Incorporating Lexical and Syntactic Knowledge for Unsupervised Cross-Lingual Transfer

Jianyu Zheng[*♠]**, Fengfei Fan**[*◇]**, Jianquan Li**[†◇]
♠Tsinghua University ◇Beijing Ultrapower Software Co.,Ltd.
zheng_jianyu@126.com {fanfengfei, lijianquan2}ultrapower.com.cn

Abstract

Unsupervised cross-lingual transfer involves transferring knowledge between languages without explicit supervision. Although numerous studies have been conducted to improve performance in such tasks by focusing on cross-lingual knowledge, particularly lexical and syntactic knowledge, current approaches are limited as they only incorporate syntactic or lexical information. Since each type of information offers unique advantages and no previous attempts have combined both, we attempt to explore the potential of this approach. In this paper, we present a novel framework called "Lexicon-Syntax Enhanced Multilingual BERT" that combines both lexical and syntactic knowledge. Specifically, we use Multilingual BERT (mBERT) as the base model and employ two techniques to enhance its learning capabilities. The *code-switching* technique is used to implicitly teach the model lexical alignment information, while a *syntactic-based graph attention network* is designed to help the model encode syntactic structure. To integrate both types of knowledge, we input code-switched sequences into both the syntactic module and the mBERT base model simultaneously. Our extensive experimental results demonstrate this framework can consistently outperform all baselines of zero-shot cross-lingual transfer, with the gains of 1.0∼3.7 points on text classification, named entity recognition (ner), and semantic parsing tasks.

Keywords: cross-lingual transfer, lexicon, syntax, code-switching, graph attention network

1. Introduction

Unsupervised cross-lingual transfer refers to the process of leveraging knowledge from one language, and applying it to another language without explicit supervision (Conneau et al., 2019). Due to the free requirement of the labeled data in target language, it is highly preferred for low-resource scenarios. Recently, unsupervised cross-lingual transfer has been widely applied in various natural language processing (NLP) tasks, such as part-of-speech (POS) tagging (Kim et al., 2017; de Vries et al., 2022), named entity recognition (NER) (Fetahu et al., 2022; Xie et al., 2018), machine reading comprehension (Hsu et al., 2019; Chen et al., 2022), and question answering (QA) (Nooralahzadeh and Sennrich, 2023; Asai et al., 2021).

The success of unsupervised cross-lingual transfer can be attributed to its ability to exploit connections across languages, which are reflected in various linguistic aspects such as lexicon, semantics, and syntactic structures. Consequently, many studies have sought to enhance models by encouraging them to learn these cross-lingual commonalities. For instance, in the lexical domain, Qin et al. (2021) utilize bilingual dictionaries to randomly replace certain words with their translations in other languages, thereby encouraging models to implicitly align representations between the source language and multiple target languages. In the area

of syntax, several works have developed novel neural architectures to guide models in encoding the structural features of languages. Ahmad et al. (2021), for example, proposes a graph neural network (GNN) to encode the structural representation of input text and fine-tune the GNN along with the multilingual BERT (mBERT) for downstream tasks. Both lexical and syntactic approaches facilitate the alignment of linguistic elements across different languages, thereby enhancing the performance of cross-lingual transfer tasks.

However, language is a highly intricate system (Ellis and Larsen-Freeman, 2009), with elements at various levels being interconnected. For example, sentences are composed of phrases, which in turn are composed of words. In cross-lingual transfer, we hypothesize that merely guiding models to focus on a single linguistic aspect is inadequate. Instead, by simultaneously directing models to learn linguistic knowledge across diverse levels, their performance can be further improved. Table 1 presents some example sentences extracted from the XNLI dataset (Conneau et al., 2018). These parallel sentence pairs demonstrate that the multilingual model makes incorrect predictions for sentence pairs in the target languages (French and German) when only one aspect of linguistic knowledge, such as lexical or syntactic knowledge, is incorporated. However, when both types of knowledge are integrated into the model, the correct prediction is obtained. Despite this, most previous studies have focused on either syntactic or lexical information alone, without considering the integra-

---

∗ Equal Contribution
† Jianquan Li is the corresponding author

8986

Lang	Premise(P)/Hypothesis(H)	Label	+Lex	+Syn	Ours
fr	**P**:Votre société charitable fournit non seulement de les services sociaux communautaires efficaces à les animaux et les personnes, mais sert également également de fourrière pour la Ville de Nashua. **H**:La société humaine est le refuge pour animaux de Nashua.	entali	contra	contra	entail
de	**P**:Ihre humane Gesellschaft erbringt nicht nur effektive gemeinschaftlich-soziale Dienstleistungen für Tiere und ihre Menschen, sondern dient auch als Zwinger der Stadt Nashua. **H**:Die Humane Society ist Nashuas Tierheim.	entail	contra	contra	entail
en	**P**:Your humane society provides not only effective community social services for animals and their people , but also serves as the pound for the City of Nashua . **H**:The humane society is Nashua's animal shelter .				

Table 1: The parallel sentence pairs in French and German from XNLI(Conneau et al., 2018), which are translated from English. Each sentence pair consist of a Premise sentence(P) and a Hypothesis sentence(H). The "Label" column indicates the relationship between each sentence pair, which can be contradiction(contra), entailment(entail) or neutral. "+Lex" and "+Syn" represent the prediction results from the multilingual models infused with lexical and syntactic knowledge, respectively. The "ours" column shows the results of integrating both types of knowledge into the model. Compared to the other two methods, our method can accurately predict the relationship between each sentence pair.

tion of both types of information.

