model with the triplet network as `Distill-XLMR-Triplet`, and the comparison model using word-to-word translation as `Finedtuned-SBERT`. For all

models, we used the `paraphrase-distilroberta-base-v2` (SBERT) and `xlm-roberta-base` (XLM-R) variants. We applied the same hyperparameters across all models to ensure a fair comparison. Specifically, the maximum sequence length for both the teacher and student models is set to 128, the batch size to 16, and the number of epochs to 5. Training was conducted on an NVIDIA RTX A6000 48 GB device.

3.2.3 Evaluation

We perform evaluation using a variety of similarity metrics over the embeddings of sentence pairs produced by each model: namely Cosine Distance, Manhattan Distance, and Euclidean Distance. For all metrics, a model is considered *correct* if the positive pair has a smaller distance than the negative pair.

Cosine distance measures the cosine of the angle between two vectors. The smaller the angle, the more similar the vectors are. The cosine distance is calculated as 1−cosine similarity.

For each triplet (anchor, positive, and negative), we calculate the cosine distance between the anchor-positive pair and the anchor-negative pair. Accuracy is determined by whether the positive pair has a smaller distance than the negative pair.

$$\text{Cosine Distance} = 1 - \frac{A \cdot B}{\|A\|\|B\|}$$

where A and B are the vectors of embeddings.

Dot product measures the projection of one vector onto another. A higher dot product indicates greater similarity between the vectors.

The dot product is calculated between the anchor-positive pair and the anchor-negative pair. A correct triplet will have a larger dot product for the anchor-positive pair compared to the anchor-negative pair.

$$\text{Dot Product} = \sum_{i=1}^{n} A_i \times B_i$$

where A and B are vectors of embeddings.

Manhattan distance, also known as L1 distance, sums the absolute differences between the corresponding components of two vectors. It measures how much one would move in a grid-like fashion from one point to another.

For each triplet, Manhattan distance is calculated between the anchor-positive pair and the anchor-negative pair. A correct triplet occurs when the distance between the anchor-positive pair is smaller than the distance between the anchor-negative pair.

$$\text{Manhattan Distance} = \sum_{i=1}^{n} |A_i - B_i|$$

where A and B are vectors of embeddings.

Euclidean distance is the straight-line distance between two points in the vector space. It is the most common way to measure the distance between two vectors.

For each triplet, the Euclidean distance is calculated between the anchor-positive pair and the anchor-negative pair. A correct triplet occurs when the anchor-positive distance is smaller than the anchor-negative distance.

$$\text{Euclidean Distance} = \sqrt{\sum_{i=1}^{n} (A_i - B_i)^2}$$

where A and B are vectors of embeddings.

For each metric, we calculate the distance between the anchor-positive embeddings and the anchor-negative embeddings. The model's accuracy for each metric is determined by how often the positive distance is smaller than the negative distance across all triplets.

The accuracy for each distance metric is calculated as:

$$\text{Accuracy} = \frac{\text{Number of correct triplets}}{\text{Total number of triplets}}$$

Cosine Accuracy: Proportion of correct triplets based on cosine distance.

Dot Accuracy: Proportion of correct triplets based on the dot product.

Manhattan Accuracy: Proportion of correct triplets based on Manhattan distance.

Euclidean Accuracy: Proportion of correct triplets based on Euclidean distance.

Max Accuracy: The best-performing metric among Cosine, Manhattan, and Euclidean distances.

The following metrics are logged:

- **Accuracy Cosine Distance:** accuracy_cos $\times$ 100
- **Accuracy Dot Product:** accuracy_dot $\times$ 100
- **Accuracy Manhattan Distance:** accuracy_manhattan $\times$ 100
- **Accuracy Euclidean Distance:** accuracy_euclidean $\times$ 100

Max Accuracy: The highest accuracy value among cosine, Manhattan, and Euclidean distance metrics.

There are currently no high-quality datasets available for evaluating Persian-English text similarity, and most translation datasets consist of very short sentences, often just one or two words. As a result, we focus on evaluating the proposed triplet datasets.

There is a significant demand for developing effective metrics for text similarity due to the inherent challenges in this area. However, the available datasets used for evaluating such metrics are often unreliable or unsuitable for many languages, especially LRLs ones like Persian. Furthermore, both the evaluation datasets and metrics are typically limited to specific languages.

For example, [Cer et al., 2017] aimed to assess systems on the task of measuring semantic equivalence between text pairs, with a particular focus on multilingual and cross-lingual contexts. Their task was divided into six subtasks, covering languages such as English, Spanish, Arabic, and Turkish, along with cross-lingual evaluations between English and these other languages. The participating systems were required to produce similarity scores ranging from 0 to 5, reflecting the extent to which the meanings of the two texts were aligned.

This task provided a standardized dataset and evaluation metrics for comparing the performance of various approaches to textual similarity. Results indicated that neural network-based models, particularly those leveraging multilingual word embeddings, performed well in both monolingual and cross-lingual settings. However, it is important to note that this task, despite being benchmarked, remains difficult, even for humans, due to the complexity of determining semantic similarity between sentences.

In our work, we employ average cosine similarity as an evaluation metric. To calculate this metric, we first compute the cosine similarity for each pair of sentences by comparing their respective embedding vectors, which are fixed-size vectors with 768 dimensions. Cosine similarity measures the angular similarity between these vectors, with values ranging from -1 to 1, where 1 indicates perfect similarity, 0 indicates no similarity, and -1 indicates complete dissimilarity.

Once the cosine similarities are calculated for all sentence pairs in the dataset, we then compute the average cosine similarity across the entire dataset to get an overall measure of how similar the sentence pairs are, on average. Additionally, we calculate the standard deviation of these cosine similarity scores to understand the variability in the similarity values. A higher standard deviation would indicate that the dataset contains a diverse range of sentence pairs with varying degrees of similarity, whereas a lower standard deviation suggests that the sentence pairs are more uniformly similar or dissimilar.

This approach gives us both a central tendency (mean cosine similarity) and a measure of spread (standard deviation) for the sentence similarity across the dataset, allowing for a more nuanced evaluation of the model's performance on the task.

Specifically, the model is tasked with computing the cosine similarity between the anchor and positive sentences, and separately between the anchor and negative sentences. The goal is for the cosine similarity between the anchor and positive sentences to be as high as possible, while the similarity between the anchor and negative sentences should be minimized. This allows us to effectively evaluate the performance of the model in distinguishing between

semantically similar and dissimilar sentence pairs.

Additionally, we calculated the average and standard deviation of sentence word lengths in each dataset to evaluate how these factors influence the model's ability to cluster similar sentences. By examining the average sentence length, we can determine the typical sentence complexity within the dataset, while the standard deviation reveals the variation in sentence lengths. Understanding these patterns helps us assess whether sentence length diversity affects the model's performance in grouping semantically similar sentences.

3.2.4 Results

To compare the Distill-XLMR and Distill-XLMR-Triplet models, you can find the results at table 3.8.

