Towards a Common Understanding of Contributing Factors for Cross-Lingual Transfer in Multilingual Language Models: A Review

Fred Philippy[1,2]* and **Siwen Guo**[1] and **Shohreh Haddadan**[1]

[1]Zortify Labs, Zortify S.A.
19, rue du Laboratoire L-1911 Luxembourg
[2]SnT, University of Luxembourg
29, Avenue J.F Kennedy L-1359 Luxembourg
{fred, siwen, shohreh}@zortify.com

Abstract

In recent years, pre-trained Multilingual Language Models (MLLMs) have shown a strong ability to transfer knowledge across different languages. However, given that the aspiration for such an ability has not been explicitly incorporated in the design of the majority of MLLMs, it is challenging to obtain a unique and straightforward explanation for its emergence. In this review paper, we survey literature that investigates different factors contributing to the capacity of MLLMs to perform zero-shot cross-lingual transfer and subsequently outline and discuss these factors in detail. To enhance the structure of this review and to facilitate consolidation with future studies, we identify five categories of such factors. In addition to providing a summary of empirical evidence from past studies, we identify consensuses among studies with consistent findings and resolve conflicts among contradictory ones. Our work contextualizes and unifies existing research streams which aim at explaining the cross-lingual potential of MLLMs. This review provides, first, an aligned reference point for future research and, second, guidance for a better-informed and more efficient way of leveraging the cross-lingual capacity of MLLMs.

1 Introduction

The objective of cross-lingual transfer is to leverage knowledge learned by a model in a source language and to transfer it to a target language. While such a process of transferring knowledge and concepts across languages seems natural for a polyglot, it is believed to be less straightforward for a language model. Nevertheless, multilingual language models (MLLMs), such as mBERT (Devlin et al., 2019), XLM (Conneau and Lample, 2019) and XLM-R (Conneau et al., 2020a) demonstrate effective cross-lingual transfer capabilities. Such a transfer ability is moderately expected from XLM, given that parallel data is leveraged through a cross-lingual transfer learning objective during pre-training. However, it is less anticipated for mBERT and XLM-R, which are pre-trained on separate monolingual corpora without any explicit cross-lingual signal. Nevertheless, the latter show a surprisingly strong cross-lingual transfer capacity on a variety of downstream tasks (Hu et al., 2020). While no apparent factors explaining the nature of this ability can be intuitively derived from the properties of MLLMs, there have been many attempts to understand this behavior. Past research has outlined and investigated various factors that may impact cross-lingual transfer performance in MLLMs, but there are still open questions due to conflicting findings across studies. In our work, we inspect findings from past research investigating the inner workings of cross-lingual transfer in MLLMs. We not only outline overlapping contributions with consensual findings but also highlight and attempt to resolve conflicts between contradictory studies. Our work is structured according to five different types of factors whose impact on cross-lingual transfer capacity has been investigated in the past:

1. Linguistic Similarity
2. Lexical Overlap
3. Model Architecture
4. Pre-Training Settings
5. Pre-Training Data.

The examination of these factors provides insight into how and why MLLMs perform differently in different contexts. This understanding contributes to the overall explainability of MLLMs, which is essential for efficiently leveraging their cross-lingual transfer capacities and improving their performance in general.

A list of all the papers surveyed in this study is provided in Appendix A.

* Research was conducted at Zortify.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 5877–5891

2 Background

2.1 Multilingual Language Models

State-of-the-art MLLMS are predominantly based on the Transformer architecture (Vaswani et al., 2017). These models aim to produce multilingual representations of text that can be used for various downstream tasks across different languages. However, MLLMs may adopt different learning objectives to achieve this goal. Some models exploit parallel data and incorporate a cross-lingual learning objective during pre-training, such as XLM (Conneau and Lample, 2019) and UniCoder (Huang et al., 2019), while other models rely on separate monolingual corpora without any explicit cross-lingual supervision, such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020a).

Despite their impressive performance, MLLMs also face several challenges and limitations, such as the imbalance in the pre-training data, the limited availability of evaluation datasets for different (low-resource) languages and the trade-off between model capacity and language coverage, known as the *curse of multilinguality*, which affects their efficiency and effectiveness.. Therefore, more research is needed to understand, improve, and develop multilingual models that can achieve a balanced and robust performance across languages. Within this line of research, cross-lingual transfer has proven to be a valuable method to leverage resources from high-resource languages to improve downstream task performance for low-resource languages.

2.2 (Zero-Shot) Cross-Lingual Transfer

In the context of MLLMs, cross-lingual transfer refers to transferring certain knowledge from one language to another. From a practical standpoint, a traditional pipeline for zero-shot cross-lingual transfer typically includes two steps: **i)** A multilingual model is fine-tuned on a labeled dataset in the source language, and **ii)** The fine-tuned model is applied to a target language without any additional fine-tuning. In a few-shot setting, a small number of labeled samples in the target language are utilized for additional fine-tuning of the model.

During recent years, a number of studies have investigated cross-lingual transfer methods (Pikuliak et al., 2021). In addition to the zero-shot transfer approach, there are some studies that apply machine translation to enable cross-lingual transfer (Conneau et al., 2018; Conneau and Lample, 2019; Conneau et al., 2020a; Hu et al., 2020). In the translate-train approach, the labeled training set is translated from the source language into the target language for the purpose of fine-tuning. Correspondingly, the translate-test approach involves translating the test set from the target language into the source language during inference. In our review, we focus on the aforementioned traditional cross-lingual transfer process to avoid making the assumption that a translation system for the source language is available. Additionally, given that machine translation is highly context-dependent and is often unreliable when dealing with unconventional and ambiguous languages, it would add external factors to our effort of trying to understand the transfer behavior of MLLMs.

