# Retrofitting Multilingual Sentence Embeddings with Abstract Meaning Representation*

**Deng Cai**[♡] **Xin Li**[♠,†] **Jackie Chun-Sing Ho**[♡] **Lidong Bing**[♠] **Wai Lam**[♡]
[♡]The Chinese University of Hong Kong
[♠]DAMO Academy, Alibaba Group
thisisjcykcd@gmail.com
{xinting.lx,l.bing}@alibaba-inc.com
{jackieho@link,wlam@se}.cuhk.edu.hk

## Abstract

We introduce a new method to improve existing multilingual sentence embeddings with Abstract Meaning Representation (AMR). Compared with the original textual input, AMR is a structured semantic representation that presents the core concepts and relations in a sentence explicitly and unambiguously. It also helps reduce surface variations across different expressions and languages. Unlike most prior work that only evaluates the ability to measure semantic similarity, we present a thorough evaluation of existing multilingual sentence embeddings and our improved versions, which include a collection of five transfer tasks in different downstream applications. Experiment results show that retrofitting multilingual sentence embeddings with AMR leads to better state-of-the-art performance on both semantic textual similarity and transfer tasks. Our codebase and evaluation scripts can be found at https://github.com/jcyk/MSE-AMR.

## 1 Introduction

Multilingual sentence embedding (MSE) aims to provide universal sentence representations shared across different languages (Hermann and Blunsom, 2014; Pham et al., 2015; Schwenk and Douze, 2017). As an important ingredient of cross-lingual and multilingual natural language processing (NLP), MSE has recently attracted increasing attention in the NLP community. MSE has been widely adopted to bridge the language barrier in several downstream applications such as bitext mining (Guo et al., 2018; Schwenk, 2018), document classification (Eriguchi et al., 2018; Singla et al., 2018; Yu et al., 2018) and natural language inference (Artetxe and Schwenk, 2019). Prior work typically borrows fixed-size embedding vectors from multilingual neural machine models (Schwenk and Douze, 2017; Yu et al., 2018) or trains siamese neural networks to align the semantically similar sentences written in different languages (Wieting et al., 2019; Yang et al., 2020; Feng et al., 2020).

Despite the recent progress, the current evaluation of multilingual sentence embeddings has focused on cross-lingual Semantic Textual Similarity (STS) (Agirre et al., 2016; Cer et al., 2017) or bi-text mining tasks (Zweigenbaum et al., 2018; Artetxe and Schwenk, 2019). Nevertheless, as pointed out by Gao et al. (2021), the evaluation on semantic similarity may not be sufficient because better performance on STS does not always indicate better embeddings for downstream tasks. Therefore, for a more comprehensive MSE evaluation, it is necessary to additionally evaluate downstream tasks, which is largely ignored in recent work (Chidambaram et al., 2019; Reimers and Gurevych, 2020; Feng et al., 2020). In this paper, we collect a set of multilingual transfer tasks and test various existing multilingual sentence embeddings. We find that different methods excel at different tasks and the conclusions drawn from the STS evaluation do not always hold in the transfer tasks and vice versa. We aim to establish a standardized evaluation protocol for future research in multilingual sentence embeddings.

To improve the quality of existing MSE models, we explore Abstract Meaning Representation (AMR) (Banarescu et al., 2013), a symbolic semantic representation, for augmenting existing neural semantic representations. Our motivation is twofold. First, AMR explicitly offers core concepts and relations in a sentence. This helps prevent learning the superficial patterns or spurious correlations in the training data, which do not generalize well to new domains or tasks (Poliak et al., 2018; Clark et al., 2019). Second, AMR reduces the variances in surface forms with the same meaning. This helps alleviate the data sparsity issue as there are

* This work was supported by Alibaba Group through the Alibaba Innovative Research (AIR) Program. † XL is the corresponding author.



*Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6456 - 6472
December 7-11, 2022 

rich lexical variations across different languages.

