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

Iacer Calixto1,2, Alessandro Raganato3, Tommaso Pasini1,*
1 Center for Data Science, New York University  
2 ILLC, University of Amsterdam  
3 Department of Digital Humanities, University of Helsinki, Finland

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.

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...}

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.

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

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...
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 \(x_{1:m} = \text{MMLM}(x_{1:m})\) be contextualised word representations for some input text \(x_{1:m}\) with \(m\) words, and computed with a pretrained MMLM. Let \(x_{n:k}(n \geq 1, k \leq m)\) be a subsequence of contextualised word representations of a single hyperlink \(x_{w:k}\) consisting of \(k - n\) words. In our working example we use a single hyperlink \(x_{w:k}\) 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 \(x_{w:k}\) 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
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 \(x_i, i \in [n, k]\) as input to the prediction head. In Concat CLS we use the concatenation \([x_i; x_{CLS}]\) of the representation of each word in the hyperlink \(x_i, i \in [n, k]\) 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 \(x_i, i \in [n, k]\) with probability \(p_r\) or the [CLS] token representation \(x_{CLS}\) 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 fine-tune 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; Miháltz 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/glossstag.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).
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-
<table>
<thead>
<tr>
<th></th>
<th>XNLI acc.</th>
<th>PAWS-X acc.</th>
<th>POS F1</th>
<th>NER F1</th>
<th>XQuAD F1 / EM</th>
<th>MLQA F1 / EM</th>
<th>TyDiQA F1 / EM</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hu et al. (2020)</td>
<td>79.2</td>
<td>86.4</td>
<td>72.6</td>
<td>65.4</td>
<td>76.6 / 60.8</td>
<td>71.6 / 53.2</td>
<td>65.1 / 45.0</td>
<td>70.1</td>
</tr>
<tr>
<td>XLM-R (Ours)</td>
<td>78.6</td>
<td>87.9</td>
<td>76.1</td>
<td>64.0</td>
<td>71.7 / 56.3</td>
<td>70.3 / 50.0</td>
<td>72.6 / 57.0</td>
<td>70.8</td>
</tr>
<tr>
<td>+MLM (Wiki EN)</td>
<td>79.1</td>
<td>87.9</td>
<td>76.4</td>
<td>62.3</td>
<td>70.6 / 55.2</td>
<td>69.3 / 50.0</td>
<td>72.7 / 56.8</td>
<td>70.4</td>
</tr>
<tr>
<td>+MLM (Wiki 15)</td>
<td>79.3</td>
<td>88.7</td>
<td>75.7</td>
<td>64.4</td>
<td>71.8 / 56.8</td>
<td>70.2 / 50.6</td>
<td>72.6 / 56.7</td>
<td>71.1</td>
</tr>
<tr>
<td>+MLM (Wiki 100)</td>
<td>79.2</td>
<td>88.0</td>
<td>76.0</td>
<td>63.4</td>
<td>71.5 / 56.5</td>
<td>70.1 / 50.5</td>
<td>73.5 / 57.4</td>
<td>70.9</td>
</tr>
<tr>
<td>Wiki EN + T</td>
<td>78.7</td>
<td>88.3</td>
<td>77.3</td>
<td>63.6</td>
<td>70.8 / 55.6</td>
<td>69.8 / 50.4</td>
<td>73.2 / 57.0</td>
<td>70.9</td>
</tr>
<tr>
<td>Wiki EN + C</td>
<td>79.0</td>
<td>87.9</td>
<td>76.9</td>
<td>63.7</td>
<td>71.5 / 55.7</td>
<td>70.3 / 50.0</td>
<td>73.0 / 57.2</td>
<td>70.9</td>
</tr>
<tr>
<td>Wiki EN + R</td>
<td>78.7</td>
<td>88.6</td>
<td>76.9</td>
<td>64.4</td>
<td>71.1 / 55.8</td>
<td>69.9 / 50.1</td>
<td>72.7 / 57.0</td>
<td>71.0</td>
</tr>
<tr>
<td>Wiki 15 + T</td>
<td>79.0</td>
<td>88.1</td>
<td>77.2</td>
<td>64.1</td>
<td>71.3 / 56.5</td>
<td>70.4 / 50.6</td>
<td>73.4 / 57.8</td>
<td>71.2</td>
</tr>
<tr>
<td>Wiki 15 + C</td>
<td>79.2</td>
<td>88.4</td>
<td>77.3</td>
<td>64.7</td>
<td>72.1 / 56.9</td>
<td>70.8 / 50.5</td>
<td>73.2 / 57.3</td>
<td>71.4</td>
</tr>
<tr>
<td>Wiki 15 + R</td>
<td>79.1</td>
<td>88.3</td>
<td>76.7</td>
<td>64.7</td>
<td>71.5 / 56.4</td>
<td>70.3 / 50.6</td>
<td>72.8 / 56.7</td>
<td>71.1</td>
</tr>
<tr>
<td>Wiki 100 + T</td>
<td>78.6</td>
<td>88.8</td>
<td>76.9</td>
<td>64.8</td>
<td>71.7 / 56.0</td>
<td>70.1 / 50.0</td>
<td>72.7 / 56.9</td>
<td>71.1</td>
</tr>
<tr>
<td>Wiki 100 + C</td>
<td>78.6</td>
<td>88.6</td>
<td>77.6</td>
<td>62.1</td>
<td>71.2 / 56.4</td>
<td>69.9 / 50.0</td>
<td>73.2 / 57.4</td>
<td>70.9</td>
</tr>
<tr>
<td>Wiki 100 + R</td>
<td>78.8</td>
<td>87.6</td>
<td>76.7</td>
<td>64.2</td>
<td>71.1 / 56.1</td>
<td>69.9 / 50.3</td>
<td>73.1 / 57.4</td>
<td>70.9</td>
</tr>
</tbody>
</table>