Metric	**distill-translation**	**Distill-XLMR-Triplet**
Cosine Accuracy	0.7975	0.9471
Dot Accuracy	0.2071	0.0702
Manhattan Accuracy	0.7741	0.9560
Euclidean Accuracy	0.7728	0.9560
Max Accuracy	0.7975	0.9560

Table 3.8: Evaluation Metrics for Distill-XLMR and Distill-XLMR-Triplet Models

These metrics were originally introduced as part of the evaluation for the monolingual SBERT model. [Reimers and Gurevych, 2019]

Distill-XLMR-Triplet is outperforming Distill-XLMR on most metrics, particularly on cosine, Manhattan, and Euclidean accuracies. This suggests that Distill-XLMR-Triplet excels at embedding similar sentences closer together and dissimilar sentences further apart, both in terms of angular and spatial distance.

Dot Accuracy is consistently low for both models. Generally, in high-dimensional spaces, dot products can become less reliable for measuring similarity because small changes in individual vector components can lead to large changes in the final dot product. In contrast, metrics like cosine similarity or Euclidean distance often provide more stable measurements

in high-dimensional embeddings.

In the following, we will present the results of investigate Distill-XLMR on similar sentences with different lengths.

Fig. 3.6 shows how average cosine similarity changes across the two models as a function of the absolute difference in length between Persian sentences and their English equivalents.

We see that as the difference in length increases, `Distill-XLMR-Triplet` performance outstrips that of `Distill-XLMR` and is is much better able to preserve the similarity of known similar sentences in the latent space. There may be many cases where the meaning of a single word in one language is expressed with a multi-word expression in the other (for instance, Persian makes much greater use of light-verb constructions than English does), and the mean pooling used in SBERT before minimizing the distance between equivalent sentences may be averaging out the representations of idiomatic expressions like these, at the expense of the expressed meaning.

The reference model appears to lose some information in the encoding process, allowing the proposed method to catch up its performance. This suggests that the multilingual model, when applied to languages with different structures and grammar rules, might tend to overlook syntactic and semantic dependencies and focus primarily on word embeddings. We provide example evidence toward this in Sec. 3.1.5.

In addition, the Fig. 3.6 illustrates that both models—one trained on Persian sentences and the other not—behave similarly when given the same inputs, helping to pinpoint why the current approach may not work effectively.

Table 3.9 provides an example test sentence pair (source and target) that `Distill-XLMR-Triplet` is able to successfully identify as equivalent and `Distill-XLMR` is not. The difference in token counts between the two sentences is in the interval of 15-20 words (60 English words compared to only 40 Persian words). In pairs whose lengths differ to this extent, `Distill-XLMR-Triplet` displays its greatest performance advantage over `Distill-XLMR`. One possible reason for this is the lack of definite articles in Persian, or single Persian words being rendered with multi-

ple English words, resulting in the English sentence in a pair or triplet having a larger word count.

Language	Sentence	Length
English	This business model led from Myriad being a startup in 1994 to being a publicly traded company with 1200 employees and about $500M in annual revenue in 2012; it also led to controversy over high prices and the inability to get second opinions from other diagnostic labs, which in turn led to the landmark "Association for Molecular Pathology v. Myriad Genetics."	60
Persian (transliterated)	*Alava bar in, ekhtelaf-e nazarha va bahs-hayi darbareye gheymat-haye balaye in sherkat va natavani dar gereftan-e nazar-e dovvom az azmayeshgah-haye tashkhisi digar pish amad, ke be nobeh khod monjar be davaye hoqouqi anjoman-e barjesteh asib-shenasi molokuli alay-he sherkat-e Myriad shod.*	40
Persian (translated)	In addition, disagreements and debates arose regarding the high prices of this company and the inability to obtain a second opinion from other diagnostic laboratories, which in turn led to a legal dispute by the prominent Association for Molecular Pathology against Myriad Genetics.	43

Table 3.9: Example test sentence pair that `Distill-XLMR-Triplet` successfully identified as equivalent and `Distill-XLMR` did not. "Length" denotes the number of whole words in the sentence, excluding punctuation (and ezāfe in the Persian transliteration).

3.2.5 Discussion

First and foremost, for effective knowledge distillation and for the student model to accurately learn the distribution of the teacher model, it is crucial that the student model mimics the teacher's training methodology when their goals are completely similar. This alignment ensures that the student captures not only the surface-level outputs but also the underlying patterns and structures that the teacher model has learned.

In this context, knowledge distillation should go beyond simple loss minimization and focus on deeper representation learning. By incorporating a triplet network approach, the student model can more thoroughly learn the distribution of sentence embeddings, which is vital for tasks involving sentence similarity and representation. The triplet network is particularly advantageous because it not only ensures that the embeddings for similar sentences (anchor and positive examples) are close together but also pushes dissimilar sentences (anchor and negative examples) further apart in the embedding space.

This method allows the student model to learn the fine-grained relationships between

sentences, effectively capturing the full distribution of the teacher model's knowledge. By leveraging triplet loss in the knowledge distillation process, the student model becomes more robust, learning nuanced sentence representations that are aligned with the teacher model's performance, even when operating with fewer parameters or less capacity. This thorough alignment is essential for achieving high performance in downstream tasks while maintaining the efficiency benefits of knowledge distillation.

To see how the Finedtuned-SBERT works, you can find a representative example in Table 3.7. It illustrates both a correct and an incorrect test case for a Persian sentence and its predicted English equivalent, chosen from a set of 50K test samples across various topics.

Although the finetuned-SBERT model maps non-English text to English, the generated sentences often don't follow grammatical rules in either language, and the relationships between words may seem meaningless. Despite this, the model can still identify a similar sentence quite effectively. In some cases, the input sentence may not make sense, yet the model produces a surprisingly accurate result.

This could be because most of the words are similar between the two sentences, leading the model to identify them as similar while overlooking the fact that their semantic meaning is incoherent. This highlights the effectiveness of word embeddings, but it also shows that semantic similarity can be overlooked in cross-lingual tasks.

This is likely what happened in the main model: it tends to map sentences with similar translated words to each other rather than sentences with similar meanings. This occurs because the model relies heavily on word embeddings and mean-pooling, focusing more on surface-level word similarities.

To address this, consider the examples provided in Table 3.10, where a simple modification significantly enhances the model's ability to interpret sentences. By chunking the input text into smaller segments of up to 20 words and then calculating the embeddings for each chunk before applying mean-pooling, the model achieves far more accurate results.

For similar sentences, where the texts meanings are nearly identical, one would expect the

cosine similarity of their embeddings to approach 1. However, the model initially provides a similarity score of only 0.68. After applying chunking and mean-pooling, the similarity score improves to 0.87, representing a substantial 0.19 increase, which demonstrates a significant improvement in the model's performance.

For dissimilar sentences, the desired outcome is for the sentence embeddings to be as close to 0 as possible, given that slight changes in the text drastically alter their meanings, often making them contradictory. However, the model struggles to distinguish these differences. By applying the proposed chunking method, the performance improves by 0.13.

This improvement can be attributed to the model's enhanced focus when it considers smaller sets of words at a time, as opposed to distributing attention evenly across all words, which may dilute the focus on key components. The attention mechanism was likely initially trained to distribute attention broadly, but in cases like these, where sentences require more focused interpretation, concentrating attention on smaller chunks proves more effective.

Last but not least, note that the key advantage of the proposed adversarial model lies in its capacity for robust training on a considerably larger English dataset. This comprehensive training contrasts with the Knowledge Distillation (KD) approach, which, due to limitations, reduces the dataset to just 500,000 pairs. The adversarial model's broader training scope allows it to better capture nuances and complexities, making it more effective in handling the intricacies of cross-lingual tasks.