On the other hand, despite that AMR is advocated to act as an interlingua (Xue et al., 2014; Hajič et al., 2014; Damonte and Cohen, 2018), little work has been done to reflect on the ability of AMR to have impact on subsequent tasks. In order to advance research in AMR and its applications, multilingual sentence embedding can be seen as an important benchmark for highlighting its ability to abstract away from surface realizations and represent the core concepts expressed in the sentence. To our knowledge, this is the first attempt to leverage the AMR semantic representation for multilingual NLP.

We learn AMR embeddings with contrastive siamese network (Gao et al., 2021) and AMR graphs derived from different languages (Cai et al., 2021). Experiment results on 10 STS tasks and 5 transfer tasks with four state-of-the-art embedding methods show that retrofitting multilingual sentence embeddings with AMR improves the performance substantially and consistently.

Our contribution is three-fold.

- We propose a new method to obtain high-quality semantic vectors for multilingual sentence representation, which takes advantage of language-invariant Abstract Meaning Representation that captures the core semantics of sentences.
- We present a thorough evaluation of multilingual sentence embeddings, which goes beyond semantic textual similarity and includes various transfer tasks in downstream applications.
- We demonstrate that retrofitting multilingual sentence embeddings with Abstract Meaning Representation leads to better performance on both semantic textual similarity and transfer tasks.

## 2 Related Work

**Universal Sentence Embeddings** Our work aims to learn universal sentence representations, which should be useful for a broad set of applications. There are two lines of research for universal sentence embeddings: unsupervised approaches and supervised approaches. Early unsupervised approaches (Kiros et al., 2015; Hill et al., 2016; Gan et al., 2017; Logeswaran and Lee, 2018) design various surrounding sentence reconstruction/prediction objectives for sentence representation learning. Jernite et al. (2017) exploit sentence-level discourse relations as supervision signals for training sentence embedding model. Instead of using the interactions of sentences within a document, Le and Mikolov (2014) propose to learn the embeddings for texts of arbitrary length on top of word vectors. Likewise, Chen (2017); Pagliardini et al. (2018); Yang et al. (2019b) calculate sentence embeddings from compositional $n$-gram features. Recent approaches often adopt contrastive objectives (Zhang et al., 2020; Giorgi et al., 2021; Wu et al., 2020; Meng et al., 2021; Carlsson et al., 2021; Kim et al., 2021; Yan et al., 2021; Gao et al., 2021) by taking different views—from data augmentation or different copies of models—of the same sentence as training examples.

On the other hand, supervised methods (Conneau et al., 2017; Cer et al., 2018; Reimers and Gurevych, 2019; Gao et al., 2021) take advantage of labeled natural language inference (NLI) datasets (Bowman et al., 2015; Williams et al., 2018), where a sentence embedding model is fine-tuned on entailment or contradiction sentence pairs. Furthermore, Wieting and Gimpel (2018); Wieting et al. (2020) demonstrate that bilingual and back-translation corpora provide useful supervision for learning semantic similarity. Another line of work focuses on regularizing embeddings (Li et al., 2020; Su et al., 2021; Huang et al., 2021) to alleviate the representation degeneration problem. Very recently, Opitz and Frank (2022) combine the strengths of AMR metrics and embedding similarities for accurate and explainable sentence similarity rating.

**Multilingual Sentence Embeddings** Recently, multilingual sentence representations have attracted increasing attention. Schwenk and Douze (2017); Yu et al. (2018); Artetxe and Schwenk (2019) propose to use encoders from multilingual neural machine translation to produce universal representations across different languages. Chidambaram et al. (2019); Wieting et al. (2019); Yang et al. (2020); Feng et al. (2020) fine-tune siamese networks (Bromley et al., 1993) with contrastive objectives using parallel corpora. Reimers and Gurevych (2020) train a multilingual model to map sentences to the same embedding space of an existing English model. Different from existing work, our work resorts to multilingual AMR, a language-agnostic disambiguated semantic representation, for performance enhancement.