In this work, we aim to enhance unsupervised cross-lingual transfer by integrating knowledge from different linguistic levels. To achieve this, we propose a framework called "Lexicon-Syntax Enhanced Multilingual BERT" ("LS-mBERT"), based on a pre-trained multilingual BERT model. Specifically, we first preprocess the input source language sequences to obtain each word's part-of-speech information and dependency relationships between words in each sentence. Then, we replace some words in the sentence with their translations from other languages while preserving the established dependency relationships. Furthermore, we employ a graph attention network(Veličković et al., 2017) to construct a syntactic module, the output of which is integrated into the attention heads of the multilingual BERT. This integration guides the entire model to focus on syntactic structural relationships. Finally, during the fine-tuning process, we simultaneously train the multilingual BERT and the syntactic module with the pre-processed text. As a result, our framework enables the multilingual BERT to not only implicitly learn knowledge related to lexical alignment but also encode knowledge about syntactic structure.

To validate the effectiveness of our framework, we conduct experiments on various tasks, including text classification, named entity recognition (ner), and semantic parsing. The experimental results show that our framework consistently outperforms all baseline models in zero-shot cross-lingual transfer across these tasks. For instance, our method achieves the improvement of 3.7 points for mTOP dataset. Our framework also demonstrates significant improvements in generalized cross-lingual transfer. Moreover, we examine the impact of important parameters, such as the replacement ratio of source words, and languages for replacement. To facilitate further research explorations, we release our code at https://github.com/Tian14267/LS_mBert.

2.   Related Work

Cross-lingual transfer is crucial in the field of natural language processing (NLP) as it enables models trained on one language to be applied to another. To enhance performance in transfer tasks, numerous studies focus on addressing the characteristics of various languages and their relationships.

2.1.   Incorporating Lexical Knowledge for Cross-lingual Transfer

A group of studies aims to incorporate lexical alignment knowledge into cross-lingual transfer research (Zhang et al., 2021a; Wang et al., 2022; Qin et al., 2021; Lai et al., 2021). For example, Zhang et al. (2021a) and Wang et al. (2022) employ bilingual dictionaries to establish word alignments and subsequently train cross-lingual models by leveraging explicit lexical associations between languages. Other methods (Qin et al., 2021; Lai et al., 2021) involve substituting a portion of words in a sentence with their equivalents from different languages, a technique commonly known as "code-switching." By increasing the diversity of input text, these approaches promote implicit alignments of language representations. However, this group of studies mainly offers insights into lexical translation across languages, while neglecting the learning of language-specific structural rules.

2.2.   Incorporating Syntactic Knowledge for Cross-lingual Transfer

Another research category focuses on integrating syntactic knowledge for cross-lingual transfer (Ahmad et al., 2021; Yu et al., 2021; Zhang et al., 2021b; He et al., 2019; Cignarella et al., 2020; Xu et al., 2022; Shi et al., 2022; Wang et al., 2021). Many studies in this group (Ahmad et al., 2021; Wang et al., 2021) develop graph neural networks

to encode syntactic structures, a category to which our work also belongs. Taking inspiration from Ahmad et al. (2021), we adopt a similar architecture, specifically using a graph attention network to encode syntactic knowledge. Other methods (Cignarella et al., 2020; Xu et al., 2022) extract sparse syntactic features from text and subsequently incorporate them into the overall model. Although these approaches consider the relationships between language elements, they frequently overlook the alignments across languages, which impedes the effective transfer of linguistic elements and rules between languages.

Consequently, we combine the strengths of these two categories of approaches. First, we replace the input sequence with translated words from other languages, which aids in guiding the entire model to acquire implicit alignment information. Then, we introduce an additional module to assist the model in encoding syntax.

3. Methodology

In this section, we provide a detailed introduction to our framework "LS-mBERT", as illustrated in Figure 1. Our objective is to enhance the cross-lingual transfer capabilities of multilingual BERT (mBERT) by incorporating both lexical and syntactic knowledge. Given an input sequence, we first pre-process it using a part-of-speech tagger and a universal parser(Section 3.1). This yields the part-of-speech tag for each word and dependency relationships among words in the sequence. To enable mBERT to implicitly encode word alignment information, we substitute some words with their translations from other languages using a code-switching technology (Section 3.2). Moreover, to guide mBERT in attending to syntactic relationships, we construct a graph attention network (GAT), introduced in Section 3.3. The output of the graph attention network is then used as input to the attention heads within BERT, effectively biasing attention information between words. Finally, to integrate both syntactic and lexical knowledge, we pass the code-switched text into both the GAT network and mBERT, which are trained simultaneously (Section 3.4).

3.1. Pre-processing Input Sequence

The initial step involves pre-processing the input data to obtain prior knowledge for subsequent training. As our framework incorporates syntactic knowledge, we opt for an off-the-shelf parser with high accuracy to process the input text. In this case, we employ the UDPipe toolkit(Straka and Straková, 2017) to parse the inputs sentences, and Stanza(Qi et al., 2020) to annotate the part-

of-speech information of each word. By utilizing both tools, given a sentence, we can obtain the dependency relationships between words and their part-of-speech information, which are then utilized to provide syntactic knowledge and enhance word representations, respectively.

3.2. Code-switching for Text (lexical knowledge)

As our objective is to improve unsupervised cross-lingual transfer, introducing explicit alignment signals would be inappropriate. Therefore, we employ an implicit strategy to guide the entire model to encode word alignment information. Inspired by the work of Qin et al. (2021), we opt for the code-switching strategy. Specifically, we first randomly select a proportion α of words within each source sentence. Then, for each selected word, we use a high-quality bilingual dictionary to substitute it with a corresponding translation from another target language. This method not only promotes the implicit alignment of representations across diverse languages within our model, but also enhances the model's robustness when processing input text.