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

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-
ambiguity on a specific domain. In Proc. of SemEval.


Eli Pocicello, Antton Gurrutxaga, Eneko Agirre, Iza- skun Aldezabal, and German Rigau. 2008. WNet-TERM: Enriching the MCR with a terminological


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):

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, zh-min-xia, zh-yue.

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 $h_l$. 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” $(x_{nk})$ refers to the article “Algorithm”, which is mapped to the ID bn:00002705n.

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 $r_l = \frac{\min(e_l, K)}{\sum_l \min(e_l, K)}$, where $e_l$ is the number of examples per language $l$ and the constant $K = 2^{17}$ leads to sampling more often
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: \([1, 10k], (10k, 40k], (40k, 50k], (50k, 70k], (70k, 250k]\).

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: \( \text{hdim} \times 10,000, \text{hdim} \times 30,000, \text{hdim} \times 10,000, \text{hdim} \times 20,000, \text{hdim} \times 180,000 \), 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.

\[
p(x_i = k) \propto \text{AdaptiveSoftmax}_k(W_t \cdot x_i + b_t),
\]

where \( \text{AdaptiveSoftmax}_k \) computes the probability of the hyperlink target \( h_t = k, x_{n:k} \) is a hyperlink consisting of words \( \{x_n, \cdots, x_k\} \), and \( W_t \) and \( b_t \) are trained parameters.

Concat CLS is computed in Equation 2.

\[
p(x_i = k) \propto \text{AdaptiveSoftmax}_k(W_c \cdot [x_i; x_{\text{CLS}}] + b_c),
\]

where \( \text{AdaptiveSoftmax}_k \) computes the probability of the hyperlink target \( h_t = k, x_{n:k} \) is a hyperlink consisting of words \( \{x_n, \cdots, x_k\} \), and \( W_c \) and \( b_c \) are trained parameters.

Replace CLS is computed in Equation 3.

\[
x = \text{sample}(x_i, x_{\text{CLS}}),
\]

\[
p(x_i = k) \propto \text{AdaptiveSoftmax}_k(W_r \cdot x + b_r),
\]

where \( \text{sample}(a, b) \) samples \( a \) or \( b \) with probability 0.9 and 0.1, respectively; \( \text{AdaptiveSoftmax}_k \) computes the probability of the hyperlink target \( h_t = k, x_{n:k} \) is a hyperlink consisting of words \( \{x_n, \cdots, x_k\} \), 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 andNavigli, 2020).