Limitations

One of the key challenges faced in this study is the unclear and subjective definition of "similarity," especially in cross-lingual contexts. Similarity is often a nuanced and culturally-dependent concept, which makes it difficult to standardize across languages and evaluators. In low-resource languages like Persian, gathering a well-scored and diverse dataset evaluated by multiple, competent annotators is particularly challenging. The lack of available datasets further complicates this process. Most of the Persian datasets available online are heavily fo-

cused on political content, limiting their applicability for evaluating general-domain models, which are critical for real-world applications outside of specific political discourse.

Another limitation arises from the way large language models (LLMs) are trained. These models rely heavily on repeated patterns of words and concepts, allowing them to learn semantic relationships over large corpora. However, expressions, idiomatic phrases, and slang—particularly those unique to specific regions—are underrepresented in training data, making them difficult to model and evaluate effectively in cross-lingual contexts. This limits the model's ability to capture the full range of linguistic nuances, especially for low-resource languages.

A major challenge in this study lies in the computational demands of knowledge distillation, particularly when using large language models as both the teacher and the student. Knowledge distillation involves transferring knowledge from a larger, more powerful model (the teacher) to a smaller model (the student). However, uploading and running two large models simultaneously—one as the teacher and one as the student—presents significant resource constraints. Even on powerful GPUs, the memory requirements for loading and fine-tuning both models together are substantial. In practice, this means that even when using modest batch sizes, such as 16, the GPU memory quickly becomes insufficient, making it extremely difficult to carry out the training process efficiently. This computational limitation hinders experimentation and prevents the model from being trained in a way that fully leverages the potential of knowledge distillation.

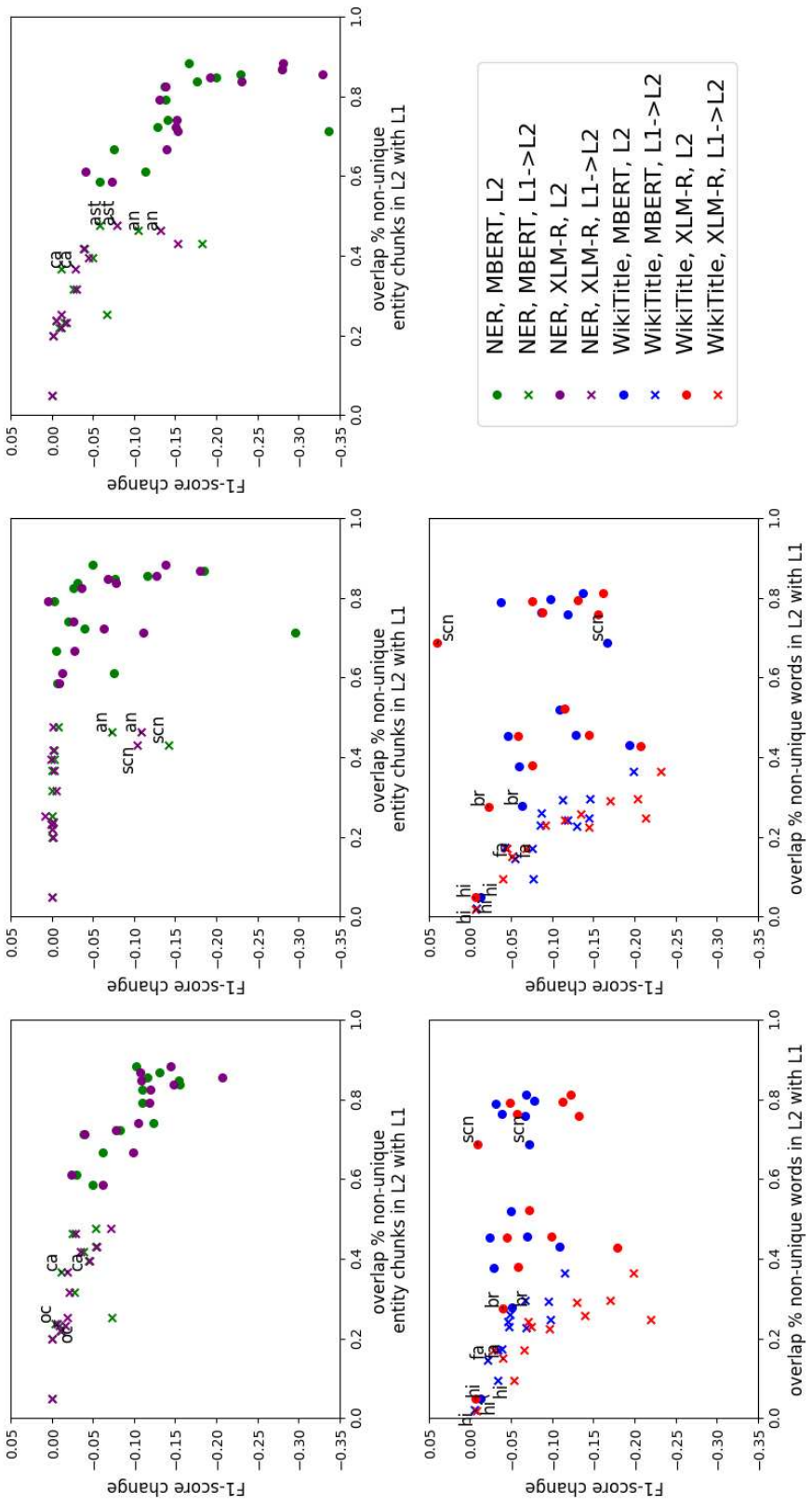

Figure 3.2: Change in F_1 score under perturbation as a function of degree of vocabulary overlap. Left to right, top to bottom: P3 for NER, P4 for NER, P5 for NER, P4 for title selection, P4 for title selection using random substitutions instead of the most cosine-similar words.

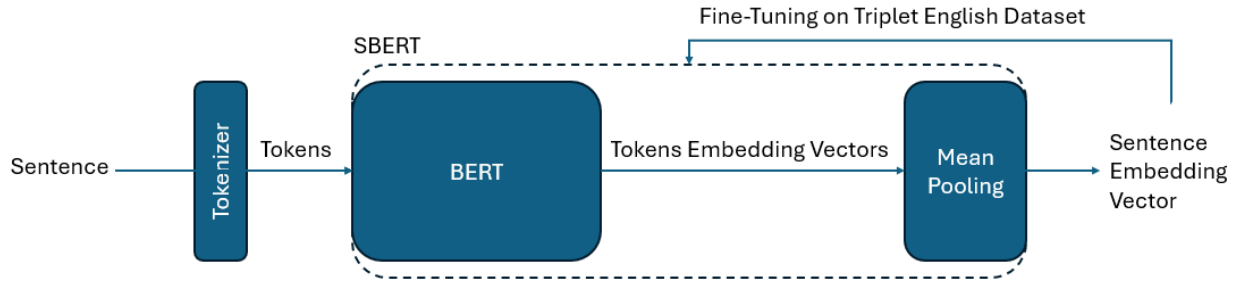

Figure 3.3: The overall architecture of the sentence transformer SBERT involves processing triplet sentences through several stages. First, the sentences are passed through a tokenizer, which breaks them down into tokens. These tokens are then embedded into vectors using the BERT model. To obtain a fixed-size sentence embedding, SBERT applies mean-pooling over the token embeddings. After generating the sentence embeddings for the triplet (anchor, positive, and negative examples), the triplet loss is computed based on the similarity between these embeddings. This loss is then backpropagated through the network, allowing it to optimize and refine the sentence embeddings for better semantic representation.