3.3. Graph Attention Network (syntactic knowledge)

To guide mBERT in acquiring syntactic knowledge better, we construct an external syntactic module by referring to the method introduced by Ahmad et al. (2021). The overview of this module is displayed in Figure 2. Given that there are n tokens in the input sequence, we first represent each token by combining its embedding representation with part-of-speech (POS) information. The representation of the i-th token can be calculated: $x_i = c_i W_c + pos_i W_{pos}$, where c_i and pos_i represent the token representation and the part-of-speech representation of the i-th token, respectively; while W_c and W_{pos} denote the token parameter matrix and the part-of-speech parameter matrix. Then, the encoded sequence $s' = [x_1, x_2, \cdots, x_n]$ is passed into the subsequent syntactic module, which is designed with a graph attention network (GAT) (Veličković et al., 2017). The GAT module comprises a total of L layers, each with m attention heads. These attention heads play a crucial role in generating representations for individual tokens by attending to neighboring tokens in the graph. Each attention in GAT operates as follows: $O = Attention(T, T, V, M)$, wherein T denotes the query and key matrices, and V represents the value matrix. Besides, M signifies the mask matrix, determining whether a pair of words in the dependency tree can attend each other. Notably, the relationships between words in the attention matrix are modeled based on the distances between
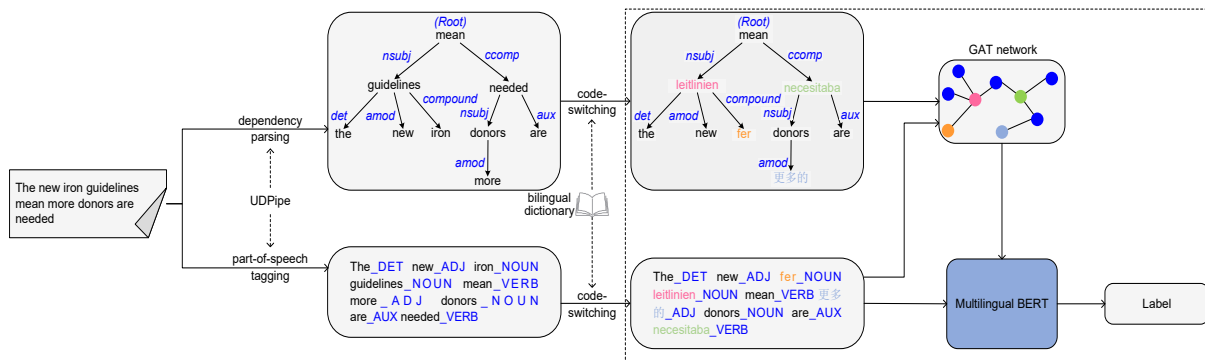
Figure 1: An overview of lexicon-syntax enhanced multilingual BERT ("LS-mBERT"). An example sentence is provided to explain how this framework works. To introduce lexical alignment knowledge, we utilize bilingual dictionaries to randomly replace some words in the sentence with the equivalent words from other languages (pink for German, green for Spanish, light blue for Chinese, and orange for French). Then, an graph attention network (GAT) is developed to encode the syntactic structure of this sentence. The output representation of GAT is sent to the attention heads in multilingual BERT for guiding them to focus on the language-specific structures.

words in the dependency tree, rather than the positional information within the word sequence. Subsequently, the resulting representations produced by all attention heads are concatenated to form the output representations for each token. Finally, the output sequence from the final layer can be denoted as $Y = [y_1, y_2, \cdots, y_n]$, where y_i represents the output representation for the i-th token. To maintain the lightweight nature of the architecture, certain elements in GAT have been excluded. Specifically, we do not employ feed-forward sublayers, residual connections, or positional representations. We found that these modifications do not result in a significant performance gap.

3.4. Summary of the Framework: Lexicon-syntax Enhanced Multilingual BERT

In this subsection, we provide an overview of our "LS-mBERT" framework, as illustrated in Figure 1. We first select multilingual BERT (mBERT) as the base model. Then, we process the input sequence using the code-switching strategy in Section 3.2, resulting in the code-switched sequence s'. It is important to note that despite some words in each sentence being replaced with other languages, the original dependency relationships between words are still preserved in s'. Next, we feed the code-switched text into both mBERT and the syntactic module (GAT), facilitating the fusion of the two types of knowledge. Furthermore, this step guides the entire model to better align different languages within the high-dimensional vector space during training. After GAT processes the code-switched sequence, the output from the final layer is utilized to bias the attention heads of mBERT. The

calculation process can be described as follows: $O = Attention(Q + YW_l^Q, K + YW_l^K, V)$, where Q, K, and V represent the query, key, and value matrices, respectively; While W_l^Q and W_l^K are new parameters to learn for biasing the query and key matrices.
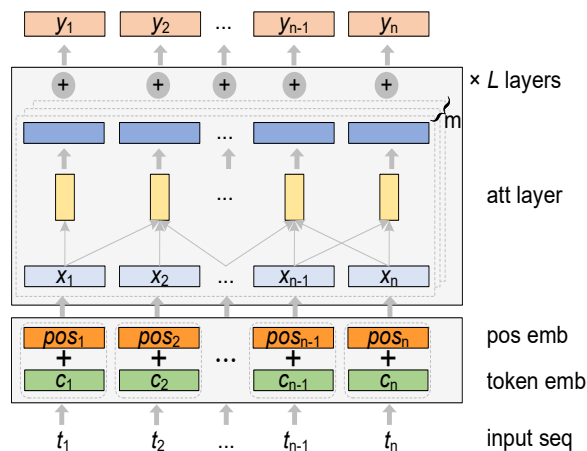

Figure 2: The architecture of graph attention network (Ahmad et al., 2021; Veličković et al., 2017). Each input token is represented by combining its token embedding and part-of-speech embedding. Each attention head within the graph attention network(GAT) generates a representation for each token embedding by attending to its neighboring tokens in the dependency graph. Next, the resulting representations are concatenated to form the output representation for each token. Finally, we can obtain the representations of the output sequence embeddings from the final layer of GAT.

