ENTITYCS: Improving Zero-Shot Cross-lingual Transfer with Entity-Centric Code Switching

Chenxi Whitehouse[1,2,*], **Fenia Christopoulou**[1], **Ignacio Iacobacci**[1]

[1]Huawei Noah's Ark Lab, London, UK
[2]City, University of London
chenxi.whitehouse@city.ac.uk
{efstathia.christopoulou, ignacio.iacoboacci}@huawei.com

Abstract

Accurate alignment between languages is fundamental for improving cross-lingual pre-trained language models (XLMs). Motivated by the natural phenomenon of code-switching (CS) in multilingual speakers, CS has been used as an effective data augmentation method that offers language alignment at word- or phrase-level, in contrast to sentence-level via parallel instances. Existing approaches either use dictionaries or parallel sentences with word-alignment to generate CS data by randomly switching words in a sentence. However, such methods can be suboptimal as dictionaries disregard semantics, and syntax might become invalid after random word switching. In this work, we propose ENTITYCS, a method that focuses on ENTITY-level Code-Switching to capture fine-grained cross-lingual semantics without corrupting syntax. We use Wikidata and the English Wikipedia to construct an entity-centric CS corpus by switching entities to their counterparts in other languages. We further propose entity-oriented masking strategies during intermediate model training on the ENTITYCS corpus for improving entity prediction. Evaluation of the trained models on four entity-centric downstream tasks shows consistent improvements over the baseline with a notable increase of 10% in Fact Retrieval. We release the corpus and models to assist research on code-switching and enriching XLMs with external knowledge[1].

1 Introduction

Cross-lingual pre-trained Language Models (XLMs) such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020a), have achieved state-of-the-art zero-shot cross-lingual transferability across diverse Natural Language Understanding (NLU) tasks. Such models have been particularly enhanced with the use of bilingual parallel sentences together with alignment methods (Yang et al., 2020; Chi et al., 2021a; Hu et al., 2021; Gritta and Iacobacci, 2021; Feng et al., 2022). However, obtaining high quality parallel data is expensive, especially for low-resource languages. Therefore, alternative data augmentation approaches have been proposed, one of which is Code Switching.

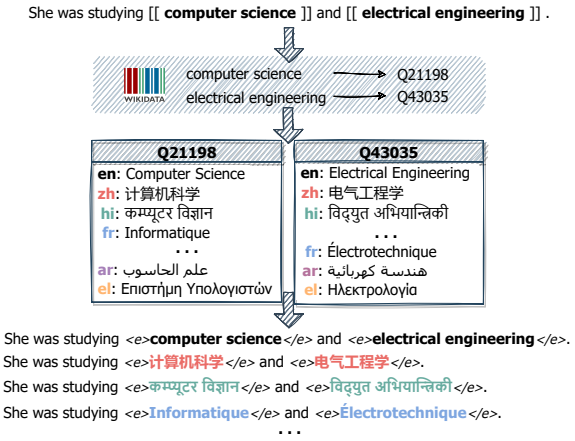


Figure 1: Illustration of generating ENTITYCS sentences from an English sentence extracted from Wikipedia. Entities in double square brackets indicate wikilinks.

Code Switching (CS) is a phenomenon when multilingual speakers alternate words between languages when they speak, which has been studied for many years (Gumperz, 1977; Khanuja et al., 2020; Doğruöz et al., 2021). Code-switched sentences consist of words or phrases in different languages, therefore they capture semantics of finer-grained cross-lingual expressions compared to parallel sentences, and have been used for multilingual intermediate training (Yang et al., 2020) and fine-tuning (Qin et al., 2020; Krishnan et al., 2021). Nevertheless, since manually creating large-scale

*Work conducted as Research Intern at Huawei Noah's Ark Lab, London, UK.

[1]Code and models available at https://github.com/huawei-noah/noah-research/tree/master/NLP/EntityCS.

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 6727 - 6743
December 7-11, 2022

CS datasets is costly and only few natural CS texts exist (Barik et al., 2019; Xiang et al., 2020; Chakravarthi et al., 2020; Lovenia et al., 2021), research has turned to automatic CS data generation. Some of those approaches generate CS data via dictionaries, usually ignoring ambiguity (Qin et al., 2020; Conneau et al., 2020b). Others, require parallel data and an alignment method to match words or phrases between languages (Yang et al., 2020; Rizvi et al., 2021). In both cases, what is switched is chosen randomly, potentially resulting in syntactically odd sentences or switching to words with little semantic content (e.g. conjunctions).

On the contrary, entities contain external knowledge and do not alter sentence syntax if replaced with other entities, which mitigates the need for any parallel data or word alignment tools. Motivated by this, we propose ENTITYCS, a Code-Switching method that focuses on ENTITIES, as illustrated in Figure 1. Resources such as Wikipedia and Wikidata offer rich cross-lingual entity-level information, which have been proven beneficial in XLMs pre-training (Jiang et al., 2020; Calixto et al., 2021; Jiang et al., 2022). We use such resources to generate an entity-based CS corpus for intermediate training of XLMs. Entities in wikilinks[2] are switched to their counterparts in other languages retrieved from the Wikidata Knowledge Base (KB), thus alleviating ambiguity.

Using the ENTITYCS corpus, we propose a series of masking strategies that focus on enhancing Entity Prediction (EP) for better cross-lingual entity representations. We evaluate the models on entity-centric downstream tasks including Named Entity Recognition (NER), Fact Retrieval, Slot Filling (SF) and Word Sense Disambiguation (WSD). Extensive experiments demonstrate that our models outperform the baseline on zero-shot cross-lingual transfer, with +2.8% improvement on NER, surpassing the prior best result that uses large amounts of parallel data, +10.0% on Fact Retrieval, +2.4% on Slot Filling, and +1.3% on WSD.

The main contributions of this work include: a) construction of an entity-level CS corpus, ENTITYCS, based solely on the English Wikipedia and Wikidata, mitigating the need for parallel data, word-alignment methods or dictionaries; b) a series of intermediate training objectives, focusing on Entity Prediction; c) improvement of zero-shot performance on NER, Fact Retrieval, Slot Filling and WSD; d) further analysis of model errors, behaviour of different masking strategies throughout training as well as impact across languages, demonstrating how our models particularly benefit non-Latin script languages.

STATISTIC	COUNT
Languages	93
English Sentences	54,469,214
English Entities	104,593,076
Average Sentence Length	23.37
Average Entities per Sentence	2
CS Sentences per EN Sentence	≤ 5
CS Sentences	231,124,422
CS Entities	420,907,878

Table 1: Statistics of the ENTITYCS Corpus.

2 Methodology

We introduce the details of the ENTITYCS corpus construction, as well as different entity-oriented masking strategies used in our experiments.

2.1 ENTITYCS Corpus Construction

Wikipedia is a multilingual online encyclopedia available in more than 300 languages[3]. Structured data of Wikipedia articles are stored in Wikidata, a multilingual document-oriented database. With more than six million articles, English Wikipedia has the potential to serve as a rich resource for generating CS data. We use the English Wikipedia and leverage entity information from Wikidata to construct an entity-based CS corpus.

To achieve this, we make use of wikilinks in Wikipedia, i.e. links from one page to another. We use the English Wikipedia dump[4] and extract raw text with WikiExtractor[5] while keeping track of wikilinks. Wikilinks are typically surrounded by square brackets in Wikipedia dump, in the format of [[*entity* | *display text*]], where *entity* is the title of the target Wikipedia page it links to, and *display text* corresponds to what is displayed in the current article. We then employ SpaCy[6] for sentence segmentation. Since we are interested in creating entity-level CS instances, we only keep sentences containing at least one wikilink. Sentences longer

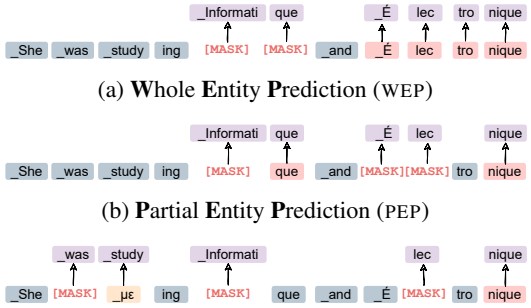
[2] https://en.wikipedia.org/wiki/Help:Link#Wikilinks_(internal_links)

[3] https://en.wikipedia.org/wiki/Wikipedia

[4] https://dumps.wikimedia.org/enwiki/latest/ (Nov 2021 version).

[5] https://github.com/attardi/wikiextractor

[6] https://spacy.io/



(a) **W**hole **E**ntity **P**rediction (WEP)

(b) **P**artial **E**ntity **P**rediction (PEP)

(c) **P**artial **E**ntity **P**rediction with **MLM** (PEP+MLM)

Figure 2: Illustration of the proposed masking strategies. Random subwords are chosen from the entire vocabulary, thus can be from different languages. (c) shows a case where "study" is replaced with a Greek subword.

Masking Strategy		Entity (%)				Non-Entity (%)			
	p	MASK	RND	SAME	p	MASK	RND	SAME
MLM	15	80	10	10	15	80	10	10
WEP	100	80	0	20	0	-	-	-
PEP_{MRS}	100	80	10	10	0	-	-	-
PEP_{MS}	100	80	0	10	0	-	-	-
PEP_{M}	100	80	0	0	0	-	-	-
+MLM WEP	50	80	0	20	15	80	10	10
+MLM PEP_{MRS}	50	80	10	10	15	80	10	10
+MLM PEP_{MS}	50	80	0	10	15	80	10	10
+MLM PEP_{M}	50	80	0	0	15	80	10	10

Table 2: Summary of the proposed masking strategies. p corresponds to the probability of choosing candidate items (entity/non-entity subwords) for masking. **MASK**, **RND**, **SAME** represent the percentage of replacing a candidate with Mask, Random or the Same item. When combining WEP/PEP with MLM (+MLM), we lower p to 50%.

than 128 words are also removed. This results in 54.5M English sentences and 104M entities in the final ENTITYCS corpus.

