Wikipedia Entities as Rendezvous across Languages: Grounding Multilingual Language Models by Predicting Wikipedia Hyperlinks

Calixto, I.; Raganato, A.; Pasini, T.

DOI
10.18653/v1/2021.naacl-main.286

Publication date
2021

Document Version
Final published version

Published in
The 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies

Wikipedia Entities as *Rendezvous* across Languages: Grounding Multilingual Language Models by Predicting Wikipedia Hyperlinks

Iacer Calixto[1,2] **Alessandro Raganato**[3] **Tommaso Pasini**[4,*]
[1]Center for Data Science, New York University [2]ILLC, University of Amsterdam
[3]Department of Digital Humanities, University of Helsinki, Finland
[4]Department of Computer Science, University of Copenhagen
iacer.calixto@nyu.edu, alessandro.raganato@helsinki.fi,
tommaso.pasini@di.ku.dk

Abstract

Masked language models have quickly become the *de facto* standard when processing text. Recently, several approaches have been proposed to further enrich word representations with external knowledge sources such as knowledge graphs. However, these models are devised and evaluated in a monolingual setting only. In this work, we propose a language-independent entity prediction task as an intermediate training procedure to ground word representations on entity semantics and bridge the gap across different languages by means of a shared vocabulary of entities. We show that our approach effectively injects new lexical-semantic knowledge into neural models, improving their performance on different semantic tasks in the zero-shot crosslingual setting. As an additional advantage, our intermediate training does not require any supplementary input, allowing our models to be applied to new datasets right away. In our experiments, we use Wikipedia articles in up to 100 languages and already observe consistent gains compared to strong baselines when predicting entities using only the English Wikipedia. Further adding extra languages lead to improvements in most tasks up to a certain point, but overall we found it non-trivial to scale improvements in model transferability by training on ever increasing amounts of Wikipedia languages.

1 Introduction

Pretrained Multilingual Masked Language Models (MMLMs) such as mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020) and their variants have achieved state-of-the-art results across diverse natural language understanding tasks. Typically, a MMLM model is pretrained on very large amounts of raw text in different languages using the masked language modelling (MLM) objective and is further finetuned on (usually limited amounts of) task data. In the zero-shot crosslingual setting, which is our focus in this paper, a MMLM is finetuned on the target task using data in a single language (e.g., English) and is evaluated on the same task but in different languages (e.g., non-English languages).

We introduce the **multilingual Wikipedia hyperlink prediction** objective to contextualise words in a text with entities and concepts from an external knowledge source by using Wikipedia articles in up to 100 languages. Hyperlink prediction is a knowledge-rich task designed to (1) inject semantic knowledge from Wikipedia entities and concepts into the MMLM token representations, and (2) with a similar motivation as the translated language modelling loss of Conneau and Lample (2019), i.e., to inject explicit language-independent knowledge into a model trained via self-supervised learning, but in our case *without parallel data*. We devise a training procedure where we mask out hyperlinks in Wikipedia articles and train the MMLM to predict the hyperlink identifier similarly to standard MLM but using a "hyperlink vocabulary" of 250k concepts shared across languages.

We use the state-of-the-art MMLM XLM-R-large (Conneau et al., 2020) and show that by adding an add-on training step using Wikipedia hyperlink prediction we consistently improve several zero-shot crosslingual natural language understanding tasks across a diverse array of languages: crosslingual Word Sense Disambiguation in 18 languages including English (XL-WSD; Pasini et al., 2021); the crosslingual Word-in-Context task (XL-WiC; Raganato et al., 2020) in 12 non-English languages; and in 7 tasks from the XTREME benchmark (Hu et al., 2020) in up to 40 languages.

Recently, Zhang et al. (2019, ERNIE) and Peters et al. (2019, KnowBERT) devised different methods to incorporate entities from external knowledge graphs into masked language model (LM) training. Since then, several works followed (Wang et al., 2021; Sun et al., 2020; Xiong et al., 2020; Yamada

*Work carried out while at the University of Rome "La Sapienza".

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3651–3661
June 6–11, 2021.

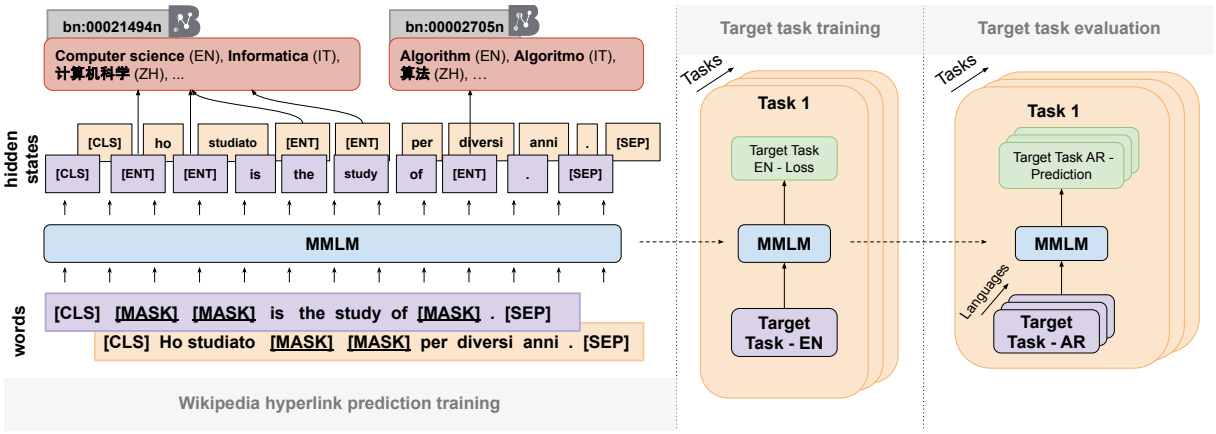

Figure 1: We finetune a pretrained MMLM using multilingual Wikipedia hyperlink prediction, then further train a model on a set of target tasks in English and evaluate on non-English data (i.e., zero-shot crosslingual setting).

et al., 2020) showing increasingly better performance than masked LMs that rely on information from raw text only. Nevertheless, all these methods were proposed for a single language[1] and cannot be easily applied to transfer learning in a zero-shot crosslingual setting.

