Cross-Lingual Ability of Multilingual Masked Language Models: A Study of Language Structure

Yuan Chai[1,*], **Yaobo Liang**[2], **Nan Duan**[2]
[1] Beihang University
[2] Microsoft Research Asia
[1]chaiyuan@buaa.edu.cn
[2]{yalia, nanduan}@microsoft.com

Abstract

Multilingual pre-trained language models, such as mBERT and XLM-R, have shown impressive cross-lingual ability. Surprisingly, both of them use multilingual masked language model (MLM) without any cross-lingual supervision or aligned data. Despite the encouraging results, we still lack a clear understanding of why cross-lingual ability could emerge from multilingual MLM. In our work, we argue that cross-language ability comes from the commonality between languages. Specifically, we study three language properties: constituent order, composition and word co-occurrence. First, we create an artificial language by modifying property in source language. Then we study the contribution of modified property through the change of cross-language transfer results on target language. We conduct experiments on six languages and two cross-lingual NLP tasks (textual entailment, sentence retrieval). Our main conclusion is that the contribution of constituent order and word co-occurrence is limited, while the composition is more crucial to the success of cross-linguistic transfer.

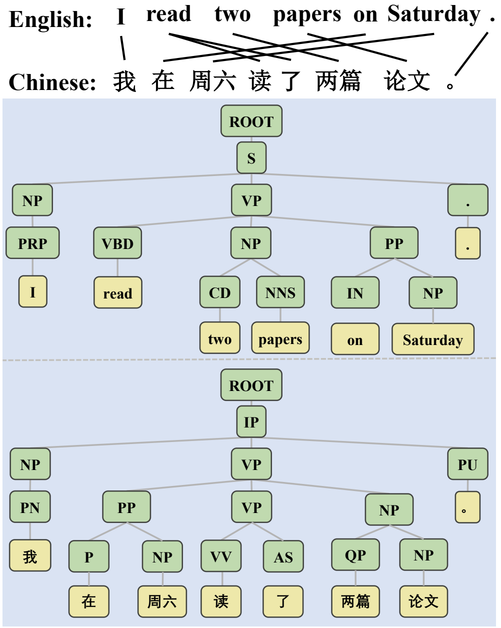

Figure 1: Example of two sentences in different languages, which are different in constituent order, but are very similar in constituent tree. Note that we simplify the constituent tree for better understanding.

1 Introduction

Zero-Shot Cross-Lingual transfer aims to build models for the target language by reusing knowledge learned from the source language. In this way, models can be efficiently implemented in multilingual as well as low-resource language scenarios. Traditionally, it is solved by a two-step pipeline (Ruder et al., 2019): a shared multilingual textual representation is first built and then supervised data from the source language is used on the top of it to train task-specific models. With the recent emergence of multilingual language models, the standard paradigm in this field has shifted to the pre-trained fine-tuning paradigm. Multilingual pre-trained language models, such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020a), have proven effective for cross-lingual transfer with better results on a large number of downstream tasks and languages (Pires et al., 2019; Conneau et al., 2020a).

The most surprising part is that mBERT and XLM-R are both trained without using any parallel corpus. Previous work (Pires et al., 2019; Wu and Dredze, 2019) attribute this success to the shared anchor words. But recent work (Conneau et al., 2020b; K et al., 2020) shows that cross-lingual transfer still could emerge even corpus of languages are from different domain, or don't share any common words. For cross-lingual model, sharing transformer encoder weight is critical, while whether having language specific word embedding or language identity marker is not important. This makes us more curious about what kind of common prop-

[*] Work done during internship at Microsoft Research Asia.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 4702 - 4712
May 22-27, 2022 ©2022 Association for Computational Linguistics

erty of languages could make cross-lingual transfer successful.

In our work, we study three language structure properties: 1) Constituent order. Specifically, we study three common constituent orders: order of verb and object, adposition and noun phrase, adjective and noun. 2) Composition. Composition means that we could combine two or several simple meanings and build a new more complicated meaning. For example, two words could form a phrase, and more composition could form a sentence recursively. 3) Word co-occurrence. We take the bag of words assumption and study the word co-occurrence in a sentence.

We use Figure 1 to better show the composition similarity between two sentences. Although the English sentence and Chinese sentence have different word order, but they are both first divided into a noun phrase and a verb phrase, and the verb phrase is then divided into three parts that are identical in meaning but inconsistent in order.

To better analyze the contribution of these three properties, we use the control variable method. Based on a successful transfer between the source and target languages, we change or remove only one structure property in the source language. We measure the importance of this property by testing the change in performance of the cross-language transfer from the modified source language to the target language. The results show that the effect of constituent order and word co-occurrence is small, while composition has a greater effect.

