EntityCS: Improving Zero-Shot Cross-Lingual Transfer with Entity-Centric Code Switching

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

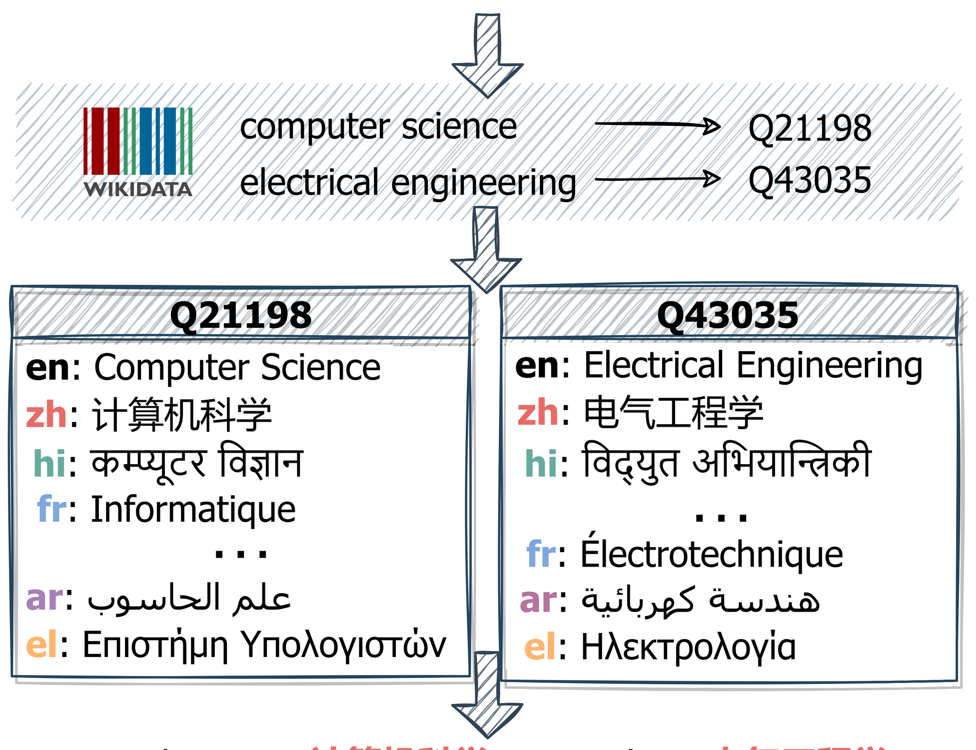
[1] City, University of London
[2] Huawei Noah's Ark Lab, London, UK

Introduction

- **Code-Switching** (CS) has proven to be an effective data augmentation method for improving cross-lingual transfer
  - Existing natural CS data usually contain only one pair of languages [6]
  - Most automatic methods use dictionaries or alignment tools which are expensive and can introduce noise [5, 8]
- We propose **EntityCS**, a method that focuses on **Entity**-level **C**ode-**S**witching to capture fine-grained cross-lingual semantics without corrupting syntax
- We construct and release an EntityCS corpus with 93 languages based on English Wikipedia and Wikidata
- We design novel **masking strategies** for **entity prediction**
- We train an XLM on the constructed EntityCS corpus with the proposed masking strategies, which shows consistent improvement on entity-centric downstream tasks

EntityCS Corpus

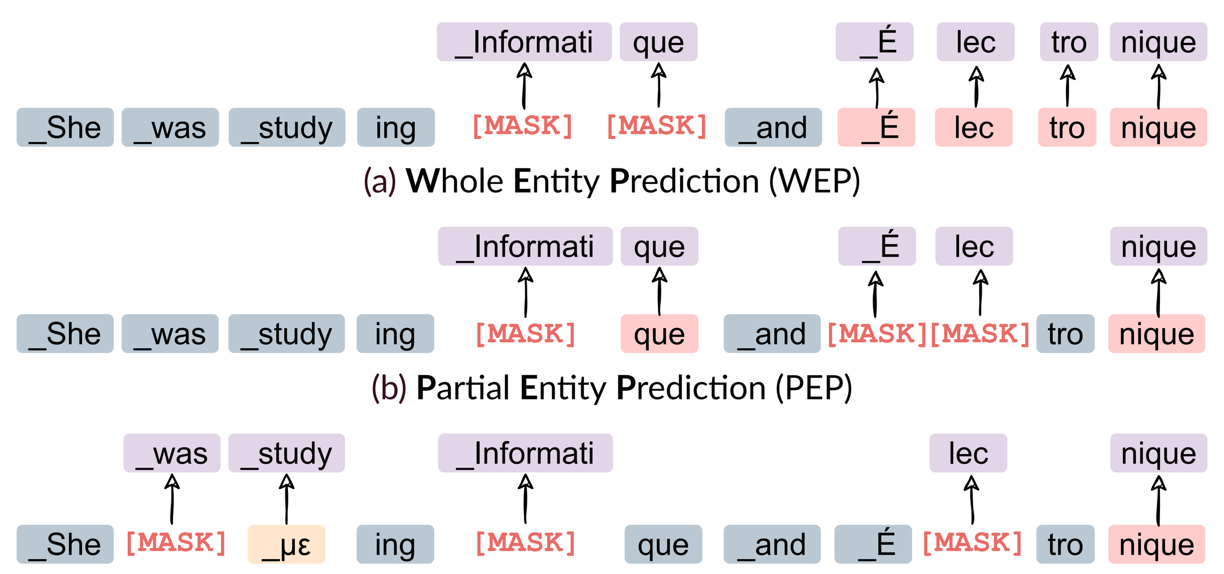
She was studying [[**computer science**]] and [[**electrical engineering**]] .