As illustrated in Figure 1, given an English sentence with wikilinks, we first map the entity in each wikilink to its corresponding Wikidata ID and retrieve its available translations from Wikidata. For each sentence, we check which languages have translations for all entities in that sentence, and consider those as candidates for code-switching. We select 92 target languages in total, which are the overlap between the available languages in Wikidata and XLM-R (Conneau et al., 2020a) (the model we use for intermediate training). We ensure all entities are code-switched to the same target language in a single sentence, avoiding noise from including too many languages. To control the size of the corpus, we generate up to five ENTITYCS sentences for each English sentence. In particular, if fewer than five languages have translations available for all the entities in a sentence, we create ENTITYCS instances with all of them. Otherwise, we randomly select five target languages from the candidates. If no candidate languages can be found, we do not code-switch the sentence. Instead, we keep it as part of the English corpus. Finally, we surround each entity with entity indicators (<e>, </e>). The statistics of the ENTITYCS corpus are summarised in Table 1. A histogram of the number of sentences and entities per language is shown in Appendix A.

2.2 Masking Strategies

To test the effectiveness of intermediate training on the generated ENTITYCS corpus, we experiment with several training objectives using an existing pre-trained language model. Firstly, we employ the conventional 80-10-10 MLM objective, where 15% of sentence subwords are considered as masking candidates. From those, we replace subwords with [MASK] 80% of the time, with Random subwords (from the entire vocabulary) 10% of the time, and leave the remaining 10% unchanged (Same). To integrate entity-level cross-lingual knowledge into the model, we propose Entity Prediction objectives, where we only mask subwords belonging to an entity. By predicting the masked entities in ENTITYCS sentences, we expect the model to capture the semantics of the same entity in different languages. Two different masking strategies are proposed for predicting entities: Whole Entity Prediction (WEP) and Partial Entity Prediction (PEP).

In WEP, motivated by Sun et al. (2019) where whole word masking is also adopted, we consider all the *words* (and consequently subwords) inside an entity as masking candidates. Then, 80% of the time we mask *every* subword inside an entity, and 20% of the time we keep the subwords intact. Note that, as our goal is to predict the entire masked entity, we do not allow replacing with Random subwords, since it can introduce noise and result in the model predicting incorrect entities. After entities are masked, we remove the entity indicators <e>, </e> from the sentences before feeding them to the model. Figure 2a shows an example of WEP.

For PEP, we also consider all entities as masking candidates. In contrast to WEP, we do not force subwords belonging to one entity to be either all masked or all unmasked. Instead, each *individual entity subword* is masked 80% of the time. For the remaining 20% of the masking candidates, we

experiment with three different replacements. First, PEP_{MRS}, corresponds to the conventional 80-10-10 masking strategy, where 10% of the remaining subwords are replaced with Random subwords and the other 10% are kept unchanged. In the second setting, PEP_{MS}, we remove the 10% Random subwords substitution, i.e. we predict the 80% masked subwords and 10% Same subwords from the masking candidates. In the third setting, PEP_{M}, we further remove the 10% Same subwords prediction, essentially predicting only the masked subwords. An example of PEP is illustrated in Figure 2b.

Prior work has proven it is effective to combine Entity Prediction with MLM for cross-lingual transfer (Jiang et al., 2020), therefore we investigate the combination of the Entity Prediction objectives together with MLM on non-entity subwords. Specifically, when combined with MLM, we lower the entity masking probability (p) to 50% to roughly keep the same overall masking percentage. Figure 2c illustrates an example of PEP combined with MLM on non-entity subwords. A summary of the masking strategies is shown in Table 2, along with the corresponding masking percentages.

3 Experimental Setup

After preparing the EntityCS corpus, we further train an XLM with WEP, PEP, MLM and the joint objectives. We use the sampling strategy proposed by Conneau and Lample (2019), where high resource languages are down-sampled and low resource languages get sampled more frequently. Since recent studies on pre-trained language encoders have shown that semantic features are highlighted in higher layers (Tenney et al., 2019; Rogers et al., 2020), we only train the embedding layer and the last two layers of the model[7] (similarly to Calixto et al. (2021)). We randomly choose 100 sentences from each language to serve as a validation set, on which we measure the perplexity every 10K training steps. Details of parameters used for intermediate training can be found in Appendix C.

3.1 Downstream Tasks

As the EntityCS corpus is constructed with code-switching at the entity-level, we expect our models to mostly improve entity-centric tasks. Thus, we choose the following datasets: WikiAnn (Pan et al., 2017) for NER, X-FACTR (Jiang et al., 2020) for Fact Retrieval, MultiATIS++ (Xu et al., 2020) and MTOP (Li et al., 2021) for Slot Filling, and XL-WiC (Raganato et al., 2020) for WSD[8]. More details on the datasets can be found in Appendix B.

After intermediate training on the EntityCS corpus, we evaluate the zero-shot cross-lingual transfer of the models on each task by fine-tuning on task-specific English training data. For NER we use the checkpoint with the lowest validation set perplexity during intermediate training. Similarly, for the probing dataset X-FACTR (only consisting of a test set), we probe models with the lowest perplexity and report the maximum accuracy score for all, single- and multi-token entities between the two proposed decoding methods (independent and confidence-based) from the original paper (Jiang et al., 2020). For MultiATIS++, MTOP, and XL-WiC datasets, we choose the checkpoints with the best performance on the English validation set[9]. For all experiments, except X-FACTR, we fine-tune models with five random seeds and report average and standard deviation.

3.2 Pre-Training Languages

Given the size of the EntityCS corpus, we primarily select a subset from the total 93 languages, that covers most of the languages used in the downstream tasks. This subset contains 39 languages, from WikiAnn, excluding Yoruba[10]. We train XLM-R-base (Conneau and Lample, 2019) on this subset, then fine-tune the new checkpoints on the English training set of each dataset and evaluate on all of the available languages.

4 Main Results

Results are reported in Table 3 where we compare models trained on the EntityCS corpus with MLM, WEP, PEP_{MS} and PEP_{MS}+MLM masking strategies. For MultiATIS++ and MTOP, we report results of training only Slot Filling (SF), as well as joint training of Slot Filling and Intent Classification (SF/Intent).

Named Entity Recognition For NER, we can see that all models with CS intermediate training show consistent improvement on WikiAnn over the

[7]Preliminary experiments where we updated the entire network revealed the model suffered from catastrophic forgetting.

[8]The result reported on the XL-WiC for prior work is our re-implementation based on https://github.com/pasinit/xlwic-runs.

[9]We observed performance drop for those tasks at later checkpoints.

[10]Yoruba is not included in the EntityCS corpus, as we only consider languages XLM-R is pre-trained on.

	MODEL	NER (F1) WIKIANN	FACT RETR. (ACC.) X-FACTR			SLOT FILLING (F1, F1/ACC.) MULTIATIS++		MTOP		WSD (ACC.) XL-WIC
			all	*single*	*multi*	*SF*	*SF / Intent*	*SF*	*SF / Intent*	
	XLM-R$^{\text{PRIOR}}$	61.8	3.5	9.4	2.6	–	–	–	–	58.0
	XLM-R$^{\text{OURS}}$	61.6 $_{0.28}$	3.5	9.4	2.6	71.8 $_{1.96}$	73.0 $_{0.70}$ / 89.1 $_{1.04}$	**73.2** $_{0.89}$	72.5 $_{0.78}$ / 86.0 $_{0.69}$	59.1 $_{1.52}$
EntityCS	MLM	63.5 $_{0.50}$	2.5	6.4	1.7	72.1 $_{2.34}$	74.0 $_{0.69}$ / 89.6 $_{1.43}$	72.8 $_{0.60}$	72.7 $_{0.31}$ / **86.3** $_{0.41}$	59.3 $_{0.44}$
EntityCS	WEP	62.4 $_{0.68}$	**6.1**	**19.4**	3.0	71.6 $_{1.20}$	71.7 $_{0.82}$ / 89.7 $_{1.25}$	72.2 $_{0.57}$	**73.0** $_{0.47}$ / 86.0 $_{0.44}$	**60.4** $_{0.97}$
EntityCS	PEP$_{\text{MS}}$	63.3 $_{0.70}$	6.0	15.0	**4.3**	73.4 $_{1.70}$	**74.4** $_{0.67}$ / **90.0** $_{0.90}$	71.5 $_{0.67}$	72.7 $_{0.64}$ / 86.1 $_{0.51}$	60.2 $_{0.85}$
EntityCS	PEP$_{\text{MS}}$+MLM	**64.4** $_{0.50}$	5.7	13.9	3.9	**74.2** $_{0.43}$	74.3 $_{0.82}$ / 89.0 $_{0.87}$	73.0 $_{0.33}$	72.5 $_{0.57}$ / 85.8 $_{0.77}$	59.8 $_{0.75}$

Table 3: Average performance across languages on the test set of downstream tasks. XLM-R$^{\text{PRIOR}}$ corresponds to previous reported results with XLM-R-base, referring to Chi et al. (2021b) for WikiAnn, Jiang et al. (2020) for X-FACTR and Raganato et al. (2020) for XL-WiC. XLM-R$^{\text{OURS}}$ shows our re-implemented results with XLM-R-base. Results (excluding X-FACTR) are averaged across five seeds with standard deviation reported as a subscript.