2 Approach

Notation Let $\boldsymbol{x}_{1:m} = \text{MMLM}(x_{1:m})$ be contextualised word representations for some input text $x_{1:m}$ with m words, and computed with a pretrained MMLM. Let $\boldsymbol{x}_{n:k}$ $(n \geq 1, k \leq m)$ be a subsequence of contextualised word representations of a single hyperlink $x_{n:k}$ consisting of $k - n$ words. In our working example we use a single hyperlink $x_{n:k}$ for simplicity, but in practice there may be multiple hyperlinks in the input $x_{1:m}$.

Data We download and preprocess Wikipedia articles in 100 languages, and extract all hyperlinks in the text. We use BabelNet (Navigli and Ponzetto, 2010) — a large multilingual knowledge base comprising WordNet, Wikipedia, and many other resources — to map Wikipedia articles in different languages about the same subject onto unique identifiers. For instance, regardless of their language all "computer science" articles are mapped to the same identifier h_t, in this case `bn:00021494n`.[2] After each article is mapped to a single identifier, we create prediction targets for every hyperlink by using the identifier of its referenced article. See Appendix A for more details.

Wikipedia Hyperlink Prediction Our main goal is to use the rich semantic knowledge contained in the multilingual Wikipedias' structure to improve language model pretraining. Our approach can be seen as intermediate-task training (Phang et al., 2018, 2020) where we use Wikipedias' hyperlinks as labelled data to further finetune a pretrained MMLM model before training it one last time in the actual target task of interest. Motivated by recent studies on pretrained language encoders demonstrating that semantic features are highlighted in higher layers (Raganato and Tiedemann, 2018; Jawahar et al., 2019; Cui et al., 2020; Rogers et al., 2021), we further train only the last two layers of the MMLM. Moreover, similarly to the MLM procedure, we replace the hyperlink tokens $x_{n:k}$ by the `[MASK]` token or by a random token 80% and 10% of the time, respectively (Devlin et al., 2019).

Since the number of Wikipedia articles is very large, we only consider the most frequent 250k referenced articles h_t as possible hyperlinks in our model and we use the *adaptive softmax* activation function to speed-up training (Grave et al., 2017). Our objective allows us to consider text-entity alignments during training only. At prediction time, instead, we simply feed the model with raw text with no need of precomputed alignments. This makes our model easy to use and to adapt to many different scenarios. For more details on the model architectures and objective, see Appendix B.

3 Experimental Setup

We use **XLM-R-large** (Conneau et al., 2020) as our MMLM, which is pretrained on a large volume

[1] Mostly English, except for Sun et al. (2020) where Mandarin is also used in a monolingual setting.

[2] `https://babelnet.org/synset?word=bn:00021494n&lang=EN`

of raw multilingual corpora using MLM training.

3.1 Models

We propose three different model architectures which differ in how the input to the hyperlink classification head is computed. In *Token* we use the vector representation of each token in the hyperlink text $\boldsymbol{x}_i, i \in [n, k]$ as input to the prediction head. In *Concat CLS* we use the concatenation $[\boldsymbol{x}_i; \boldsymbol{x}_{\texttt{CLS}}]$ of the representation of each word in the hyperlink $\boldsymbol{x}_i, i \in [n, k]$ with the `[CLS]` token representation as input to the prediction head. Finally, in *Replace CLS* the input to the prediction head is the representation of each word in the hyperlink $\boldsymbol{x}_i, i \in [n, k]$ with probability p_r or the `[CLS]` token representation $\boldsymbol{x}_{\texttt{CLS}}$ with probability $1 - p_r$. More details on the architectures in Appendix B.1.

3.2 Methodology

We follow a sequential, three steps approach to training and evaluating our models. We first finetune the pretrained MMLM on the Wikipedia hyperlink prediction task, then finetune again this time on the target-task training data in English, and finally evaluate the model on non-English target-task evaluation data in a zero-shot crosslingual setting (see Figure 1). We use Wikipedia articles in different sets of languages (Section 3.3) and experiment with many diverse target tasks (Section 3.4).

3.3 Wikipedia Languages

We experiment using only English (*Wiki EN*), 15 different languages (*Wiki 15*), or 100 Wikipedia languages (*Wiki 100*). By doing that, i) we include a monolingual albeit resource-rich baseline (*Wiki EN*), ii) we investigate the impact of including a varied mixture of languages from different families (*Wiki 15*), and iii) we also experiment if going massively multilingual has a noticeable impact on crosslingual transferability (*Wiki 100*).

3.4 Target Tasks

Word Sense Disambiguation We follow the zero-shot crosslingual setting of Pasini et al. (2021, XL-WSD), which includes 17 languages plus English, i.e., we train on the English SemCor (Miller et al., 1993) dataset merged with the Princeton WordNet Gloss corpus[3] and test on all available languages (Miller et al., 1993; Raganato et al., 2017; Edmonds and Cotton, 2001; Snyder and Palmer, 2004; Pradhan et al., 2007; Navigli et al., 2007; Agirre et al., 2010; Navigli et al., 2013; Moro and Navigli, 2015; Pociello et al., 2008; Simov and Osenova, 2010; Benítez et al., 1998; Huang et al., 2010; Raffaelli et al., 2008; Pedersen et al., 2009; Postma et al., 2016; Vider and Orav, 2002; Guinovart, 2011; Mihăltz et al., 2008; Isahara et al., 2008; Yoon et al., 2009; Fišer et al., 2012).