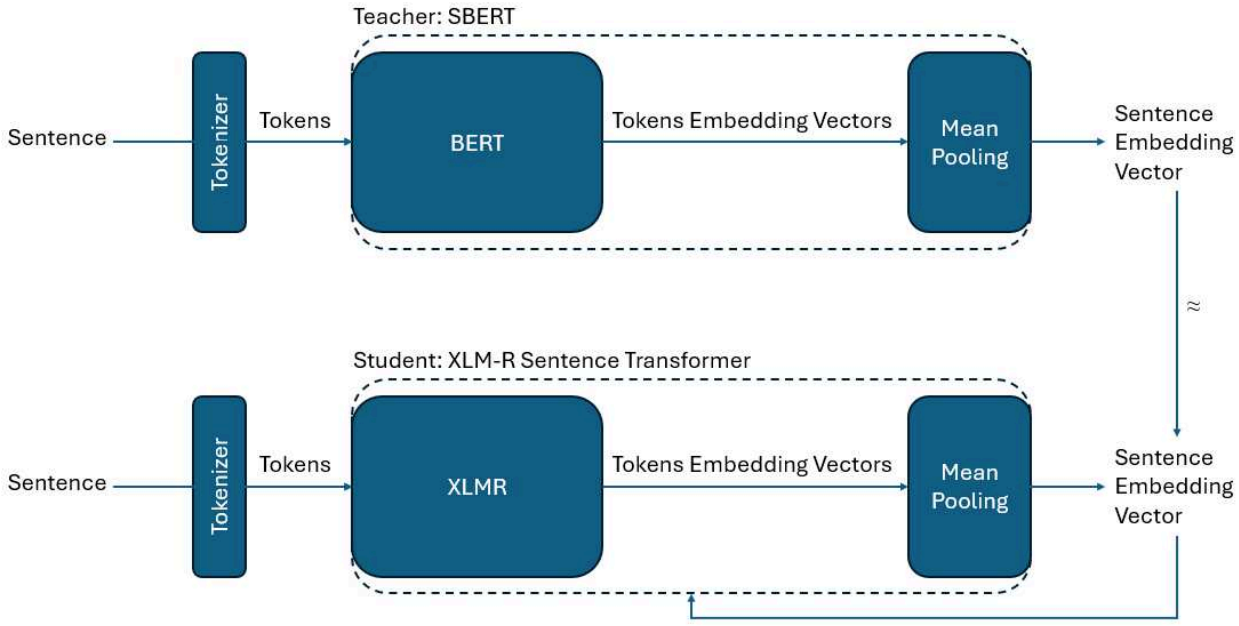

Figure 3.4: In this architecture, the English sentence is passed through SBERT to generate a high-quality sentence embedding, while the non-English sentence is processed through XLM-R with mean-pooling applied over its token embeddings to produce a comparable sentence vector. The key objective of this setup is to minimize the difference between the English and non-English sentence embeddings, effectively training the XLM-R model to produce embeddings for non-English sentences that are as semantically meaningful as those generated by SBERT for English sentences.

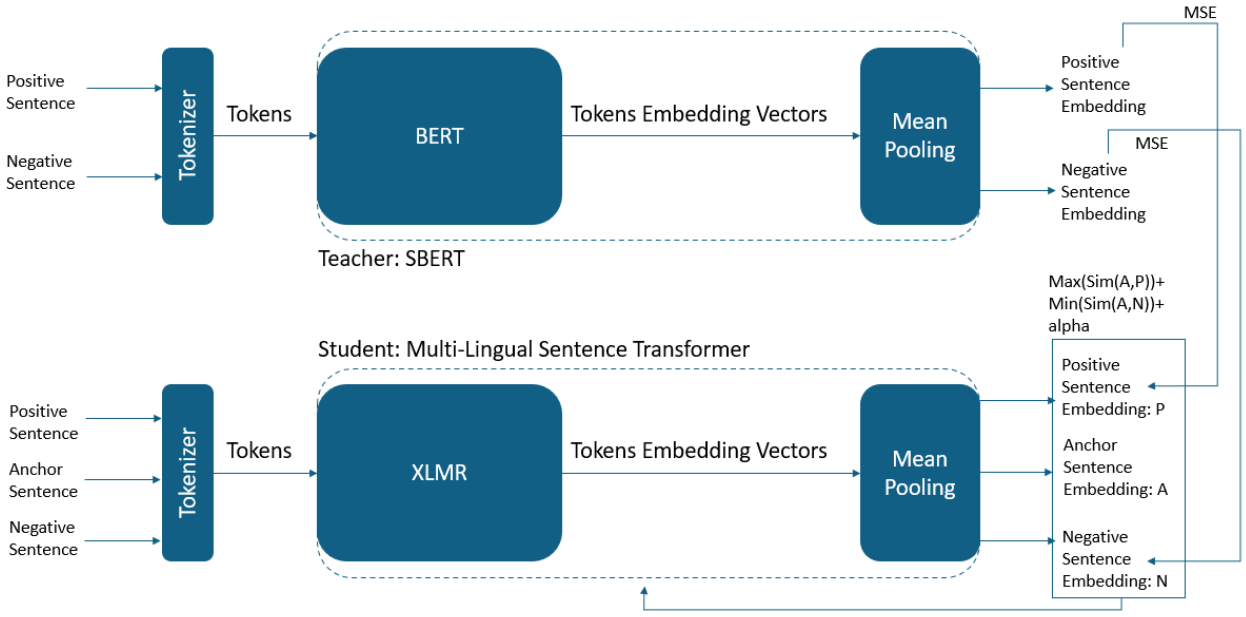

Figure 3.5: The English sentences are processed through the SBERT model, which generates high-quality sentence embeddings using a BERT backbone. This provides a strong semantic representation of the English text. Simultaneously, all three sentences (anchor, positive, and negative) — including both English and non-English sentences — are passed through the XLM-R model, a multilingual sentence transformer. The XLM-R model, acting as the student, is trained using a combination of two key losses: triplet loss and knowledge distillation loss.