4. Experiments

4.1. Experimental Settings

As above mentioned, we use UDPipe (Straka and Straková, 2017) and Stanza (Qi et al., 2020) for parsing sentences and obtaining words' part-of-speech information in all languages, and employ MUSE (Lample et al., 2018) as the bilingual dictionary for word substitution. For all tasks, we identify the optimal parameter combinations by searching within the candidate sets. The learning rate is set to 2e-5, utilizing $AdamW$ as the optimizer. The batch size is 64, and the maximum length for input sequences is 128 tokens. For code-switching, we vary the replacement ratio (α) from 0.3 to 0.7 with a step of 0.1. For the GAT network, we adopt the identical parameter values as employed in the work of Ahmad et al. (2021). Specifically, we set L to 4 and k to 4.

4.2. Tasks

Our framework is evaluated on the following tasks, using English as the source language. Some statistics are summarized in Table 2, along with the detailed descriptions provided below.

Text Classification. Text Classification is a task that assigns predefined categories to open-ended text. In our experiment, we utilize two publicly available dataset: XNLI and PAWS-X. In XNLI (Conneau et al., 2018), models need to predict whether a given pair of sentences is entailed, contradicted, or neutral; In PAWS-X (Yang et al., 2019), models are required to determine whether two given sentences or phrases convey the same meaning. When implementing the two tasks, to establish connections between the dependency trees of the two sentences, we introduce two edges from the $[CLS]$ token to the root nodes. Subsequently, we apply the code-switching technique to randomly replace certain words in the sentence pairs.

Named Entity Recognition. Named Entity Recognition (NER) is a task that involves the automatic identification and categorization of named entities. In our experiment, we employ the Wikiann (Pan et al., 2017) dataset. Wikiann consists of Wikipedia articles annotated with person, location, organization, and other tags in the IOB2 format. Our method is evaluated across 15 languages. To ensure that the models can obtain complete entity information, we exclusively substitute words that do not constitute named entities during the code-switching process.

Task-oriented Semantic Parsing. In this task, the models are required to determine the intent of the utterance and then fill the relevant slots. The dataset for the experiment is mTOP (Li et al., 2021), which is an almost parallel corpus, containing 100k examples in total across 6 languages. Our experiments cover 5 languages.

4.3. Baselines

We choose the following methods as baselines to compare:

- **mBERT.** We exclusively utilize the multilingual BERT model to perform zero-shot cross-lingual transfer for these tasks.

- **mBERT+Syn.** A graph attention network (GAT) is integrated with multilingual BERT, and these two components are jointly trained for all tasks.

- **mBERT+Code-switch.** The multilingual BERT model is fine-tuned with the code-switched text across various languages.

5. Results and analysis

5.1. Cross-Lingual Transfer Results

The main experimental results are displayed in Table 3. Our method consistently demonstrates superior performance across all tasks compared to other baselines. This indicates our method's effectiveness for cross-lingual transfer, achieved through the incorporation of lexical and syntactic knowledge. Especially for the tasks Wikiann and mTOP, our method exhibits a significant improvement, with an increase of 2.2 and 3.7 points, respectively, when compared to the baseline with the best performance. In addition, since code-switching technique blends words from various language, we calculate the results across the languages excluding English, as shown in the column "AVG/en" in Table 3. We find that the performance gap between our method and each baseline in most tasks becomes wider. This also indicates that our method can more effectively align non-English languages within the same vector space implicitly.

For each task, we discover most of languages can gain improvement by using our method, as compared to the top-performing baseline. Specifically, 84.6% (11/13), 100.0% (7/7), 80.0% (12/15) and 100.0% (5/5) languages demonstrate improvement in XNLI, PAWS-X, Wikiann and mTOP respectively. Furthermore, our method also provides improvement for non-alphabetic languages in many tasks, such as Chinese, Japan and Korean. This reflects that our method can be effectively generalized into various target languages, even in cases where significant differences exist between the source and target languages.

Task	Dataset	|Train|	|Dev|	|Test|	|Lang|	Metric
Classification	XNLI	392K	2.5K	5K	13	Accuracy
Classification	PAWS-X	49K	2K	2K	7	Accuracy
NER	Wikiann	20K	10K	1-10K	15	F1
Semantic Parsing	mTOP	15.7K	2.2K	2.8-4.4K	5	Exact Match

Table 2: Evaluation datasets. |Train|, |Dev| and |Test| delegate the numbers of examples in the training, validation and testing sets, respectively. |Lang| is the number of target languages we use in each task.