Word-in-Context We use the crosslingual Word-in-Context dataset (XL-WiC; Raganato et al., 2020) with data in 12 diverse languages. The task is to predict whether an ambiguous word that appears in two different sentences share the same meaning. We finetune the model on the English WiC (Pilehvar and Camacho-Collados, 2019) dataset and evaluate on the 12 XL-WiC languages.

XTREME The XTREME (Hu et al., 2020) evaluation suite contains diverse tasks in up to 40 different languages. We perform crosslingual evaluation on: question answering (XQuAD; MLQA; TyDiQA; Artetxe et al., 2020; Lewis et al., 2020; Clark et al., 2020), natural language inference (XNLI; Conneau et al., 2018), paraphrase detection (PAWS-X; Yang et al., 2019), part-of-speech tagging (POS; Nivre et al., 2018), and named entity recognition (NER; Pan et al., 2017). As standard in the two unsupervised sentence retrieval tasks, BUCC (Zweigenbaum et al., 2018), and Tatoeba (Artetxe and Schwenk, 2019), XLM-R is tested considering the output of its 14-th layer, which, however, is not tuned during our intermediate task. We therefore do not report results on these tasks.[4]

Task Architectures Across all the tasks, we finetune transformer-based models by adding a classification head for each task.[5]

4 Results and Discussion

Results on XL-WSD and XL-WiC tasks (Tables 1 and 2) suggest that our models have a better grasp of word-level semantics than XLM-R, which does not have explicit semantic signals during its pretraining. This is consistent across languages and hyperlink prediction architectures, also when compared to the baseline XLM-R additionally finetuned using MLM training on in-domain Wikipedia data. Our best models outperform the baselines in both tasks by several points. Interestingly, training on

[3] http://wordnetcode.princeton.edu/glosstag.shtml

[4] More details in Appendix B.2.

[5] Details in Appendix B.1.1 (XL-WSD), B.1.2 (XL-WiC), and B.1.3 (XTREME).

	WIKI EN					WIKI 15				WIKI 100			
	XLM-R	**+MLM**	**+T**	**+C**	**+R**	**+MLM**	**+T**	**+C**	**+R**	**+MLM**	**+T**	**+C**	**+R**
EN$_{ALL}$	77.7	77.4	76.8	78.4	78.5	77.6	78.5	**78.7**	78.5	77.4	78.4	78.6	78.3
BG	72.0	71.9	72.1	72.6	70.8	71.7	73.3	**73.5**	73.1	71.8	72.9	73.2	73.4
CA	50.0	49.5	52.7	52.9	50.8	49.9	54.0	53.7	54.3	50.1	54.6	52.7	**54.7**
DA	80.6	80.4	81.7	81.7	79.9	80.6	82.4	82.5	82.4	80.7	82.4	**82.8**	82.1
DE	83.2	83.6	83.6	84.1	83.9	83.3	83.6	**85.2**	83.1	83.4	84.1	83.1	83.3
ES	75.9	76.8	78.2	78.0	75.2	76.9	78.4	**79.1**	78.2	77.3	78.2	78.1	78.5
ET	66.1	66.2	66.6	67.2	65.9	66.6	67.7	**68.4**	68.3	66.7	68.3	68.2	68.0
EU	47.2	46.3	47.7	49.0	44.4	46.4	48.7	49.2	49.4	46.1	49.7	48.7	**50.3**
FR	83.9	83.9	84.2	84.4	83.4	83.9	**84.7**	84.1	84.6	83.6	83.4	84.1	84.1
GL	66.3	65.6	67.3	68.2	63.5	66.1	69.7	69.0	**70.2**	65.3	69.3	68.7	**70.2**
HR	72.3	72.7	73.9	74.0	72.2	72.8	74.3	74.2	74.5	72.9	74.5	74.1	**74.8**
HU	67.6	68.6	70.7	70.5	67.7	68.3	71.5	71.4	**72.1**	68.8	72.0	71.1	**72.1**
IT	77.7	78.9	78.7	78.8	77.1	78.8	79.3	79.4	79.1	78.5	**79.7**	79.5	79.5
JA	61.9	62.3	67.1	67.9	65.0	62.4	68.9	68.3	**69.5**	62.3	69.0	67.1	68.4
KO	64.2	63.6	64.8	64.5	64.9	63.6	65.5	**65.7**	65.9	63.4	64.8	65.6	65.1
NL	59.2	59.8	60.5	60.5	58.3	59.7	61.6	61.2	**62.0**	59.8	61.2	61.0	61.4
SL	68.4	67.2	68.9	68.6	67.0	67.4	69.1	67.9	69.0	67.8	68.4	69.5	**69.6**
ZH	51.6	52.0	55.9	56.2	56.2	52.2	56.6	**56.8**	56.5	52.5	56.4	56.0	55.9
Avg.	65.7	65.8	67.7	68.0	66.2	65.9	68.7	68.6	**68.8**	66.0	68.6	68.3	68.7

Table 1: We report F-1 performance on the XL-WSD dataset. **Avg.** is the micro-average across all languages but English. **+MLM** is the baseline model XLM-R which we continued training with the MLM objective only, whereas **+T**, **+C**, **+R** are the *Token*, *Concat CLS* and *Replace CLS* models, respectively.

