# AM²ɪCᴏ: Evaluating Word Meaning in Context across Low-Resource Languages with Adversarial Examples

**Qianchu Liu[1], Edoardo M. Ponti[1 2], Diana McCarthy[1], Ivan Vulić[1], Anna Korhonen[1]**

[1] Language Technology Lab, TAL, University of Cambridge, UK
[2] Mila/McGill University, Montreal, Canada
{ql261,ep490,iv250,alk23}@cam.ac.uk
diana@dianamccarthy.co.uk

## Abstract

Capturing word meaning in context and distinguishing between correspondences and variations across languages is key to building successful multilingual and cross-lingual text representation models. However, existing multilingual evaluation datasets that evaluate lexical semantics "in-context" have various limitations. In particular, **1)** their language coverage is restricted to high-resource languages and skewed in favor of only a few language families and areas, **2)** a design that makes the task solvable via superficial cues, which results in artificially inflated (and sometimes super-human) performances of pretrained encoders, and **3)** no support for cross-lingual evaluation. In order to address these gaps, we present AM²ɪCᴏ (**A**dversarial and **M**ultilingual **M**eaning **i**n **Co**ntext), a wide-coverage cross-lingual and multilingual evaluation set; it aims to faithfully assess the ability of state-of-the-art (SotA) representation models to understand the identity of word meaning in cross-lingual contexts for 14 language pairs. We conduct a series of experiments in a wide range of setups and demonstrate the challenging nature of AM²ɪCᴏ. The results reveal that current SotA pretrained encoders substantially lag behind human performance, and the largest gaps are observed for low-resource languages and languages dissimilar to English.

## 1 Introduction

Pretrained language models (LMs) such as BERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) offer a natural way to distinguish different word meanings in context without performing explicit sense disambiguation. This property of "meaning contextualization" is typically evaluated either via standard entity linking (Rao et al., 2013; Shen et al., 2014) and Word Sense Disambiguation (WSD) tasks (Navigli, 2009; Moro et al., 2014; Raganato et al., 2017) or, recently, via the Word-in-Context (WiC) evaluation paradigm (Pilehvar and Camacho-Collados, 2019; Raganato et al., 2020).

Although monolingual evaluation in English is still predominant, a need has been recognized to construct similar resources for other languages to support cross-lingual evaluation and model diagnostics. This includes multilingual and cross-lingual WSD benchmarks (Navigli and Ponzetto, 2012; Navigli et al., 2013; Scarlini et al., 2020; Barba et al., 2020, *inter alia*), cross-lingual entity linking (Tsai and Roth, 2016; Raiman and Raiman, 2018; Upadhyay et al., 2018) and, most recently, multilingual WiC (termed XL-WiC) spanning 12 languages (Raganato et al., 2020).

This most recent WiC evaluation approach is particularly attractive as **1)** it bypasses the dependence on modeling predefined ontologies (entity linking) and explicit sense inventories (WSD), and **2)** it is framed as a simple binary classification task: for a target word $w$ appearing in two different contexts $c_1$ and $c_2$, the system must decide whether $w$ conveys the same meaning in both contexts, or not.

However, the current WiC evaluation still allows ample room for improvement: **1)** current language coverage is limited, and biased towards resource-rich Indo-European languages; **2)** coverage of lexical concepts, due to their paucity in language-specific WordNets, is also limited; **3)** XL-WiC is a monolingual resource available in different languages, i.e., it does not support cross-lingual assessments. Further, **4)** the current WiC datasets offer low human upper bounds and inflated (even super-human) performance for some languages.[1] This is due to superficial cues where **5)** many examples in the current WiC datasets can be resolved relying either on the target word alone without any context or on the context alone, which eludes evaluation honing in on the interplay between target words and their corresponding contexts.

In order to address these limitations and provide

---

[1] In turn, this might give a false impression that some languages in the task are 'solved' by the current pretrained LMs.

a more comprehensive evaluation framework, we present **AM$^2$ICo** (**A**dversarial and **M**ultilingual **M**eaning **in** **Co**ntext), a novel multilingual and cross-lingual WiC task and resource. It covers a typologically diverse set of 15 languages, see Table 2. Based on Wikipedia in lieu of WordNet, AM$^2$ICo covers a wider set of ambiguous words which especially complements WiC on the long tail of entity names and adds challenges in generalization for a larger vocabulary than a restricted set of common words. More importantly, the use of Wikipedia enables WiC evaluation on low-resource languages (e.g., Basque, Georgian, Bengali, Kazakh). We also improve the WiC resource design; it now **1)** includes adversarial examples and careful data extraction procedures to prevent the models from backing off to superficial clues, **2)** results in a more challenging benchmark with truer and much wider gaps between current SotA pretrained encoders and human capability (see §2.3), and **3)** enables cross-lingual evaluation and analysis.

The ample and diverse data in AM$^2$ICo enables a wide spectrum of experiments and analyses in different scenarios. We evaluate SotA pretrained encoders, multilingual BERT and XLM-R, both off-the-shelf using a metric-based approach (i.e., without any task adaptation) and after task-specific fine-tuning. With fine-tuned models, we investigate zero-shot cross-lingual transfer as well as transfer from multiple source languages. In general, our results across these diverse scenarios firmly indicate a large gap between human and system performance across the board, which is even more prominent when dealing with resource-poor languages and languages dissimilar to English, holding promise to guide modeling improvements in the future.

In the hope that AM$^2$ICo will be a challenging and valuable diagnostic and evaluation asset for future work in multilingual and cross-lingual representation learning, we release the data along with the full guidelines at `https://github.com/cambridgeltl/AM2iCo`.