3 Factors That Affect Cross-Lingual Transfer

3.1 Linguistic Similarity

The hypothesis that linguistic similarity correlates with cross-lingual transfer performance has been examined repeatedly. With regard to quantifying such a relationship, we observe two main approaches: **i)** synthetically modifying a specific linguistic feature of a natural language and observing the impact on transfer performance by controlling the magnitude of the modification; and **ii)** using linguistic similarity metrics to capture the similarity between two natural languages.

Two established linguistic similarity metrics which are commonly used for this purpose are: the World Atlas of Language Structures (WALS)[1] (Dryer and Haspelmath, 2013), a database of structural properties of languages, and `lang2vec`[2], a tool providing vectors that represent linguistic properties of languages based on the URIEL (Littell et al., 2017) database. An alternative metric for evaluating linguistic similarity is eLinguistics[3] (Beaufils and Tomin, 2020), which is based on the comparison of consonants in word pairs. Table 1 lists papers that have investigated the impact of linguistic similarity, along with the linguistic components that were studied and the metrics used.

[1] `https://wals.info/`

[2] `lang2vec` enables querying the URIEL database. It extracts vectors which encode different linguistic components for each language. This, in turn, allows to quantify the similarity or dissimilarity between languages.

[3] `http://www.elinguistics.net/`

Paper	Task	Model	Lang. type	Features	Metric
Lin et al. (2019)	DP, EL, MT, POS	/	NL	GEN, GEO, INV, PHON, SYN	lang2vec
Pires et al. (2019)	NER, POS	mBERT	NL	SYN	WALS
Tran and Bisazza (2019)	DP	mBERT	NL	SYN	lang2vec
Dufter and Schütze (2020)	SR, WA, WT	BERT (small)	SL	SYN	/
K et al. (2020)	NER, NLI	Bilingual BERT	NL/SL	SYN, UniFreq	/
Lauscher et al. (2020)	DP, POS, NER, NLI, QA	mBERT, XLM-R	NL	SYN, PHON, INV, GEN, GEO	lang2vec
Dolicki and Spanakis (2021)	NER, NLI, POS	XLM-R	NL	GEN, GEO, SYN	lang2vec, WALS
Srinivasan et al. (2021)	NER, NLI, POS	mBERT, XLM-R	NL	ALL	lang2vec, WALS
Ahuja et al. (2022)	DC, NER, NLI, POS, QA	mBERT, XLM-R	NL	ALL, GEN, GEO, PHON, SYN	lang2vec, WALS
Deshpande et al. (2022)	NER, NLI, POS, QA	Bilingual RoBERTa (small)	SL	SYN	/
de Vries et al. (2022)	POS	XLM-R Base	NL	FAM, SYN, WS, WST	/
Eronen et al. (2022)	DC	mBERT, XLM-R	NL	ALL	eLinguistics, WALS
Wu et al. (2022)	AJ, SA, SS, NLI	English RoBERTa	SL	SYN	/

Table 1: List of studies investigating linguistic features that impact cross-lingual transfer. The **Lang. type** column indicates the type of language that has been used. We use the following abbreviations. **NL**: Natural Languages, **SL**: Synthetic languages. The **Features** column indicates which linguistic features have been investigated. We use the following abbreviations. **ALL**: Aggregated language distance of multiple linguistic features, **GEN**: Genetic distance, **GEO**: Geographical distance, **INV**: Inventory, **PHON**: Phonology, **SYN**: Syntax, **UniFreq**: Unigram Frequency, **WS**: Writing system, **WST**: Writing system type. The **Metrics** column indicates which type of metric has been used to measure language similarity between natural languages. The abbreviations of the **Task** column can be found in Table 2 in Appendix A.

Is Word Order Important? The impact of word order[4], or more generally, syntax, has been extensively investigated in the past. Based on experiments with different settings, its positive effect on cross-lingual transfer has been confirmed for Dependency Parsing (DP) (e.g., Lin et al., 2019; Lauscher et al., 2020), Named Entity Recognition (NER) (e.g., Dolicki and Spanakis, 2021; Deshpande et al., 2022; Ahuja et al., 2022), Part-Of-Speech Tagging (POS) (e.g., Ahuja et al., 2022; de Vries et al., 2022; Deshpande et al., 2022), Natural Language Inference (NLI) (e.g., K et al., 2020; Lauscher et al., 2020; Ahuja et al., 2022) and Question Answering (QA) (e.g., Deshpande et al., 2022; Ahuja et al., 2022; Lauscher et al., 2020). Furthermore, Dufter and Schütze (2020) sought to validate these findings on a representation level by evaluating cross-lingual transfer on word translation, word retrieval and sentence retrieval.

Despite the common findings stated above, there are contradictions in the results of a number of studies in which different experimental settings are used. Wu et al. (2022) and Deshpande et al. (2022) investigated the impact of word order by isolating it from other factors. In both works, language variants were created by randomly permutating, inversing, or consistently adapting word order to a different language via a dependency tree. A common finding has been that reversed or randomized word order deteriorates cross-lingual transfer performance significantly more than adapting the word order to a different language. This makes it hard to compare the aforementioned findings to results from Dufter and Schütze (2020) and K et al. (2020) who solely evaluated on language variants with

[4] *Word order* describes the degree of similarity between the source and target language in terms of elements like subject-object-verb, subject-verb and object-verb order.

reversed or randomly permuted word order, respectively. Even if both latter works found evidence that word order impacts transfer performance, it is important to consider that Wu et al. (2022) and Deshpande et al. (2022) have comparable findings in similar settings but observed a less significant effect when switching to a more structured syntactic modification.