		BG	**DA**	**ET**	**FA**	**HR**	**JA**	**KO**	**NL**	**ZH**	**DE**	**FR**	**IT**	**Avg.**
XLM-R		61.8	65.2	62.6	65.8	66.9	61.7	65.6	69.2	68.3	61.1	58.8	62.2	64.1
+MLM (Wiki EN)		63.0	69.9	69.7	73.6	71.3	63.7	69.5	72.4	71.5	65.1	62.3	62.5	67.9
+MLM (Wiki 15)		64.1	67.8	68.5	73.0	70.8	62.7	66.8	72.9	69.8	64.1	61.5	64.5	67.2
+MLM (Wiki 100)		65.5	67.9	68.5	76.3	69.1	60.8	71.1	70.5	68.3	61.7	59.7	61.2	66.7
Wiki EN	**+ T**	65.3	69.6	65.6	77.4	69.4	63.2	67.9	72.6	70.5	65.4	62.4	64.2	67.8
	+ C	66.6	69.0	68.7	74.9	**74.3**	**65.9**	69.5	72.9	70.8	67.1	63.4	**66.6**	69.1
	+ R	**68.4**	68.4	69.0	75.4	73.0	65.3	68.4	73.0	69.6	66.3	62.4	64.9	68.7
Wiki 15	**+ T**	64.6	67.5	64.1	75.8	68.9	62.7	71.0	70.3	67.2	63.8	61.6	65.0	66.9
	+ C	65.0	69.5	68.7	75.3	69.6	64.3	69.9	73.4	70.1	65.6	61.9	62.5	68.0
	+ R	67.4	68.0	64.4	73.3	72.1	63.4	65.1	69.5	67.1	63.3	59.8	61.7	66.2
Wiki 100	**+ T**	66.7	69.7	**70.5**	**78.5**	67.9	64.8	**72.3**	**74.3**	70.9	**67.2**	**64.0**	65.7	**69.4**
	+ C	61.1	64.1	65.6	71.3	66.9	60.1	68.0	67.9	66.9	59.5	57.9	59.1	64.0
	+ R	65.0	**70.3**	68.2	73.0	72.1	62.0	68.5	71.7	**72.3**	65.3	61.8	63.2	67.8

Table 2: Accuracy scores on the crosslingual Word-in-Context (XL-WiC) test set.

15 languages tends to slightly outperform training on all 100 languages on XL-WSD, but on XL-WiC results with our best models trained on 100 languages outperforms all other configurations most of the time by a reasonable margin. These results corroborate our hunch that the intermediate task injects semantic knowledge within the neural model.

In Table 3, we confirm that our models preserve the sentence-level comprehension capabilities of the underlying XLM-R architecture and that it performs either comparably or favourably to the baselines in the XTREME benchmark, across target tasks and languages.

Training on the English Wikipedia only can be surprisingly effective at times (Tables 2 and 3), and training on 100 languages shows more consistent improvements only on XL-WiC but fails to lead to similar improvements on other tasks. We note that performance on XL-WSD is similar when using 15 or 100 languages, while our evaluation using XTREME shows that performance is slightly worse when using 100 languages compared to using 15 languages only. We conjecture this could be due to the fact we finetune only the last two layers of XLM-R (see Appendix B), so the model retains most of the multilingual knowledge it learned dur-

		XNLI *acc.*	PAWS-X *acc.*	POS *F1*	NER *F1*	XQuAD *F1 / EM*	MLQA *F1 / EM*	TyDiQA *F1 / EM*	Avg.
Hu et al. (2020)		79.2	86.4	72.6	65.4	76.6 / 60.8	71.6 / 53.2	65.1 / 45.0	70.1
XLM-R (Ours)		78.6	87.9	76.1	64.0	71.7 / 56.3	70.3 / 50.0	72.6 / 57.0	70.8
+MLM (Wiki EN)		79.1	87.9	76.4	62.3	70.6 / 55.2	69.3 / 50.0	72.7 / 56.8	70.4
+MLM (Wiki 15)		79.3	88.7	75.7	64.4	71.8 / 56.8	70.2 / 50.6	72.6 / 56.7	71.1
+MLM (Wiki 100)		79.2	88.0	76.0	63.4	71.5 / 56.5	70.1 / 50.5	73.5 / 57.4	70.9
Wiki EN	**+ T**	78.7	88.3	77.3	63.6	70.8 / 55.6	69.8 / 50.4	73.2 / 57.0	70.9
	+ C	79.0	87.9	76.9	63.7	71.5 / 55.7	70.3 / 50.0	73.0 / 57.2	70.9
	+ R	78.7	88.6	76.9	64.4	71.1 / 55.8	69.6 / 50.1	72.7 / 57.0	71.0
Wiki 15	**+ T**	79.0	88.1	77.2	64.1	71.3 / 56.5	70.4 / **50.6**	**73.4 / 57.8**	71.2
	+ C	**79.2**	88.4	77.3	64.7	**72.1 / 56.9**	**70.8** / 50.5	73.2 / 57.3	**71.4**
	+ R	79.1	88.3	76.7	64.7	71.5 / 56.4	70.3 / **50.6**	72.8 / 56.7	71.1
Wiki 100	**+ T**	78.6	**88.8**	76.9	**64.8**	71.7 / 56.0	70.1 / 50.0	72.7 / 56.9	71.1
	+ C	78.6	88.6	**77.6**	62.1	71.2 / 56.4	69.9 / 50.0	73.2 / 57.4	70.9
	+ R	78.8	87.6	76.7	64.2	71.1 / 56.1	69.9 / 50.3	73.1 / 57.4	70.9

Table 3: Results on different target tasks of the XTREME benchmark.

ing pretraining (Liu et al., 2019; Hao et al., 2019).