Tasks	Methods	en	ar	bg	de	el	es	fr	hi	ru	tr	ur	vi	zh	ko	nl	pt	ja	AVG / en	AVG
XNLI (Conneau et al., 2018)	mBERT	80.8	64.3	68.0	70.0	65.3	73.5	73.4	58.9	67.8	60.9	57.2	69.3	67.8	-	-	-	-	66.4	67.5
	mBERT+Syn	**81.6**	65.4	69.3	70.7	66.5	74.1	73.2	60.5	68.8	**62.4**	58.7	69.9	69.3	-	-	-	-	67.4	68.5
	mBERT+code-switch	80.9	64.2	70.0	71.5	67.1	73.7	73.2	61.6	68.9	58.6	57.8	69.9	70.0	-	-	-	-	67.2	68.3
	our method	81.3	**65.8**	**71.3**	**71.8**	**68.3**	**75.2**	**74.2**	**62.8**	**70.7**	61.1	58.8	**71.8**	**70.8**	-	-	-	-	**68.6**	**69.5**
PAWS-X (Yang et al., 2019)	mBERT	94.0	-	-	85.7	-	87.4	87.0	-	-	-	-	-	77.0	69.6	-	-	73.0	80.2	81.7
	mBERT+Syn	93.7	-	-	86.2	-	89.5	88.7	-	-	-	-	-	78.8	75.5	-	-	75.9	82.7	83.9
	mBERT+code-switch	92.4	-	-	85.9	-	87.9	88.3	-	-	-	-	-	80.2	78.0	-	-	78.0	83.4	84.3
	our method	**93.8**	-	-	**87.2**	-	**89.6**	**89.4**	-	-	-	-	-	**81.8**	**79.0**	-	-	**80.0**	**84.6**	**85.6**
Wikiann(Pan et al., 2017)	mBERT	83.7	36.1	76.0	75.2	68.0	75.8	79.0	65.0	63.9	69.1	38.7	71.0	-	58.9	81.3	79.0	-	66.9	68.1
	mBERT+Syn	84.1	34.6	76.9	75.4	68.2	**76.0**	79.1	64.0	64.2	68.7	38.0	**73.1**	-	58.0	81.7	79.5	-	67.0	68.1
	mBERT+code-switch	82.4	39.2	77.1	75.2	68.2	71.0	78.0	66.1	64.2	72.4	41.3	69.2	-	59.9	81.3	78.9	-	67.3	68.3
	our method	**84.5**	**41.4**	**78.9**	**77.3**	**70.2**	75.3	**80.3**	**67.6**	63.9	**73.1**	**46.8**	72.6	-	**62.2**	**81.8**	**80.8**	-	**69.4**	**70.5**
mTOP(Li et al., 2021)	mBERT	81.0	-	-	28.1	-	40.2	38.8	9.8	-	-	-	-	-	-	-	-	-	29.2	39.6
	mBERT+Syn	81.3	-	-	30.0	-	43.0	41.2	11.5	-	-	-	-	-	-	-	-	-	31.4	41.4
	mBERT+code-switch	82.3	-	-	40.3	-	47.5	48.2	16.0	-	-	-	-	-	-	-	-	-	38.0	46.8
	our method	**83.5**	-	-	**44.5**	-	**54.2**	**51.7**	**18.8**	-	-	-	-	-	-	-	-	-	**47.3**	**50.5**

Table 3: The experimental results on four tasks. The best results in each task are highlighted in bold. The baselines include "mBERT", "mBERT+Syn" and "mBERT+codeswitch". They delegate "only using mBERT", "using mBERT with a syntactic module (GAT)" and "mBERT with the code-switching technique" for cross-lingual transfer. The results of "mBERT" is from Hu et al. (2020). For "mBERT+Syn" and "mBERT+code-switch", we adopt open-source code of the work of Ahmad et al. (2021) and Qin et al. (2021) to reproduce these experiments, and report the results. The evaluation metrics are F1 value for the NER task, Accuracy for classification tasks, and Exact Match for semantic parsing. The "AVG" column means the average performance across all language for each method, while the "AVG /en" indicates the average performance on the languages excluding English.

5.2. Generalized Cross-Lingual Transfer Results

In practical scenarios, cross-lingual transfer could involve any language pair. For example, in a cross-lingual question-answering (QA) task, the context passage may be in German, while the multilingual model is required to answer the question in French. Considering on this, we conduct zero-shot cross-lingual transfer experiments within a generalized setting. Since PAWS-X and mTOP are completely parallel, we evaluate the performance of our method and "mBERT" baseline on generalized cross-lingual transfer tasks using the two dataset. The experimental results are illustrated in Figure 3.

For both classification and semantic parsing benchmarks, we have observed improvements among most language pairs. This reflects that our method is very effective for generalized cross-lingual transfer. Furthermore, when English is included in the language pair, there is a substantial enhancement in performance. Specifically, when English serves as the source language, the average performance of target languages is increased over 10% and 3% in mTOP and PAWS-X dataset, respectively. This reflects the effectiveness of the code-switching in aligning other languages with English. For the PAWS-X dataset, we find that some non-Indo-European languages such as

Japanese, Korean, and Chinese can achieve improvements, even when the source languages belong to the Indo-European language family, including English, Spanish, French, and German. It reflects that syntactic knowledge can effectively narrow the gap of language structures for this task, especially for the language pairs without close linguistic relationships.

6. Analysis and Discussion

6.1. Impact on Languages

We investigate whether our method can improve the performance of specific languages or language groups. As shown in Figure 4, we display the performance improvement of our method by comparing the "mBERT" baseline. We find that almost languages can obtain benefits from our method. Particularly, when the target language, such as German, Spanish and French, belongs to the Indo-European language family, the improvement is very significant. Furthermore, the performance in the mTOP task is improved significantly by our method among all languages. This may be because that our method consider both syntax and lexicon simultaneously, which is beneficial for the semantic parsing task.
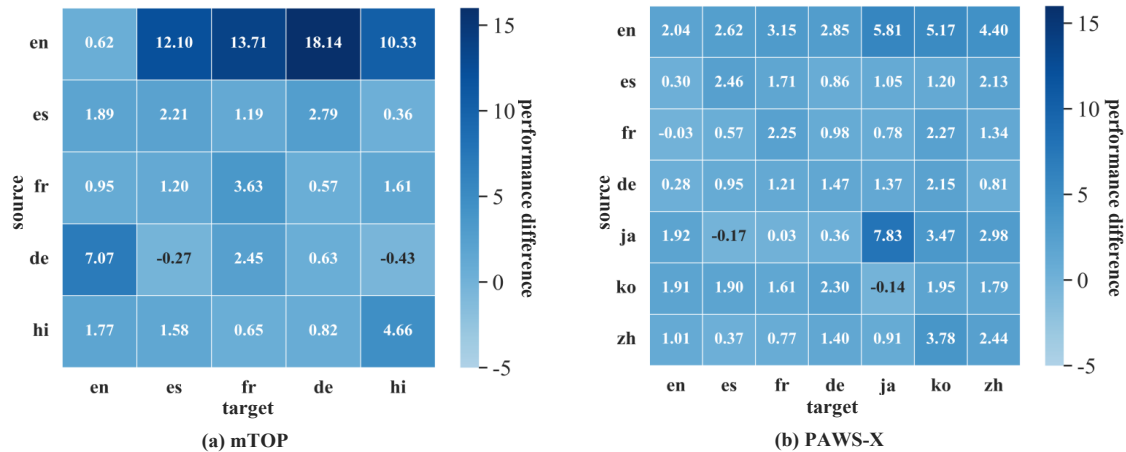
	en	es	fr	de	hi
en	0.62	12.10	13.71	18.14	10.33
es	1.89	2.21	1.19	2.79	0.36
fr	0.95	1.20	3.63	0.57	1.61
de	7.07	-0.27	2.45	0.63	-0.43
hi	1.77	1.58	0.65	0.82	4.66