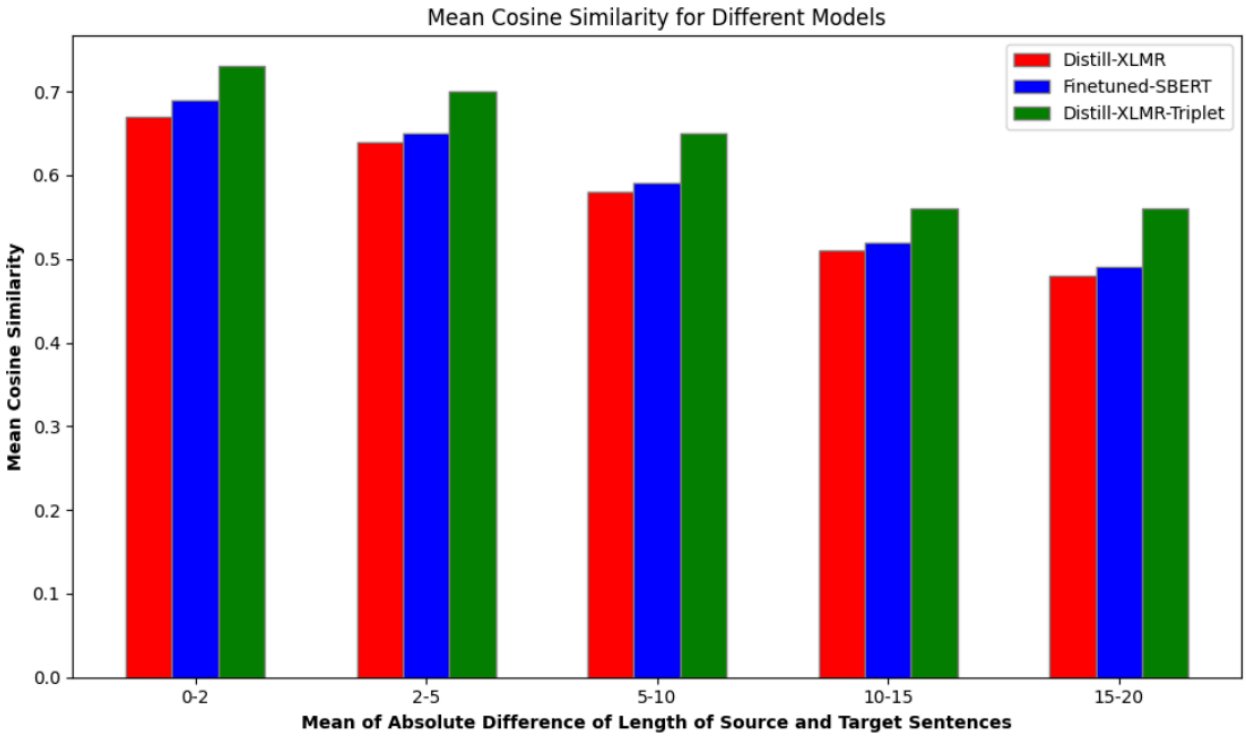

Figure 3.6: Average of cosine similarity vs. absolute difference in length between source (Persian) and target (English) sentences.

Category	Details
Similar Sentences	
Persian Sentence	پیشرفت سریع تکنولوژی ارتباطات را آسان‌تر کرده اما هم‌زمان مشکلاتی در حفظ تعاملات انسانی واقعی ایجاد کرده است.
Translated Sentence	The rapid advancement of technology has made communication easier, but at the same time, it has created challenges in maintaining genuine human interactions.
English Sentence	The rapid advancements in technology over the past few decades have fundamentally reshaped and revolutionized the way we communicate and interact with one another. With the advent of new tools, platforms, and devices, it has become remarkably easier to establish instant connections with people from all corners of the globe, breaking down traditional barriers of distance and time. Whether through social media, messaging apps, video calls, or emails, the ability to engage with friends, family, and colleagues, no matter how far away they may be, is now at our fingertips. However, despite these incredible improvements in connectivity, these technological innovations have also introduced a series of new and complex challenges. One of the most significant of these challenges is the difficulty in maintaining genuine, authentic human interactions in a world that is increasingly dominated by digital communication. As our lives become more intertwined with technology, the richness of face-to-face conversations and the emotional depth of personal relationships can sometimes be overshadowed by the convenience of virtual interactions. In essence, while technology has brought people closer together in a practical sense, it has also, paradoxically, created a gap in the quality of our connections, leading to concerns about the loss of meaningful human bonds in this ever-evolving digital landscape.
Distill-XLMR Cosine similarity	0.68
Solution Cosine similarity	0.87
Dissimilar Sentences	
Persian Sentence	پیشرفت سریع تکنولوژی نحوه ارتباطات ما را دگرگون نکرده است بلکه تنها ارتباط با افراد درسراسر جهان را آسان‌تر ساخته است؛ ولی نمی توان گفت مشکلات جدیدی به وجود آورده است.
Translated Sentence	The rapid advancement of technology has not transformed the way we communicate, but has only made it easier to connect with people across the world. However, it cannot be said that it has created new challenges.
English Sentence	The rapid advancements in technology have transformed how we communicate, making it easier to connect with people across the globe, yet at the same time, they have introduced new challenges in maintaining genuine human interactions.
Distill-XLMR Cosine similarity	0.82
Solution Cosine similarity	0.69

Table 3.10: Chunking-example: A comparison of similar and dissimilar sentence pairs with their respective Cosine Similarities before and after chunking.

Chapter 4

Conclusion and Future Work

In first stream [Manafi and Krishnaswamy, 2024], we have presented a set of adversarial perturbations to test the ability of language models to generalize from higher-resourced languages to lower-resourced languages in a cross-lingual transfer zero-shot setting. Our experiments are performed in a language-agnostic manner for both NER and title selection tasks. To our knowledge, this is the first time such an experimental set has been performed with an explicit focus on LRLs and cross-lingual transfer from HRLs. We conducted evaluations on 21 languages, encompassing both high and LRLs, employing two widely recognized multilingual models, mBERT and XLM-R. Results exhibit variations across different languages, influenced by their linguistic structures and similarities. Our core findings can be summarized as follows:

- There is a pronounced effect of vocabulary overlap on NER performance. Perturbing named entities so that the test data contains only non-overlapping words has a statistically very significant impact on model performance.
- Although models utilizing cross-lingual transfer typically exhibit lower numerical performance than models trained in a native LRL setting, they are often somewhat more robust to certain types of perturbations of the input.
- Title selection, as a proxy for document classification, in LRLs appears to heavily rely on word memorization.

These proposed test sets have the potential for further exploration, particularly in challenging tokenizers directly. For example, the Persian examples discussed in Sec. 3.1.5 suggest

that, although BPE tokenization methods should help LRL performance by not biasing token vocabulary toward frequent tokens in a specific language [Sennrich et al., 2016], similarity between subword tokens may be overvalued when optimizing the embedding space. This, among other factors, motivates the need for an equitable consideration of lower-resource languages in building NLP models [Joshi et al., 2020].

Our results show that MLLMs exhibit some level of "mutual intelligibility" between individual languages; a model can "get by" in a language like Asturian if it knows a similar language like Spanish. However, a risk of MLLMs is a systematic encoding of biases toward HRLs, which also has implications for representations of minority and regional languages. Our own datasets, due to availability of online resources, are still biased toward Indo-European languages. A multilingual capacity is not necessarily enough to handle an arbitrary input language [Virtanen et al., 2019, Scheible et al., 2020, Tanvir et al., 2021, Nath et al., 2023], and performance is sensitive to minor changes to the input, even without perturbations in the latent space [Narasimhan et al., 2022].

Finally, this research has been conducted on encoder models. The reasons for this are manifold but center around the fact that encoder models, being older and smaller, typically demand fewer computational resources, allowing us to perform more experiments. Additionally, unlike currently-touted SOTA decoder models like GPT-4, most encoder model weights and processing pipelines are freely available on platforms like HuggingFace [Wolf et al., 2019], meaning that we can directly access the embedding spaces to inform our perturbation techniques. Nonetheless, our findings could influence directions for probing and prompt engineering for generative models that exhibit multilingual capability. Most open-weight generative models (e.g., LLaMA 2 [Touvron et al., 2023]) are not multilingual; those that are, such as ChatGPT/GPT-4 [Qin et al., 2023], remain closed. However, since our techniques are general, they could be applied to open-source multilingual generative models like XGLM [Lin et al., 2021b]. We do note that multilingual generative models still do not necessarily all the languages we study here, which further indicates a resource deficiency for

many languages when it comes to SOTA generative NLP.