On the other hand, Lauscher et al. (2020) and Ahuja et al. (2022) obtained results containing evidence that word order may be more important for mBERT than for XLM-R. A possible explanation for this finding is that mBERT encodes more syntactic knowledge than XLM-R, as shown by Zheng and Liu (2022).

Which Other Linguistic Features Affect Cross-Lingual Transfer? In addition to examining the effect of similar word order, some research has also focused on the impact of other linguistic characteristics. Srinivasan et al. (2021) measured general language similarity by aggregating multiple lang2vec vectors. They observed a high, medium and low importance of language similarity for cross-lingual transfer in POS, QA and NLI, respectively. Their observation holds for both mBERT and XLM-R. By evaluating on a document classification task, Eronen et al. (2022) observed a medium correlation between the cross-lingual transfer performance of both models and an aggregation of WALS features.

On a more detailed level, low **geographical distance**[5] between languages has been found to be beneficial for cross-lingual transfer on several occasions (Lin et al., 2019; Lauscher et al., 2020; Dolicki and Spanakis, 2021; Ahuja et al., 2022). Similarly, low **genetic distance**[6] has also been shown to positively affect cross-lingual transfer (Lin et al., 2019; Lauscher et al., 2020; Dolicki and Spanakis, 2021; de Vries et al., 2022; Eronen et al., 2022). However, it has not been selected as a predictive feature in the Lasso regression performed by Ahuja et al. (2022). Low **phonological distance**[7] has been demonstrated to be more important for token-level tasks (NER, POS, DP, QA) than for sentence-level tasks (NLI, MT) (Lin et al., 2019; Lauscher et al., 2020; Ahuja et al., 2022). **Inventory features**[8] have been shown to be of low importance when selecting a suitable transfer language (Lin et al., 2019; Lauscher et al., 2020).

Furthermore, K et al. (2020) investigated the utility of the hypothesis that similar words have a similar frequency in their respective language (Zipf's law). The authors assessed cross-lingual transfer using a synthetic target language, which has a similar unigram frequency but no other explicit commonality. Although its utility in combination with additional factors has not been evaluated, unigram frequency has been found to be unable to ensure a successful transfer between languages as a standalone feature.

Conclusion In previous research, syntax has been suggested as potentially the most important linguistic contributor for better cross-lingual transfer. However, we hypothesize that its impact may be overestimated when assessed by randomly permutating or inversing word order, since such syntactic modifications are unlikely to occur in natural languages. Besides syntax, other linguistic features, such as geographical, genetic and phonological similarity, have been identified as potential linguistic contributors as well. In addition, we emphasize the importance of investigating the distinct interplay of different linguistic features.

3.2 Lexical Overlap

Since lexical overlap may intuitively create a potential connection between closely related languages and therefore possibly explain the varying transfer performance across language pairs, its impact has been investigated on many occasions. Lexical overlap merely specifies the amount of shared words or subwords between a language pair. Typically, it is calculated as the percentage of unique words or subwords common to the vocabularies of both the source and target languages. There are various approaches to quantify lexical overlap between languages. A common corpus-based method is to divide the number of shared words or subwords between two monolingual corpora by the total number of unique words or subwords in both corpora. Two further metrics that aim to quantify lexical overlap are ezGlot[9] (Kovacevic et al., 2022) and the normalized Levenshtein distance (LDND) (Wichmann et al., 2010).

[5] **Geographical distance** is based on the orthodromic distance between languages' primary locations.

[6] **Genetic distance** between two languages measures their degree of common ancestry.

[7] **Phonological distance** measures the difference of phonological properties between languages.

[8] **Inventory features** describe a language's phonetic, phonological, and morphological components.

[9] https://www.ezglot.com/

Does High Lexical Overlap Improve Cross-Lingual Transfer? While many studies have found a positive correlation between lexical overlap and cross-lingual transfer performance (Wu and Dredze, 2019; Patil et al., 2022; de Vries et al., 2022), other studies do not support the existence of such a positive correlation (Pires et al., 2019; Tran and Bisazza, 2019; K et al., 2020; Conneau et al., 2020b).

Pires et al. (2019), Tran and Bisazza (2019) and Wu and Dredze (2019) applied the traditional cross-lingual zero-shot transfer evaluation pipeline (see Section 2.2) on different tasks and natural languages. Besides showcasing the cross-lingual capacity of mBERT, their objective was to measure the impact of lexical overlap on this ability. Despite the similarities of their experiments, their findings are not all consistent. Based on the experiments on POS and DP in more than 16 languages, Pires et al. (2019) and Tran and Bisazza (2019) have found that cross-lingual transfer performance is largely independent of lexical overlap. Wu and Dredze (2019), on the other hand, derived a correlation between transfer performance and lexical overlap from results on more tasks but fewer languages.

de Vries et al. (2022) evaluated cross-lingual transfer performance across languages with different writing systems. They found that a shared writing system and thus a higher lexical overlap (measured by LDND) contribute to better cross-lingual transfer. However, they also showed that cross-script transfer is not impossible. Such a finding clearly supports the hypothesis that lexical overlap should not be seen as a self-contained factor. Based on these findings, it becomes evident that a more detailed analysis of the impact of lexical overlap is needed. Such detailed analyses would provide additional clarification on the apparent contradictions among past contributions.