(a) mTOP

	en	es	fr	de	ja	ko	zh
en	2.04	2.62	3.15	2.85	5.81	5.17	4.40
es	0.30	2.46	1.71	0.86	1.05	1.20	2.13
fr	-0.03	0.57	2.25	0.98	0.78	2.27	1.34
de	0.28	0.95	1.21	1.47	1.37	2.15	0.81
ja	1.92	-0.17	0.03	0.36	7.83	3.47	2.98
ko	1.91	1.90	1.61	2.30	-0.14	1.95	1.79
zh	1.01	0.37	0.77	1.40	0.91	3.78	2.44

(b) PAWS-X

Figure 3: Results for generalized zero-shot cross-lingual transfer on mTOP and PAWS-X. We report the performance differences between our method and "mBERT" baseline across all languages.
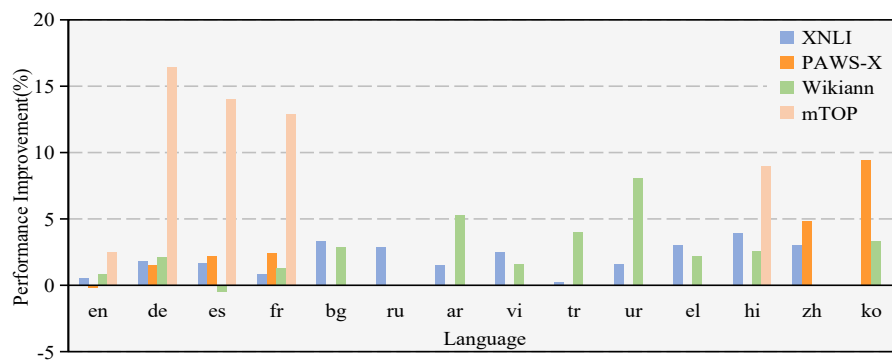

Figure 4: Performance improvements for XNLI, PAWS-X, Wikiann, and mTOP across languages. The languages in x-axis are grouped by language families: IE.Germanic (en, de), IE.Romance (es, fr), IE.Slavic (bg, ru), Afro-asiatic (ar), Austro-asiatic (vi), Altaic (tr, ur), IE.Greek (el), IE.Indic (hi), Sino-tibetan (zh), Korean (ko).

6.2. Representation Similarities across Languages

To evaluate the effectiveness of our method in aligning different languages, we employ the representation similarity between languages as the metric. Specifically, we utilize the testing set of XNLI (Conneau et al., 2018) as the dataset, which consists of parallel sentences across multiple languages. Then we take the vector of [CLS] token from the final layer of our model, as well as the vectors from two baselines ("mBERT+Syn" and "mBERT+code-switch") for each sentence. Following Libovický et al. (2019), the centroid vector for representing each language is calculated by averaging these sentence representations. Finally, we adopt cosine similarity as the indicator to assess the degree of alignment between English and each target language.

Figure 5 illustrates the similarities between languages by using our method and the other two baselines. It can be easily found that our

method outperforms the other two baselines in aligning language representations. This suggests that infusing two types of knowledge is indeed effective in reducing the disparities in language typologies, which improve cross-lingual transfer performance. In addition, we observe that "mBERT+code-switch" performs better than "mBERT+Syn", which reflects that lexical knowledge is more useful than syntactic knowledge for this task.

6.3. Impact of Code-switching

The replacement ratio α for code-switching is an important hyper-parameter in our method. Hence, we explore its impact on mTOP and PAWS-X, by varying α from 0 to 0.9 in increments of 0.1, shown in Figure 6. When α is set to 0, it represents the results of the baseline "mBERT+Syn". As α increases, more source words are substituted with their equivalent words from other languages. The performance improvement certificates the ef-
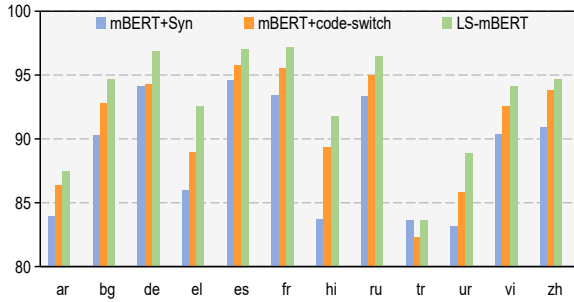
Figure 5: The similarities between languages. We first calculate the centroid representation for each language following Libovickỳ et al. (2019). Then we adopt cosine similarity to evaluate the similarity between English and each target language.