## 2  AM$^2$ICo: Cross-Lingual Word-in-Context Evaluation

**Task Definition.** AM$^2$ICo is a standard binary classification task on pairs of word in context instances. Each pair consists of a target word with its context in English and a target word with its context in a target language. Formally, each dataset of AM$^2$ICo spans a set of $N$ examples $\hat{x}_i$,

$i = 1, \ldots, N$ for a language pair. Each example $\hat{x}_i$ is in fact a pair of items $\hat{x}_i = (x_{i,src}, x_{i,trg})$, where the item $x_{i,src}$ is provided in the source language $L_{src}$ and the item $x_{i,trg}$ is in the target language $L_{trg}$. The item $x_{i,src}$ in turn is another pair $x_{i,src} = (w_{i,src}, c_{i,src})$; it contains a target word $w_{i,src}$ from $L_{src}$ and its (wider) context $c_{i,src}$ (also in $L_{src}$) in which that word appears, see Table 1; the same is valid for $x_{i,trg}$. The classification task is then to judge whether the words $w_{i,src}$ and $w_{i,trg}$ occurring in the respective contexts $c_{i,src}$ and $c_{i,trg}$ have the same sense/meaning (i.e., whether they refer to the same entity/concept), or not.

**Final Resource.** The full AM$^2$ICo resource comprises datasets for 14 language pairs, where English is paired with 14 target languages. For brevity, in the rest of the paper we refer to the dataset of each language pair simply with the $L_{trg}$ language code (e.g., ZH instead of EN-ZH); languages and codes are provided in Table 2.

As illustrative examples, we show a positive pair (label 'T') and a negative pair (label 'F') from the ZH AM$^2$ICo dataset in Table 1 (Examples 1 and 2). In the positive example, both target words 'Apollo' and '阿波罗' in their contexts refer to the same concept: the Apollo spaceflight program. In the negative example, the Chinese target word '阿波罗' refers to the Apollo aircraft, but the English target word 'Apollo' now refers to the Greek God.

In what follows we describe the creation of AM$^2$ICo. We also demonstrate the benefits of AM$^2$ICo and its challenging nature.

### 2.1  Data Creation

Wikipedia is a rich source of disambiguated contexts for multiple languages. The availability of Wikipedia's cross-lingual links provides a direct way to identify cross-lingual concept correspondence. The items $x_{i,src}$ and $x_{i,trg}$ are then extracted by taking the surrounding (sentential) context of a hyperlinked word in a Wikipedia article. We balance the context length by (i) discarding items longer than 100 words, and (ii) adding preceding and following sentences to the context for sentences shorter than 30 words. Using the Wikipedia dumps of our 15 languages (see Table 2), we create monolingual items $x$ for each language. We select only *ambiguous* target words ($w$-s), that is, words

| no. | English $x_{i,src}$ | Chinese $x_{i,trg}$ | Label |
|---|---|---|---|
| 1 | Bill Kaysing ( July 31 , 1922 – April 21 , 2005 ) was an American writer who claimed that the six **Apollo** Moon landings between July 1969 and December 1972 were hoaxes , and so a founder of the Moon hoax movement . | 泰坦系列导弹的发射任务结束后，LC-16被移交给NASA用做双子座计划的航天员训练及 **阿波罗** 中飞船服务舱的静态试车。[...] *(After the launch of the Titan missiles, LC-16 was handed over to NASA for the training of the astronauts in the Gemini program and the static test run of the service module of the spacecraft in* **Apollo** *[...])* | T |
| 2 | Nearer the house , screening the service wing from view , is a Roman triumphal arch , the " Temple of **Apollo** " , also known ( because of its former use a venue for cock fighting ) as " Cockpit Arch " , which holds a copy of the famed Apollo Belvedere . | 阿波罗-联盟测试计划中，美国的 **阿波罗** 航天器和苏联的联盟航天器在地球轨道中对接。... *(In the Apollo-Soyuz test plan, America's* **Apollo** *spacecraft and the Soviet Union's Soyuz spacecraft are docked in the Earth orbit [...])* | F |
| 3 | Bill Kaysing ( July 31 , 1922 – April 21 , 2005 ) was an American writer who claimed that the six **Apollo** Moon landings between July 1969 and December 1972 were hoax es , and so a founder of the Moon hoax movement . | 泰坦系列导弹的发射任务结束后，LC-16被移交给 **NASA** 用做双子座计划的航天员训练及阿波罗中飞船服务舱的静态试车。[...] *(After the launch of the Titan missiles, LC-16 was handed over to* **NASA** *for the training of the astronauts in the Gemini program and the static test run of the service module of the spacecraft in Apollo [...])* | F |

Table 1: Positive (1), negative (2) and adversarial negative examples (3) from ZH AM²ICO. Target words are provided in boldface and with a gray background. Translations of the ZH items are provided in italic.

| | DE | RU | JA | ZH | AR | KO | FI | TR | ID | EU | KA | BN | KK | UR |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| train | 50,000 | 28,286 | 16,142 | 13,154 | 9,622 | 7,070 | 6,322 | 3,904 | 1,598 | 978 | – | – | – | – |
| dev | 500 | 500 | 500 | 500 | 500 | 500 | 500 | 500 | 500 | 500 | 500 | 332 | 276 | 108 |
| test | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 700 | 400 | 400 |

Table 2: Data sizes for AM²ICO across 14 language pairs. We also provide larger dev and test sets for DE and RU spanning 5,000 and 10,000 examples, respectively. **EN**=English; **DE**=German; **RU**=Russian; **JA**=Japanese; **ZH**=Chinese; **AR**=Arabic; **KO**=Korean; **FI**=Finnish; **TR**=Turkish; **ID**=Indonesian; **EU**=Basque; **KA**=Georgian; **BN**=Bengali, **KK**=Kazakh; **UR**=Urdu.

that link to at least two different Wikipedia pages.[2]

For each word, we then create monolingual positive examples by pairing two items (i.e., word-context pairs) $x_i$ and $x_j$ in which the same target word $w$ is linked to the same Wikipedia page, signaling the same meaning. In a similar fashion, monolingual negative examples are created by pairing two items where the same target word $w$ is linked to two different Wikipedia pages. We ensure that there is roughly an equal number of positive and negative examples for each target word.

Now, each monolingual example (i.e., pair of items) $\hat{x}$ contains the same word occurring in two different contexts. In order to create a cross-lingual dataset, we leverage the Wikipedia cross-lingual links; we simply (i) replace one of the two items from each English pair with an item in the target

language, and (ii) replace one of the two items from each target language pair with an English item, where the cross-lingual replacements point to the same Wikipedia page as indicated by the cross-lingual Wiki links. Through this procedure, the final datasets cover a sufficient (and roughly comparable) number of examples containing ambiguous words both in English and in $L_{trg}$. We also rely on data selection heuristics that improve the final data quality, discussed in §2.2 and §2.3.