MODEL	AR	HE	VI	ID	JV	MS	TL	EU	ML	TA	TE	AF	NL	EN	DE	EL	BN	HI	MR	UR
XLM-R$^{\text{OURS}}$	44.6	51.9	68.3	48.6	59.6	63.3	72.5	**61.2**	63.2	54.3	49.3	76.3	80.7	83.4	75.4	74.2	67.9	68.3	61.8	55.8
PEP$_{\text{MS}}$	49.6	53.0	70.0	58.5	**62.0**	64.9	**75.7**	59.8	63.3	**57.7**	52.1	76.4	80.9	83.8	75.1	76.3	72.5	70.1	**66.8**	61.5
PEP$_{\text{MS}}$+MLM	**51.5**	**54.0**	**70.9**	**61.1**	59.3	**69.9**	74.6	59.3	**66.3**	57.6	**54.8**	**77.9**	**81.5**	**84.2**	**75.5**	**77.1**	**74.6**	**70.7**	66.3	**65.9**
	FA	FR	IT	PT	ES	BG	RU	JA	KA	KO	TH	SW	YO	MY	ZH	KK	TR	ET	FI	HU
XLM-R$^{\text{OURS}}$	47.6	78.0	78.2	78.9	76.2	77.3	63.9	22.9	66.4	48.8	4.3	**68.3**	45.4	52.7	27.7	44.2	76.9	72.4	75.6	76.9
PEP$_{\text{MS}}$	**55.6**	78.8	78.5	78.6	75.8	78.0	66.4	21.3	67.0	50.2	**4.6**	66.9	44.7	55.2	26.9	48.9	77.4	**73.4**	76.6	77.8
PEP$_{\text{MS}}$+MLM	54.2	**79.5**	**78.9**	**80.1**	**78.2**	**79.6**	**67.7**	**23.2**	**68.2**	**52.1**	4.0	66.4	**48.4**	**56.1**	**29.8**	**52.0**	**78.6**	71.9	**76.8**	**78.8**

Table 4: F1-score per language on the WikiAnn test set. Results are averaged across five seeds.

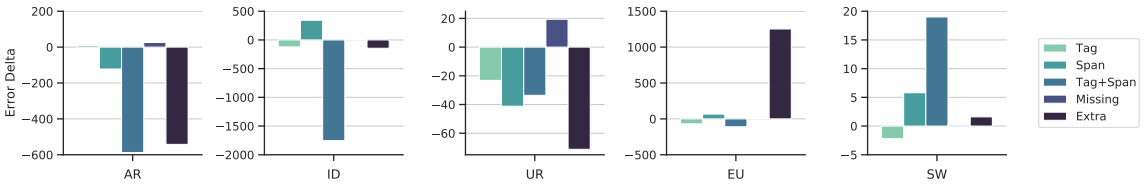


Figure 3: Error Delta (lower is better) for different types of errors in the WikiAnn test set between vanilla XLM-R-base and PEP$_{\text{MS}}$+MLM. We show error count differences for AR, ID and UR, the three languages with the largest F1-score improvement, as well as EU and SW, the two languages that under-perform the baseline.

baseline, with PEP$_{\text{MS}}$+MLM having +2.8% absolute improvement. This also outperforms XLM-Align (Chi et al., 2021b) with 63.7 F1-score, which uses a large amount of parallel data (more results can be found in Appendix D). The classic MLM objective seems equally effective with PEP, possibly because parts of entities are chosen as masking candidates in both. WEP on the other hand, results in lower performance, indicating that full entity prediction from the remaining context is more challenging. We report performance per language for PEP$_{\text{MS}}$ and PEP$_{\text{MS}}$+MLM in Table 4[11]. We can see that almost all languages benefit from training on the EntityCS corpus. In our best setting PEP$_{\text{MS}}$+MLM, AR, ID, and UR show the largest improvement (around +10% in AR and ID). EU and SW on the other hand, result in worse performance compared to the baseline.

As such, we take PEP$_{\text{MS}}$+MLM, and further analyse typical NER errors including five categories: Tag, Span, Tag+Span, Missing Extraction and Extra Extraction. Missing Extraction occurs when the prediction fails to identify an entity, while Extra Extraction represents errors when a non-entity is wrongly predicted as an entity. We select EU and SW (lower F1-score than the baseline)[12], AR, ID, and UR (languages with the largest improvement), and show a delta bar plot in Figure 3. Compared with the baseline, AR, ID, and UR improve consistently Tag+Span and Extra Extraction. All but ID, improve Span detection while most languages perform worse in Missing Extraction. EU

[11]Per-language results on WikiAnn for other models are reported in Table 9 in Appendix D.

[12]We do not consider Thai due to its erroneous tokenization in WikiAnn where everything is tokenized into individual *characters*.

and SW on the other hand, result in slightly worse Span and Extra Extraction errors.

We further investigate the reasons for this behaviour. In ID, we observe that around 80% of the Span errors are due to additionally identifying the token *"ALIH"* (means "moving", "changing" in English) as the start of an entity. For example, for the input [*"ALIH"*, *"Indofood"*, *"Sukses"*, *"Makmur"*], the gold entity is *"ORG: Indofood Sukses Makmur"*, whereas the model predicts *"ORG: ALIH Indofood Sukses Makmur"*. This also occurs in XLM-R as 68% of the span errors. Consultation with a native speaker reveals this may be an inaccuracy of the dataset (*"ALIH"* should not appear before the actual entities in WikiAnn). As for EU, lower overlap between WikiAnn entities and Wikipedia (only 47% vs 57% on average across all WikiAnn languages), might explain the prediction of additional entities not contained in the dataset.

Fact Retrieval For X-FACTR, all models trained with Entity Prediction outperform the baseline, whereas MLM is worse than vanilla XLM-R as expected. For single-token classification, WEP achieves the best results with +10% gain over XLM-R. On the other hand in PEP_{MS}, we mask part of the entity subwords for prediction, therefore it is best when predicting multi-token entities. Notably, models trained on large parallel data such as InfoXLM (Chi et al., 2021a) and XLM-Align (Chi et al., 2021b), perform poorly with 3% and 5% single-token accuracy, respectively, and <1% multi-token accuracy as they focus on alignment at the sentence-level (see Appendix D). Results per language for X-FACTR are available in Table 12 of Appendix D.

Slot Filling In SF-only training, the best performing model is PEP_{MS}+MLM, where we achieve +2.4% gain over XLM-R on MultiATIS++, also competitive with the best result from XLM-Align (74.4, see Appendix D). In contrast, no improvements can be observed in MTOP over the baseline. Manual inspection of the dataset reveals that this discrepancy can be attributed to domain differences. MultiATIS++ contains entities such as city names, whereas MTOP consists of dialogues with a personal assistant, e.g. setting up reminders, thus fewer entities occur and limit the benefits of entity-centric CS training. When jointly optimising SF and Intent Classification, models also improve over the baseline on SF (+1.4% for MultiATIS++ and +0.5% for MTOP), however with lower gains. We speculate that the additional intent labels offer complementary information to the task, minimising the impact of external information.

MODEL	LATIN SCRIPT						NON LATIN SCRIPT			
	ES	DE	FR	PT	TR	*avg*	ZH	JA	HI	*avg*
XLM-R$^{\text{OURS}}$	81.5	79.8	74.8	76.5	43.0	71.1	77.2	56.8	50.6	61.5
MLM	78.8	78.0	74.4	74.6	39.7	69.1	76.4	70.3	61.5	69.4
PEP_{MS}	79.3	79.7	75.3	76.2	45.3	71.1	77.8	69.0	62.9	69.9
PEP_{MS}+MLM	81.3	81.4	78.2	76.1	42.1	**71.8**	78.8	68.8	65.8	**71.1**

Table 5: F1-score (average across five seeds) for languages with Latin and Non-Latin script on MultiATIS++ test set when using SF-only training.

We then categorise languages in MultiATIS++ based on whether they have the same script as English (Latin), and investigate their performance on SF-only training. From Table 5 we can see that the models trained on the ENTITYCS corpus demonstrate notable improvements on languages with non-Latin scripts, +9.6% on average. This indicates that with entity-focused training, models capture information that is especially useful for languages with different scripts than English. More results on MultiATIS++ can be found in Table 10 of Appendix D.

Word Sense Disambiguation For XL-WiC we observe less improvement across tasks, where WEP performs best with +1.3% over the baseline. This behaviour can be attributed to the nature of the task, as in our entity-based training objectives, disambiguation is assumed already addressed and treated as implicit external information. Notably, when testing XLM-Align that uses parallel data, we observe that it does not improve ambiguous word-level semantics across languages (57% accuracy).

5 Analysis

We conduct further analysis on different masking strategies, their impact across languages and training steps when performing intermediate training on the ENTITYCS corpus. We primarily focus on WikiAnn, as it contains the largest amount of languages from the datasets we evaluate on.

5.1 Performance vs Pre-training Languages

For WikiAnn, X-FACTR and MultiATIS++, we additionally train MLM, WEP and PEP_{MS}+MLM by varying the amount of languages in the ENTITYCS corpus. We experiment with using English (no code-switching), the subset of 39 languages (as mentioned in subsection 3.2) and all 93 languages.

MODEL		WIKIANN	X-FACTR			MULTIATIS++	
			all	*single*	*multi*	*SF*	*SF / Intent*
XLM-R$^{\text{OURS}}$		61.6	3.5	9.4	2.6	70.6	73.0 / 88.9
MLM	EN	61.0	1.1	2.7	0.7	71.5	72.1 / 89.6
	39	**63.5**	2.6	6.4	1.7	72.5	**73.8 / 90.2**
	93	63.3	**2.7**	**6.8**	**1.8**	**72.7**	73.4 / 89.6
WEP	EN	61.9	3.3	8.5	1.6	**71.8**	72.2 / **91.1**
	39	**62.4**	**6.1**	**19.4**	**3.0**	71.1	71.7 / 89.7
	93	59.4	5.8	18.6	2.7	70.4	**72.9** / 90.3
PEP_{MS} +MLM	EN	61.2	2.7	6.6	1.6	71.3	72.3 / 90.7
	39	**64.4**	**5.7**	**13.9**	**3.9**	**73.4**	**74.4** / 90.0
	93	63.6	5.5	13.2	3.8	72.8	72.7 / **90.8**

Table 6: Results (average over five seeds) with different number of pre-training languages.