Does the Impact of Lexical Overlap on Transfer Performance Depend on Other Linguistic Features? With the intention of a more fine-grained investigation, K et al. (2020) and Conneau et al. (2020b) have conducted experiments in a controlled setup by synthetically adjusting the amount of lexical overlap. In both cases, no significant correlation between lexical overlap and transfer performance was observed. Patil et al. (2022) used similar configurations but differentiated between high- and low-resource settings. In contrast to previous findings, they observed a positive correlation between subword overlap and transfer performance. Furthermore, they concluded that this correlation increases when the source language has a smaller pre-training corpus.

Deshpande et al. (2022) took this a step further by transferring exclusively from synthetic English to English. This allowed them to isolate the impact of lexical overlap and control interactions with other linguistic features. From their experiments, it can be concluded that lexical overlap matters most when the word orders of the source and target languages differ. This finding explains the results of K et al. (2020) and Conneau et al. (2020b) who only used language pairs of similar word order and did not observe a high impact of lexical overlap on transfer performance. The only language pair in their experiments with dissimilar word order was English-Hindi, which has small lexical overlap by default due to their different scripts. Consequently, further reducing the overlap is, as observed in their results, not expected to impact transfer performance. Moreover, this potentially explains the aforementioned findings of Pires et al. (2019) and Tran and Bisazza (2019) who performed their experiments on a subset of languages for which word order and lexical overlap are strongly correlated. In both studies, language pairs with low lexical overlap were most likely also differing in their word order, while language pairs with higher lexical overlap tended to have similar word order. Pires et al. (2019), unfortunately, did not provide exact transfer performance values. However, in line with our aforementioned observations, in their study a correlation between transfer performance and lexical overlap could be observed in language pairs with low lexical overlap and thus dissimilar word order. This correlation decreases as lexical overlap increases and thus word order becomes mostly similar.

Does the Impact of Lexical Overlap on Transfer Performance Depend on the Type of Downstream Task? Lin et al. (2019), Srinivasan et al. (2021) and Ahuja et al. (2022) trained predictors to predict the cross-lingual transfer performance of a given language model for a variety of downstream tasks. Lexical overlap between source and target languages was selected as one of the predictor variables. By comparing the feature importance values of lexical overlap, clear differences across different types of downstream tasks emerged. While Lin et al. (2019) and Srinivasan et al. (2021) observed

high feature importance values of lexical overlap for syntactic tasks like POS, NER and DP, and lower feature importance values for the semantic-oriented task of NLI, Ahuja et al. (2022) found the opposite.

Given the minor but numerous differences among studies, providing a thorough explanation of the aforementioned contradictory findings is challenging. One notable distinction among the three similar contributions is the use of tree-based methods, specifically Gradient-Boosted Decision Trees and XGBoost, by Lin et al. (2019) and Srinivasan et al. (2021), respectively, and the use of Lasso Regression, a type of linear regression, by Ahuja et al. (2022). Given that tree-based models are able to capture nonlinear relationships between the dependent and independent variables while Lasso Regression can only describe such a relationship linearly, the latter method might attribute higher feature importance to linearly related predictors compared to predictors that have a more significant but nonlinear impact on the dependent variable. A recent study by Patankar et al. (2022) provides evidence in support of our hypothesis.

Conclusion We found evidence that lexical overlap is particularly important when the pre-training corpus for the source language is small or when the word order between the source and target languages is dissimilar. However, we conclude that lexical overlap is not a sufficient standalone factor to explain cross-lingual transfer. We also observed in experiment results in the literature that cross-lingual transfer is feasible between languages with different scripts (and thus zero lexical overlap), which further supports our conclusion. We recommend that future experiments take a closer look at the interaction between lexical overlap and further contributing factors. Moreover, future experiments may be set up in a way to provide additional insight into task-specific differences that are currently not fully understood.

3.3 Model Architecture

Model architecture may be crucial to the success of cross-lingual transfer because it determines how a model processes and represents information. Therefore, it is closely connected to the model's capacity to learn and capture knowledge. An ill-suited architecture could potentially hinder the model's ability to transfer knowledge from one language to another.

Which Model Architecture Components Can Affect Transfer Performance? K et al. (2020) provided one of the first investigations on the impact of model architecture on cross-lingual transfer. In their study, they focused on three main architectural components of Transformer-based models: **i)** network depth, **ii)** number of attention heads, **iii)** number of model parameters. They found that an increased network depth (i.e., more hidden layers), with a fixed number of model parameters, leads to better cross-lingual transfer. Increasing the number of model parameters with a fixed number of hidden layers had a similar but less significant impact. The number of attention heads, on the other hand, were found to be irrelevant for cross-lingual transfer performance. In their experiments, satisfactory transfer performance could even be achieved with only a single attention head.

Conneau et al. (2020b) trained a bilingual BERT model where all parameters are shared, and compared the transfer performance to the case where the embedding layer and/or up to the first six Transformer layers are separated for both languages. In the experiments on NLI, DP, and NER for three different natural language pairs, they observed that the transfer performance decreases when fewer layers are shared. This finding led the authors to hypothesize that a limited model capacity requires the model to use its parameters more efficiently by aligning the representations of semantically similar text across different languages, instead of creating separate embedding spaces for different languages. This hypothesis was confirmed by Dufter and Schütze (2020) who observed a degradation of mBERTs cross-lingual transfer ability by purposely overparameterizing the model. On the other hand, the authors referred to the "curse of multilinguality" (Conneau et al., 2020a) which states that, for a fixed model size, the number of languages a model can cover until its overall performance starts to decrease is limited. This can be alleviated by expanding the model capacity, i.e., by increasing the number of parameters, but as mentioned previously, too many parameters could deteriorate cross-lingual transfer performance.