Finally, in each cross-lingual dataset we reserve 1,000 examples for testing, 500 examples as dev data; the rest is used for training. The exception are 4 resource-poor languages, where all the data examples are divided between dev and test. All data portions in all datasets are balanced. We ensure zero overlap between train, dev, and test portions. The final AM²ICO statistics are given in Table 2.

**Human Validation.** We employ human annotators to assess the quality of AM²ICO. For each dataset,

---

[2]To avoid rare words that are potentially unknown to non-experts, we retain only words that are among the top 200k words by frequency in each respective Wikipedia.

we recruit two annotators who each validate a random sample of 100 examples, where 50 examples are shared between the two samples and are used to compute inter-rater agreement.[3]

## 2.2 Data Selection Heuristics

One critical requirement of AM$^2$iCo is ensuring a high human upper bound. In the initial data creation phase, we observed several sources of confusion among human raters, typically related to some negative pairs being frequently labeled as positive; we identified two causes of this discrepancy and then mitigated it through data selection heuristics.

First, some common monosemous words still might get linked to multiple different Wikipedia pages, thus creating confusing negative pairs. For instance, some pronouns (e.g., 'he', 'it') and common nouns (e.g., 'daughter', 'son') may link to different entities as a result of coreference resolution. However, truly ambiguous words are typically directly defined in Wikipedia Disambiguation pages. We thus keep only the negative pairs that link to separate entries found in the Wikipedia Disambiguation pages. The second issue concerns concept granularity, as Wikipedia sometimes makes too fine-grained distinctions between concepts: e.g., by setting up separate pages for a country's name in different time periods.[4] We mitigate this issue by requiring that the negative pairs do not share common or parent Wikipedia categories.

The application of these heuristics during data creation (see §2.1) yields a substantial boost in human performance: e.g., the scores increase from 74% to 88% for ZH, and from 76% to 94% for DE.

## 2.3 Adversarial Examples

Another requirement is assessing to which extent models can grasp the meaning of a target word based on the (complex) *interaction* with its context. However, recently it was shown that SotA pretrained LMs exploit superficial cues while solving language understanding tasks due to spurious correlations seeping into the datasets (Gururangan et al., 2018; Niven and Kao, 2019). This hinders generalizations beyond the particular datasets and makes the models brittle to minor changes in the

---

[3]The annotators were recruited via two crowdsourcing platforms, Prolific and Proz, depending on target language coverage. The annotators were native speakers of the target language, fluent in English, and with an undergraduate degree.

[4]E.g., 'China' can be linked to the page 'Republic of China (1912–1949)' and to the page 'Empire of China (1915–1916)'.

| | WiC | XL-WiC | MCL-WiC | AM$^2$iCo |
|---|---|---|---|---|
| examples (mean) | 7,466 | **14,510** | 3600 | 13,074 |
| examples (median) | 7,466 | 1,676 | 2000 | **8,570** |
| word types (mean) | 4,130 | 7,255 | 2766 | **9,868** |
| word types (median) | 4,130 | 1,201 | 2072 | **8,520** |
| context length | 17 | 22.7 | 26.13 | **53.5** |
| human accuracy | 80 | 81.8 | - | **90.6** |
| human agreement | 80 | -[5] | 94.2 | 88.4 |
| languages | 1 | 12 | 5 | **15** |
| language families | 1 | 5 | 3 | **10** |

Table 3: Comparison of the most salient data statistics of AM$^2$iCo versus WiC, XL-WiC and MCL-WiC.

input space (Jia and Liang, 2017; Iyyer et al., 2018). As verified later in §4, we found this to be the case also for the existing WiC datasets: just considering the target word and neglecting the context (or vice versa) is sufficient to achieve high performance.

To remedy this issue, we already ensured that models could not rely solely on target words in §2.1 by including both positive and negative examples for each ambiguous word in different contexts. Further, we now introduce adversarial negative examples in AM$^2$iCo to penalize models that rely only on context without considering target words. To create such negative examples, we sample a positive pair $x_i$ and instead of the original target word $w_i$, we take another related word $\tilde{w}_i$ in the same context $c_i$ as the new target word $w_i$.

We define the related word as a hyperlinked mention sharing the same parent Wiki category as the original target word: e.g., in Table 1 we change the target word '阿波罗' (Apollo) from Example 1 into the related word '*NASA*', resulting in Example 3. Both words share a common Wiki parent category, 美国国家航空航天局 (NASA). The contexts of both examples deal with spaceships; hence, only a fine-grained understanding of lexical differences between the target words warrants the ability to recognize *Apollo* as identical to '阿波罗' but different from '*NASA*'. Overall, adversarial examples amount to roughly $1/4$ of our dataset.

## 2.4 Data Statistics and Language Coverage

We summarize the main properties of AM$^2$iCo while comparing against previous word-in-context datasets WiC, XL-WiC and MCL-WiC in Table 3. More detailed per-language scores are listed in Table 4. First, we emphasize the accrued reliability of AM$^2$iCo, as both human accuracy and interannotator agreement are substantially higher than with WiC and XL-WiC (i.e., rising by ~10 points).

---

[5]There is only one annotator for most languages in XL-WiC. Therefore no agreement score can be computed.

Second, for a comparable overall dataset size we increase the number of examples and word types in resource-poor languages. If we consider their median across languages, AM$^2$iCo has 8,570 and 8,520, respectively, around four times more than XL-WiC (1,676 and 1,201) and MCL-WiC (2000 and 2072). XL-WiC are heavily skewed towards a small number of languages, namely German and French, and provides large datasets in those languages. MCL-WiC only offers training data for English. In contrast, AM$^2$iCo provides a more balanced representation of its languages. Third, in AM$^2$iCo we deliberately include longer contexts. While the data in WiC, XL-WiC and MCL-WiC are derived from concise dictionary examples, AM$^2$iCo data reflect natural text where key information may be spread across a much wider context.