We also hypothesise that the English Wikipedia size (in number of words) and quality (in coverage of our hyperlink vocabulary) may also be a reason why training solely on English already brings large gains in transfer to other tasks. For comparison, the English Wikipedia is the one with the most data, i.e., about 73M hyperlinks, where the second highest resource language is German with only about 28M hyperlinks (see Table 4 in Appendix B). Regarding the coverage of our hyperlink vocabulary with 250k entries, the English Wikipedia covers over 249k hyperlink types at least 10 times, whereas the second highest coverage is for the French Wikipedia, which covers over 142k hyperlink types at least 10 times. We plan on investigating the effect of the size and coverage of hyperlinks further in future work.

Limitations Finally, we highlight that: (1) We report results using single model runs, therefore we have no estimates of the variance of these models; (2) We lack a more thorough hyperparameter search to further consolidate our results. In both cases, the reason we made such choices is because of the high cost of training large models such as XLM-R large.

5 Conclusions and Future work

We presented a multilingual Wikipedia hyperlink prediction intermediate task to improve the pretraining of contextualised word embedding models. We trained three model variants on different sets of languages, finding that injecting multilingual semantic knowledge consistently improves performance on several zero-shot crosslingual tasks. As future work, we plan to devise a solution to allow crosslingual transferability to scale more efficiently with the number of languages. Finally, we will investigate the impact on resource-poor vs resource-rich languages, and the effect of the size and coverage of hyperlinks in model transferability.

Acknowledgments

We would like to thank Clara Vania and Sam Bowman for comments on early versions of this work, and our three anonymous reviewers for their helpful comments and feedback.

IC has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 838188. TP and AR gratefully acknowledge the support of the ERC Consolidator Grants MOUSSE No. 726487, and FoTran No. 771113 under the European Union's Horizon 2020 research and innovation programme. AR also thanks the CSC - IT Center for Science (Finland) for the computational resources.

References

Eneko Agirre, Oier Lopez de Lacalle, Christiane Fellbaum, Shu-Kai Hsieh, Maurizio Tesconi, Monica Monachini, Piek Vossen, and Roxanne Segers. 2010. SemEval-2010 task 17: All-words word sense dis-

ambiguation on a specific domain. In *Proc. of SemEval*.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the ACL 2019*.

Giuseppe Attardi. 2015. Wikiextractor. https://github.com/attardi/wikiextractor.

Laura Benítez, Sergi Cervell, Gerard Escudero, Mònica López, German Rigau, and Mariona Taulé. 1998. Methods and tools for building the Catalan WordNet. *Proc. of ELRA Workshop on Language Resources for European Minority Languages*.

Michele Bevilacqua and Roberto Navigli. 2020. Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2854–2864, Online. Association for Computational Linguistics.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

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

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pages 7059–7069. Curran Associates, Inc.

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

Leyang Cui, Sijie Cheng, Yu Wu, and Yue Zhang. 2020. Does bert solve commonsense task via commonsense knowledge? *arXiv preprint arXiv:2008.03945*.

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.

Philip Edmonds and Scott Cotton. 2001. SENSEVAL-2: Overview. In *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 1–5. Association for Computational Linguistics.

Darja Fišer, Jernej Novak, and Tomaž Erjavec. 2012. SloWNet 3.0: development, extension and cleaning. In *Proc. of 6th International Global Wordnet Conference*.

Edouard Grave, Armand Joulin, Moustapha Cissé, David Grangier, and Hervé Jégou. 2017. Efficient softmax approximation for gpus. In *International Conference on Machine Learning (ICML)*, pages 1302–1310. PMLR.

Xavier Gómez Guinovart. 2011. Galnet: WordNet 3.0 do galego. *Linguamática*, 3(1).

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2019. Visualizing and understanding the effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4143–4152, Hong Kong, China. 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. *Proceedings of the 37th International Conference on Machine Learning*.

Chu-Ren Huang, Shu-Kai Hsieh, Jia-Fei Hong, Yun-Zhu Chen, I-Li Su, Yong-Xiang Chen, and Sheng-Wei Huang. 2010. Chinese WordNet: Design, Implementation and Application of an Infrastructure for Cross-Lingual Knowledge Processing. *Journal of Chinese Information Processing*, 24(2).

Hitoshi Isahara, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama, and Kyoko Kanzaki. 2008. Development of the Japanese WordNet. In *Sixth International conference on Language Resources and Evaluation*.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual 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 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Márton Miháltz, Csaba Hatvani, Judit Kuti, György Szarvas, János Csirik, Gábor Prószéky, and Tamás Váradi. 2008. Methods and Results of the Hungarian WordNet Project. In *Proc. of The Fourth Global WordNet Conference*.

George A Miller, Claudia Leacock, Randee Tengi, and Ross T Bunker. 1993. A semantic concordance. In *Proceedings of the workshop on Human Language Technology*, pages 303–308.

Andrea Moro and Roberto Navigli. 2015. SemEval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 288–297, Denver, Colorado.

Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. SemEval-2013 task 12: Multilingual word sense disambiguation. In *Proc. of SemEval*.

Roberto Navigli, Kenneth C. Litkowski, and Orin Hargraves. 2007. SemEval-2007 task 07: Coarse-grained English all-words task. In *Proc. of SemEval*.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden. Association for Computational Linguistics.

Joakim Nivre, Mitchell Abrams, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Mohammed Attia, et al. 2018. Universal dependencies 2.2.

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.

Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. XL-WSD: An extra-large and cross-lingual evaluation framework for word sense disambiguation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Bolette S. Pedersen, Sanni Nimb, Jørg Asmussen, Nicolai Hartvig Sørensen, Lars Trap-Jensen, and Henrik Lorentzen. 2009. DanNet: the challenge of compiling a wordnet for Danish by reusing a monolingual dictionary. *Language Resources and Evaluation*, 43.

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.

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

Jason Phang, Thibault Févry, and Samuel R Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

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.