Statistics	Count
Languages	93
English Sentences	54M
English Entities	105M
Ave. Sentence Length	23.4
Ave. Entities per Sentence	2
CS Sentences per EN Sentence	<=5
CS Sentences	231M
CS Entities	421M

She was studying <e>计算机科学</e> and <e>电气工程学</e>.
She was studying <e>कम्प्यूटर विज्ञान</e> and <e>विद्युत अभियान्त्रिकी</e>.
She was studying <e>**Informatique**</e> and <e>**Électrotechnique**</e>.
She was studying <e>**computer science**</e> and <e>**electrical engineering**</e>.
...

Masking Strategies for Entity Prediction



(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)

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

- We propose **WEP** (predict every subword in an entity), **PEP** (predict partial subwords in an entity) and their combination with Masked Language Modelling (**MLM**)
- WEP is useful for predicting entire entities (single-token entity prediction), PEP benefits multi-token entity prediction, MLM helps especially when context is important
- p: probability of choosing candidate items (entity/non-entity subwords) for masking. When combining WEP/PEP with MLM, we lower p to 50%

Main Results

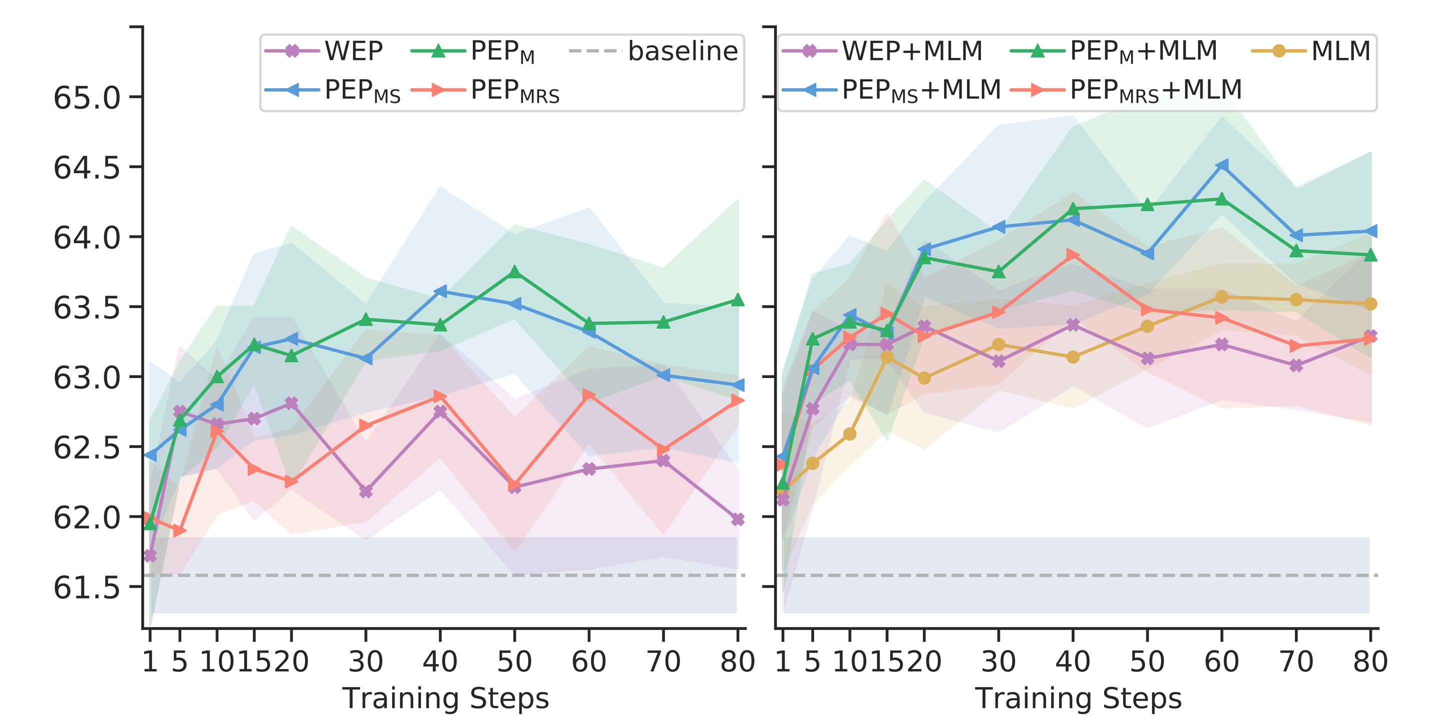
Model	NER (F1) WikiAnn [4]	Fact Retr. (Acc.) X-FACTR [3] *all*	*single*	*multi*	Slot Filling (F1, F1/Acc.) MultiATIS++ [7] *SF*	*SF / Intent*	WSD (Acc.) XL-WiC [1]
XLM-R^{PRIOR}	61.8 [2]	3.5	9.4	2.6 [3]	–	–	58.0 [1]
XLM-R^{OURS}	61.6 $_{0.28}$	3.5	9.4	2.6	71.8 $_{1.96}$	73.0 $_{0.70}$ / 89.1 $_{1.04}$	59.1 $_{1.52}$
MLM	63.5 $_{0.50}$	2.5	6.4	1.7	72.1 $_{2.34}$	74.0 $_{0.69}$ / 89.6 $_{1.43}$	59.3 $_{0.44}$
WEP	62.4 $_{0.68}$	**6.1**	**19.4**	3.0	71.6 $_{1.20}$	71.7 $_{0.82}$ / 89.7 $_{1.25}$	**60.4** $_{0.97}$
PEP_{MS}	63.3 $_{0.70}$	6.0	15.0	**4.3**	73.4 $_{1.70}$	**74.4** $_{0.67}$ / **90.0** $_{0.90}$	60.2 $_{0.85}$
PEP_{MS}+MLM	**64.4** $_{0.50}$	5.7	13.9	3.9	**74.2** $_{0.43}$	74.3 $_{0.82}$ / 89.0 $_{0.87}$	59.8 $_{0.75}$