Our selection of languages is guided by the recent initiatives to cover a typologically diverse language sample (Ponti et al., 2020). In particular, AM$^2$iCo covers 15 languages, more than XL-WiC (12 languages) and MCL-WiC (5 languages). Diversity can be measured along multiple axes, such as family, geographic areas, and scripts (Ponti et al., 2019). AM$^2$iCo includes 10 language families, namely: Afro-Asiatic (1 language), Austronesian (1), Basque (1), Indo-European (5), Japonic (1), Kartvelian (1), Koreanic (1), Sino-Tibetan (1), Turkic (2), Uralic (1). This provides a more balanced sample of the cross-lingual variation compared to XL-WiC (5 families) and MCL-WiC (3 families). Regarding geography, in addition to the areas covered by XL-WiC and MCL-WiC (mostly Europe and Eastern Asia), we also represent South-East Asia (with ID), the Middle East (TR), the Caucasus (KA), the Indian subcontinent (UR and BN), as well as central Asia (KK). Finally, AM$^2$iCo also introduces scripts that were absent in other datasets, namely the Georgian alphabet and the Bengali script (a Northern Indian abugida), for a total of 8 distinct scripts.

## 3 Experimental Setup

We now establish a series of baselines on AM$^2$iCo to measure the gap between current SotA models and human performance.

**Pretrained Encoders.** Multilingual contextualized representations $\mathbf{e}_i \in \mathbb{R}^d$ for each target word are obtained via BASE variants of cased multilingual BERT (mBERT, Devlin et al., 2019) and

XLM-R[6] (Conneau et al., 2020), available in the HuggingFace repository (Wolf et al., 2020).

**Classification.** Given two contextualized representations $\mathbf{e}_{i,src}$ and $\mathbf{e}_{i,trg}$ for a pair of target words, two setups to make prediction are considered: the first, **metric**-based, is a non-parametric setup. In particular, we follow Pilehvar and Camacho-Collados (2019) and score the distance $\delta$ between the representations via cosine similarity. A threshold $t$ from the development set is set via grid search across 0.02 intervals in the $[0, 1]$ interval. Therefore, if $\delta(\mathbf{e}_{i,src}, \mathbf{e}_{i,trg}) \geq t$ the pair is classified as negative, and positive otherwise. On the other hand, the **fine-tuning** setup is parametric: following Raganato et al. (2020), we train a logistic regression classifier that takes the concatenation of the contextualized representations $[\mathbf{e}_{i,src} \oplus \mathbf{e}_{i,trg}]$ as input.[7] The entire model (both the encoder and the classifier) is then fine-tuned to minimize the cross-entropy loss of the training set examples with Adam (Kingma and Ba, 2015). We perform grid search for the learning rate in $[5e-6, 1e-5, 3e-5]$, and train for 20 epochs selecting the checkpoint with the best performance on the dev set.

**Cross-lingual Transfer.** In addition to **supervised** learning, we also carry out cross-lingual transfer experiments where data splits may belong to different language pairs. The goal is transferring knowledge from a source *language pair $\ell_s$* to a target *language pair $\ell_t$*. To simulate different scenarios of data paucity, in the fine-tuning setup we consider: 1) **zero-shot** transfer, where train and development sets belong to $\ell_s$ and the test set to $\ell_t$; 2) **zero-shot + TLD**[8] transfer, which is similar except for the dev set given in $\ell_t$; 3) on top of zero-shot + TLD, we provide a small amount of training examples in $\ell_t$, which we denote as **few-shot** transfer; 4) finally, in the **joint multilingual** setup, we train a single model on the concatenation of the train sets for all language pairs and select the hyper-parameters on the development set of $\ell_t$.

## 4 Results and Discussion

**Metric-based vs Fine-tuning.** We report the results for the supervised learning setting (where all data splits belong to the same language pair) in Ta-

---

[6]Due to limited resources, we leave experiments with large variants of the models for future research.

[7]The first WordPiece is always used if a word is segmented into multiple WordPieces.

[8]TLD stands for "Target Language Development (set)"

|  |  | DE | RU | JA | ZH | AR | KO | FI | TR | ID | EU | KA | BN | KK | UR |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Mtr | MBERT | 67.1 | 65.0 | 62.3 | 65.8 | 63.9 | 62.1 | 61.7 | 57.1 | **66.3** | **64.1** | **60.4** | **60.0** | **59.2** | **58.8** |
| | XLM-R | 65.0 | 63.1 | 56.7 | 56.7 | 58.4 | 57.5 | 64.1 | 62.4 | 65.7 | 62.9 | 58.3 | 56.2 | 58.0 | 55.5 |
| Ft | MBERT | **80.0** | 77.4 | 73.9 | **71.0** | **67.4** | **68.2** | **71.6** | **69.3** | 64.6 | 62.2 | – | – | – | – |
| | XLM-R | 77.4 | **76.1** | **75.9** | 68.9 | 65.9 | 65.3 | 68.4 | 64.4 | 54.6 | 55.8 | – | – | – | – |
| Hm | accuracy | 93.5 | 89.5 | 93.0 | 87.5 | 93.5 | 93.5 | 90.5 | 90.5 | 91.5 | 92.5 | 90.0 | 89.5 | 85.5 | 88.0 |
| | agreement | 90.0 | 78.0 | 90.0 | 94.0 | 100.0 | 92.0 | 88.0 | 96.0 | 92.0 | 84.0 | 94.0 | 80.0 | 80.0 | 80.0 |

Table 4: Accuracy of MBERT and XLM-R on AM$^2$ICO in a supervised learning setting. We report metric-based classification (MTR) results, as well as the scores in the fine-tuning setup (FT). The third group of rows (HM) displays human performance, in terms of both accuracy and inter-rater agreement. Results for the larger test sets for DE and RU are reported in Table 9 in the Appendix.

ble 4. The metric-based approach achieves consistent scores across all languages, fluctuating within the range $[57.1, 67.1]$ for MBERT and $[55.5, 65.0]$ for XLM-R. This indicates that the pretrained encoder alone already contains some relevant linguistic knowledge, to a certain degree. In comparison, fine-tuning yields more unequal results, being more data-hungry. In particular, it performs worse than the metric-based approach on languages with small training data size (e.g., ID and EU in Table 4), whereas it surpasses the metric-based approach on languages with abundant examples (e.g., DE, RU).