Eli Pociello, Antton Gurrutxaga, Eneko Agirre, Izaskun Aldezabal, and German Rigau. 2008. WN-TERM: Enriching the MCR with a terminological

dictionary. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Marten Postma, Emiel van Miltenburg, Roxane Segers, Anneleen Schoen, and Piek Vossen. 2016. Open Dutch WordNet. In *Proc. of the Eight Global Wordnet Conference*.

Sameer Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. SemEval-2007 task-17: English lexical sample, SRL and all words. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 87–92, Prague, Czech Republic. Association for Computational Linguistics.

Yada Pruksachatkun, Phil Yeres, Haokun Liu, Jason Phang, Phu Mon Htut, Alex Wang, Ian Tenney, and Samuel R. Bowman. 2020. jiant: A software toolkit for research on general-purpose text understanding models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 109–117, Online.

Ida Raffaelli, Marko Tadić, Božo Bekavac, and Željko Agić. 2008. Building Croatian WordNet. In *Fourth global wordnet conference (gwc 2008)*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110, Valencia, Spain. Association for Computational Linguistics.

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.

Alessandro Raganato and Jörg Tiedemann. 2018. An analysis of encoder representations in transformer-based machine translation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 287–297, Brussels, Belgium.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Kiril Simov and Petya Osenova. 2010. Constructing of an Ontology-based Lexicon for Bulgarian. In *Proc. of LREC*.

Benjamin Snyder and Martha Palmer. 2004. The English all-words task. In *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 41–43, Barcelona, Spain. Association for Computational Linguistics.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8968–8975.

Kadri Vider and Heili Orav. 2002. Estonian WordNet and Lexicography. In *Proc. of the Eleventh International Symposium on Lexicography*.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021. Kepler: A unified model for knowledge embedding and pre-trained language representation. *TACL*.

Wenhan Xiong, Jingfei Du, William Yang Wang, and Veselin Stoyanov. 2020. Pretrained encyclopedia: Weakly supervised knowledge-pretrained language model. In *International Conference on Learning Representations*.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online. Association for Computational Linguistics.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. 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 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Ae-Sun Yoon, Soon-Hee Hwang, Eun-Ryoung Lee, and Hyuk-Chul Kwon. 2009. Construction of Korean WordNet. *Journal of KIISE: Software and Applications*, 36(1).

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2018. Overview of the third bucc shared task: Spotting parallel sentences in comparable corpora.

In *Proceedings of 11th Workshop on Building and Using Comparable Corpora*, pages 39–42.

A Wikipedia Data Details

We download the Wikipedia dump from January 11, 2020, and preprocess it using the WikiExtractor script (Attardi, 2015). We download Wikipedia articles for the following 100 languages as in Conneau and Lample (2019):[6] *af*, *als*, *am*, *ang*, *an*, *ar*, *arz*, *ast*, *az*, *bar*, *be*, *bg*, *bn*, *br*, *bs*, *ca*, *ceb*, *ckb*, *cs*, *cy*, *da*, *de*, *el*, *en*, *eo*, *es*, *et*, *eu*, *fa*, *fi*, *fr*, *fy*, *ga*, *gan*, *gl*, *gu*, *he*, *hi*, *hr*, *hu*, *hy*, *ia*, *id*, *is*, *it*, *ja*, *jv*, *ka*, *kk*, *kn*, *ko*, *ku*, *la*, *lb*, *lt*, *lv*, *mk*, *ml*, *mn*, *mr*, *ms*, *my*, *nds*, *ne*, *nl*, *nn*, *no*, *oc*, *pl*, *pt*, *ro*, *ru*, *scn*, *sco*, *sh*, *si*, *simple*, *sk*, *sl*, *sq*, *sr*, *sv*, *sw*, *ta*, *te*, *th*, *tl*, *tr*, *tt*, *uk*, *ur*, *uz*, *vi*, *war*, *wuu*, *yi*, *zh_classical*, *zh*, *zh_min_nan*, *zh_yue*.

Language sets used for training We finetune MMLM models on the Wikipedia hyperlink prediction task using articles in different sets of languages to investigate the impact of multilingualism. *Wiki EN* includes only articles in English (*en*); *Wiki 15* includes articles in *bg*, *da*, *de*, *en*, *es*, *et*, *eu*, *fa*, *fr*, *hr*, *it*, *ja*, *ko*, *nl*, *zh*; finally, *Wiki 100* includes articles in all 100 languages listed above.

Rationale *Wiki EN* is a monolingual albeit resource-rich baseline. In *Wiki 15*, we explore the impact of including languages with different amounts of data and from a mixture of different language families. In *Wiki 100*, we wish to see if going massively multilingual has a noticeable impact on our models' crosslingual transferability.

Hyperlink extraction We use BabelNet (Navigli and Ponzetto, 2010) — a large multilingual knowledge base comprising WordNet, Wikipedia, and many other resources — to map Wikipedia articles in different languages about the same subject onto unique identifiers. For instance, all "computer science" articles (e.g., *Ciencias de la computación* in Spanish, *Computer science* in English, *Informatik* in German, etc.) are mapped to the same identifier h_t, in this case `bn:00021494n`.[7] After each article is mapped to a single identifier, we create prediction targets for every hyperlink by using the identifier of its referenced article. For example, in Figure 3 the text "algorithmic processes" ($x_{n:k}$)

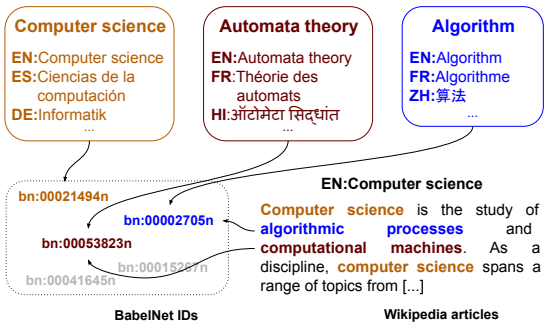

Figure 3: We show Wikipedia articles in different languages, which topics include **computer science**, **automata theory**, and **algorithm**, being mapped to BabelNet IDs. Articles on the same topic, regardless of their language, are mapped to the same identifier. Bottom-right: we show a part of the English article on **Computer science** and show two example hyperlinks and their targets h_t.

refers to the article "Algorithm",[8] which is mapped to the ID `bn:00002705n`[9] (h_t).