B.1.2 XL-WiC

We follow Raganato et al. (2020) and add a binary classification head on top of the pretrained MMLM, 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
<table>
<thead>
<tr>
<th>Language Code</th>
<th>Language</th>
<th># training links</th>
<th>size</th>
<th>Language Code</th>
<th>Language</th>
<th># training links</th>
<th>size</th>
</tr>
</thead>
<tbody>
<tr>
<td>AF</td>
<td>Afrikaans</td>
<td>524,682</td>
<td>37M</td>
<td>KO</td>
<td>Korean</td>
<td>2,955,253</td>
<td>191M</td>
</tr>
<tr>
<td>ALS</td>
<td>Tosk Albanian</td>
<td>203,333</td>
<td>13M</td>
<td>KU</td>
<td>Kurdish</td>
<td>91,047</td>
<td>46M</td>
</tr>
<tr>
<td>AM</td>
<td>Amharic</td>
<td>35,586</td>
<td>2.1M</td>
<td>LA</td>
<td>Latin</td>
<td>791,760</td>
<td>29M</td>
</tr>
<tr>
<td>ANG</td>
<td>Anglo-Saxon</td>
<td>315,250</td>
<td>530K</td>
<td>LB</td>
<td>Luxembourgish</td>
<td>224,604</td>
<td>13M</td>
</tr>
<tr>
<td>AN</td>
<td>Aragonese</td>
<td>8,376</td>
<td>12M</td>
<td>LT</td>
<td>Lithuanian</td>
<td>1,276,418</td>
<td>76M</td>
</tr>
<tr>
<td>AR</td>
<td>Arabic</td>
<td>6,342,628</td>
<td>343M</td>
<td>LV</td>
<td>Latvian</td>
<td>748,963</td>
<td>46M</td>
</tr>
<tr>
<td>ARZ</td>
<td>Egyptian Arabic</td>
<td>1,738,581</td>
<td>35M</td>
<td>MK</td>
<td>Macedonian</td>
<td>983,105</td>
<td>63M</td>
</tr>
<tr>
<td>AST</td>
<td>Asturian</td>
<td>971,410</td>
<td>77M</td>
<td>ML</td>
<td>Malayalam</td>
<td>332,530</td>
<td>37M</td>
</tr>
<tr>
<td>AZ</td>
<td>Azerbaijani</td>
<td>786,016</td>
<td>53M</td>
<td>MN</td>
<td>Mongolian</td>
<td>90,807</td>
<td>63M</td>
</tr>
<tr>
<td>BAR</td>
<td>Bavarian</td>
<td>78,614</td>
<td>4.4M</td>
<td>MR</td>
<td>Marathi</td>
<td>169,347</td>
<td>13M</td>
</tr>
<tr>
<td>BE</td>
<td>Belarusian</td>
<td>1,138,871</td>
<td>82M</td>
<td>MS</td>
<td>Malay</td>
<td>1,242,850</td>
<td>54M</td>
</tr>
<tr>
<td>BG</td>
<td>Bulgarian</td>
<td>2,340,267</td>
<td>158M</td>
<td>MY</td>
<td>Burmese</td>
<td>48,285</td>
<td>5.9M</td>
</tr>
<tr>
<td>BN</td>
<td>Bengali</td>
<td>549,982</td>
<td>56M</td>
<td>NDS</td>
<td>Low Saxon</td>
<td>168,053</td>
<td>12M</td>
</tr>
<tr>
<td>BR</td>
<td>Breton</td>
<td>318,303</td>
<td>15M</td>
<td>NE</td>
<td>Nepali</td>
<td>66,667</td>
<td>5.8M</td>
</tr>
<tr>
<td>BS</td>
<td>Bosnian</td>
<td>596,758</td>
<td>34M</td>
<td>NL</td>
<td>Dutch</td>
<td>10,647,696</td>
<td>551M</td>
</tr>
<tr>
<td>CA</td>