**XLM-R vs MBERT.** Table 4 also reveals that XLM-R is more sensitive to train data size than MBERT, often falling behind in both the metric-based and fine-tuning setups, especially for resource-poorer languages. These findings are in line with what Vulić et al. (2020) report for Multi-SimLex, which are grounded on lexical semantics similarly to AM$^2$ICO. However, they contradict the received wisdom from experiments in other multilingual sentence-level tasks (Ponti et al., 2020; Conneau et al., 2020), where XLM-R outperforms MBERT in cross-lingual transfer. While the exact causes go beyond the scope of this work, we speculate that the two encoders excel in separate aspects of semantics, the lexical and the sentence level.

**Effect of Data Size on Fine-Tuning.** To further investigate the effect of train data size on fine-tuning, we perform an in-depth analysis on some selected languages (DE, RU and JA). Note that we use the larger dev and test sets for DE and RU for this experiment. We study how performance changes as we vary the number of training examples from 500 to the full set. The results in Figure 1 indicate that, while fine-tuning starts lower than the metric-based baseline, it grows steadily and begins to take the lead from around 2,500 train examples.

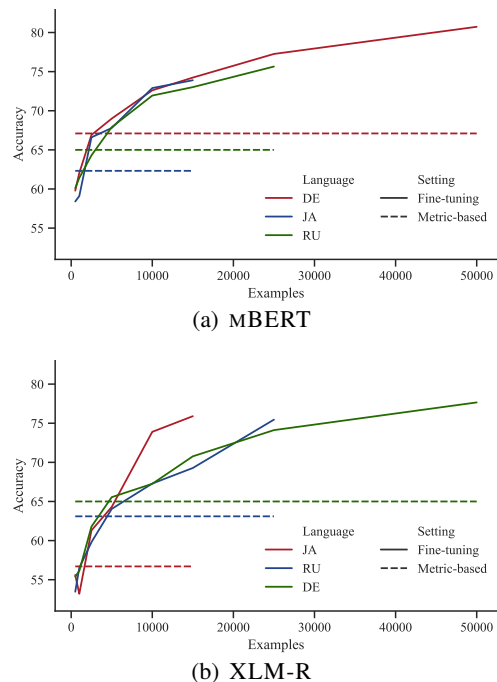**Zero-shot Transfer.** The results are presented in



(a) MBERT



(b) XLM-R

Figure 1: Impact of train data size (for fine-tuning) on performance in DE, RU (larger test sets used), and JA from AM$^2$ICO. X axis is in the log scale.

Table 5. We select the training data of each of the five languages with most data (DE, RU, JA, ZH, AR) in turn for source-language fine-tuning. Subsequently, we report the average prediction performance across all remaining 9 target languages.

First, we note that the TLD variant for hyperparameter selection does not yield gains. Second, the best choice of a source language appears to be German across the board, achieving an average score of 71.2 with MBERT and 72.0 with XLM-R. Nevertheless, this is simply due to its ample number of examples (50k). In fact, when controlling for this variable by equalizing the total size of each train split to 10k, see the bottom half of Table 5, all source languages perform comparably.

Breaking down the average results into individ-

| $\ell_s$ | Zero-shot | | +TLD | |
|---|---|---|---|---|
| | MBERT | XLM-R | MBERT | XLM-R |
| DE (all) | 71.2 | **72.0** | 71.5 | 71.7 |
| RU (all) | 71.1 | 69.8 | 71.0 | 69.9 |
| JA (all) | 68.1 | 61.9 | 68.6 | 63.2 |
| ZH (all) | 66.2 | 60.3 | 66.6 | 62.1 |
| AR (all) | 67.7 | 61.8 | 67.1 | 62.1 |
| DE (10k) | 65.4 | 62.4 | 65.9 | 62.5 |
| RU (10k) | 66.4 | 64.7 | 66.1 | 64.6 |
| JA (10k) | 67.5 | 61.3 | 67.2 | 61.7 |
| ZH (10k) | 66.0 | 62.8 | 65.8 | 62.1 |
| AR (10k) | **67.7** | 61.8 | 67.1 | 62.1 |

Table 5: Zero-shot transfer from 5 high-resource source languages to the remaining 9 languages in AM$^2$ICO. The parentheses contain (approximate) train data sizes.
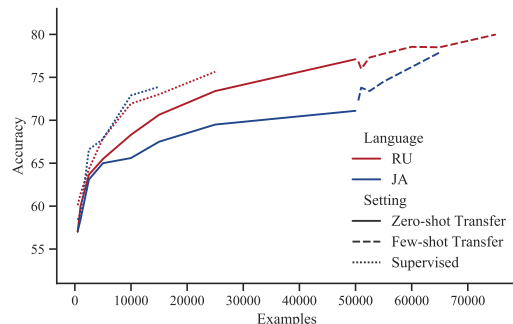


Figure 2: MBERT performance in two target languages, RU (larger test set) and JA, across different amounts of training examples under different settings: supervised learning (examples directly from $L_{trg}$) few-shot transfer (examples from DE $L_{src}$ plus $L_{trg}$ examples), and zero-shot transfer (+TLD) from DE.

ual languages in Table 6 (top section), however, reveals an even more intricate picture. In particular, the best source language for KK is RU and for JA is ZH, rather than DE. This can be explained by the fact that these pairs share their scripts, Cyrillic and Kanji / Hanzi, respectively, at least in part. This indicates that a resource-leaner but related language might sometimes be a more effective option as source language than a resource-rich one. It is also noteworthy that zero-shot transfer from DE outperforms supervised learning in most languages, except for those both resource-rich and distant (JA, ZH and AR).