The main contributions are summarized as follows: (1) We analyze the source of cross-linguistic ability from shared properties in language structure and propose three candidate answers. (2) We used the control variable method, modifying only the target property in the corpus and keeping the other settings identical, thus better quantifying the contribution of the studied properties. (3) Our experiments clearly show that constituent order and word co-occurrence make very limited contributions to cross-lingual ability, while composition is the key to cross-lingual transfer.

2 Study Design

In this section, we introduce the design of our study, including the three language structure properties, and the overall setup. We also detail the pre-training and fine-tuning settings for better reproduction.

2.1 Dissecting Language Structure

Constituent Order In constituent tree, the constituents in a grammar rule are often ordered. For example in English, "`S->NP VP`" means that we should put the noun phrase at beginning of sentence and put the verb phrase after it. There are many linguistic studies to summarize and compare the constituent order of different languages, such as WALS (Dryer and Haspelmath, 2013). We mainly study three WALS features, 83A (Order of Object and Verb), 85A (Order of Adposition and Noun), and 87A (Order of Adjective and Noun).

Composition Composition means to combine two or several meanings and build a new more complicated meaning. As shown in Figure 1, "two" and "papers" could form a new meaning "two papers". And we could further combine it with "read" to form "read two papers". To better dissect the language structure, the composition in our study doesn't have order. By recursively combining meanings, we could express infinite meanings with finite words. The combination process forms an unordered tree.

Word Co-Occurrence In our paper, we study the word co-occurrence at the sentence level. Some words often co-occur in a context window or a sentence. As shown in Figure 1, the word co-occurrence of sentences with same meaning may also be similar in different languages, which may be a source of cross-lingual ability.

The natural language sentences of most languages are composed of a list of ordered words. But different languages may have different word order. We argue that most research about word order, for example research in WALS, are studying constituent order. The term "word order" hypothesis that any word could have any neighboring word. But "constituent order" hypothesis that some words should always group together and form a constituent and the order between groups are the object of study. The words from two constituents are unlikely to be neighboring words. Based on this, we dissect the "word order" into two concepts "constituent order" and "composition". "composition" is the rules to group words to phrase, clause and sentence. If we remove all the word order information, we will only have a set of unordered words and we name this feature as word co-occurrence. Bag-of-Word assumption only takes word co-occurrence information and has achieved great success in topic modeling and word embedding. We also hope to study its influence to cross-lingual ability of MLM.

2.2 Overall Setup

Bilingual Pre-training Following previous studies (Conneau et al., 2020b; K et al., 2020), our experiments were done on the corpus of only two languages, source (English) and target (multiple languages). By involving only one pair of languages, we can ensure that the performance of a given target language is only affected by the source language, without worrying about interference from a third language. In our work, we select English as source language because it has best constituent parser and most of cross-lingual benchmarks only have English training data.

Only Modify Source Language We believe the source of cross-lingual ability is the commonality between languages, and it can be destroyed by modifying either language in the pair. We decide to only modify source language and leave target language unmodified. This makes results on target language comparable to each other and ensures that changes in the results only come from changes of language property rather than modifications in the target language. By keeping the other settings the same and modifying only the source language, we exclude the interference of extraneous factors. This setup follows the control variable method and allows a more precise quantification.

Consistent in Pre-training and Evaluation We study the different commonality by creating a new language. So we make modifications both in pre-training and downstream evaluation. This consistency could help to generalize our conclusions to new languages beyond the 100 human natural languages. For example, the new languages could be other modalities like image, audio and video. Or programming languages like Python, Java and Lisp. We may meet extra-terrestrial someday and could access their unlabeled textual corpus. We still hope cross-lingual research could help us to understand their languages.

2.3 Multilingual Masked Language Model

Our multilingual masked language model pre-training follows the standard setup such as mBERT, XLM-R. Specifically, we mask 15% of the input tokens, of which 80% are replaced with mask tokens, 10% keep the original words, and 10% are randomly replaced with sampled words from the multilingual vocabulary. The training objective is to recover the masked tokens. We use the entire Wikipedia for each language as pre-training data and the model parameters are shared across languages. Unlike standard multilingual pre-trained models, the vocabulary in our experiments is not shared across languages. To remove confounding factors, our vocabulary is learned individually on each language using BPE (Sennrich et al., 2016), as (Conneau et al., 2020b; K et al., 2020) have demonstrated that sharing vocabulary has a very limited effect on cross-lingual transfer. Note the softmax prediction layer shared across languages is still preserved.

Implementation Details We use base size model in each experiment, which is a Transformer (Vaswani et al., 2017) with 12 layers, 12 heads, and GELU activation functions. The vocabulary size is 32k for each language, the embedding dimension is 768, the hidden dimension of the feed-forward layer is 3072, and the dropout rate is 0.1. We use the Adam optimizer and the polynomial decay learning rate scheduler with 3×10^{-4} learning rate and 10k linear warm-up steps during training. We train each model with 8 NVIDIA 32GB V100 GPUs and use total batch size 2048 with gradient accumulation strategy. We stop pre-training at 160k steps and evaluate the pre-trained model on downstream tasks every 8k steps and report the best result.