<td>Catalan</td>
<td>6,180,563</td>
<td>395M</td>
<td>NL</td>
<td>Norwegian Nynorsk</td>
<td>983,245</td>
<td>54M</td>
</tr>
<tr>
<td>CEB</td>
<td>Cebuano</td>
<td>15,029,079</td>
<td>178M</td>
<td>NO</td>
<td>Norwegian</td>
<td>4,095,644</td>
<td>227M</td>
</tr>
<tr>
<td>CKB</td>
<td>Central Kurdish</td>
<td>88,593</td>
<td>6.1M</td>
<td>OC</td>
<td>Occitan</td>
<td>562,718</td>
<td>22M</td>
</tr>
<tr>
<td>CS</td>
<td>Czech</td>
<td>4,697,945</td>
<td>341M</td>
<td>PL</td>
<td>Polish</td>
<td>10,753,690</td>
<td>685M</td>
</tr>
<tr>
<td>CY</td>
<td>Welsh</td>
<td>561,936</td>
<td>23M</td>
<td>PT</td>
<td>Portuguese</td>
<td>10,065,298</td>
<td>581M</td>
</tr>
<tr>
<td>DA</td>
<td>Danish</td>
<td>2,273,079</td>
<td>135M</td>
<td>RO</td>
<td>Romanian</td>
<td>2,376,428</td>
<td>129M</td>
</tr>
<tr>
<td>DE</td>
<td>German</td>
<td>28,064,840</td>
<td>2.1G</td>
<td>RU</td>
<td>Russian</td>
<td>15,691,268</td>
<td>1.4G</td>
</tr>
<tr>
<td>EL</td>
<td>Greek</td>
<td>1,611,904</td>
<td>166M</td>
<td>SCN</td>
<td>Sicilian</td>
<td>64,902</td>
<td>3.4M</td>
</tr>
<tr>
<td>EN</td>
<td>English</td>
<td>73,084,305</td>
<td>4.9G</td>
<td>SCO</td>
<td>Scots</td>
<td>174,304</td>
<td>9.7M</td>
</tr>
<tr>
<td>EO</td>
<td>Esperanto</td>
<td>2,330,837</td>
<td>110M</td>
<td>SH</td>
<td>Serbo-Croatian</td>
<td>3,076,574</td>
<td>113M</td>
</tr>
<tr>
<td>ES</td>
<td>Spanish</td>
<td>19,125,611</td>
<td>1.2G</td>
<td>SI</td>
<td>Sinhala</td>
<td>26,321</td>
<td>3.5M</td>
</tr>
<tr>
<td>ET</td>
<td>Estonian</td>
<td>1,417,295</td>
<td>82M</td>
<td>SIMPLE</td>
<td>Simple English</td>
<td>1,260,400</td>
<td>57M</td>
</tr>
<tr>
<td>EU</td>
<td>Basque</td>
<td>1,743,033</td>
<td>71M</td>
<td>SK</td>
<td>Slovak</td>
<td>1,455,414</td>
<td>84M</td>
</tr>
<tr>
<td>FA</td>
<td>Persian</td>
<td>3,429,725</td>
<td>185M</td>
<td>SL</td>
<td>Slovenian</td>
<td>1,343,091</td>
<td>78M</td>
</tr>
<tr>
<td>FI</td>
<td>Finnish</td>
<td>3,748,928</td>
<td>252M</td>
<td>SQ</td>
<td>Albanian</td>
<td>280,756</td>
<td>18M</td>
</tr>
<tr>
<td>FR</td>
<td>French</td>
<td>23,415,178</td>
<td>1.5G</td>
<td>SR</td>
<td>Serbian</td>
<td>3,218,656</td>
<td>194M</td>
</tr>
<tr>
<td>FY</td>
<td>Western Frisian</td>
<td>423,234</td>
<td>26M</td>
<td>SV</td>
<td>Swedish</td>
<td>21,025,833</td>
<td>475M</td>
</tr>
<tr>
<td>GA</td>
<td>Ga</td>
<td>183,946</td>
<td>11M</td>
<td>SW</td>
<td>Swahili</td>