**Few-shot Transfer.** To study the differences between training on $\ell_s$ and $\ell_t$ with controlled train data size, we plot the model performance on two target languages (RU and JA) as a function of the amount of available examples across different transfer conditions in Figure 2. Comparing supervised learning (based on target language data) with zero-shot learning (based on DE data), it emerges how the former is always superior if the number of examples is the same. However, zero-shot learning may eventually surpass the peak performance of supervised learning by taking advantage of a larger pool of examples: this is the case in RU, but not in JA. This illustrates a trade-off between quality (in-domain but possibly scarce data) and quantity (abundant but possibly out-of-domain data).

Few-shot learning combines the desirable properties of both approaches. After pre-training a model on DE, it can be adapted on a small amount of target-language examples. Performance continues to grow with more shots; with as few as 1k JA examples it is comparable to supervised learning on 15k examples. Few-shot learning thus not only achieves the highest scores, but also leverages costly target-language data in a sample-efficient fashion.

**Joint Multilingual Learning.** The results are shown in Table 6. We observe a substantial boost in performance across all the languages compared to both zero-shot transfer from any individual language and supervised learning (cf. Table 4), including high-resource languages such as DE and RU. Low-resource languages enjoy the most copious gains: with MBERT as the encoder, UR improves by 4.8 points, KK by 4.3, BN by 5.7, and KA by 6.3. However, this is still insufficient to equalize performances across the board, as the latter group of languages continues to lag behind: between DE and UR remains a gap of 7.2 points. We speculate that the reason behind this asymmetry is the fact that in addition to being resource-poor, UR, KK, BN, and KA are also typologically distant from languages where most of the examples are concentrated. Overall, these findings suggest that leveraging multiple sources is better than a single one by virtue of the transfer capabilities of massively multilingual encoders, as previously demonstrated (Wu and Dredze, 2019; Ponti et al., 2021).

**Adversarial Examples.** Finally, we investigate whether the inclusion of adversarial examples (see §2.3) makes AM$^2$ICO less likely to be solved by relying on superficial clues. In Table 7, we compare the performance of MBERT trained on the full input (which we label FULL) with two adversarial baselines. We implement the TGT variant by inputting only the target word to the classification model and the CTX variant by replacing the target word with a '[MASK]' token. We perform analysis

| $\ell_s$ | model | DE | RU | JA | ZH | AR | KO | FI | TR | ID | EU | KA | BN | KK | UR |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **Zero-shot Transfer (+TLD)** | | | | | | | | | | | | | |
| DE | MBERT | - | **75.1** | 71.1 | **68.9** | 64.8 | 71.1 | **78** | 75.7 | 75.4 | **74.2** | 70.2 | 68 | 62.8 | **68.5** |
| | XLM-R | - | **77.5** | 70.1 | 67.7 | 67.4 | 70.2 | **77.2** | 76.6 | 74.8 | 70.7 | **71.9** | **70.6** | 66.5 | **67** |
| RU | MBERT | **74.4** | - | **71.7** | 68.4 | **65.1** | 71 | 74.8 | 72.6 | 74.5 | 71.2 | 69.3 | **70.3** | 67 | 67.8 |
| | XLM-R | **72.4** | - | 68.2 | 64.4 | **69.5** | **72.3** | 73.7 | 71.2 | 69.7 | 68.8 | 69.6 | 68.3 | **69** | 66.7 |
| JA | MBERT | 70.9 | 70.7 | - | **68.9** | 63.2 | **72.5** | 70.9 | 72.5 | 70.1 | 67.7 | 65.3 | 67.9 | 65.8 | 65 |
| | XLM-R | 64.7 | 65.7 | - | **70.4** | 63.2 | 67.1 | 66.5 | 64.9 | 62.5 | 62.5 | 61.7 | 66.7 | 57.8 | 59.3 |
| ZH | MBERT | 71.5 | 69.2 | 68.7 | - | 61.6 | 70.6 | 68.6 | 70.8 | 66.9 | 68 | 64.9 | 65.6 | 61.5 | 62.8 |
| | XLM-R | 66.5 | 67.1 | **72.1** | - | 62.6 | 62.9 | 66.2 | 63.8 | 63.4 | 61.6 | 59.8 | 61.1 | 59.8 | 60.3 |
| AR | MBERT | 69.2 | 67.6 | 68.5 | 66.3 | - | 67.1 | 68.6 | 69.5 | 68.2 | 67.6 | 66.9 | 68.1 | 63.8 | 63.8 |
| | XLM-R | 62 | 63 | 62.2 | 62.0 | - | 61.9 | 62.2 | 63.6 | 63.9 | 61.3 | 62.1 | 64.9 | 60.8 | 58.3 |
| | | **Joint Multilingual Learning** | | | | | | | | | | | | | |
| all | MBERT | **80.4** | **82.1** | 78.2 | 75.2 | 73.3 | 75.8 | **81.2** | **80.6** | 78.4 | **75.9** | 76.5 | **76** | 71.3 | **73.3** |
| | XLM-R | 79.4 | 80.9 | **79.4** | **76.1** | **73.6** | 76 | **81.2** | 80.5 | 77.9 | 74.2 | **77.7** | 73 | **74.5** | 72.5 |

Table 6: Results for zero-shot transfer from a single source (top section) and joint transfer from multiple sources (bottom section) in AM²ICO. The best scores for each setup are in bold. Results for the larger testsets of DE an RU are reported in Table 10.

| Dataset | $\ell$ | Full | Ctx | Tgt | Hm |
|---|---|---|---|---|---|
| **WiC** | EN | 67.1 | 65.1 | 55.2 | 80.0 |
| **XL-WiC** | DE | 81.0 | 75.0 | 80.0 | 74.0 |
| **MCL-WiC** | EN | 83.3 | 82.6 | 53.4 | 96.8 |
| **AM²ICO-A** | DE | 84.2 | 83.6 | 50.0 | 93.0 |
| **AM²ICO +A** | DE | 80.0 | 73.6 | 66.9 | 93.5 |
| **AM²ICO-A** | ZH | 79.8 | 78.2 | 49.0 | 86.0 |
| **AM²ICO +A** | ZH | 71.0 | 66.0 | 61.6 | 87.5 |

Table 7: The impact of adversarial examples on MBERT performance. +A indicates the presence of adversarial examples, -A their absence. CTX represents training a model on the context only, TGT on the target word only, and FULL on the whole input. HM stands for human accuracy.