2.4 Downstream Cross-lingual Evaluation

We consider Cross-lingual Natural Language Inference (XNLI) dataset (Conneau et al., 2018) and Tatoeba dataset (Artetxe and Schwenk, 2019) in XTREME benchmark (Hu et al., 2020) to evaluate performance.

XNLI is a standard cross-lingual textual entailment dataset, which asks whether a premise sentence entails, contradicts, or is neutral toward a hypothesis sentence in the same language. We use the **zero-shot cross-lingual transfer** setting, where we first fine-tune the pre-trained model with source (English) language and then directly test the model with target language. XNLI is a three-category classification task which uses accuracy as its metric. The three categories in the test set are uniformly distributed, so the score of random guesses is 33.33%.

Tatoeba is a cross-lingual sentence retrieval dataset which consists of up to 1,000 English-aligned sentence pairs covering 122 languages. Tatoeba uses the source to target Top-1 accuracy as its metric. Note that Tatoeba only has test set so we use the pre-trained model directly without fine-tuning.

Evaluation Details For XNLI, the task-specific layer is a two layer linear mapping with tanh function between them, which takes the `[cls]` token as input. We use the Adam optimizer and linear decay learning rate scheduler with 7×10^{-6} learning rate and 12.5k linear warmup steps during fine-tuning. We fine-tune each model with batch size 32 for 10 epochs and evaluate on the English dev set every 3k steps to select the best model. We report the result on average of four random seed. For Tatoeba, we use the average pooling subword representation (excluding special token) of sentences at the 8-th layer as sentence representations following XTREME settings (Hu et al., 2020). Evaluation is done by finding the nearest neighbor for each sentence in the other language according to cosine similarity.

3 Constituent Order

Previous work(Pires et al., 2019) has argued that the cross-lingual transfer performance between languages with same constituent order is 10%-20% better than languages with different constituent order. So we further conduct control variate experiments to study the influence of constituent order. First we introduce the constituent order we studied and experiment setup. Then we analyze the effects of constituent order through the results. Our main conclusion is that the contribution of constituent order is about 1%.

3.1 Constituent Order Modification

Following (Naseem et al., 2012; Pires et al., 2019), we use a subset of order-related features from WALS to study constituent order. Specifically, we examine:

- **Order of Object and Verb.** Corresponding to 83A in WALS and grammar "`VP->VB NP`" in the constituent tree. Two orders are defined in WALS, OV for Object-Verb order and VO for Verb-Object order. English is an OV language and we change it to VO by changing the grammar to "`VP->NP VB`". Note that we consider all tags starting with `VB` (`VBZ`, `VBD`) as `VB`.
- **Order of Adposition and Noun Phrase.** Corresponding to 85A in WALS and grammar "`PP->IN NP`" in the constituent tree. Two orders are defined in WALS, Prepositions (Pre) for Preposition-Noun Phrase order and Postpositions (Post) for Noun Phrase-Postposition order. English is a Prepositions language and we change it to Postpositions by changing the grammar to "`PP->NP IN`".
- **Order of Adjective and Noun.** Corresponding to 87A in WALS and grammar "`NP->JJ NN`" in the constituent tree. Two orders are defined in WALS, AN for Adjective-Noun order and NA for Noun-Adjective order. English is a AN language and we change it to NA by changing the grammar to "`NP->NN JJ`".

Specifically, we use the Constituency Parsing tool in Stanford's CoreNLP (Manning et al., 2014) to obtain the constituent trees. For each order-related feature, we filter out the parent node and children nodes satisfy the feature's grammar. For example, the grammar for order of object and verb is "`VP->VB NP`". We filter out the parent node, whose constituent label is VP, with and only with two children nodes, whose constituent labels are VB and NP respectively. Then we will change the order of the two children nodes. After we recursively check and modify all tree nodes, we in-order traverse the tree and get the sentence with constituent order modified. In Figure 2, we show examples of modifying constituent order.

We select Spanish, Russian, Hindi, Turkish, Thai, and Vietnamese as target languages, considering the variance of script, typological features and pre-training resources.

Unlike the analysis of correlations between constituent order and results of target languages (Pires et al., 2019), we follow the principle of control variables and modify the constituent order directly in the corpus. In this way, we can ensure that the differences in results come from constituent order modifications only.