<td>317,669</td>
<td>11M</td>
</tr>
<tr>
<td>GAN</td>
<td>Gan Chinese</td>
<td>8,742</td>
<td>425K</td>
<td>TA</td>
<td>Tamil</td>
<td>691,010</td>
<td>54M</td>
</tr>
<tr>
<td>GL</td>
<td>Galician</td>
<td>1,632,786</td>
<td>108M</td>
<td>TE</td>
<td>Telugu</td>
<td>307,488</td>
<td>26M</td>
</tr>
<tr>
<td>GU</td>
<td>Gujarati</td>
<td>256,284</td>
<td>6.6M</td>
<td>TH</td>
<td>Thai</td>
<td>862,265</td>
<td>82M</td>
</tr>
<tr>
<td>HE</td>
<td>Hebrew</td>
<td>6,256,536</td>
<td>396M</td>
<td>TL</td>
<td>Tagalog</td>
<td>218,323</td>
<td>13M</td>
</tr>
<tr>
<td>HI</td>
<td>Hindi</td>
<td>546,648</td>
<td>45M</td>
<td>TR</td>
<td>Turkish</td>
<td>2,336,668</td>
<td>150M</td>
</tr>
<tr>
<td>HR</td>
<td>Croatian</td>
<td>1,825,455</td>
<td>112M</td>
<td>TT</td>
<td>Tatar</td>
<td>499,022</td>
<td>15M</td>
</tr>
<tr>
<td>HU</td>
<td>Hungarian</td>
<td>3,785,965</td>
<td>275M</td>
<td>UK</td>
<td>Ukrainian</td>
<td>7,949,672</td>
<td>562M</td>
</tr>
<tr>
<td>HY</td>
<td>Armenian</td>
<td>1,639,934</td>
<td>124M</td>
<td>UR</td>
<td>Urdu</td>
<td>526,498</td>
<td>29M</td>
</tr>
<tr>
<td>IA</td>
<td>Interlingua</td>
<td>51,882</td>
<td>2.4M</td>
<td>UZ</td>
<td>Uzbek</td>
<td>308,536</td>
<td>9.3M</td>
</tr>
<tr>
<td>ID</td>
<td>Indonesian</td>
<td>3,504,017</td>
<td>159M</td>
<td>VI</td>
<td>Vietnamese</td>
<td>4,877,318</td>
<td>221M</td>
</tr>
<tr>
<td>IS</td>
<td>Icelandic</td>
<td>252,888</td>
<td>17M</td>
<td>WAR</td>
<td>Waray</td>
<td>4,738,778</td>
<td>46M</td>
</tr>
<tr>
<td>IT</td>
<td>Italian</td>
<td>15,407,079</td>
<td>1011M</td>
<td>WUW</td>
<td>Wu Chinese</td>
<td>47,388</td>
<td>4.4M</td>
</tr>
<tr>
<td>JA</td>
<td>Japanese</td>
<td>13,318,170</td>
<td>949M</td>
<td>YI</td>
<td>Yiddish</td>
<td>85,374</td>
<td>5.1M</td>
</tr>
<tr>
<td>JUV</td>
<td>Javanese</td>
<td>213,822</td>
<td>11M</td>
<td>ZH CLASSICAL</td>
<td>Classical Chinese</td>
<td>7,654,040</td>
<td>2.9M</td>
</tr>
<tr>
<td>KA</td>
<td>Georgian</td>
<td>894,180</td>
<td>70M</td>
<td>ZH</td>
<td>Chinese</td>
<td>38,104</td>
<td>499M</td>
</tr>
<tr>
<td>KK</td>
<td>Kazakh</td>
<td>660,065</td>
<td>41M</td>
<td>ZH MIN NAM</td>
<td>Min Nan</td>
<td>1,101,622</td>
<td>12M</td>
</tr>
<tr>
<td>KN</td>
<td>Kannada</td>
<td>114,724</td>
<td>16M</td>
<td>ZH YUE</td>
<td>Cantonese</td>
<td>344,432</td>
<td>17M</td>
</tr>
</tbody>
</table>

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.