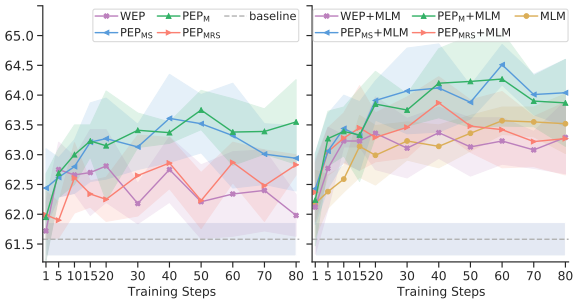


Figure 4: F1-score comparison on WikiAnn test set (average across five seeds) as a function of the number of training steps (in *ten* thousands) with various masking objectives. EP only strategies are on the left, EP + MLM strategies are on the right.

From Table 6, we see that as expected, using English sentences only does not improve the average performance across languages (only Intent Classification accuracy and WikiAnn with WEP increase over the baseline). However, models trained on English only can benefit English performance, e.g. with an average of +23.1% gain on single- and +5.6% on multi-token predictions in X-FACTR over baseline XLM-R using WEP. When trained with all 93 languages, models with all masking strategies show positive performance over XLM-R, but the majority of results show that training on fewer languages has overall the best performance. This indicates that including a wider selection of languages does not necessarily contribute to better results and scaling to too many languages still remains non-trivial. However, note that the subset of 39 languages covers most languages in those downstream tasks. As such, in cases where more languages are tested, increasing the number of pre-training languages may result in better performance.

5.2 Performance Across Training

Figure 4 shows a comparison of using different training objectives on the EntityCS corpus, as a function of the number of training steps on the WikiAnn test set. From the figure, we observe most masking strategies reach a plateau after the middle of training. Comparing to the objectives that include MLM training, there is a clear increase in performance across the board, further proving that joint training of entities and non-entities is not only beneficial performance-wise but also results in smoother learning curves. Notably, all objectives surpass the baseline during the entire training.

We can also see that the gains from CS intermediate training do not come from additional training data. To prove this, one can look at Table 6, where WikiAnn F1-score trained on English-only sentences (no CS) shows that extra English-only training data does not improve over the XLM-R baseline (61.6). This observation can be combined with Figure 4 at step 200K, which corresponds to the steps required for training on the English-only sentences, i.e. the amount of training data that all models see is the same. We can see that all masking strategies achieve an F1-score above 62.3. This shows that the NER performance gain can be attributed to the design of the EntityCS corpus and the training objectives.

5.3 Random and Same Subword Prediction

We further investigate the impact of Random subword substitution and Same subword prediction of the masking candidates when performing PEP. Comparing PEP_{MRS}, PEP_{MS} and PEP_{M} in Figure 4, we can see that including Random subword substitution in PEP results in worse performance, while further removing predicting Same subwords does not have a significant effect. Intuitively, Random subword substitution can lead to predicting an incorrect entity and weaken what a model learns, for instance by making the model predict a wrong entity from some remaining (randomly replaced) subwords. On the other hand, predicting the Same subword is an easier task of copying the input entity to the output. As a result, models seem to neither gain nor lose performance with or without it. These observations are supported by recent findings (Wettig et al., 2022), although focused on monolingual settings. We also observe a similar behaviour when combining PEP with MLM.

5.4 Entity Masking Percentage

To check the impact of the entity masking candidates percentage during training, we increase the masking probability (p) from 50% to 80% and 100% for the best model, PEP_{MS}+MLM, and test its effect on the WikiAnn test set. We observe that further increasing the masking percentage results in a performance drop from $64.4_{\pm 0.50}$ F1-score to $63.8_{\pm 0.67}$ at 80%, and $64.0_{\pm 0.43}$ at 100%. We speculate that, masking too many subwords makes the task of entity prediction from the remaining context more challenging. The close performance between percentages though, can be explained by the fact that only two entities exist per sentence on average, as shown in Table 1.

6 Related Work

Cross-Lingual Pre-Training Most existing XLMs use parallel data to improve multilingual contextualised word representations for different languages (Ouyang et al., 2021; Luo et al., 2021; Chi et al., 2021b). Adapters have been applied to improve zero-shot and few-shot cross-lingual transfer by simply training a handful of model parameters (Pfeiffer et al., 2020; Ansell et al., 2021). In addition, meta-learning techniques (Nooralahzadeh et al., 2020; Tarunesh et al., 2021) have proven highly effective for fast adaption to new languages (Dou et al., 2019). Compared to those approaches, our method aims to further improve cross-lingual transferability via intermediate training on an entity-based CS corpus created from Wikipedia wikilinks, requiring no parallel data.

Code Switching Methods based on code-switching have been successfully applied on cross-lingual model pre-training and fine-tuning on various NLU tasks such as NER (Priyadharshini et al., 2020; Liu et al., 2021), Part-of-Speech Tagging (Ball and Garrette, 2018), Machine Translation (Srivastava and Singh, 2020), Intent Classification and Slot Filling (Krishnan et al., 2021), as well as on code-switched datasets (Rizal and Stymne, 2020; Prasad et al., 2021).

A major challenge when studying CS is the lack of training data (Gupta et al., 2020). Existing CS corpora mostly include English and one other language, e.g. English-Chinese, English-Hindi, English-Spanish, extracted from social media platforms (Barik et al., 2019; Xiang et al., 2020; Chakravarthi et al., 2020; Lovenia et al., 2021). Thus, automatic methods for generating CS data in multiple languages have recently been proposed.

Qin et al. (2020) and Conneau et al. (2020b) create CS data from downstream task datasets by randomly switching individual words to a target language using translations from bilingual dictionaries, at the cost of introducing ambiguity errors or switching to words without important content. Krishnan et al. (2021) use CS to improve Intent Classification and Slot Filling. Rather than switching individual words, they obtain phrase information from the slot labels and generate phrase-level CS sentences via automatic translations. Yang et al. (2020) create CS sentences by randomly substituting source phrases with their target equivalents in parallel sentences after obtaining word alignments. Jiang et al. (2020) select a subset of Wikipedia sentences in four languages that contain multilingual entities from X-FACTR, and create CS sentences by switching entities from English to non-English entities *and* vice versa, via Wikidata translations. Our work shares several similarities, while the main differences include that we use only the English Wikipedia, create an entity-level English-to-other languages CS corpus including 93 languages in total and test multiple entity prediction training objectives on four downstream tasks.

Knowledge Integration into Language Models Large Pre-trained Languages Models (PLM) lack explicit grounding to real world entities and relations, making it challenging to recover factual knowledge (Bender et al., 2021). Most current research on knowledge integration into PLMs focuses on monolingual models. KnowBert (Peters et al., 2019) integrates knowledge into BERT (Devlin et al., 2019) by identifying entity spans in the input text and uses an entity linker to retrieve entity embeddings from a KB. KEPLER (Wang et al., 2021b) jointly optimises MLM and knowledge embeddings with supervision from a KB. K-Adapter (Wang et al., 2021a) integrates learnable factual and linguistic knowledge adapters to PLMs by training them in a multi-task setting on relation and dependency-tree prediction. These models show improvements over the baseline on various tasks including relation classification, entity typing and word sense disambiguation.

Integrating multilingual knowledge into XLMs has also recently been addressed. Jiang et al. (2022) train a model with two knowledge-related tasks, en-

tity prediction and object entailment. They use WikiData description embeddings in one language (English and non-English) to predict an entity in a target language as a classification task, preserving an entity vocabulary. Calixto et al. (2021) use Wikipedia articles in 100 languages together with BabelNet (Navigli and Ponzetto, 2012), a multilingual sense-inventory for WSD, by predicting the WikiData ID of each entity. Another work taking advantage of entities by Ri et al. (2022) uses dedicated multi-lingual entity embeddings on 24 languages and outperforms word-based pre-trained models in various cross-lingual transfer tasks.

7 Conclusions

In this work, we improve zero-shot transfer on entity-oriented tasks via entity-level code-switching. We make use of the English Wikipedia and the Wikidata KB to construct an ENTITYCS corpus by replacing entities in wikilinks with their counterparts in other languages. We further propose entity-oriented training objectives to improve entity predictions. Evaluation of the models on five datasets reveals consistent improvements on NER, Fact Retrieval and WSD over the baseline and prior work that uses large amounts of parallel data, as well as competitive results on Slot Filling.

Interestingly, we found that replacing masking candidates with Random subwords in the conventional masking strategy is harmful for entity prediction, while different masking strategies are optimal for different downstream tasks. Specifically, Whole Entity Prediction performs best when emphasis is given on single-token factual knowledge. On the other hand, entity typing and multi-token factual retrieval benefit from Partial Entity Prediction, i.e. when the model is given incomplete information on entities and is required to fill-in the gaps. Concurrently jointly predicting non-entity and entity subwords can improve tasks where the entire input context is important for prediction. In addition, our method is particularly beneficial for languages with non-Latin script.

Our corpus construction process is generic and can be scaled to many more languages, while the proposed masking strategies can be used with any existing language model. Future work will focus on code-switching beyond entities, such as verbs and phrases, as well as using other sources beyond Wikipedia and Wikidata.

Limitations

An important limitation of the work is that before code-switching an entity, its morphological inflection is not checked. This can lead to potential errors as the form of the CS entity might not agree with the surrounding context (e.g. plural). There should be few cases as such, as we are only switching entities. However, this should be improved in a later version of the corpus. Secondly, the diversity of languages used to construct the ENTITYCS corpus is restricted to the overlap between the available languages in WikiData and XLM-R pre-training. This choice was for a better comparison between models, however it is possible to extend the corpus with more languages that XLM-R does not cover.