The second stream focused on the intricate task of creating reliable cross-lingual sentence embeddings, particularly for LRLs like Persian, using HRLs models like SBERT as a foundation. The challenge of representing sentences, which are more complex and context-dependent than words, becomes even more difficult in a cross-lingual context where differences in grammar, structure, and cultural nuances between languages complicate the training process. While word embeddings are effective for lexical-level tasks, sentence embeddings must capture not only the meanings of individual words but also their relationships, syntactic structure, and broader semantic roles within sentences. This task becomes significantly harder when dealing with multiple languages, especially when LRLs lack sufficient training data for building dedicated sentence Transformers.

In our study, we set out to explore the limitations of translation-based approaches for knowledge transfer and propose a more robust method using a cross-lingual triplet dataset. Instead of relying solely on translations, we generated a dataset of Persian anchor sentences paired with English positive and negative sentences. The key difference between our approach and previous work is the use of triplet loss, which allows the model to learn both the similarities and differences between sentences. This approach is especially effective for tasks where subtle distinctions between sentence meanings are important, such as paraphrase detection or semantic similarity tasks. By incorporating negative examples (sentences that are dissimilar to the anchor), the model is encouraged to learn more discriminative sentence embeddings.

Our results showed that the triplet loss method significantly outperformed previous approaches in cross-lingual tasks, particularly in identifying semantically dissimilar sentences. Traditional knowledge distillation methods that rely on binary classification (similar or dissimilar) tend to struggle in cases where the degree of similarity is more nuanced. For example, a translated sentence may not be identical to its original but may still convey a similar meaning. Our triplet-based approach accounts for these nuances, making it more adaptable

to real-world NLP tasks that require a more sophisticated understanding of sentence-level meaning.

However, one of the limitations of the sentence Transformers was its performance on similar sentences with different lengths. Sentence Transformers often rely on mean-pooling or other aggregation methods to generate a fixed-size sentence embedding, which can result in the loss of important syntactic and semantic information, especially for longer or more complex sentences. To address this, we experimented with chunking the input sentences into smaller segments before generating embeddings. This method allowed the model to focus on smaller groups of words at a time, resulting in more accurate embeddings for longer sentences. While this technique improved the model's performance, further research is needed to fully address the challenges posed by longer and more complex sentences.

Our contributions are as follow:

- Dataset Quality and Availability are Key: The availability of high-quality, large-scale triplet datasets was critical to the success of our approach. By generating a large dataset of cross-lingual triplet sentence pairs from Persian and English Wikipedia pages, we ensured the model could learn effective representations. However, future research should focus on creating more diverse and larger datasets to improve generalizability across different languages and tasks.

- The incorporation of triplet loss, as opposed to binary labeling approaches, significantly improves the model's ability to distinguish between semantically similar and dissimilar sentences in cross-lingual tasks. By training on triplet datasets with an anchor, positive, and negative sentence structure, the model learns more nuanced sentence embeddings that are effective across languages.

- While our model improved sentence embeddings in cross-lingual contexts, it struggled with longer, more complex sentences due to the inherent limitations of aggregation techniques like mean-pooling and distributed attention mechanism. The chunking

method we introduced helped mitigate this issue to some extent, but further refinement is necessary for handling long-form text.

Sentence embedding Transformers typically utilize **encoders** instead of **generators** due to the fundamental differences in their architectures and purposes. Encoders are more suitable for generating fixed-length vector representations that capture the meaning of a sentence, while generators are designed for sequence generation tasks. Below are the key reasons why encoders are preferred for sentence embeddings:

1. Focus on Input Representation

Encoders are designed to process an input sentence and convert it into a meaningful, fixed-size embedding. They focus on understanding the content and semantics of the input, which is essential for sentence embeddings. The goal is to capture the meaning of a sentence in a dense vector.

On the other hand, generators are designed to produce output sequences (such as in text generation), which does not align with the task of creating a static sentence embedding.

2. Bidirectional Context

Encoders in models like BERT process the entire input sentence bidirectionally. This means that both the left and right contexts of each token are considered, which is important for generating high-quality sentence embeddings.

Generators, especially in auto-regressive models like GPT, process input in a left-to-right manner, which may not fully capture the bidirectional context required for sentence embeddings.

3. Fixed-Length Output

Sentence embedding requires a fixed-length vector, regardless of the length of the input sentence. Encoders naturally produce embeddings of fixed size for any input. This is necessary for tasks like comparison, clustering, or classification.

Generators, on the other hand, produce variable-length sequences, which are not ideal for embedding tasks that require a consistent vector size.

To recap, Encoders are ideal for sentence embedding Transformers because they focus on understanding the entire meaning and structure of a sentence and can generate fixed-length embeddings. These embeddings can be used for semantic tasks such as similarity comparison, clustering, or retrieval. In contrast, generators are more suited for tasks involving sequence generation, such as text generation or machine translation.

That said, there are still significant challenges that need to be addressed in future work. One important direction for future research is the development of more advanced aggregation techniques that can better capture the meaning of longer sentences without sacrificing important details. Current sentence embedding methods, while effective for shorter and moderately complex sentences, often struggle when dealing with longer, more intricate structures. This is due to the tendency of these models to either over-simplify the content or lose critical context in the process of reducing the sentence into a fixed-length vector.

Advanced aggregation techniques could address this by incorporating hierarchical representations, allowing the model to maintain and emphasize key phrases, dependencies, and relationships within the sentence. Additionally, future work should focus on improving the ability of these models to handle diverse sentence structures across different languages, particularly in LRLs, where nuanced meaning is often harder to capture. By ensuring that important semantic components are preserved, these techniques could lead to more robust and contextually aware sentence embeddings, enhancing the performance of downstream tasks such as machine translation, summarization, and information retrieval.

Additionally, there is a need for larger and more diverse cross-lingual datasets to further improve the performance of multilingual sentence Transformers. Our study focused on Persian and English, but the methods we developed could be extended to other LRLs, provided that sufficient training data is available.

Moreover, as multilingual NLP models become more widely used, issues of fairness and

bias will become increasingly important. During our experiments, we observed that the model occasionally struggled with idiomatic expressions and cultural references unique to Persian. Ensuring that models are able to handle these linguistic nuances without introducing bias or unfairness will be critical in creating equitable NLP systems. Further research into bias detection and mitigation in cross-lingual models will be necessary to ensure that these models are fair and effective across all languages.

The results from both "Cross-Lingual Transfer Robustness to Lower-Resource Languages on Adversarial Datasets" and "Challenges of Cross-Lingual Transfer in Sentence Transformers from HRLs to LRLs" highlight the intricate and often non-trivial nature of transferring knowledge between HRLs and LRLs. Despite advancements in cross-lingual transfer learning techniques, there remain significant challenges in ensuring robustness and fairness, especially when applied to adversarial datasets or when embedding techniques like sentence Transformers are used.