Wu et al. (2022) demonstrated the importance of a well-trained embedding layer for cross-lingual transfer. When the embedding layer is reinitialized before fine-tuning, the performance on the GLUE benchmark (Wang et al., 2018) decreases by 40%. More specifically, Deshpande et al. (2022) found

that the cross-lingual alignment of the static token embeddings used by the embedding layer is crucial for satisfactory cross-lingual transfer performance.

Conclusion There is evidence to suggest that an overparameterized model might create language-specific sub-spaces and therefore struggle to provide cross-lingual representations. Concurrently, models with fewer parameters are required to use their parameters more efficiently and thus align representations across languages more easily. Therefore, we strongly suggest to explore how the trade-off between languages and parameters affects cross-linguality in MLLMs.

Furthermore, one contribution has revealed evidence that for a fixed number of parameters, model depth can be more important than the number of attention heads. However, it is not well studied yet how model architecture components and data-specific components (e.g., dataset size, number of languages) interact to impact cross-lingual transfer performance.

3.4 Pre-Training Settings

Given that MLLMs are able to perform zero-shot cross-lingual transfer, their cross-lingual capacity has to emerge during pre-training as they are not exposed to any task-specific data in the target language during fine-tuning. Therefore, investigating factors related to the pre-training process could lead to a better understanding of the cross-lingual capacity of MLLMs as well as how to further improve it.

Which Pre-Training Components Contribute to the Cross-Lingual Capabilities of MLLMs? Devlin et al. (2019) introduced the Next Sentence Prediction (NSP) objective to pre-train language models in combination with the Masked Language Model (MLM) objective. However, the usefulness of NSP for downstream tasks has been debated on several occasions (Yang et al., 2019; Conneau and Lample, 2019; Liu et al., 2019; Joshi et al., 2020). K et al. (2020) probed its impact on cross-lingual transfer performance. By removing NSP from the pre-training process, performance improved for both NER and NLI. This finding is particularly remarkable for NLI as this task is considered to be closely related to NSP, as both tasks involve the classification of sentence pairs. Furthermore, they also found that training on subwords rather than words or characters provides more cross-lingual capacity to the model. Lastly, it has been shown that adding a language identity marker to the input during pre-training does not significantly improve cross-lingual transfer performance. This outcome may suggest that MLLMs automatically learn language-specific information (Wu and Dredze, 2019; Liu et al., 2020) or that such additional input is not necessary for their cross-lingual capability. Furthermore, Liu et al. (2020) showed that pre-training on longer input sequences helps MLLMs to achieve better cross-lingual transfer abilities, especially when pre-trained on large corpora.

Apart from the learning objective, the impact of tokenizers and their vocabulary on a model's cross-lingual potential have been examined as well. Artetxe et al. (2020) evaluated transfer performance of bilingual and multilingual BERT models pre-trained with different vocabulary settings on four different downstream task datasets. In multilingual settings, they found that increased joint vocabulary size[10] leads to improved cross-lingual transfer performance. Furthermore, in the context of bilingual models, cross-lingual transfer performance is enhanced when disjoint subword vocabularies[11] are utilized instead of a joint subword vocabulary for both languages. That said, it is unclear how well disjoint vocabularies would perform when scaling the model to more languages.

Ahuja et al. (2022) also studied the effect of tokenizers on cross-lingual transfer. They quantify tokenizer quality by applying two metrics introduced by Rust et al. (2021), namely the tokenizer's *fertility* and its proportion of continued words. Both features are included in their cross-lingual transfer performance prediction model. By looking at the feature importance values, it became clear that cross-lingual transfer performance depends significantly more on a high-quality tokenizer for POS, NER and QA than for Document Classification (DC) and Sentence Retrieval (SR). Such a finding aligns with the fact that the former downstream tasks operate to a greater extent on token level than the latter ones.

Conclusion Previous studies have identified a number of pre-training components which may enable an improved cross-lingual transfer capacity of MLLMs. Some examples include removing NSP from the pre-training learning objective, a larger

[10]Experiments were conducted with vocabulary sizes of 32k, 64k, 100k, and 200k.

[11]A joint vocabulary of 32k subwords was compared to two separate vocabularies, each with 32k subwords, for each language.

vocabulary size and a high-quality multilingual tokenizer.

3.5 Pre-Training Data

MLLMs, such as mBERT, are able to learn cross-lingual representations during pre-training without having been specifically designed to do so. This may happen as a result of the model's exposure to multiple languages during the pre-training phase. However, the impact of the pre-training corpus on this self-learned ability is not yet fully comprehended.

Does the Pre-training Corpus Size Influence a Model's Cross-Lingual Transfer Ability? Lauscher et al. (2020), Srinivasan et al. (2021) and Ahuja et al. (2022) found that the size of the pre-training target language corpora correlates strongly with the transfer performance of mBERT and XLM-R for high-level tasks (NLI & QA) and less for low-level tasks (DP, POS, NER).

Liu et al. (2020) performed a more controlled experiment by comparing two multilingual BERT models pre-trained on different amounts of data from 15 languages. When trained on a small corpus of 200k sentences per language, mBERT showed poor zero-shot cross-lingual transfer performance, with results only comparable to those of non-contextualized word embedding models such as GloVe (Pennington et al., 2014) and Word2Vec (Mikolov et al., 2013) that were also trained on the same amount of data. Increasing the pre-training corpus size to 1000k sentences per language resulted in significantly improved transfer performance of mBERT, while both non-contextualized word embedding models did not demonstrate such an enhancement in transfer performance.