We also acknowledge that the proposed approach was mainly evaluated on entity-centric tasks. Evaluation on more general natural language understanding tasks such as Natural Language Inference is feasible and the impact of the method on them should be explored. In addition, we only CS from English to other languages, keeping the context in English. This was because fine-tuning is performed on English sentences alone, and we wanted to avoid using raw training sentences from non-English languages (which are naturally much more limited). Nevertheless, we believe that CS from non-English articles to English is an important next step, where prior work has shown promising results. Finally, we did experiments only with base-sized models for speed, though improvements should stand for large models as well.

Acknowledgements

We would like to thank Muhammad Idham Habibie for assisting us with the error analysis on Indonesian, as well as the Huawei Noah's Ark London NLP team for providing feedback on an earlier version of the manuscript. We also acknowledge the MindSpore team for providing technical support[13].

References

Alan Ansell, Edoardo Maria Ponti, Jonas Pfeiffer, Sebastian Ruder, Goran Glavaš, Ivan Vulić, and Anna Korhonen. 2021. MAD-G: Multilingual adapter generation for efficient cross-lingual transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4762–4781, Punta Cana,

[13] https://www.mindspore.cn/en, https://github.com/mindspore-ai

Dominican Republic. Association for Computational Linguistics.

Kelsey Ball and Dan Garrette. 2018. Part-of-speech tagging for code-switched, transliterated texts without explicit language identification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3084–3089, Brussels, Belgium. Association for Computational Linguistics.

Anab Maulana Barik, Rahmad Mahendra, and Mirna Adriani. 2019. Normalization of Indonesian-English code-mixed Twitter data. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 417–424, Hong Kong, China. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.

Iacer Calixto, Alessandro Raganato, and Tommaso Pasini. 2021. Wikipedia entities as rendezvous across languages: Grounding multilingual language models by predicting Wikipedia hyperlinks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3651–3661, Online. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020. Corpus creation for sentiment analysis in code-mixed Tamil-English text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 202–210, Marseille, France. European Language Resources association.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021a. 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.

Zewen Chi, Li Dong, Bo Zheng, Shaohan Huang, Xian-Ling Mao, Heyan Huang, and Furu Wei. 2021b. Improving pretrained cross-lingual language models via self-labeled word alignment. 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 3418–3430, Online. Association for Computational Linguistics.

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

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

Alexis Conneau, 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.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping. *arXiv preprint arXiv:2002.06305*.

A. Seza Doğruöz, Sunayana Sitaram, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2021. A survey of code-switching: Linguistic and social perspectives for language technologies. 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 1654–1666, Online. Association for Computational Linguistics.

Zi-Yi Dou, Keyi Yu, and Antonios Anastasopoulos. 2019. Investigating meta-learning algorithms for low-resource natural language understanding 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 1192–1197, Hong Kong, China. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Milan Gritta and Ignacio Iacobacci. 2021. XeroAlign: Zero-shot cross-lingual transformer alignment. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 371–381, Online. Association for Computational Linguistics.

John J Gumperz. 1977. The sociolinguistic significance of conversational code-switching. *RELC journal*, 8(2):1–34.

Deepak Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2020. A semi-supervised approach to generate the code-mixed text using pre-trained encoder and transfer learning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2267–2280, Online. Association for Computational Linguistics.

Junjie Hu, Melvin Johnson, Orhan Firat, Aditya Siddhant, and Graham Neubig. 2021. Explicit alignment objectives for multilingual bidirectional encoders. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3633–3643, Online. Association for Computational Linguistics.

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 Generalization. *ArXiv*, abs/2003.11080.

Xiaoze Jiang, Yaobo Liang, Weizhu Chen, and Nan Duan. 2022. Xlm-k: Improving Cross-lingual Language Model Pre-training with Multilingual Knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10840–10848.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020. X-FACTR: Multilingual factual knowledge retrieval from pretrained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5943–5959, Online. Association for Computational Linguistics.

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020. GLUECoS: An evaluation benchmark for code-switched NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3575–3585, Online. Association for Computational Linguistics.

Jitin Krishnan, Antonios Anastasopoulos, Hemant Purohit, and Huzefa Rangwala. 2021. Multilingual code-switching for zero-shot cross-lingual intent prediction and slot filling. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 211–223, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Linlin Liu, Bosheng Ding, Lidong Bing, Shafiq Joty, Luo Si, and Chunyan Miao. 2021. MulDA: A multilingual data augmentation framework for low-resource cross-lingual NER. 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 5834–5846, Online. Association for Computational Linguistics.

Holy Lovenia, Samuel Cahyawijaya, Genta Indra Winata, Peng Xu, Yan Xu, Zihan Liu, Rita Frieske, Tiezheng Yu, Wenliang Dai, Elham J. Barezi, and Pascale Fung. 2021. ASCEND: A spontaneous Chinese-English Dataset for Code-switching in Multi-turn Conversation. *CoRR*, abs/2112.06223.

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.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artif. Intell.*, 193:217–250.

Farhad Nooralahzadeh, Giannis Bekoulis, Johannes Bjerva, and Isabelle Augenstein. 2020. Zero-shot cross-lingual transfer with meta learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4547–4562, Online. 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.

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A.

Smith. 2019. Knowledge enhanced contextual word representations. 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 43–54, Hong Kong, China. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. 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 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Archiki Prasad, Mohammad Ali Rehan, Shreya Pathak, and Preethi Jyothi. 2021. The effectiveness of intermediate-task training for code-switched natural language understanding. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 176–190, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Mani Vegupatti, and John P. McCrae. 2020. Named Entity Recognition for Code-Mixed Indian Corpus using Meta Embedding. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 68–72.

Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. 2020. CoSDA-ML: Multi-Lingual Code-Switching Data Augmentation for Zero-Shot Cross-Lingual NLP. In *IJCAI*.

Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2020. XL-WiC: A multilingual benchmark for evaluating semantic contextualization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7193–7206, Online. Association for Computational Linguistics.

Ryokan Ri, Ikuya Yamada, and Yoshimasa Tsuruoka. 2022. mLUKE: The power of entity representations in multilingual pretrained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7316–7330, Dublin, Ireland. Association for Computational Linguistics.

Arra'Di Nur Rizal and Sara Stymne. 2020. Evaluating word embeddings for Indonesian–English code-mixed text based on synthetic data. In *Proceedings of the The 4th Workshop on Computational Approaches to Code Switching*, pages 26–35, Marseille, France. European Language Resources Association.

Mohd Sanad Zaki Rizvi, Anirudh Srinivasan, Tanuja Ganu, Monojit Choudhury, and Sunayana Sitaram. 2021. GCM: A toolkit for generating synthetic code-mixed text. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 205–211, Online. Association for Computational Linguistics.

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

Vivek Srivastava and Mayank Singh. 2020. PHINC: A parallel Hinglish social media code-mixed corpus for machine translation. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 41–49, Online. Association for Computational Linguistics.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. ERNIE: Enhanced Representation through Knowledge Integration. *ArXiv*, abs/1904.09223.

Ishan Tarunesh, Sushil Khyalia, Vishwajeet Kumar, Ganesh Ramakrishnan, and Preethi Jyothi. 2021. Meta-learning for effective multi-task and multilingual modelling. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3600–3612, Online. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Shyam Upadhyay, Manaal Faruqui, Gokhan Tür, Hakkani-Tür Dilek, and Larry Heck. 2018. (almost) Zero-Shot Cross-Lingual Spoken Language Understanding. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6034–6038.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Jianshu Ji, Guihong Cao, Daxin

Jiang, and Ming Zhou. 2021a. K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1405–1418, Online. Association for Computational Linguistics.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021b. Kepler: A unified model for knowledge embedding and pre-trained language representation. *Trans. Assoc. Comput. Linguistics*, 9:176–194.

Alexander Wettig, Tianyu Gao, Zexuan Zhong, and Danqi Chen. 2022. Should you mask 15% in masked language modeling? *arXiv preprint arXiv:2202.08005*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Rong Xiang, Mingyu Wan, Qi Su, Chu-Ren Huang, and Qin Lu. 2020. Sina Mandarin alphabetical words:a web-driven code-mixing lexical resource. 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 833–842, Suzhou, China. Association for Computational Linguistics.

Weijia Xu, Batool Haider, and Saab Mansour. 2020. End-to-end slot alignment and recognition for cross-lingual NLU. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5052–5063, Online. 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 *AAAI*.

A Language Distribution of the ENTITYCS Corpus

Figure 5 shows the histogram of code-switched entities and sentences in the ENTITYCS corpus for each languages, except English.

B Datasets

We evaluate our models on the following datasets.

WikiAnn (Pan et al., 2017) is a cross-lingual name tagging and linking dataset based on Wikipedia articles, where named entities are annotated as location (`LOC`), organisation (`ORG`) and person (`PER`) tags following the IOB2 format. The original dataset contains 282 languages. We evaluate our models on the 40 languages from WikiAnn that are included in the XTREME benchmark (Hu et al., 2020).

X-FACTR (Jiang et al., 2020) is a multilingual fact retrieval benchmark similar to LAMA (Petroni et al., 2019). It probes factual knowledge stored in pre-trained language models by prompt-based fill-in-the-blank cloze queries, covering 23 languages. X-FACTR includes both single- and multi-token entities, and two decoding methods (independent and confidence-based) are proposed.

MultiATIS++ (Xu et al., 2020) is an expansion of the Multilingual ATIS (Upadhyay et al., 2018) dataset, which includes nine languages (English, Spanish, German, French, Portuguese, Chinese, Japanese, Hindi and Turkish) from four language families (Indo-European, Sino-Tibetan, Japonic and Altaic). It contains dialogues in a single domain, Air Travel Information services. While processing the dataset, we noticed 14 examples in the test set do not have matching number of tokens and slot labels, which we ignore during evaluation.

MTOP (Li et al., 2021) is a Multilingual Task-Oriented Parsing dataset that includes six languages from 11 domains that are related to interactions with a personal assistant. We use the standard flat labels as reported in Li et al. (2021).