The robustness study illustrates that while cross-lingual models can generalize well to LRLs in certain scenarios, they may struggle when faced with adversarial conditions where language-specific nuances come into play. This suggests that simplistic assumptions about language similarity can be problematic. On the other hand, the challenges identified in transferring sentence embeddings from HRLs to LRLs further underline the fact that current models may fail to capture the rich syntactic and semantic structures inherent in many LRLs, leading to suboptimal performance and even misrepresentation of meaning.

In general, languages differ significantly, and we must avoid oversimplifying the problem to basic knowledge transfer. Each language brings unique syntactic, semantic, and morphological traits that can significantly affect how knowledge is transferred. For example, linguistic diversity, script differences, and cultural nuances add complexity to the transfer process. Therefore, we must design transfer learning approaches that account for all these distinct attributes rather than treating them as basic technical challenges.

Ultimately, the interplay between robustness and fairness in cross-lingual transfer learning

is central to creating truly universal models that serve diverse linguistic communities. As cross-lingual systems evolve, addressing these challenges holistically will be key to their broader application and success across languages, especially those that are underrepresented in the training data.

Bibliography

Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.365. URL https://aclanthology.org/2020.emnlp-main.365.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, 2020a.

Jiaxi Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual bert? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, 2019.

Shijie Wu and Mark Dredze. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. 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 833–844, 2019.

Vaidehi Patil, Partha Talukdar, and Sunita Sarawagi. Overlap-based vocabulary generation improves cross-lingual transfer among related languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 219–233, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.18. URL https://aclanthology.org/2022.acl-long.18.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.

Yann LeCun and Yoshua Bengio. Convolutional networks for images, speech, and time series. In *The handbook of brain theory and neural networks*, pages 255–258. MIT Press, 1995.

Jonas Gehring, Michael Auli, David Grangier, and Yann N Dauphin. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1243–1252, 2017.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *OpenAI blog*, 1(8):1–12, 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V

Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Yoon Kim, Carl Denton, Luong Hoang, and Alexander M Rush. Structured attention networks. *arXiv preprint arXiv:1702.00887*, 2017.

Kyunghyun Cho, Bart Van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734. Association for Computational Linguistics, 2014.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. OpenAI technical report, 2019.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

Wilson L Taylor. "cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433, 1953.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://aclanthology.org/D14-1162.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *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, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410.

ZS Harris. Distributional structure, 1954.

JR Firth. A synopsis of linguistic theory 1930-1955. *Studies in Linguistic Analysis, special volume/Blackwell*, 1957.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013. URL https://arxiv.org/abs/1301.3781.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019a. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. Unsupervised learning of sentence embeddings using compositional n-gram features. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 528–540, 2018.

Zhongtao Miao, Qiyu Wu, Kaiyan Zhao, Zilong Wu, and Yoshimasa Tsuruoka. Enhancing cross-lingual sentence embedding for low-resource languages with word alignment. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Findings of the Association for*

Computational Linguistics: NAACL 2024, pages 3225–3236, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.204. URL https://aclanthology.org/2024.findings-naacl.204.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR, 2015.

Mikel Artetxe and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, 2019. doi: 10.1162/tacl_a_00288. URL https://aclanthology.org/Q19-1038.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT sentence embedding. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.62. URL https://aclanthology.org/2022.acl-long.62.

Ali Sabet, Prakhar Gupta, Jean-Baptiste Cordonnier, Robert West, and Martin Jaggi. Robust cross-lingual embeddings from parallel sentences. *arXiv preprint arXiv:1912.12481*, 2019.

Koustava Goswami, Sourav Dutta, Haytham Assem, Theodorus Fransen, and John P. McCrae. Cross-lingual sentence embedding using multi-task learning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9099–9113, Online and Punta Cana, Dominican Republic, November 2021. Association

for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.716. URL https://aclanthology.org/2021.emnlp-main.716.

Yaushian Wang, Ashley Wu, and Graham Neubig. English contrastive learning can learn universal cross-lingual sentence embeddings. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9122–9133, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.621. URL https://aclanthology.org/2022.emnlp-main.621.

Hassan Sajjad, Firoj Alam, Fahim Dalvi, and Nadir Durrani. Effect of post-processing on contextualized word representations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3127–3142, 2022.

Sadaf Ghaffari and Nikhil Krishnaswamy. Grounding and distinguishing conceptual vocabulary through similarity learning in embodied simulations. In *Proceedings of the 15th International Conference on Computational Semantics*, 2023.

Nikhil Krshnaswamy, William Pickard, Brittany Cates, Nathaniel Blanchard, and James Pustejovsky. The voxworld platform for multimodal embodied agents. In *LREC proceedings*, volume 13, 2022.

James J Gibson. The theory of affordances. *Hilldale, USA*, 1(2):67–82, 1977.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, 2019b.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NeurIPS Deep Learning and Representation Learning Workshop*, 2015.

Yinhan Liu, Neil Houlsby, Jason Wei, Tal Schuster, and Orhan Firat. Crosslingual generalization through multitask finetuning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 324–330, Online, 2020. Association for Computational Linguistics.

Julian Eisenschlos, Sebastian Ruder, Piotr Czapla, Dani Yogatama, Wang Ling, and Phil Blunsom. Multifit: Efficient multi-lingual language model fine-tuning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 110–120. Association for Computational Linguistics, 2019.

Liang Xu, Fei Wu, Xing Wang, Wei Gao, and Lei Li. Effective fine-tuning methods for cross-lingual adaptation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1900–1913. Association for Computational Linguistics, 2021.

Xiangyu Wang, Daniel Kondratyuk, and Georg Heigold. Cross-lingual lemmatization and morphology tagging with two-stage multilingual bert fine-tuning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 722–731. Association for Computational Linguistics, 2020.

Sandipan Dandapat, Priyanka Biswas, Monojit Choudhury, and Kalika Bali. Complex linguistic annotation–no easy way out! a case from bangla and hindi pos labeling tasks. In *Proceedings of the third linguistic annotation workshop (LAW III)*, pages 10–18, 2009.

Marta Sabou, Kalina Bontcheva, and Arno Scharl. Crowdsourcing research opportunities: lessons from natural language processing. In *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies*, pages 1–8, 2012.

Karën Fort. *Collaborative annotation for reliable natural language processing: Technical and sociological aspects*. John Wiley & Sons, 2016.

Emily M Bender. On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, 6, 2011.

Edoardo Maria Ponti, Helen O'horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601, 2019.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, 2020.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL https://aclanthology.org/N18-1202.

Radford Alec, Narasimhan Karthik, Salimans Tim, and S Ilya. Improving language understanding with unsupervised learning. *Citado*, 17:1–12, 2018.

Steven Cao, Nikita Kitaev, and Dan Klein. Multilingual alignment of contextual word representations. *arXiv preprint arXiv:2002.03518*, 2020.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, et al. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, volume 57, 2019.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, 2017.

Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*, 2019.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, 2020.

Shijie Wu, Alexis Conneau, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. Emerging cross-lingual structure in pretrained language models. *arXiv preprint arXiv:1911.01464*, 2019.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. *arXiv preprint arXiv:1910.11856*, 2019.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, 2020.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling BERT for natural language understanding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP): Findings*, pages 4163–4174, 2020. doi: 10.18653/v1/2020.findings-emnlp.372.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled

version of bert: smaller, faster, cheaper and lighter. In *Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing*, 2019.