Lin et al. (2019) found that the ratio between the pre-training data corpus size of the transfer and target language is an important factor for successful cross-lingual transfer for POS but less so for MT and DP. However, the size of the target language pre-training corpus is not examined as a distinct feature in their work, making it more challenging to compare their findings with those mentioned previously.

Does the Source of the Pre-training Corpus Affect Cross-Lingual Transfer Performance? Dufter and Schütze (2020) found that cross-lingual transfer performance decreases when the respective monolingual pre-training corpora come from the same domain but are not parallel (e.g., by pre-training on different parts of the same corpus from a given domain). Conneau et al. (2020b) obtained similar results for monolingual pre-training corpora from different domains (e.g., Wikipedia vs. Common Crawl). Deshpande et al. (2022) found that pre-training on corpora from different domains has a more significant negative impact on cross-lingual transfer performance than pre-training on non-parallel corpora from the same domain. Interestingly, Conneau et al. (2020b) and Deshpande et al. (2022) found that the negative effect of different pre-training corpora sources on cross-lingual transfer performance is the most significant for NER. A potential explanation could be that in both cases, the NER dataset consists of Wikipedia text which was also used as the pre-training corpus in their baseline experiments. To the best of our knowledge, there is no research available on the impact of using a shared source for pre-training and task-specific data in the cross-lingual transfer context.

Conclusion Target language pre-training corpus size and comparable corpora sources across languages have been identified as two crucial factors for enhanced cross-lingual transfer capabilities in MLLMs. However, pre-training corpus size of the target language has been shown to be more important for higher-level than for lower-level tasks.

4 Related Work

Recently, numerous studies have investigated how to leverage the cross-lingual potential of MLLMs for better transfer among languages. Pikuliak et al. (2021) conducted a survey on existing cross-lingual transfer paradigms but did not investigate the components that are responsible for their inner workings. Doddapaneni et al. (2021), in their survey on pre-trained MLLMs, commented on various factors that affect cross-lingual transfer. Since they discussed a wide range of topics, they could not investigate in depth the findings from the studies that examined these factors. After the publication of that work, many studies have further investigated various factors that impact transfer performance and have helped to resolve some of the conflicts among past contributions.

Malkin et al. (2022) introduced a *Linguistic Blood Bank* that shows that not all languages transfer equally well among each other. This emphasizes the need for a clearer understanding of the

underlying factors that contribute to this imbalance. On a related note, Turc et al. (2021) found that English is not the overall best source language for cross-lingual transfer, despite its dominance in the pre-training corpus.

Hence, automating the process of selecting a source language for cross-lingual transfer has been pursued on many occasions (Lin et al., 2019; Lauscher et al., 2020; Srinivasan et al., 2021; Dolicki and Spanakis, 2021). These attempts focused on creating meta-models[12] which aim to predict the most suitable source language for a given use-case based on some of the factors from Section 3.

By incorporating typological features, Ansell et al. (2021), Lee et al. (2022) and Chronopoulou et al. (2023) enhanced the performance of adapters for low-resource languages. However, our survey reveals that adapters and other methods could benefit from more than just typological factors when dealing with low-resource scenarios.

5 Discussion

Building on previous research, our study investigated various factors that impact cross-lingual transfer performance. We examined a range of factors, including language-related factors as well as factors related to the models and training data. One of the existing challenges is the presence of contradictory findings from past studies. To better understand these discrepancies, we outlined possible explanations that could account for these differences, including the varying implementation details of experiments and evaluation methods.

One of the key variations among the various studies is the use of synthetic and natural languages. Synthetic languages can be created with a controlled level of variation by manipulating specific linguistic features. However, they may not capture the full range of complexity found in natural languages, which may limit their usefulness in drawing conclusions that apply to real-world settings.

While we acknowledge the value of the efficiency of using transfer performance prediction models to automate the selection of transfer languages, the accuracy of relying on feature importance values to make conclusions about the individual impact of specific factors on cross-lingual transfer performance cannot be taken as an absolute.

Our survey results show that all the factors we examined affect cross-lingual transfer in different ways and settings. Although the interaction of factors has only been investigated in a limited number of past studies, our findings suggest that some factors can influence the importance of others. Additionally, there is evidence suggesting that there are task-specific differences, for example, the pre-training corpus size being more important for higher-level tasks and lexical overlap, and word order being more important for lower-level tasks. Therefore, we strongly encourage future research to examine the full range of interactions among different factors as well as the underlying reasons for task-specific divergences.

Given that especially linguistic features have been shown to have a strong impact on cross-lingual transfer performance, we suggest that future research could examine whether languages are indeed the most suitable basis for constructing multilingual models. Instead of focusing on the distribution of languages in the pre-training corpus, it might be more efficient to focus on the distribution of linguistic features. One possible approach is to cluster texts according to their syntactic complexity or their morphological diversity, irrespective of their language affiliation. This would enable the development of a model that could potentially better transfer to languages that were absent in the pre-training corpus but which share linguistic features with the languages that the model has seen during pre-training.

In addition, we advocate for the development of more multilingual downstream task datasets that encompass a wider and more diverse range of languages, as this would enable a more comprehensive and robust assessment of cross-lingual transfer capabilities across various language models and approaches. Furthermore, we urge more investigation on the influence of the aforementioned factors on generative models, as this area remains relatively unexplored despite the current prominence of GPT-like models.