3.2 Effect of Constituent Order

Table 1 and Table 2 show the results on XNLI and Tatoeba. With the results of modifying three features and test on six different target languages, we can draw the following three conclusions:

Modifying constituent order barely affects source language. As shown in 1, modifying constituent order in the source language barely change

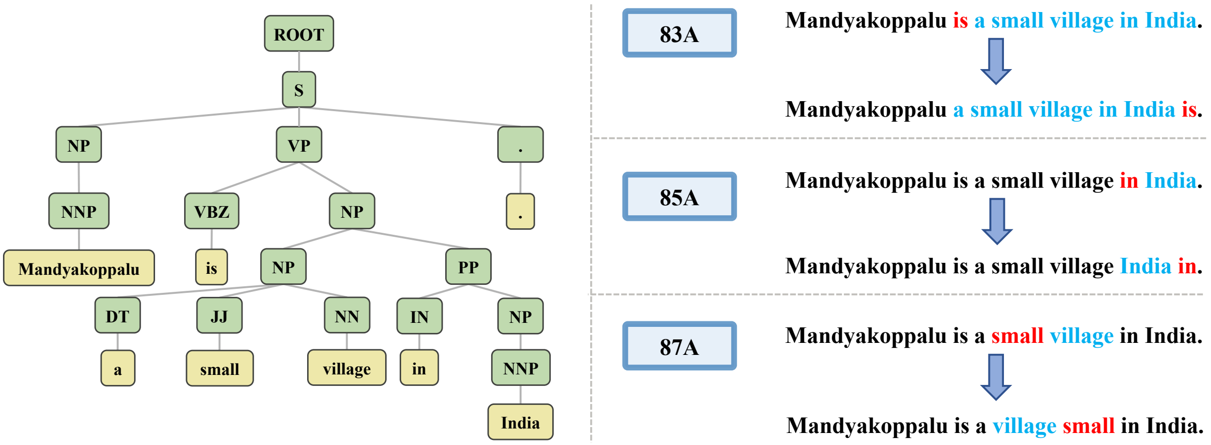
ROOT
S
NP
VP
.
NNP
VBZ
NP
.
Mandyakoppalu
is
NP
PP
DT
JJ
NN
IN
NP
a
small
village
in
NNP
India
83A
Mandyakoppalu is a small village in India.
Mandyakoppalu a small village in India is.
85A
Mandyakoppalu is a small village in India.
Mandyakoppalu is a small village India in.
87A
Mandyakoppalu is a small village in India.
Mandyakoppalu is a village small in India.

Figure 2: Example of constituent tree and constituent order modification based on WALS features. 83A, 85A and 87A represent order of object and verb, order of adposition and noun phrase and order of adjective and noun.

source	ru (VO, Pre, AN)		hi (OV, Post, AN)		tr (OV, Post, AN)	
	source	target	source	target	source	target
en	83.93	73.90	83.83	69.46	83.92	72.48
en-OV	83.81	74.03	83.24	69.66	83.81	73.27
en-Post	83.56	74.09	83.68	70.20	83.93	72.84
en-NA	83.64	73.85	83.54	69.56	84.07	72.88
source	es (VO, Pre, NA)		th (VO, Pre, NA)		vi (VO, Pre, NA)	
	source	target	source	target	source	target
en	83.97	78.11	83.29	70.21	83.87	74.55
en-OV	83.81	78.76	83.20	70.25	82.76	74.78
en-Post	83.55	78.22	83.80	70.31	83.64	74.34
en-NA	84.01	78.18	84.01	71.34	83.94	74.86

Table 1: Comparison of XNLI results before and after modification of constituent order. "en" represents the unmodified English, en-OV represents the modification of English from Verb-Object order to Object-Verb, and the same for the others. We marked its original constituent order after each target language.

source	ru	hi	tr
en	78.1	60.6	58.9
en-OV	73.6	66.9	64.5
en-Post	70.1	65.0	60.4
en-NA	77.7	59.8	59.9
source	es	th	vi
en	82.2	59.9	76.0
en-OV	78.1	54.6	68.7
en-Post	80.4	51.8	72.7
en-NA	76.8	52.6	78.5

Table 2: Comparison of Tatoeba results before and after modification of constituent order.

its XNLI results (basically 0.3%). This means that our modifications do not affect the overall meaning of the language. The modified language is still a reasonable language for both humans and models.

Changing source language's constituent order to same as target language could improve cross-lingual transfer. In Tables 1 and 2, we find that modifying constituent order achieves consistent gains on most low-resource languages. For example, modifying 83A in Turkish achieves gains 0.79% on XNLI and 5.6% on Tatoeba , 85A in Hindi gains 0.74% on XNLI and 4.4% on Tatoeba , 87A in Vietnamese gains 0.31% on XNLI and 2.5% on Tatoeba . However, this pattern is not very stable in high-resource languages. For example, 87A in Spanish gains only 0.08% on XNLI and decreases by 5.4% on Tatoeba instead.