fectiveness of code-switching technique. Notably, when about half of the words are replaced (0.5 for PAWS-X and 0.4 for mTOP), the performance reaches their peaks. After that, both tasks experience a decline in performance. This decline might be because the expression of meaning and sentence structure are influenced severely as too many words are replaced. Therefore, it is a optimal choice to set α between 0.4 to 0.5 for code-switching.
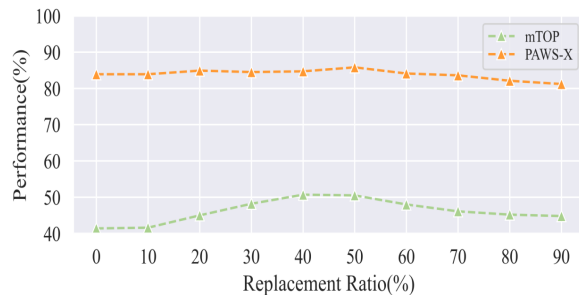

Figure 6: Performance on mTOP and PAWS-X with different replacement ratio α in code-switching.

Furthermore, we investigate whether the choice of the replacement language in code-switching impacts our model's performance. We select mTOP and PAWS-X as the testing tasks. In code-switching, we devise three different measures for language replacement: "Exclusively replacing with the target language", "Replacing with languages from the same language family as the target language"; and "Replacing with languages selected randomly". The experimental results are illustrated in Figure 7. We can easily observe that "Exclusively replacing with the target language" performs best, while "Replacing with randomly selected languages" yields the poorest results. Hence, this also underscores the importance of selecting languages closely related to each target language for

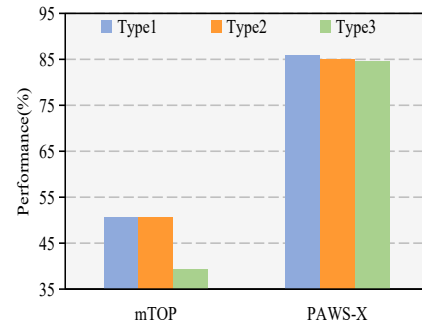substitution when employing the code-switching technique.

Figure 7: Performance on mTOP and PAWS-X with different replacement languages in code-switching. The source language for both tasks is English, and the results are averaged across all target languages excluding English. "Type1" represents the replacement with the target language; "Type2" represents the replacement with languages from the same language family as the target language; "Type3" represents the replacement with randomly selected languages.

6.4. Performance with XLM-R

To validate the universality of our method, we substitute multilingual BERT with XLM-R in our framework. XLM-R is a more robust multilingual pre-trained model known for its exceptional cross-lingual transfer capabilities. Subsequently, we test its performance on the PAWX-S dataset, and the experimental results are displayed in Table 4.

In Table 4, we also observe that our framework outperforms the other three baselines. This indicates that integrating lexical and syntactic knowledge is beneficial for enhancing performance, irrespective of the base model employed. Notably, our framework only achieves the slight performance improvement when utilizing XLM-R as the base model compared to employing multilingual BERT. It may be because that the base model, XLM-R, adopt larger corpus during pre-training, resulting in preserving richer language information. Consequently, XLM-R itself has possessed superior cross-lingual transfer capabilities. The assistance by incorporating external linguistic knowledge appears to be relatively minor in comparison.

6.5. Limitations and Challenges

In our study, we adopt a bilingual dictionary, such as MUSE (Lample et al., 2018), to substitute words in other languages. However, we randomly choose a target language word when there exist multiple translations for a source language word. This approach, although convenient, neglect the

Task	Methods	en	ar	bg	de	el	es	fr	hi	ru	tr	ur	vi	ko	nl	pt	AVG
PAWS-X	XLM-R	84.2	48.5	80.5	77.0	77.8	76.1	79.8	67.5	70.4	76.0	54.2	78.5	59.1	83.3	79.3	72.8
	XLM-R+Syn	83.5	46.4	80.1	76.0	78.9	77.6	79.1	72.1	70.6	76.1	55.3	77.6	59.0	83.1	79.2	73.0
	XKLM-R+code-switch	83.4	46.8	81.7	78.2	79.2	71.1	78.6	72.9	70.6	77.2	57.9	76.0	58.2	83.6	80.0	73.0
	our method	83.1	44.9	82.7	76.8	78.4	76.9	79.6	71.1	70.1	76.6	60.4	78.2	58.1	83.5	79.7	73.3

Table 4: Results for PAWS-X with XLM-R.

context of the source language word, potentially leading to inaccurate translations. This also highlights us to explore more precise word alignment methods in the future.

Furthermore, the tasks we have evaluated are quite limited, with some of them involving only a few languages. In the future, we will extend our method to more cross-lingual tasks. Meanwhile, we also develop dataset for these tasks to support more languages.

7. Conclusion

In this paper, we present a framework called "lexicon-syntax enhanced multilingual BERT" ("LS-mBERT"), which infuses lexical and syntactic knowledge to enhance cross-lingual transfer performance. Our method employs code-switching technology to generate input text mixed in various languages, enabling the entire model to capture lexical alignment information during training. Besides, a syntactic module consisting of a graph attention network (GAT) is introduced to guide mBERT in encoding language structures. The experimental results demonstrate that our proposed method outperforms all the baselines across different tasks, which certificates the effectiveness of integrating both types of knowledge into mBERT for improving cross-lingual transfer. In the future, we plan to incorporate different linguistic knowledge into large language models (LLMs) to further enhance cross-lingual transfer performance.

8. Acknowledgements

9. Bibliographical References

Wasi Ahmad, Haoran Li, Kai-Wei Chang, and Yashar Mehdad. 2021. Syntax-augmented multilingual bert for cross-lingual transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4538–4554.

Akari Asai, Jungo Kasai, Jonathan H Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021. Xor qa: Cross-lingual open-retrieval question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 547–564.