Limitations

One potential limitation of this review is our selection bias which may affect the representativeness of the included papers. Another limitation is the potential differences in methodologies across the

[12] In this context, the objective of a meta-model is to predict the performance of other models.

papers we reviewed, which makes it difficult to draw generalizable conclusions. Different studies use different experimental settings and methods for measuring feature importance, which could also impact the comparability of the findings across the included studies. Furthermore, we acknowledge the potential publication bias which might lead to an overestimation of the impact of different factors, as studies with statistically significant results may be more likely to be published than those with non-significant results.

Ethics Statement

We have carefully reviewed the relevant literature to ensure that all research included in this review has been conducted in accordance with ethical guidelines. We have also attempted to present a fair and accurate representation of the current state of research on this topic. We hope that this review will contribute to the ongoing debate about the factors impacting cross-lingual transfer performance, with the ultimate goal of ensuring that low-resource languages can equally benefit from the use of multilingual language models. We believe that it is important for all languages and communities to have equal access to the benefits and opportunities provided by the advances in natural language processing, and we hope that our review will serve as a useful resource in this regard.

References

Kabir Ahuja, Shanu Kumar, Sandipan Dandapat, and Monojit Choudhury. 2022. Multi Task Learning For Zero Shot Performance Prediction of Multilingual Models. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5454–5467, Dublin, Ireland. Association for Computational Linguistics.

Alan Ansell, Edoardo Maria Ponti, Jonas Pfeiffer, Sebastian Ruder, Goran Glavaš, Ivan Vulić, and Anna Korhonen. 2021. MAD-G: Multilingual Adapter Generation for Efficient Cross-Lingual Transfer. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 4762–4781, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the Cross-lingual Transferability of Monolingual Representations. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4623–4637, Online. Association for Computational Linguistics.

Vincent Beaufils and Johannes Tomin. 2020. Stochastic approach to worldwide language classification: the signals and the noise towards long-range exploration. preprint, SocArXiv.

Alexandra Chronopoulou, Dario Stojanovski, and Alexander Fraser. 2023. Language-Family Adapters for Low-Resource Multilingual Neural Machine Translation. In Proceedings of the The Sixth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2023), pages 59–72, Dubrovnik, Croatia. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. Unsupervised Cross-lingual Representation Learning at Scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual Language Model Pretraining. In Advances in Neural Information Processing Systems, volume 32. Curran Associates, Inc.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating Cross-lingual Sentence Representations. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Emerging Cross-lingual Structure in Pretrained Language Models. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6022–6034, Online. Association for Computational Linguistics.

Wietse de Vries, Martijn Wieling, and Malvina Nissim. 2022. Make the Best of Cross-lingual Transfer: Evidence from POS Tagging with over 100 Languages. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7676–7685, Dublin, Ireland. Association for Computational Linguistics.

Ameet Deshpande, Partha Talukdar, and Karthik Narasimhan. 2022. When is BERT Multilingual? Isolating Crucial Ingredients for Cross-lingual Transfer. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3610–3623, Seattle, United States. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of

Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sumanth Doddapaneni, Gowtham Ramesh, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2021. A Primer on Pretrained Multilingual Language Models. ArXiv:2107.00676 [cs].

Błażej Dolicki and Gerasimos Spanakis. 2021. Analysing The Impact Of Linguistic Features On Cross-Lingual Transfer. ArXiv:2105.05975 [cs].

Matthew S. Dryer and Martin Haspelmath. 2013. WALS Online. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Philipp Dufter and Hinrich Schütze. 2020. Identifying Elements Essential for BERT's Multilinguality. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4423–4437, Online. Association for Computational Linguistics.

Juuso Eronen, Michal Ptaszynski, Fumito Masui, Masaki Arata, Gniewosz Leliwa, and Michal Wroczynski. 2022. Transfer language selection for zero-shot cross-lingual abusive language detection. Information Processing & Management, 59(4):102981.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalisation. In Proceedings of the 37th International Conference on Machine Learning, pages 4411–4421. PMLR.

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A Universal Language Encoder by Pre-training with Multiple Cross-lingual Tasks. 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 2485–2494, Hong Kong, China. Association for Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving Pre-training by Representing and Predicting Spans. Transactions of the Association for Computational Linguistics, 8:64–77.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-Lingual Ability of Multilingual BERT: An Empirical Study. In Proceedings of the 8th International Conference on Learning Representations (ICLR 2020).

Lazar Kovacevic, Vladimir Bradic, Gerard de Melo, Sinisa Zdravkovic, and Olga Ryzhova. 2022. Ezglot.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. 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, Online. Association for Computational Linguistics.

Jaeseong Lee, Seung-won Hwang, and Taesup Kim. 2022. FAD-X: Fusing Adapters for Cross-lingual Transfer to Low-Resource Languages. In Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 57–64, Online only. Association for Computational Linguistics.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing Transfer Languages for Cross-Lingual Learning. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Chi-Liang Liu, Tsung-Yuan Hsu, Yung-Sung Chuang, and Hung-Yi Lee. 2020. A Study of Cross-Lingual Ability and Language-specific Information in Multilingual BERT. ArXiv:2004.09205 [cs].

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv:1907.11692 [cs].

Dan Malkin, Tomasz Limisiewicz, and Gabriel Stanovsky. 2022. A Balanced Data Approach for Evaluating Cross-Lingual Transfer: Mapping the Linguistic Blood Bank. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4903–4915, Seattle, United States. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. ArXiv:1301.3781 [cs].

Shantanu Patankar, Omkar Gokhale, Onkar Litake, Aditya Mandke, and Dipali Kadam. 2022. To Train or Not to Train: Predicting the Performance of Massively Multilingual Models. In Proceedings of the First Workshop on Scaling Up Multilingual Evaluation, pages 8–12, Online. Association for Computational Linguistics.