The overall effect of modifying constituent order to cross-lingual transfer is limited. No matter what modification is made, the results on six different target languages showed very limited changes (basically within 1% on XNLI and 8% on Tatoeba). This further suggests that constituent order has limited effect on cross-lingual transfer. In other words, constituent order is not the key component of language structure. As for the magnitude of the variation, it is slightly higher on Tatoeba than on XNLI. We believe there are two main reasons. First, the average sentence length of Tatoeba is lower than XNLI, so the effect of modification will be magnified. Second, Tateoba doesn't have training data and the zero-shot evaluations are highly unstable. For example, (Phang et al., 2020) achieved more than 20% gains by fine-tuning the model on XNLI at first.

3.3 Comparison to Previous Work

The "conflict" conclusion between our work and Pires et al. 2019 is because of the difference of experiment design. In the experiment about object and verb order, Pires et al. 2019 change the source language to totally different language, and test on target languages. For example train on English (VO) or Hindi (OV), and test on French (VO). And they found the transfer from VO to VO is much better than transfer from OV to VO. While our experiments use modified English. We argue that the verb and object order isn't the only difference of source language in their experiment. For example, most of Europe languages are VO and most of Central Asia languages are OV. Languages in the same region are more similar than languages in different region. Our work conduct control variate experiments and could analysis the importance of constituent order better.

4 Composition: The Key to Zero-Shot Cross-Lingual Transfer

In this section, we study the contribution of constituent order, composition and word co-occurrence respectively. We first present how to completely remove constituent order and composition step-by-step from the corpus, and then analyze the results. Subsequently, by controlling the rate of composition retention, we further quantified its contribution to cross-lingual transfer.

4.1 Language Ablation of Removing Constituent Order and Composition

First, we introduce several experiments settings:

Constituent Shuffle: Removing Constituent Order. When removing the constituent order, we should be careful to keep the composition untouched. As shown in Figure 3, we shuffle the children nodes of same intermediate node in the constituent tree, while preserving the parent-children relation between nodes unchanged. By comparing its results with the baseline, we can quantify the contribution of constituent order.

Word Shuffle: Removing Constituent Order and Composition. To further remove the composition, we randomly shuffle the words in sentence. This "Word Shuffle" operation will remove constituent order and composition together. By comparing it with the results of "constituent shuffle", we can quantify the contribution of composition.

Baselines Without Pre-training: Removing Constituent Order, Composition and Word Co-occurence. We also provide a "Without Pre-training" baseline in XNLI and "Word Embedding Average" baseline in Tatoeba to quantify the contribution of word co-occurrence by comparing with "Word Shuffle". On XNLI, "Without Pre-training" represents a Transformer model with same structure as pre-trained model but with random initialized weights. Then we fine-tune it with source language and test on target languages. Because Tatoeba doesn't have any training data, we use the average of word embedding as a baseline. The word embedding is extracted from the embedding layer of "word shuffle" setting. The performance of word embedding average baseline still credits to word-occurrence but not to pre-training.

Second, to quantify the modification degree, we define two metrics. Inversion Ratio is the number of inverse pairs in the modified sentence, which normalized by the number of total word pairs in the sentence. Word Move Distance is the average distance of each word moved in sentence, which normalized by length of each sentence. As shown in Table 3, the sentences after constituent shuffle and word shuffle are almost identical in two metrics and both much higher than sentences modifying local constituent order. This shows that constituent shuffle also makes lots of word order modifications and has high randomness.

source type	IR (%)	WMD (%)
en-OV	3.04	2.71
en-Post	4.26	3.66
en-AN	0.49	0.35
Constituent Shuffle	50.44	31.1
Word Shuffle	51.89	33.54

Table 3: Inversion Ratio (IR) and Word Move Distance (WMD) of the modifications used in the experiments.

4.2 Contribution of Each Part

We present the results in Table 4 and Table 5. we can draw the following three conclusions:

On monolingual pre-training and fine-tuning,

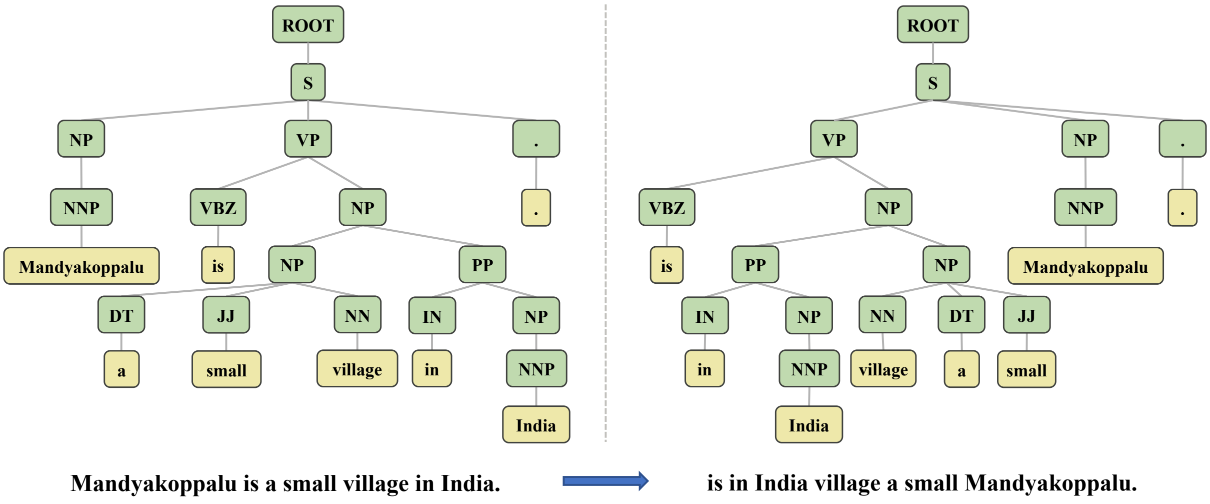
ROOT
S
NP
VP
.
NNP
VBZ
NP
.
Mandyakoppalu
is
NP
PP
DT
JJ
NN
IN
NP
a
small
village
in
NNP
India
Mandyakoppalu is a small village in India.
is in India village a small Mandyakoppalu.

Figure 3: Example of constituent shuffle by disordering children of each intermediate node in constituent tree.

source type	es		ru		hi	
	source	target	source	target	source	target
Original Language	83.97	78.11	83.93	73.90	83.83	69.46
Constituent Shuffle	81.88	76.42	81.40	71.25	81.97	70.18
Word Shuffle	70.45	39.84	71.34	35.14	70.37	39.16
Without Pre-training	56.67	34.27	56.43	35.20	56.71	34.44
Random Guess	33.33	33.33	33.33	33.33	33.33	33.33

Table 4: Comparison of XNLI results before and after removing constituent order and composition.

source type	es	ru	hi
Original Language	82.2	78.1	60.6
Constituent Shuffle	68.3	49.4	55.2
Word Shuffle	37.0	10.6	15.7
Word Embedding Average	32.3	8.5	13.5
Random Guess	0.1	0.1	0.1

Table 5: Comparison of Tatoeba results before and after removing constituent order and composition.

pre-training without composition still achieve good results. We can observe that whether removing constituent order or removing composition in the source language, it still shows meaningful results (much higher than random guess) on XNLI. This illustrates that textual entailment of monolingual languages can have good performance relying only on word co-occurrence.

Cross-lingual transfer works with composition and doesn't work without composition. When the constituent order is removed, only a limited performance loss (within 3%) is shown on both the source and target languages, and it is almost constant on the performance gap between source and target languages. This shows again that the contribution of constituent order to cross-language transfer is very limited and it is not a critical component of the language structure. However, when composition is removed, the cross-lingual transfer results on target language are only slightly higher than random guess. This clearly shows that composition is the key to cross-lingual transfer. As for word co-occurrence, it only contributes 5% on XNLI and 10% to 15% on Tatoeba. These results show that it does make some contribution but the contribution is very limited. Relying on word co-occurrence alone is not enough for a reasonable cross-lingual performance.

Removing constituent order and keeping composition may improve cross-lingual transfer. We observe an interesting result in Table 4. There is about 2% drop on both Spanish and Russian after removing constituent order. However, the results show a 0.7% improvement on Hindi while English dropped 2%. We think this is because that model relies on every possible feature to solve English task but only relies on the commonality between language to achieve cross-lingual transfer. The model will use constituent order and composition feature to solve XNLI in unmodified English but only could use composition feature in constituent shuffled English. For languages with similar constituent order to English, more language features may lead to better performance. But for languages