Raphael Tang, Yao Lu, Lin Liu, Lili Mou, Olga Vechtomova, and Jimmy Lin. Distilling task-specific knowledge from bert into simple neural networks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2019.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://arxiv.org/abs/2003.10555.

Yanan Liang, Yeyun Wu, Wei Li, Hongcheng Wang, Yuwei Zhang, Jian-Guang Lou, Daxin Jiang, Ming Zhou, and Shuai Wang. Xglue: A new benchmark dataset for cross-lingual pre-training, understanding, and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, 2020. doi: 10.18653/v1/2020.emnlp-main.485.

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. Patient knowledge distillation for BERT model compression. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4323–4332, 2019. doi: 10.18653/v1/D19-1441.

Abhijnan Nath, Shadi Manafi Avari, Avyakta Chelle, and Nikhil Krishnaswamy. Okay, let's do this! modeling event coreference with generated rationales and knowledge distillation. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3931–3946, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.218. URL https://aclanthology.org/2024.naacl-long.218.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies (NAACL-HLT), pages 3543–3556, 2019. doi: 10.18653/v1/N19-1357.

Erik Tjong Kim Sang and Fien De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147, 2003.

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Martha Palmer, Nianwen Xue, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, Robert Belvin, et al. Ontonotes release 4.0. *LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium*, 2011.

Kyle Gorman and Steven Bedrick. We need to talk about standard splits. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 2786. NIH Public Access, 2019.

Anders Søgaard, Sebastian Ebert, Jasmijn Bastings, and Katja Filippova. We need to talk about random splits. *arXiv preprint arXiv:2005.00636*, 2020.

Rob van der Goot. We need to talk about train-dev-test splits. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4485–4494, 2021.

Gabriel Bernier-Colborne and Philippe Langlais. Hardeval: Focusing on challenging tokens to assess robustness of ner. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1704–1711, 2020.

Asahi Ushio and Jose Camacho-Collados. T-ner: an all-round python library for transformer-based named entity recognition. *arXiv preprint arXiv:2209.12616*, 2022.

Emma Strubell, Patrick Verga, David Belanger, and Andrew McCallum. Fast and accurate entity recognition with iterated dilated convolutions. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2670–2680, 2017.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, 2020.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8: 842–866, 2020.

Ji Gao, Jack Lanchantin, Mary Lou Soffa, and Yanjun Qi. Black-box generation of adversarial text sequences to evade deep learning classifiers. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 50–56. IEEE, 2018.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Semantically equivalent adversarial rules for debugging nlp models. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 856–865, 2018.

Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, 2017.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1875–1885, 2018.

Zhengli Zhao, Dheeru Dua, and Sameer Singh. Generating natural adversarial examples. *arXiv preprint arXiv:1710.11342*, 2017.

Paul Michel, Xian Li, Graham Neubig, and Juan Pino. On evaluation of adversarial perturbations for sequence-to-sequence models. 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 3103–3114, 2019.

Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan Boyd-Graber. Trick me if you can: Human-in-the-loop generation of adversarial examples for question answering. *Transactions of the Association for Computational Linguistics*, 7:387–401, 2019a.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. Universal adversarial triggers for attacking and analyzing nlp. 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)*, 2019b.

Bin Liang, Hongcheng Li, Miaoqiang Su, Pan Bian, Xirong Li, and Wenchang Shi. Deep text classification can be fooled. *arXiv preprint arXiv:1704.08006*, 2017.

Matthias Blohm, Glorianna Jagfeld, Ekta Sood, Xiang Yu, and Ngoc Thang Vu. Comparing attention-based convolutional and recurrent neural networks: Success and limitations in machine reading comprehension. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 108–118, 2018.

Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. Textbugger: Generating adversarial text against real-world applications. *arXiv preprint arXiv:1812.05271*, 2018.

Pasquale Minervini and Sebastian Riedel. Adversarially regularising neural nli models to integrate logical background knowledge. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 65–74, 2018.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8018–8025, 2020.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. Bert-attack: Adversarial attack against bert using bert. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6193–6202, 2020.

Wee Chung Gan and Hwee Tou Ng. Improving the robustness of question answering systems to question paraphrasing. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 6065–6075, 2019.

Bill Yuchen Lin, Wenyang Gao, Jun Yan, Ryan Moreno, and Xiang Ren. Rockner: A simple method to create adversarial examples for evaluating the robustness of named entity recognition models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3728–3737, 2021a.

Sowmya Vajjala and Ramya Balasubramaniam. What do we really know about state of the art ner? In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5983–5993, 2022.

Akshay Srinivasan and Sowmya Vajjala. A multilingual evaluation of ner robustness to adversarial inputs. *arXiv preprint arXiv:2305.18933*, 2023.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. Semeval-2022 task 11: Multilingual complex named entity recognition (multiconer). In *Proceedings of the 16th international workshop on semantic evaluation (SemEval-2022)*, pages 1412–1437, 2022.

Abhijnan Nath, Sina Mahdipour Saravani, Ibrahim Khebour, Sheikh Mannan, Zihui Li, and Nikhil Krishnaswamy. A generalized method for automated multilingual loanword detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4996–5013, 2022.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th*

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1946–1958, 2017.

Holger Schwenk and Xian Li. A corpus for multilingual document classification in eight languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1560.

Giuseppe Attardi. Wikiextractor. https://github.com/attardi/wikiextractor, 2015.

Liat Ein Dor, Yosi Mass, Alon Halfon, Elad Venezian, Ilya Shnayderman, Ranit Aharonov, and Noam Slonim. Learning thematic similarity metric from article sections using triplet networks. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 49–54, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2009. URL https://aclanthology.org/P18-2009.

Jörg Tiedemann. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf.

Mohammad Abdous, Poorya Piroozfar, and Behrouz MinaeiBidgoli. Pests: Persian_english cross lingual corpus for semantic textual similarity. *Language Resources and Evaluation*, pages 1–21, 2024.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, 2020b.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin

Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In Steven Bethard, Marine Carpuat, Marianna Apidianaki, Saif M. Mohammad, Daniel Cer, and David Jurgens, editors, *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2001. URL https://aclanthology.org/S17-2001.

Shadi Manafi and Nikhil Krishnaswamy. Cross-lingual transfer robustness to lower-resource languages on adversarial datasets. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4174–4184, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.372.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 1715. Association for Computational Linguistics, 2016.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. Multilingual is not enough: Bert for finnish. *arXiv preprint arXiv:1912.07076*, 2019.

Raphael Scheible, Fabian Thomczyk, Patric Tippmann, Victor Jaravine, and Martin Boeker. Gottbert: a pure german language model. *arXiv preprint arXiv:2012.02110*, 2020.

Hasan Tanvir, Claudia Kittask, Sandra Eiche, and Kairit Sirts. Estbert: A pretrained

language-specific bert for estonian. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 11–19, 2021.

Abhijnan Nath, Sheikh Mannan, and Nikhil Krishnaswamy. AxomiyaBERTa: A phonologically-aware transformer model for Assamese. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11629–11646, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.739. URL https://aclanthology.org/2023.findings-acl.739.

Sharan Narasimhan, Suvodip Dey, and Maunendra Desarkar. Towards robust and semantically organised latent representations for unsupervised text style transfer. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 456–474, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.34. URL https://aclanthology.org/2022.naacl-main.34.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv e-prints*, pages arXiv–1910, 2019.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*, 2023.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig,

Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*, 2021b.