XL-WiC (Raganato et al., 2020) is a cross-lingual word disambiguation dataset (Word in Context), formed as a binary classification problem. Given a target word and two contexts, the goal is to identify if the word is used in the same sense in both contexts. The dataset contains both nouns and verbs as target words, covers 12 languages and was created as an extension to the English WiC dataset (Pilehvar and Camacho-Collados, 2019).

C Hyper-Parameter Settings

C.1 Intermediate Training

We use 8 NVIDIA V100 32GB GPUs for training our models on the ENTITYCS corpus, with the Hugging Face library (Wolf et al., 2020). During fine-tuning, all models were ran on a single Nvidia

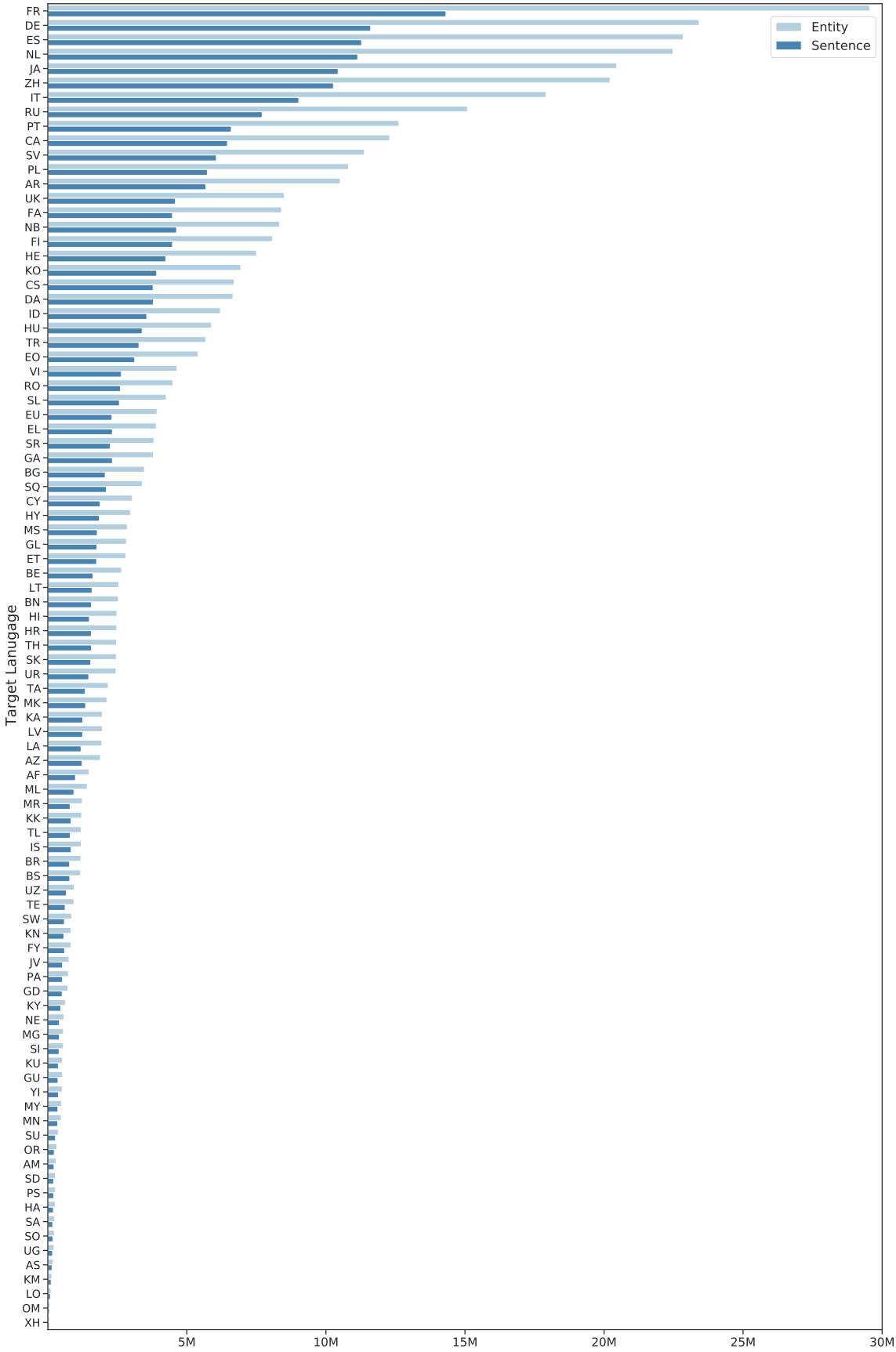


Figure 5: Number of Code-Switched Entities and Sentences in the ENTITYCS corpus.

V100 32GB GPU. We set the batch size to 16 and gradient accumulation steps to 2, resulting in an effective batch size of 256. For speedup, we employ half precision (fp16) in the experiments. In each batch we allow examples from multiple languages, based on the sampling strategy followed by Conneau and Lample (2019). We train for a single epoch with a maximum learning rate $5e^{-5}$ and linear decay scheduler, no warmup or weight decay, gradient clipping equal to 1.0, and early stopping if perplexity does not drop after 20 consecutive evaluations (we evaluate every 10K training steps).

C.2 Downstream Tasks

For downstream tasks, we evaluate models on the English validation set five times per epoch following Dodge et al. (2020). For fine-tuning XLM-R-base on WikiAnn, MultiATIS++, MTOP and XL-WiC, we fix the number of training epochs to 10, gradient clipping to 1.0 and maximum sequence length to 128. We select the batch size from $\{8, 32\}$, learning rate from $\{1e^{-5}, 2e^{-5}, 3e^{-5}, 4e^{-6}, 5e^{-6}, 6e^{-6}\}$, and warm up ratio from $\{0, 0.1\}$. The best hyper-parameters per task are reported in Table 7.

D Additional Results

We report per-language results in the following tables. WikiAnn results can be found in Table 9. X-FACTR results in Table 12, MultiATIS++ Slot Filling-only training in Table 10, and XL-WiC in Table 11.

Comparison with InfoXLM and XLM-Align described in the paper is also summarised in Table 8. These results are not included in the main table as InfoXLM and XLM-Align use parallel data, therefore the comparison is not fair.

Parameter	WIKI-ANN	MULTIATIS++		MTOP		XL-WIC
		SF	*Joint*	*SF*	*Joint*	
Learning Rate	$1e^{-5}$	$3e^{-5}$	$3e^{-5}$	$2e^{-5}$	$3e^{-5}$	$1e^{-5}$
Warmup Ratio	0.1	0.0	0.0	0.1	0.1	0.0
Batch Size	8	8	8	8	8	8

Table 7: Best hyper-parameters used for the datasets.

	Model	NER (F1)	FACT RETR. (ACC.)			SLOT FILLING (F1)		WSD (ACC.)
		WIKIANN	X-FACTR			MULTIATIS++	MTOP	XL-WIC
			all	*single*	*multi*	*SF*	*SF*	
	XLM-R$^{\text{OURS}}$	61.6 $_{0.28}$	3.5	9.4	2.6	70.6 $_{1.55}$	72.3 $_{0.98}$	59.1 $_{1.52}$
	INFOXLM (Chi et al., 2021a)	62.8	1.1	3.3	0.6	73.9 $_{1.95}$	74.7 $_{0.30}$	56.9 $_{0.81}$
	XLM-ALIGN (Chi et al., 2021b)	63.7	1.5	5.0	1.0	**74.4** $_{0.29}$	**74.9** $_{0.36}$	56.9 $_{1.22}$
EntityCS	WEP	62.4 $_{0.68}$	**6.1**	**19.4**	3.0	71.6 $_{1.20}$	73.2 $_{0.89}$	**60.4** $_{0.97}$
EntityCS	PEP$_{\text{MS}}$	63.3 $_{0.70}$	6.0	15.0	**4.3**	73.4 $_{1.70}$	71.5 $_{0.67}$	60.2 $_{0.85}$
EntityCS	PEP$_{\text{MS}}$+MLM	**64.4** $_{0.50}$	5.7	13.9	3.9	74.2 $_{0.43}$	73.0 $_{0.33}$	59.8 $_{0.75}$

Table 8: Comparison with models using parallel data.