Nuo Chen, Linjun Shou, Ming Gong, and Jian Pei. 2022. From good to best: Two-stage training for cross-lingual machine reading comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10501–10508.

Alessandra Teresa Cignarella, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, Paolo Rosso, and Farah Benamara. 2020. Multilingual irony detection with dependency syntax and neural models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1346–1358.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

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.

Wietse de Vries, Martijn Wieling, and Malvina Nissim. 2022. Make the best of cross-lingual trans-

fer: 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.

Nick C Ellis and Diane Larsen-Freeman. 2009. *Language as a complex adaptive system*, volume 11. John Wiley & Sons.

Besnik Fetahu, Anjie Fang, Oleg Rokhlenko, and Shervin Malmasi. 2022. Dynamic gazetteer integration in multilingual models for cross-lingual and cross-domain named entity recognition. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2777–2790.

Xu Han, Yuqi Luo, Weize Chen, Zhiyuan Liu, Maosong Sun, Zhou Botong, Hao Fei, and Suncong Zheng. 2022. Cross-lingual contrastive learning for fine-grained entity typing for low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2241–2250.

Shexia He, Zuchao Li, and Hai Zhao. 2019. Syntax-aware multilingual semantic role labeling. 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 5350–5359.

Tsung-Yuan Hsu, Chi-Liang Liu, and Hung-Yi Lee. 2019. Zero-shot reading comprehension by cross-lingual transfer learning with multi-lingual language representation model. 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 5933–5940.

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 *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

Joo-Kyung Kim, Young-Bum Kim, Ruhi Sarikaya, and Eric Fosler-Lussier. 2017. Cross-lingual

transfer learning for pos tagging without cross-lingual resources. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2832–2838.

Siyu Lai, Hui Huang, Dong Jing, Yufeng Chen, Jinan Xu, and Jian Liu. 2021. Saliency-based multi-view mixed language training for zero-shot cross-lingual classification. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 599–610.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.

Anna Langedijk, Verna Dankers, Phillip Lippe, Sander Bos, Bryan Cardenas Guevara, Helen Yannakoudakis, and Ekaterina Shutova. 2022. Meta-learning for fast cross-lingual adaptation in dependency parsing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8503–8520.

Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. 2021. Mtop: A comprehensive multilingual task-oriented semantic parsing benchmark. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2950–2962.

Peiyao Li, Zhengkun Zhang, Jun Wang, Liang Li, Adam Jatowt, and Zhenglu Yang. 2023. Across: An alignment-based framework for low-resource many-to-one cross-lingual summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2458–2472.

Yunlong Liang, Fandong Meng, Chulun Zhou, Jinan Xu, Yufeng Chen, Jinsong Su, and Jie Zhou. 2022. A variational hierarchical model for neural cross-lingual summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2088–2099.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2019. How language-neutral is multilingual bert? *arXiv preprint arXiv:1911.03310*.

Ercong Nie, Sheng Liang, Helmut Schmid, and Hinrich Schütze. 2022. Cross-lingual retrieval augmented prompt for low-resource languages. *arXiv preprint arXiv:2212.09651*.

Farhad Nooralahzadeh and Rico Sennrich. 2023. Improving the cross-lingual generalisation in visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13419–13427.

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.

Kunxun Qi, Hai Wan, Jianfeng Du, and Haolan Chen. 2022. Enhancing cross-lingual natural language inference by prompt-learning from cross-lingual templates. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1910–1923.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. 2021. Cosda-ml: multi-lingual code-switching data augmentation for zero-shot cross-lingual nlp. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3853–3860.

Freda Shi, Kevin Gimpel, and Karen Livescu. 2022. Substructure distribution projection for zero-shot cross-lingual dependency parsing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6547–6563.

Milan Straka and Jana Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies*, pages 88–99.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Dong Wang and Thomas Fang Zheng. 2015. Transfer learning for speech and language processing. In *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1225–1237. IEEE.

Xinyi Wang, Sebastian Ruder, and Graham Neubig. 2022. Expanding pretrained models to thousands more languages via lexicon-based adaptation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Ziyun Wang, Xuan Liu, Peiji Yang, Shixing Liu, and Zhisheng Wang. 2021. Cross-lingual text classification with heterogeneous graph neural network. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 612–620.

Linjuan Wu, Shaojuan Wu, Xiaowang Zhang, Deyi Xiong, Shizhan Chen, Zhiqiang Zhuang, and Zhiyong Feng. 2022. Learning disentangled semantic representations for zero-shot cross-lingual transfer in multilingual machine reading comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 991–1000.

Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A Smith, and Jaime G Carbonell. 2018. Neural cross-lingual named entity recognition with minimal resources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 369–379.

Liyan Xu, Xuchao Zhang, Bo Zong, Yanchi Liu, Wei Cheng, Jingchao Ni, Haifeng Chen, Liang Zhao, and Jinho D Choi. 2022. Zero-shot cross-lingual machine reading comprehension via inter-sentence dependency graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11538–11546.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. Paws-x: A cross-lingual adversarial dataset for paraphrase identification. *arXiv preprint arXiv:1908.11828*.

Sicheng Yu, Hao Zhang, Yulei Niu, Qianru Sun, and Jing Jiang. 2021. Cosy: Counterfactual syntax for cross-lingual understanding. 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 577–589.

Tong Zhang, Long Zhang, Wei Ye, Bo Li, Jinan Sun, Xiaoyu Zhu, Wen Zhao, and Shikun Zhang. 2021a. Point, disambiguate and copy: Incorporating bilingual dictionaries for neural machine translation. 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 3970–3979.

Zhisong Zhang, Emma Strubell, and Eduard Hovy. 2021b. On the benefit of syntactic supervision for cross-lingual transfer in semantic role labeling. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6229–6246.