In Table 4 we show detailed per-language statistics for the Wikipedia data used in our experiments, including the size of the datasets and the number of hyperlinks appearing in the articles (this count already includes only the hyperlinks in our hyperlinks vocabulary of 250k types).

B Hyperparameters, Training Procedure and Model Architectures

We use XLM-R-large (Conneau et al., 2020), which has an encoder with 24 layers and a hidden state size 1024. We finetune XLM-R-large using AdamW (Kingma and Ba, 2015; Loshchilov and Hutter, 2018) with learning rate 0.00005, no weight decay, and batch size 16. We train on minibatches with maximum sequence length 256, gradient norm set to 1.0, and for 300k model updates. When finetuning XLM-R on Wikipedia hyperlink prediction, we only update the last two layers of the model.

Training data sampling We sample batches of training data from each of the languages available, i.e., depending on the experiment these can be English only, 15 languages, or 100 languages. We sample with probability $r_l = \frac{min(e_l, K)}{\sum(min(e_l, K)}$, where e_l is the number of examples per language l and the constant $K = 2^{17}$ leads to sampling more often

[6] https://github.com/facebookresearch/XLM

[7] https://babelnet.org/synset?word=bn:00021494n&lang=EN

[8] https://en.wikipedia.org/wiki/Algorithm

[9] https://babelnet.org/synset?word=bn:00002705n&lang=EN

from resource-poor languages (Raffel et al., 2020).

Adaptive softmax We collect hyperlink targets h_t from across Wikipedia articles in all the 100 languages available, sort these hyperlinks from most to least frequent, and keep only the top 250k hyperlink targets h_t. Since hyperlink frequencies follow a natural Zipfian distribution, we use the adaptive softmax activation (Grave et al., 2017) to predict hyperlinks. We bin hyperlink mentions from most to least frequent, i.e. the most frequent h_t is ranked 1st and the least frequent h_t is ranked 250k-th. We use five bins, which include hyperlinks with ranks in the following intervals: $[1, 10\text{k}]$, $(10\text{k}, 40\text{k}]$, $(40\text{k}, 50\text{k}]$, $(50\text{k}, 70\text{k}]$, $(70\text{k}, 250\text{k}]$.

The adaptive softmax activation is efficient to compute because: (1) we use one matrix multiplication for each bin, drastically reducing the number of parameters; and (2) the latter bins are only computed in case there is at least one entry in the minibatch with a target in that bin. The five-weight matrices that parameterise each bin in our adaptive softmax layer have sizes: $\texttt{hdim} \times 10,000$, $\texttt{hdim} \times 30,000$, $\texttt{hdim} \times 10,000$, $\texttt{hdim} \times 20,000$, $\texttt{hdim} \times 180,000$, respectively. Since bins are constructed so that the least frequent hyperlinks are added to the latter bins, we rarely need to compute them. This is especially important in case of the last bin, which is the most costly to compute (and is rarely used).

B.1 Model Architectures

We refer the reader for the mathematical notation in Section 2 Approach. The Wikipedia hyperlink prediction head for a single hyperlink using each of our models is shown below. **Token** is computed in Equation 1.

$$p(x_i = k) \propto \text{AdaptiveSoftmax}_k(W_t \cdot \boldsymbol{x}_i + b_t), \quad (1)$$

where AdaptiveSoftmax_k computes the probability of the hyperlink target $h_t = k$, $\boldsymbol{x}_{n:k}$ is a hyperlink consisting of words $\{\boldsymbol{x}_n, \cdots, \boldsymbol{x}_k\}$, and W_t and b_t are trained parameters.

Concat CLS is computed in Equation 2.

$$p(x_i = k) \propto \text{AdaptiveSoftmax}_k(W_c \cdot [\boldsymbol{x}_i; \boldsymbol{x}_{\text{CLS}}] + b_c), \quad (2)$$

where AdaptiveSoftmax_k computes the probability of the hyperlink target $h_t = k$, $\boldsymbol{x}_{n:k}$ is a hyperlink consisting of words $\{\boldsymbol{x}_n, \cdots, \boldsymbol{x}_k\}$, and W_c and b_c are trained parameters.

Replace CLS is computed in Equation 3.

$$\boldsymbol{x} = \text{sample}(\boldsymbol{x}_i, \boldsymbol{x}_{\text{CLS}}), \quad (3)$$
$$p(x_i = k) \propto \text{AdaptiveSoftmax}_k(W_r \cdot \boldsymbol{x} + b_r),$$

where $\text{sample}(a, b)$ samples a or b with probability 0.9 and 0.1, respectively; AdaptiveSoftmax_k computes the probability of the hyperlink target $h_t = k$, $\boldsymbol{x}_{n:k}$ is a hyperlink consisting of words $\{\boldsymbol{x}_n, \cdots, \boldsymbol{x}_k\}$, and W_r and b_r are trained parameters.

B.1.1 XL-WSD

We freeze the pretrained MMLM model weights and simply add a trained classification head on top of the pretrained MMLM. We compute representations for each subword as the sum of the last 4 layers of the model, and for each word as the average of its subword representations (Bevilacqua and Navigli, 2020).