- Average performance across languages on entity-centric tasks
  - PEP_{MS}+MLM shows the best performance on NER and Slot Filling
  - WEP is mostly beneficial for single-token fact retrieval (+10%)

MultiATIS++	Latin Script ES	DE	FR	PT	TR	*avg*	Non Latin Script ZH	JA	HI	*avg*
XLM-R^{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**

- Improvement over Latin vs. Non-Latin Script
  - We compare the performance on MultiATIS++ for demonstration
  - On average, non-Latin script languages show more improvement

Comparing Training Objectives in NER



- F1-score comparison on WikiAnn test set as a function of the number of training steps (in 10K) with various masking objectives
  - Random token replacement hurts performance
  - MLM is essential for improving NER (Left: Entity Prediction (EP) only strategies; Right: EP+MLM strategies)

Conclusions

- Our constructed EntityCS corpus and the proposed intermediate training objectives can **improve zero-shot cross-lingual transfer** of XLMs on entity-centric downstream tasks
- Our approach demonstrates **salient improvement** on languages with **Non-Latin script** compared with Latin script
- Different masking strategies are optimal under different entity prediction tasks and settings

References

[1] Raganato Alessandro et al. "XL-WiC: A Multilingual Benchmark for Evaluating Semantic Contextualization". In: *EMNLP*. 2020.
[2] Chi Zewen et al. "Improving Pretrained Cross-Lingual Language Models via Self-Labeled Word Alignment". In: *ACL-IJNLP*. 2021.
[3] Jiang Zhengbao et al. "X-FACTR: Multilingual Factual Knowledge Retrieval from Pretrained Language Models". In: *EMNLP*. 2020.
[4] Pan Xiaoman et al. "Cross-lingual Name Tagging and Linking for 282 Languages". In: *ACL*. 2017.
[5] Qin Libo et al. "CoSDA-ML: Multi-Lingual Code-Switching Data Augmentation for Zero-Shot Cross-Lingual NLP". In: *IJCAI*. 2020.
[6] Xiang Rong et al. "Sina Mandarin Alphabetical Words:A Web-driven Code-mixing Lexical Resource". In: *AACL-IJNLP*. 2020.
[7] Xu Weijia et al. "End-to-End Slot Alignment and Recognition for Cross-Lingual NLU". In: *EMNLP*. 2020.
[8] Yang Jian et al. "Alternating Language Modeling for Cross-Lingual Pre-Training". In: *AAAI*. 2020.