Model	AR	HE	VI	ID	JV	MS	TL	EU	ML	TA	TE	AF	NL	EN	DE	EL	BN	HI	MR	UR
XLM-R$^{\text{OURS}}$	44.6	51.9	68.3	48.6	59.6	63.3	72.5	61.2	63.2	54.3	49.3	76.3	80.7	83.4	75.4	74.2	67.9	68.3	61.8	55.8
MLM	50.7	53.7	72.7	56.4	59.2	68.4	75.1	58.4	65.1	58.1	53.0	76.3	80.9	84.2	75.2	76.3	73.9	69.9	64.5	67.0
WEP	49.9	52.4	69.8	57.4	60.1	66.7	74.0	60.1	60.8	56.1	48.2	76.5	80.3	83.8	74.7	74.5	70.8	67.5	61.1	60.7
PEP$_{\text{MRS}}$	47.1	52.6	69.8	56.0	60.1	62.4	74.8	56.1	61.6	56.1	50.9	77.9	81.4	83.8	75.4	74.8	69.6	68.3	64.1	48.7
PEP$_{\text{M}}$	47.7	52.9	68.9	59.1	63.1	65.5	76.3	60.0	64.0	57.5	51.6	76.8	80.9	83.9	75.1	76.5	73.0	69.6	65.8	63.3
WEP+MLM	50.3	53.2	69.8	60.8	60.7	69.8	74.5	59.2	64.8	57.2	51.7	76.4	80.9	84.1	75.4	75.2	72.1	68.9	63.9	58.6
PEP$_{\text{MRS}}$+MLM	46.7	53.6	69.6	64.0	60.2	69.2	74.3	57.5	65.6	55.8	52.3	77.5	81.3	84.1	75.5	76.2	72.2	68.4	64.8	60.1
PEP$_{\text{M}}$+MLM	52.4	53.5	70.4	60.3	59.7	69.2	75.5	58.4	66.6	58.1	54.5	77.7	81.2	84.1	75.3	76.2	74.2	70.0	67.1	64.5
	FA	FR	IT	PT	ES	BG	RU	JA	KA	KO	TH	SW	YO	MY	ZH	KK	TR	ET	FI	HU
XLM-R$^{\text{OURS}}$	47.6	78.0	78.2	78.9	76.2	77.3	63.9	22.9	66.4	48.8	4.3	68.3	45.4	52.7	27.7	44.2	76.9	72.4	75.6	76.9
MLM	51.6	79.0	78.6	79.5	77.6	78.6	67.2	22.7	66.1	50.8	2.5	65.1	42.9	55.7	29.7	50.7	77.8	71.4	76.2	78.1
WEP	50.9	77.6	77.7	77.3	74.1	78.7	66.3	20.7	64.8	52.0	2.5	65.8	50.4	52.6	26.1	52.1	75.5	71.9	75.8	76.6
PEP$_{\text{MRS}}$	53.0	78.9	78.6	78.7	77.1	78.6	67.3	21.9	63.6	51.4	3.7	66.2	45.9	54.6	26.6	49.1	78.0	72.8	77.2	77.7
PEP$_{\text{M}}$	55.3	78.5	78.4	78.6	74.3	78.2	67.2	21.0	66.7	50.0	5.0	66.8	52.3	56.9	26.7	48.4	77.6	73.2	76.6	77.3
WEP+MLM	53.4	78.1	78.3	79.0	74.9	78.3	66.6	23.0	65.8	50.4	2.1	63.9	44.9	56.6	29.1	51.0	75.2	71.8	76.9	77.2
PEP$_{\text{MRS}}$+MLM	54.3	79.2	79.1	80.0	76.4	79.1	67.6	23.6	66.1	50.9	2.4	66.6	41.4	54.5	31.1	51.8	78.4	71.4	77.1	78.7
PEP$_{\text{M}}$+MLM	50.0	79.9	78.9	79.7	78.4	79.3	68.2	22.7	67.7	51.1	3.3	64.5	43.7	56.6	29.3	51.7	78.2	72.0	76.8	78.6

Table 9: F1-score per language on the WikiAnn test set. Results are averaged across five seeds.

Model	EN	ES	DE	HI	FR	PT	ZH	JA	TR
XLM-R$^{\text{OURS}}$	95.6 $_{0.15}$	81.5 $_{0.71}$	79.8 $_{2.04}$	50.6 $_{5.35}$	74.8 $_{1.90}$	76.5 $_{1.14}$	77.2 $_{2.06}$	56.8 $_{4.99}$	43.0 $_{2.72}$
MLM	95.6 $_{0.16}$	78.8 $_{2.88}$	78.0 $_{2.56}$	61.5 $_{7.26}$	74.4 $_{3.18}$	74.6 $_{1.39}$	76.4 $_{1.81}$	70.3 $_{2.00}$	39.7 $_{4.09}$
WEP	95.7 $_{0.15}$	79.9 $_{1.34}$	80.3 $_{0.58}$	52.7 $_{4.15}$	75.6 $_{0.87}$	76.3 $_{0.63}$	78.1 $_{1.43}$	60.7 $_{7.07}$	40.4 $_{4.40}$
PEP$_{\text{MS}}$	95.3 $_{0.06}$	79.3 $_{2.60}$	79.7 $_{2.28}$	62.9 $_{2.30}$	75.3 $_{2.10}$	76.2 $_{1.60}$	77.8 $_{1.30}$	69.0 $_{4.90}$	45.3 $_{2.50}$
PEP$_{\text{MS}}$+MLM	95.6 $_{0.10}$	81.3 $_{1.90}$	81.4 $_{0.90}$	65.8 $_{2.20}$	78.2 $_{0.30}$	76.1 $_{1.00}$	78.8 $_{0.60}$	68.8 $_{3.30}$	42.1 $_{3.30}$

Table 10: F1-score (average across five seeds) on MultiATIS++ Slot Filling-only training.

Model	BG	DA	ET	FA	HR	JA	KO	NL	ZH	DE	FR	IT
XLM-R$^{\text{ours}}$	57.5 $_{1.03}$	60.6 $_{2.06}$	61.7 $_{3.23}$	62.4 $_{1.05}$	61.7 $_{2.93}$	54.0 $_{1.56}$	62.4 $_{1.99}$	61.5 $_{1.94}$	56.4 $_{3.83}$	57.7 $_{1.58}$	56.4 $_{1.40}$	57.1 $_{1.35}$
MLM	59.3 $_{0.85}$	59.0 $_{0.72}$	60.6 $_{1.15}$	63.5 $_{1.41}$	62.3 $_{1.87}$	52.6 $_{1.26}$	63.1 $_{1.38}$	62.3 $_{0.76}$	52.4 $_{1.03}$	57.2 $_{0.54}$	56.6 $_{0.33}$	58.0 $_{1.09}$
WEP	59.0 $_{1.84}$	61.3 $_{1.13}$	62.2 $_{0.74}$	64.9 $_{1.06}$	63.7 $_{2.40}$	54.7 $_{2.69}$	64.6 $_{0.74}$	63.8 $_{0.77}$	55.2 $_{3.30}$	59.6 $_{1.04}$	57.0 $_{0.98}$	59.1 $_{1.03}$
PEP$_{\text{MS}}$	59.4 $_{0.93}$	60.7 $_{1.03}$	64.4 $_{1.72}$	63.5 $_{1.51}$	64.2 $_{1.85}$	53.6 $_{2.60}$	64.6 $_{2.88}$	63.9 $_{0.79}$	52.8 $_{3.01}$	59.5 $_{1.79}$	57.4 $_{0.64}$	58.8 $_{1.25}$
PEP$_{\text{MS}}$+MLM	59.7 $_{1.04}$	60.9 $_{1.20}$	63.9 $_{1.02}$	63.1 $_{1.63}$	63.8 $_{1.89}$	53.2 $_{2.18}$	62.1 $_{3.10}$	63.0 $_{1.01}$	53.2 $_{2.24}$	59.0 $_{0.90}$	57.3 $_{0.58}$	58.3 $_{1.00}$

Table 11: XL-WiC test set accuracy (average across five seeds) across languages.