| | DE | RU | JA | ZH | AR |
|---|---|---|---|---|---|
| Adversarial | 73.7 | 66.5 | 65.3 | 64.7 | 55.0 |
| Non-adversarial | 83.1 | 77.9 | 76.5 | 75.5 | 70.5 |

Table 8: MBERT performance on adversarial and non-adversarial examples across five datasets in AM²ICO.

across AM²ICO, WiC, XL-WiC and MCL-WiC.[9]

In previous datasets, at least one of the adversarial baselines reaches performance close to the FULL model: in WiC (EN), CTX has a gap of only 2 points. In XL-WiC (DE), TGT is only 1 point away from FULL. In MCL-WiC (EN), the gap between CTX and FULL is even below 1 point. This would also be the case in AM²ICO were it not for the extra adversarial examples (rows +A): by virtue of this change, the distance between FULL and the best adversarial baseline is 6.4 points in DE and 5.0 in ZH. Therefore, it is safe to conclude that a higher score on AM²ICO better reflects a deep semantic understanding by the model. Moreover, the last column of Table 7 also includes reference human accuracy. While the best baseline in XL-WiC even surpasses the human upper bound, the addition of adversarial examples in AM²ICO combined

with higher human accuracy drastically increases the gap between the two: 13.5 points in DE and 16.5 in ZH. Overall, this results in a much more challenging evaluation benchmark.

In addition, we report separate results on these adversarial examples, and compare with model performance on non-adversarial examples for DE, RU, JA, ZH and AR, within the supervised setting (Table 8). It is clear that the adversarial examples pose a much greater challenge for the model with overall much lower scores. This is expected as for adversarial examples the models must have a finer-grained and more accurate understanding of both the target word semantics and the surrounding (sentential) context.

## 5 Related Work

**Cross-Lingual Evaluation of Word Meaning in Context.** Going beyond readily available sense inventories required for WSD-style evaluations, the comprehensive benchmarks for evaluating word meaning in context cross-lingually are still few and far between. XL-WiC (Raganato et al., 2020) extends the original English WiC framework of Pilehvar and Camacho-Collados (2019) to 12 other languages, but supports only monolingual evaluation, and suffers from issues such as small gaps between human and system performance. The SemEval-

---

[9]We select languages with sufficient and comparable training data: ZH for AM²ICO, DE for AM²ICO and XL-WiC and EN for MCL-WiC.

2021 shared task MCL-WiC does focus on cross-lingual WiC, but covers only five high-resource languages from three language families (English, French, Chinese, Arabic, Russian). Both XL-WiC and MCL-WiC mainly focus on common words and do not include less frequent concepts (e.g., named entities). Further, their language coverage and data availability are heavily skewed towards Indo-European languages.

There are several other 'non-WiC' datasets designed to evaluate cross-lingual context-aware lexical representations. Bilingual Contextual Word Similarity (BCWS) (Chi and Chen, 2018) challenges a model to predict graded similarity of cross-lingual word pairs given sentential context, one in each language. In the Bilingual Token-level Sense Retrieval (BTSR) task (Liu et al., 2019), given a query word in a source language context, a system must retrieve a meaning-equivalent target language word within a target language context.[10] However, both BCWS and BTSR are again very restricted in terms of language coverage: BCWS covers only one language pair (EN-ZH), while BTSR contains two pairs (EN-ZH/ES). Further, they provide only test data: as such, they can merely be used as general intrinsic probes for pretrained models, but cannot support fine-tuning experiments and cannot fully expose the relevance of information available in pretrained models for downstream applications. This is problematic as intrinsic tasks in general do not necessarily correlate well with downstream performance (Chiu et al., 2016; Glavaš et al., 2019).

**AM$^2$ıCo vs. Entity Linking.** Our work is related to the entity linking (EL) task (Rao et al., 2013; Cornolti et al., 2013; Shen et al., 2014) similarly to how the original WiC (based on WordNet knowledge) is related to WSD. EL systems must map entities in context to a predefined knowledge base (KB). While WSD relies on the WordNet sense inventory, the EL task focuses on KBs such as Wikipedia and DBPedia. When each entity mention is mapped to a unique Wiki page, this procedure is termed wikification (Mihalcea and Csomai, 2007). The cross-lingual wikification task (Ji et al., 2015; Tsai and Roth, 2016) grounds multilingual mentions to English Wikipedia pages. Similar to WSD, EL evaluation is tied to a specific KB. It thus faces similar limitations of WSD in terms of restricting

meanings and their distinctions to those predefined in the inventory. In comparison, AM$^2$ıCo leverages Wikipedia only as a convenient resource for extracting the examples, similar to how the original WiC work leverages WordNet. AM$^2$ıCo itself is then framed on natural text, without requiring the modeling of the KBs. Also, in comparison with EL, AM$^2$ıCo provides higher data quality and a more challenging evaluation of complex word-context interactions, achieved by a carefully designed data extraction and filtering procedure.

# 6 Conclusion

We presented AM$^2$ıCo, a large-scale and challenging multilingual benchmark for evaluating word meaning in context (WiC) across languages. AM$^2$ıCo is constructed by leveraging multilingual Wikipedias, and subsequently validated by humans. It covers 15 typologically diverse languages and a vocabulary substantially larger than all previous WiC datasets. As such, it provides more comprehensive and reliable quality estimates for multilingual encoders. Moreover, AM$^2$ıCo includes adversarial examples: resolving such examples requires genuine lexical understanding, as opposed to relying on spurious correlations from partial input. Finally, AM$^2$ıCo offers the possibility to perform cross-lingual evaluation, pairing context between different languages. Moreover, we explored the impact of language relatedness on model performance by transferring knowledge from multiple source languages. We established a series of baselines on AM$^2$ıCo, based on SotA multilingual models, revealing that the task is far from being 'solved' even with abundant training data. All models struggle especially when transferring to distant and resource-lean target languages. We hope that AM$^2$ıCo will guide and foster further research on effective representation learning across different languages.