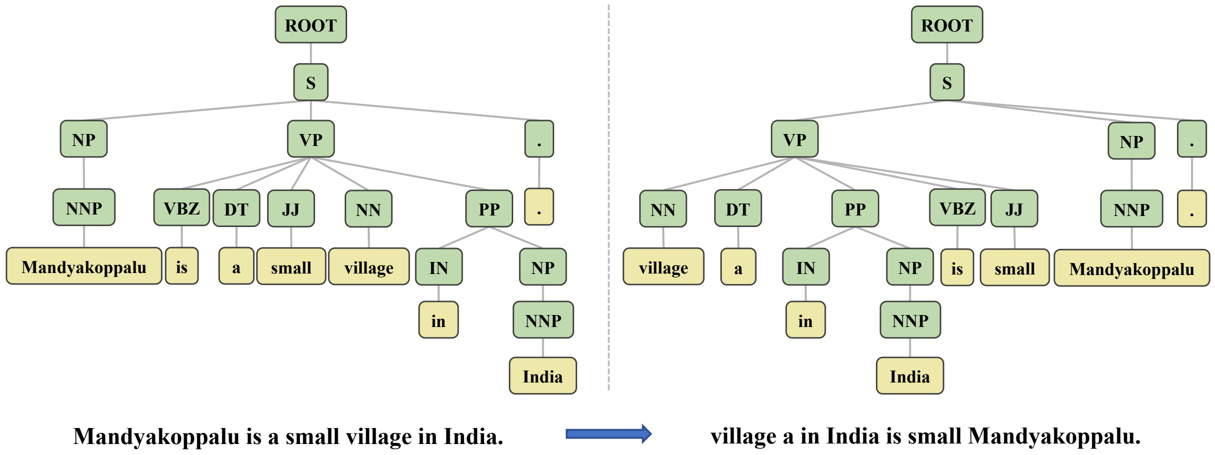
ROOT
S
NP
VP
.
NNP
VBZ
DT
JJ
NN
PP
Mandyakoppalu
is
a
small
village
IN
NP
in
NNP
India
Mandyakoppalu is a small village in India.
village a in India is small Mandyakoppalu.

Figure 4: Example of removing intermediate nodes and disordering on the constituent tree.

with different constituent order to English, only rely on composition will lead to better generalization ability. This further shows that constituent order is not key to cross-lingual transfer, and composition is the most important commonality between all languages.

4.3 Detailed Analyze Composition

To further quantify the effect of composition, we remove it in different degrees. As shown in Figure 4, we randomly remove the ratio α of intermediate nodes in the constituent tree. For each removed node, all its children are connected to its father. Note that an intermediate node is defined as a non-root node with more than one children. We show Spanish results only due to space limitation.

α		0	0.25	0.5	0.75	1
XNLI	source	82.36	81.44	81.09	80.66	70.45
	target	76.61	76.19	75.22	74.17	39.84
Tatoeba		83.1	69.8	62.9	54.8	47.7

Table 6: Comparison of XNLI and Tatoeba results on Spanish at different ratio in removing composition similarity.

In Table 6, We observe that when we remove 75% of the composition, the results on XNLI are still higher than when we completely remove it. While on Tatoeba, there is a significant decrease in the results as more compostion is removed. We argue this is due to the difference in sentence length, which is much higher in XNLI than Tatoeba. Even with 75% removed, the absolute value of retained composition is still much higher in XNLI. This result shows that only a certain ratio composition is required for reasonable performance, which shows again that composition is crucial for cross-linguistic transfer.

5 Related Work

Multilingual Pre-Training mBERT and XLM-R train multilingual MLM without using any parallel corpus and show strong cross-lingual ability. mBERT is an extension of BERT which is pre-trained on Wikipedia data over 100 languages to learn a language-invariant feature space shared across multiple languages. XLM-R (Conneau et al., 2020a) is trained on 2.5T data over 100 languages extracted from Common Crawl (Wenzek et al., 2020), which demonstrates the effect of the model trained on a large-scale corpus. Results of XLM-R on a large number of downstream cross-lingual tasks show that a large-scale training corpus can significantly improve the performance of multilingual models.

Other methods use parallel corpus in multilingual pre-training. XLM (Conneau and Lample, 2019) introduces a Translation Language Model (TLM) based on the parallel corpus which shows significant improvement on downstream tasks. Unicoder (Huang et al., 2019) introduces a multitask learning framework to learn cross-lingual representations with monolingual and parallel corpora, achieving further gains. ALM (Yang et al., 2020) allows the model to learn cross-lingual code-switch sentences, enhancing the transfer ability. Recent studies INFOXLM (Chi et al., 2021), HICTL (Wei et al., 2021), VECO (Luo et al., 2021) and ERNIE-M (Ouyang et al., 2021) use contrastive learning, back translation and other tricks further enhancing the performance of the multilingual model. Our work focus on studying the model only with multilingual MLM, and leave the study of works with parallel data as future work.

Probing Multilingual MLM mBERT and XLM-R have successfully achieved excellent cross-lingual transfer performance without using any parallel corpus. Researchers have wondered what the source of this cross-lingual ability is. (Pires et al., 2019) examines the zero-shot cross-lingual transfer performance on NER (Pan et al., 2017) and part-of-speech (POS) tagging. They believe this success comes from the shared anchor words between languages. Not coincidentally, similar conclusion is reached by (Wu and Dredze, 2019). However, this conclusion is proven inaccurate by (Conneau et al., 2020b; K et al., 2020). Their experiments show that the model still learns cross-lingual transfer ability on the corpus without anchor words at all. Besides, (Conneau et al., 2020b; K et al., 2020; Artetxe et al., 2020; Libovický et al., 2020; Muller et al., 2021) have analyzed the cross-lingual ability of multilingual masked language models in terms of language similarity, shared model parameters cross languages, model structure, training objectives, language marker. The results suggest that structure similarity and shared parameters between languages are crucial for cross-lingual transfer. In this paper, we focus on analyzing language structure. We decompose it into constituent order, composition, and word co-occurrence and study the effect of each part separately.