MODEL				AVG	EN	FR	NL	ES	RU	ZH	HE	TR	KO	VI	EL	MR	JA	HU	BN	CEB	WAR	TL	SW	PA	MG	ILO
XLM-R^{OURS}		IND	A	3.5	8.2	4.7	4.4	6.5	5.3	4.6	2.5	3.1	5.1	8.5	6.3	2.7	2.3	0.9	0.1	1.4	1.2	2.8	3.7	0.2	1.9	0.1
			S	9.4	15.2	11.3	11.0	13.4	14.4	11.9	12.3	4.0	16.7	14.2	27.3	19.5	9.2	2.2	0.0	1.7	1.3	5.1	5.6	5.8	3.7	0.4
			M	2.1	3.3	2.3	2.6	3.3	3.8	4.5	2.2	2.5	2.6	5.1	2.9	1.1	2.1	0.2	0.1	1.0	1.1	1.4	1.9	0.0	1.6	0.0
		CONF	A	3.3	4.4	2.9	2.7	4.3	5.5	5.3	3.0	3.0	5.6	9.5	7.3	3.4	4.4	0.9	0.1	1.2	1.1	2.3	2.9	0.6	1.8	0.5
			S	7.5	5.2	4.4	3.6	4.9	14.2	11.8	11.4	3.9	15.9	12.6	25.6	18.9	8.8	2.0	0.0	1.4	1.4	4.4	4.3	5.8	3.5	0.5
			M	2.6	3.9	2.3	2.7	4.2	4.1	5.2	2.7	2.4	3.4	7.0	4.3	2.07	4.2	0.3	0.1	1.0	1.1	1.3	1.9	0.4	1.5	0.5
MLM	39	IND	A	2.3	2.1	3.7	2.9	3.9	2.9	1.9	3.4	1.2	5.0	4.6	4.2	3.6	0.3	0.7	0.0	2.1	1.0	1.4	5.2	0.0	0.0	0.1
			S	6.4	5.1	8.7	6.4	9.4	6.0	8.3	8.7	3.1	16.6	9.1	19.3	17.9	2.5	1.8	0.6	2.9	1.1	4.4	8.3	0.3	0.5	0.1
			M	1.3	0.9	2.0	2.0	1.9	2.0	1.8	1.9	0.6	2.3	2.4	2.8	2.1	0.2	0.4	0.0	1.8	1.0	0.5	2.2	0.0	0.0	0.1
		CONF	A	2.5	2.5	3.6	2.9	4.3	2.6	2.0	4.8	1.1	5.7	6.3	5.2	4.2	0.4	0.6	0.1	2.0	1.0	1.2	5.2	0.0	0.0	0.1
			S	5.9	4.9	7.6	5.9	9.0	4.4	7.6	7.4	2.5	16.1	8.5	17.2	16.7	2.5	1.6	0.6	2.8	1.1	3.9	7.8	0.3	0.5	0.0
			M	1.7	1.8	2.3	2.2	2.6	2.3	1.9	3.4	0.5	3.4	4.6	4.2	2.9	0.3	0.4	0.0	1.7	1.0	0.4	2.4	0.0	0.0	0.1
WEP	EN	IND	A	3.3	18.2	6.1	6.0	5.8	1.1	0.4	0.4	1.1	0.5	8.0	3.5	0.4	0.6	3.7	0.0	3.5	0.6	5.0	4.2	0.1	1.7	1.6
			S	8.5	38.3	16.4	18.7	14.9	4.4	3.4	1.4	5.6	2.7	16.8	7.3	4.1	2.5	8.5	0.0	6.9	2.6	9.7	10.3	0.0	6.9	5.4
			M	1.6	9.4	2.7	2.9	2.9	0.6	0.3	0.4	0.3	0.3	3.5	1.2	0.1	0.5	2.6	0.0	1.5	0.3	1.5	1.9	0.1	1.1	0.5
		CONF	A	3.1	16.2	6.4	5.6	5.4	1.1	0.3	0.4	1.1	0.5	7.6	3.4	0.4	0.6	3.6	0.0	3.4	0.5	4.5	4.1	0.1	1.3	1.7
			S	7.9	35.8	15.9	17.2	13.3	4.5	2.7	1.6	5.4	2.4	15.8	7.3	4.1	2.5	8.2	0.0	6.6	2.5	8.6	7.8	0.0	6.4	5.4
			M	1.5	7.5	3.3	2.9	2.9	0.6	0.3	0.4	0.2	0.3	3.6	1.1	0.1	0.5	2.6	0.0	1.5	0.2	1.5	2.0	0.1	1.0	0.6
	39	IND	A	6.1	15.6	9.1	11.5	10.5	2.8	6.7	3.7	3.2	6.7	13.2	7.9	4.0	4.6	6.7	0.9	4.3	2.1	7.4	7.2	0.0	2.3	3.3
			S	19.4	36.4	24.1	30.3	25.6	14.3	18.5	34.7	12.2	31.5	23.4	36.0	29.8	17.8	18.5	6.1	8.5	5.0	16.9	21.3	0.0	5.4	9.3
			M	3.0	7.2	3.9	4.9	4.6	1.5	6.3	2.6	1.0	2.9	7.3	3.9	1.5	4.1	2.5	0.0	2.2	1.4	1.8	3.3	0.0	1.4	0.6
		CONF	A	4.9	12.1	8.2	9.6	8.8	2.4	3.1	3.3	2.9	5.9	9.3	7.4	3.5	1.9	5.6	0.8	4.1	1.7	6.8	5.7	0.0	1.8	3.3
			S	17.4	32.6	22.9	26.5	23.4	12.2	16.7	32.4	11.2	28.3	19.3	34.3	27.1	15.9	16.0	5.6	8.2	4.7	14.9	17.2	0.0	5.1	9.2
			M	2.1	4.6	3.3	3.6	3.0	1.2	2.6	2.3	0.8	2.7	3.9	3.7	1.4	1.5	1.8	0.0	2.1	1.0	1.9	2.0	0.0	1.0	0.7
	93	IND	A	5.8	13.9	7.6	10.1	11.2	2.8	7.2	2.9	2.9	5.8	13.6	8.1	4.4	3.2	7.2	0.6	3.1	2.4	6.8	6.6	1.0	2.5	3.2
			S	18.5	34.5	20.0	28.9	25.3	14.0	20.1	26.0	13.0	28.6	25.4	35.0	25.6	17.3	18.2	4.9	6.7	7.2	13.6	17.6	11.6	5.7	8.3
			M	2.7	6.6	3.0	4.7	5.2	1.3	6.8	2.3	0.9	2.5	7.6	4.3	2.1	2.7	3.1	0.0	1.8	0.9	1.3	1.4	0.2	1.3	0.4
		CONF	A	4.6	11.3	6.4	8.6	9.1	2.2	2.7	2.5	2.7	4.9	10.5	7.2	3.5	1.8	6.1	0.6	2.8	2.0	6.2	5.1	0.8	1.4	2.5
			S	16.3	31.5	18.4	26.3	22.3	11.8	18.0	24.3	12.2	25.5	22.3	31.3	20.8	14.5	16.4	4.4	6.1	6.6	11.6	15.3	8.9	2.4	7.6
			M	1.8	4.7	2.1	3.5	3.4	0.9	2.3	2.0	0.8	2.1	4.9	3.5	1.5	1.4	2.2	0.0	1.6	0.7	1.1	0.9	0.1	0.6	0.4
PEP$_{MS}$	39	IND	A	4.7	15.1	6.9	11.0	9.6	5.0	3.8	3.2	2.0	7.3	9.0	5.5	3.0	3.3	1.9	0.2	3.3	1.5	5.9	5.5	0.0	0.7	0.6
			S	15.0	35.2	18.6	29.4	22.0	16.7	15.7	19.4	8.7	29.3	19.2	30.2	24.5	19.9	4.6	1.7	6.4	2.5	10.3	12.8	0.0	1.1	1.2
			M	2.4	7.1	2.4	4.5	4.2	2.1	3.5	2.6	0.7	3.3	4.4	3.5	0.5	2.7	1.0	0.0	2.0	1.2	2.4	2.8	0.0	0.5	0.4
		CONF	A	6.0	15.7	8.1	12.5	11.7	5.7	6.9	5.2	2.9	9.2	14.0	6.3	5.1	6.7	3.4	0.4	3.4	1.5	6.4	5.6	0.0	0.5	0.5
			S	13.1	31.9	17.1	27.1	19.6	12.1	13.6	17.7	7.6	26.1	16.0	27.1	21.4	16.9	3.9	1.6	5.3	2.2	9.4	9.0	0.0	0.8	1.1
			M	4.3	10.0	4.5	7.1	8.6	4.0	6.7	4.7	2.0	6.2	11.8	4.9	3.3	6.4	2.7	0.2	2.4	1.3	3.9	3.8	0.0	0.2	0.3
PEP$_{MS}$+MLM	EN	IND	A	2.6	16.8	5.0	5.2	4.9	1.5	0.2	0.6	0.2	0.6	6.3	3.0	0.4	0.6	1.0	0.0	1.2	0.4	4.5	2.3	0.6	1.8	0.5
			S	6.5	35.5	13.4	15.5	12.3	6.2	2.0	4.7	0.5	3.3	13.0	8.7	3.1	3.0	1.9	0.0	2.5	0.5	7.0	5.3	0.7	4.5	0.4
			M	1.2	6.6	2.2	2.7	2.3	1.2	0.2	0.5	0.1	0.3	2.8	0.7	0.1	0.5	0.3	0.0	0.6	0.4	1.4	1.0	0.5	1.3	0.5
		CONF	A	2.7	18.0	5.3	5.1	4.6	1.5	0.6	1.0	0.2	0.9	6.3	3.0	0.6	0.9	0.9	0.0	1.5	0.3	4.6	2.1	0.6	1.6	0.4
			S	5.7	33.1	12.0	13.3	10.2	5.8	2.0	4.4	0.5	2.7	11.1	7.3	3.1	3.0	1.7	0.0	1.7	0.3	6.7	4.6	0.6	1.0	0.3
			M	1.6	10.4	3.0	3.2	2.6	1.2	0.5	0.9	0.1	0.8	3.8	1.2	0.3	0.8	0.3	0.0	1.1	0.3	1.8	1.1	0.5	1.2	0.5
	39	IND	A	4.9	14.9	9.7	10.5	10.5	7.3	5.5	4.4	1.3	7.0	9.4	5.8	2.0	2.6	1.6	0.0	2.8	0.9	4.5	6.5	0.0	0.3	0.4
			S	13.9	34.8	24.2	27.7	23.1	20.2	16.0	17.5	5.8	28.6	19.3	25.2	13.5	15.3	4.0	2.4	5.5	1.7	8.3	11.2	0.0	0.3	0.7
			M	2.4	6.3	3.4	4.6	4.2	3.8	5.2	2.7	0.5	3.0	4.6	4.1	0.6	2.1	0.7	0.0	1.7	0.8	1.9	3.1	0.0	0.3	0.2
		CONF	A	5.7	16.3	10.5	11.0	11.3	7.9	7.0	6.2	1.3	9.2	14.2	6.6	2.8	4.2	1.5	0.3	3.4	1.3	4.1	6.6	0.0	0.3	0.3
			S	12.0	30.6	21.6	24.8	20.1	16.6	15.7	16.0	3.0	27.1	15.8	20.8	12.3	12.8	3.4	0.0	4.9	2.0	7.0	9.5	0.0	0.2	0.4
			M	3.9	9.4	5.4	6.2	6.7	5.0	6.7	4.5	0.8	6.0	11.8	5.6	1.5	3.8	0.7	0.3	2.8	1.3	2.6	4.5	0.0	0.3	0.2
	93	IND	A	4.5	14.1	8.6	9.1	8.8	5.8	5.0	3.1	0.8	6.3	9.4	5.7	1.8	1.5	1.5	0.1	2.9	1.7	4.0	5.2	0.7	2.3	0.4
			S	13.2	32.7	22.0	24.4	20.3	18.6	17.0	12.8	3.8	26.9	19.4	25.9	12.2	12.0	3.4	0.9	5.0	2.1	9.7	9.5	8.1	3.1	0.9
			M	2.3	5.9	3.3	4.4	3.8	3.1	4.7	2.6	0.4	2.8	4.4	3.8	0.4	1.1	0.7	0.0	2.2	1.5	1.2	1.3	0.0	1.9	0.2
		CONF	A	5.5	15.7	9.5	9.8	9.3	6.2	6.9	5.0	0.8	8.1	13.5	7.3	2.8	3.0	1.4	0.1	2.6	1.8	4.4	6.8	2.0	2.6	0.4
			S	11.9	31.3	19.2	22.0	16.8	15.7	16.3	12.5	3.5	25.0	16.6	23.9	11.3	8.6	2.8	0.6	4.3	2.0	9.0	9.4	7.5	2.7	0.8
			M	3.8	8.7	5.2	6.0	5.5	4.2	6.6	4.5	0.5	5.4	10.9	5.9	1.7	2.7	0.7	0.1	2.3	1.8	2.3	3.7	1.5	2.6	0.2

Table 12: X-FACTR results.