---

[10]BTSR could be seen as a contextualized version of the standard bilingual lexicon induction task (Mikolov et al., 2013; Søgaard et al., 2018; Ruder et al., 2019, *inter alia*).

# References

Edoardo Barba, Luigi Procopio, Niccolò Campolungo, Tommaso Pasini, and Roberto Navigli. 2020. Mu-LaN: Multilingual label propagation for word sense disambiguation. In *Proceedings of IJCAI 2020*, pages 3837–3844.

Ta-Chung Chi and Yun-Nung Chen. 2018. CLUSE: Cross-lingual unsupervised sense embeddings. In *Proceedings of EMNLP 2018*, pages 271–281.

Billy Chiu, Anna Korhonen, and Sampo Pyysalo. 2016. Intrinsic evaluation of word vectors fails to predict extrinsic performance. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 1–6.

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 ACL 2020*, pages 8440–8451.

Marco Cornolti, Paolo Ferragina, and Massimiliano Ciaramita. 2013. A framework for benchmarking entity-annotation systems. In *Proceedings of WWW 2013*, pages 249–260.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of ACL 2019*, pages 710–721.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of NAACL-HLT 2018*, pages 107–112.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of NAACL-HLT 2018*, pages 1875–1885.

Heng Ji, Joel Nothman, Ben Hachey, and Radu Florian. 2015. Overview of TAC-KBP2015 tri-lingual entity discovery and linking. In *Proceedings of TAC 2015*.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of EMNLP 2017*, pages 2021–2031.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of ICLR 2015*.

Qianchu Liu, Diana McCarthy, Ivan Vulić, and Anna Korhonen. 2019. Investigating cross-lingual alignment methods for contextualized embeddings with token-level evaluation. In *Proceedings of CoNLL 2019*, pages 33–43.

Rada Mihalcea and Andras Csomai. 2007. Wikify! Linking documents to encyclopedic knowledge. In *Proceedings of CIKM 2007*, pages 233–242.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint, CoRR*, abs/1309.4168.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the ACL*, 2:231–244.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2):1–69.

Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. Semeval-2013 task 12: Multilingual word sense disambiguation. In *Proceedings of SEMEVAL 2013*, pages 222–231.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Joining forces pays off: Multilingual joint word sense disambiguation. In *Proceedings of EMNLP-CoNLL 2012*, pages 1399–1410.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of ACL 2019*, pages 4658–4664.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of NAACL-HLT 2019*, pages 1267–1273.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of EMNLP 2020*, pages 2362–2376.

Edoardo Maria Ponti, Helen O'horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601.

Edoardo Maria Ponti, Ivan Vulić, Ryan Cotterell, Marinela Parovic, Roi Reichart, and Anna Korhonen. 2021. Parameter space factorization for zero-shot learning across tasks and languages. *Transactions of the ACL*, 9:410–428.

Alessandro Raganato, José Camacho-Collados, and Roberto Navigli. 2017. Word sense disambiguation:

A unified evaluation framework and empirical comparison. In *Proceedings of EACL 2017*, pages 99–110.

Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2020. XL-WiC: A multilingual benchmark for evaluating semantic contextualization. In *Proceedings of EMNLP 2020*, pages 7193–7206.

Jonathan Raiman and Olivier Raiman. 2018. Deep-Type: Multilingual entity linking by neural type system evolution. In *Proceedings of AAAI 2018*, pages 5406–5413.

Delip Rao, Paul McNamee, and Mark Dredze. 2013. Entity linking: Finding extracted entities in a knowledge base. In *Multi-source, multilingual information extraction and summarization*, pages 93–115. Springer.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020. Sense-annotated corpora for word sense disambiguation in multiple languages and domains. In *Proceedings of LREC 2020*, pages 5905–5911.

Wei Shen, Jianyong Wang, and Jiawei Han. 2014. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of ACL 2018*, pages 778–788.

Chen-Tse Tsai and Dan Roth. 2016. Cross-lingual wikification using multilingual embeddings. In *Proceedings of NAACL-HLT 2016*, pages 589–598.

Shyam Upadhyay, Nitish Gupta, and Dan Roth. 2018. Joint multilingual supervision for cross-lingual entity linking. In *Proceedings of EMNLP 2018*, pages 2486–2495.

Ivan Vulić, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, et al. 2020. Multi-SimLex: A large-scale evaluation of multilingual and cross-lingual lexical semantic similarity. *Computational Linguistics*, 46(4):847–897.

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 EMNLP 2020: System Demonstrations*, pages 38–45.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of EMNLP-IJCNLP 2019*, pages 833–844.

## A   Results on the larger test set in DE and RU

Table 9 and Table 10 list results for the larger test sets of DE and RU in AM$^2$ICO.

|  |  | DE large | RU large |
|---|---|---|---|
| MTR | MBERT | 66.1 | 65.3 |
| | XLM-R | 64.5 | 63.2 |
| FT | MBERT | 80.7 | 75.7 |
| | XLM-R | 77.7 | 75.5 |

Table 9: Accuracy of MBERT and XLM-R on larger test sets of DE and RU in AM$^2$ICO in a supervised learning setting. We report metric-based classification (MTR) results, as well as the scores in the fine-tuning setup (FT).

| $\ell_s$ | model | DE large | RU large |
|---|---|---|---|
| **Zero-shot Transfer (+TLD)** | | | |
| DE | MBERT | - | 77.1 |
| | XLM-R | - | 76.4 |
| RU | MBERT | 76.1 | - |
| | XLM-R | 72.5 | - |
| JA | MBERT | 72.3 | 72.2 |
| | XLM-R | 65.5 | 66.0 |
| ZH | MBERT | 70.8 | 70.0 |
| | XLM-R | 64.7 | 64.4 |
| AR | MBERT | 69.1 | 68.9 |
| | XLM-R | 62.9 | 63.7 |
| **Joint Multilingual Learning** | | | |
| all | MBERT | 81.9 | 81.2 |
| | XLM-R | 80.5 | 80.7 |

Table 10: AM$^2$ICO transfer results for the larger test sets of DE and RU. Performance from zero-shot transfer from a single source is in the top section; joint transfer from multiple sources is in the bottom section.