B.1.2 XL-WiC

We follow Raganato et al. (2020) and add a binary classification head on top of the pretrained MMLM model, which takes as input the concatenation of the target words' embedding in the two contexts. We use the output of the 24-th layer as the target words' representation.

B.1.3 XTREME

We use the Jiant library (Pruksachatkun et al., 2020) to carry out the evaluation on XTREME. We use the output of the 24-th layer as the input token representations so as to better measure the impact of our intermediate training on the XTREME tasks.

B.2 XTREME Sentence Retrieval Tasks

BUCC (Zweigenbaum et al., 2018), and Tatoeba (Artetxe and Schwenk, 2019) are two unsupervised tasks requiring, given a sentence in a language L to retrieve its closest sentence in another language L'. XTREME baselines use the average of the 14-th layer outputs to represent the sentence.[10] Since our intermediate training procedure only tunes the last two layers, the output of the 14-th layer would be the exact same of the plain XLM-R baseline. For this reason, we did not report the results in both tasks.

[10] https://github.com/nyu-mll/jiant/blob/master/guides/tasks/task_specific.md

Language Code	Language	# training links	size	Language Code	Language	# training links	size
AF	Afrikaans	524,682	37M	KO	Korean	2,955,253	191M
ALS	Tosk Albanian	203,333	13M	KU	Kurdish	91,047	4,6M
AM	Amharic	35,586	2,1M	LA	Latin	791,760	29M
ANG	Anglo-Saxon	315,250	530K	LB	Luxembourgish	224,604	13M
AN	Aragonese	8,376	12M	LT	Lithuanian	1,276,418	76M
AR	Arabic	6,342,628	343M	LV	Latvian	748,963	46M
ARZ	Egyptian Arabic	1,738,581	35M	MK	Macedonian	983,105	63M
AST	Asturian	971,410	77M	ML	Malayalam	332,530	37M
AZ	Azerbaijani	786,016	53M	MN	Mongolian	90,807	6,3M
BAR	Bavarian	78,614	4,4M	MR	Marathi	169,347	13M
BE	Belarusian	1,138,871	82M	MS	Malay	1,242,850	54M
BG	Bulgarian	2,340,267	158M	MY	Burmese	48,285	5,9M
BN	Bengali	549,982	56M	NDS	Low Saxon	168,053	12M
BR	Breton	318,303	15M	NE	Nepali	66,667	5,8M
BS	Bosnian	596,758	34M	NL	Dutch	10,647,696	551M
CA	Catalan	6,180,563	395M	NN	Norwegian Nynorsk	983,245	54M
CEB	Cebuano	15,029,079	178M	NO	Norwegian	4,095,644	227M
CKB	Central Kurdish	88,593	6,1M	OC	Occitan	562,718	22M
CS	Czech	4,697,945	341M	PL	Polish	10,753,690	685M
CY	Welsh	561,936	23M	PT	Portuguese	10,065,298	581M
DA	Danish	2,273,079	135M	RO	Romanian	2,376,428	129M
DE	German	28,064,840	2,1G	RU	Russian	15,691,268	1,4G
EL	Greek	1,611,904	166M	SCN	Sicilian	64,902	3,4M
EN	English	73,084,305	4,9G	SCO	Scots	174,304	9,7M
EO	Esperanto	2,330,837	110M	SH	Serbo-Croatian	3,076,574	113M
ES	Spanish	19,125,611	1,2G	SI	Sinhala	26,321	3,5M
ET	Estonian	1,417,295	82M	SIMPLE	Simple English	1,260,400	57M
EU	Basque	1,743,033	71M	SK	Slovak	1,455,414	84M
FA	Persian	3,429,725	185M	SL	Slovenian	1,343,091	78M
FI	Finnish	3,748,928	252M	SQ	Albanian	280,756	18M
FR	French	23,415,178	1,5G	SR	Serbian	3,218,656	194M
FY	Western Frisian	423,234	26M	SV	Swedish	21,025,833	475M
GA	Ga	183,946	11M	SW	Swahili	317,669	11M
GAN	Gan Chinese	8,742	425K	TA	Tamil	691,010	54M
GL	Galician	1,632,786	108M	TE	Telugu	307,488	26M
GU	Gujarati	256,284	6,6M	TH	Thai	862,265	82M
HE	Hebrew	6,256,536	396M	TL	Tagalog	218,323	13M
HI	Hindi	546,648	45M	TR	Turkish	2,336,668	150M
HR	Croatian	1,825,455	112M	TT	Tatar	499,022	15M
HU	Hungarian	3,785,965	275M	UK	Ukrainian	7,949,672	562M
HY	Armenian	1,639,954	124M	UR	Urdu	526,498	29M
IA	Interlingua	51,882	2,4M	UZ	Uzbek	308,536	9,3M
ID	Indonesian	3,504,017	159M	VI	Vietnamese	4,877,318	221M
IS	Icelandic	252,888	17M	WAR	Waray	4,738,778	46M
IT	Italian	15,407,079	1011M	WUU	Wu Chinese	47,388	4,4M
JA	Japanese	13,318,170	949M	YI	Yiddish	85,374	5,1M
JV	Javanese	213,822	11M	ZH CLASSICAL	Classical Chinese	7,654,040	2,9M
KA	Georgian	894,180	70M	ZH	Chinese	38,104	499M
KK	Kazakh	660,065	41M	ZH MIN NAN	Min Nan	1,101,622	12M
KN	Kannada	114,724	16M	ZH YUE	Cantonese	344,432	17M

Table 4: Data statistics: total number of hyperlinks appearing in articles in Wikipedia in a given language, and size of the dataset for each language. K stands for kilobyte, M for megabyte, and G for gigabyte.