Vaidehi Patil, Partha Talukdar, and Sunita Sarawagi. 2022. 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. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matúš Pikuliak, Marián Šimko, and Mária Bieliková. 2021. Cross-lingual learning for text processing: A survey. Expert Systems with Applications, 165:113765.

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

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How Good is Your Tokenizer? On the Monolingual Performance of Multilingual Language Models. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 3118–3135, Online. Association for Computational Linguistics.

Anirudh Srinivasan, Sunayana Sitaram, Tanuja Ganu, Sandipan Dandapat, Kalika Bali, and Monojit Choudhury. 2021. Predicting the Performance of Multilingual NLP Models. ArXiv:2110.08875 [cs].

Ke Tran and Arianna Bisazza. 2019. Zero-shot Dependency Parsing with Pre-trained Multilingual Sentence Representations. In Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019), pages 281–288, Hong Kong, China. Association for Computational Linguistics.

Iulia Turc, Kenton Lee, Jacob Eisenstein, Ming-Wei Chang, and Kristina Toutanova. 2021. Revisiting the Primacy of English in Zero-shot Cross-lingual Transfer. ArXiv:2106.16171 [cs].

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Søren Wichmann, Eric W. Holman, Dik Bakker, and Cecil H. Brown. 2010. Evaluating linguistic distance measures. Physica A: Statistical Mechanics and its Applications, 389(17):3632–3639.

Shijie Wu and Mark Dredze. 2019. 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, Hong Kong, China. Association for Computational Linguistics.

Zhengxuan Wu, Isabel Papadimitriou, and Alex Tamkin. 2022. Oolong: Investigating What Makes Crosslingual Transfer Hard with Controlled Studies. ArXiv:2202.12312 [cs].

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In Advances in Neural Information Processing Systems, volume 32. Curran Associates, Inc.

Jianyu Zheng and Ying Liu. 2022. Probing language identity encoded in pre-trained multilingual models: a typological view. PeerJ Computer Science, 8:e899.

A Appendix

Paper	Task	Model	Lang. type	Factor
Lin et al. (2019)	DP, MT, POS, EL	/	NL	LO, LS, PTD
Pires et al. (2019)	POS, NER	mBERT	NL	LO, LS
Tran and Bisazza (2019)	DP	mBERT	NL	LO, LS
Wu and Dredze (2019)	DC, NER, DP, NLI, POS	mBERT	NL	LO
Artetxe et al. (2020)	NLI, DC, QA	Bilingual BERT, mBERT	NL	PTS
Conneau et al. (2020b)	NLI, NER, DP	Bilingual BERT	NL/SL	LO, MA, PTD
Dufter and Schütze (2020)	WA, WT, SR	BERT (small)	SL	LS, MA, PTD
Lauscher et al. (2020)	DP, POS, NER, NLI, QA	mBERT, XLM-R	NL	LS, PTD
Liu et al. (2020)	NLI	mBERT	NL	PTD, PTS
K et al. (2020)	NLI, NER	Bilingual BERT	NL/SL	LO, LS, MA, PTS
Dolicki and Spanakis (2021)	NLI, NER, POS	XLM-R	NL	LS
Srinivasan et al. (2021)	NLI, NER, POS	mBERT, XLM-R	NL	LO, LS, PTD
Wu et al. (2022)	SA, AJ, SS, NLI	English RoBERTa	SL	LS, MA
Ahuja et al. (2022)	DC, NLI, POS, NER, QA	mBERT, XLM-R	NL	LO, LS, PTD, PTS
de Vries et al. (2022)	POS	XLM-R Base	NL	LO, LS
Deshpande et al. (2022)	NLI, NER, POS, QA	Bilingual RoBERTa (small)	SL	LO, LS, MA, PTD
Eronen et al. (2022)	DC	mBERT, XLM-R	NL	LS
Patil et al. (2022)	NER, POS, DC, NLI	mBERT (12 languages)	NL/SL	LO

Table 2: List of studies investigating factors that impact cross-lingual transfer. The **Task** column indicates the downstream tasks that experiments have been performed on. We use the following abbreviation: **AJ**: Acceptability Judgement, **DC**: Document Classification, **DP**: Dependency Parsing, **EL**: Entity Linking, **LID**: Language Identification, **LS**: Language Similarity, **MTQE**: Machine Translation Quality Estimation, **NER**: Named Entity Recognition, **NLI**: Natural Language Inference, **POS**: Part-of-speech tagging, **QA**: Question Answering, **SA**: Sentiment Analysis, **SR**: Sentence Retrieval, **SS**: Sentence similarity, **WA**: Word Alignment, **WT**: Word Translation. The **Model** column indicates the models that were employed in the experiments of each paper. The **Factor** column indicates the factors for cross-lingual transfer ability that have been investigated. We use the following abbreviations: **LO**: Lexical Overlap, **LS**: Language Similarity, **MA**: Model Architecture, **PTS**: Pre-Training Settings, **PTD** : Pre-Training Data. The **Lang. type** column indicates the type of language that has been used. We use the following abbreviations. **NL**: Natural Languages, **SL**: Synthetic languages

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations section

☒ A2. Did you discuss any potential risks of your work?
During literature review no potential risks could be identified. Additionally, our review focuses on the potential benefits of different factors rather than on the risks.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Introduction (Section 1)

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
No response.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☒ Did you run computational experiments?

Left blank.

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No response.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
No response.

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
No response.

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
No response.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.