Word Order in Machine Translation and Masked Language Model Finding the appropriate word ordering in target languages significantly influences the machine translation quality for statistical machine translation (Tillmann, 2004; Chiang, 2007), neural machine translation (Kawara et al.; Zhao et al., 2018) and non-autoregressive neural machine translation (Ran et al., 2019). This is because the input and output sentence for machine translation have different order and the evaluation metrics also consider the output word order. While our study is different because we only focus on classification tasks and their outputs don't need to consider word order.

Ji et al. 2021 shows that adapting word order could get about 1% gain. While our composition reordering also could get about 1% gain. But all these gains still show that constituent order is not important for cross-lingual transfer because that removing composition will lead to more than 30% difference. Sinha et al. 2021 shows that word order is not important for English monolingual pre-training. After removing composition, our experiments also show that the performance on source languages won't drop a lot. But the performance on target language will drop more than 30%. This proves that composition is not the key for English monolingual pre-training but the key for cross-lingual transfer.

6 Conclusion

In this paper, we study the source of cross-lingual ability in the multilingual masked language model in the view of language structure. We study three language structure properties: constituent order, composition and word co-occurrence. The experiments are conducted using control variable method. we create an artificial language by modifying property in source language. We quantify the contribution of these three properties separately through cross-language transfer performance changes from the modified language to the target language. The results show that the contribution of constituent order and word co-occurrence are very limited, while composition is actually the key to cross-lingual transfer. How to use this finding to enhance pre-trained multilingual language models and improve performance on cross-lingual NLP tasks will be our focus for future work.

References

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.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

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 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7057–7067.

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.

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.

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

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, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, 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.

Tao Ji, Yong Jiang, Tao Wang, Zhongqiang Huang, Fei Huang, Yuanbin Wu, and Xiaoling Wang. 2021. Word reordering for zero-shot cross-lingual structured prediction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4109–4120.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual BERT: an empirical study. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yuki Kawara, Chenhui Chu, and Yuki Arase. Recursive neural network based preordering for english-to-japanese machine translation. *ACL 2018*, page 21.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674, Online. Association for Computational Linguistics.

Fuli Luo, Wei Wang, Jiahao Liu, Yijia Liu, Bin Bi, Songfang Huang, Fei Huang, and Luo Si. 2021. VECO: Variable and flexible cross-lingual pre-training for language understanding and generation. 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 3980–3994, Online. Association for Computational Linguistics.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. 2021. First align, then predict: Understanding the cross-lingual ability of multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2214–2231, Online. Association for Computational Linguistics.

Tahira Naseem, Regina Barzilay, and Amir Globerson. 2012. Selective sharing for multilingual dependency parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 629–637, Jeju Island, Korea. Association for Computational Linguistics.

Xuan Ouyang, Shuohuan Wang, Chao Pang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021.

ERNIE-M: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 27–38, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. 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, Vancouver, Canada. Association for Computational Linguistics.

Jason Phang, Iacer Calixto, Phu Mon Htut, Yada Pruksachatkun, Haokun Liu, Clara Vania, Katharina Kann, and Samuel Bowman. 2020. English intermediate-task training improves zero-shot cross-lingual transfer too. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 557–575.

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.

Qiu Ran, Yankai Lin, Peng Li, and Jie Zhou. 2019. Guiding non-autoregressive neural machine translation decoding with reordering information. *arXiv preprint arXiv:1911.02215*.

Sebastian Ruder, Ivan Vulic, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *J. Artif. Intell. Res.*, 65:569–631.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. 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)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2888–2913.

Christoph Tillmann. 2004. A unigram orientation model for statistical machine translation. In *Proceedings of HLT-NAACL 2004: Short Papers*, pages 101–104.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Xiangpeng Wei, Rongxiang Weng, Yue Hu, Luxi Xing, Heng Yu, and Weihua Luo. 2021. On learning universal representations across languages. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

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.

Jian Yang, Shuming Ma, Dongdong Zhang, Shuangzhi Wu, Zhoujun Li, and Ming Zhou. 2020. Alternating language modeling for cross-lingual pre-training. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9386–9393. AAAI Press.

Yang Zhao, Jiajun Zhang, and Chengqing Zong. 2018. Exploiting pre-ordering for neural machine translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).*