**Evaluation of Sentence Embeddings** Traditionally, the mainstream evaluation for assessing the

quality of *English-only* sentence embeddings is based on the Semantic Textual Similarity (STS) tasks and a suite of downstream classification tasks. The STS tasks (Agirre et al., 2012, 2013, 2014, 2015, 2016; Marelli et al., 2014; Cer et al., 2017) calculate the embedding distance of sentence pairs and compare them with the human-annotated scores for semantic similarity. The classification tasks (e.g., sentiment analysis) from SentEval (Conneau and Kiela, 2018) take sentence embeddings as fixed input features to a logistic regression classifier. These tasks are commonly used to benchmark the transferability of sentence embeddings on downstream tasks. For *multilingual* sentence embeddings, most previous work has focused on cross-lingual STS (Agirre et al., 2016; Cer et al., 2017) and the relevant bi-text mining tasks (Zweigenbaum et al., 2018; Artetxe and Schwenk, 2019). The evaluation on downstream transfer tasks has been largely ignored (Chidambaram et al., 2019; Reimers and Gurevych, 2020; Feng et al., 2020). Nevertheless, as pointed out in Gao et al. (2021) in English scenarios, better performance on semantic similarity tasks does not always indicate better embeddings for transfer tasks. For a more comprehensive evaluation, in this paper, we collect a set of multilingual transfer tasks and test various existing multilingual sentence embeddings. We aim to establish a standardized evaluation protocol for future research in multilingual sentence embeddings.

## 3 Preliminaries

### 3.1 Contrastive Siamese Network

Siamese network (Bromley et al., 1993) has attracted considerable attention for self-supervised representation learning. It has been extensively adopted with contrastive learning (Hadsell et al., 2006) for learning dense vector representations of images and sentences (Reimers and Gurevych, 2019; Chen et al., 2020). The core idea of contrastive learning is to pull together the representations of semantically close objects (images or sentences) and repulse the representations of negative pairs of dissimilar ones. Recent work in computer vision (Caron et al., 2020; Grill et al., 2020; Chen and He, 2021; Zbontar et al., 2021) has demonstrated that negative samples may not be necessary. A similar observation was made in NLP by Zhang et al. (2021) who adopted the BYOL framework (Grill et al., 2020) for sentence representation learning. In this work, we adopt the framework in (Gao et al., 2021) with in-batch negatives (Chen et al., 2017; Henderson et al., 2017). Formally, we assume a set of training examples $\mathcal{D} = \{(x_i, x_i^+, x_i^-)\}_{i=1}^N$, where $x_i^+$ and $x_i^-$ are semantically close and semantically irrelevant to $x_i$, respectively. The training is done with stochastic mini-batches. Each mini-batch consists of $M$ examples and the training objective is defined as:

$$\ell_i = -\log \frac{e^{s(x_i,x_i^+)/\tau}}{\sum_{j=1}^M e^{s(x_i,x_j^-)/\tau} + \sum_{j=1}^M e^{s(x_i,x_j^+)/\tau}} \quad (1)$$

where $s(\cdot,\cdot)$ measures the similarity of two objects and $\tau$ is a scalar controlling the temperature of training. As seen, other objects in the same mini-batch (i.e., $\{x_j^-\}_{j\neq i}$ and $\{x_j^+\}_{j\neq i}$) are treated as negatives for $x_i$. More concretely, $s(\cdot,\cdot)$ computes the cosine similarity between the representations of two objects:

$$s(x_i, x_j) = \frac{\mathbf{h}_i^\top \mathbf{h}_j}{\|\mathbf{h}_i\| \cdot \|\mathbf{h}_j\|}$$

where $\mathbf{h}_i$ and $\mathbf{h}_j$ are obtained from a neural encoder $f_\theta(\cdot)$: $\mathbf{h} = f_\theta(x)$. The model parameters $\theta$ are then optimized using the contrastive learning objective.

### 3.2 Multilingual AMR Parsing

AMR (Banarescu et al., 2013) is a broad-coverage semantic formalism originally designed for English. The accuracy of AMR parsing has been greatly improved in recent years (Cai and Lam, 2019, 2020a; Bevilacqua et al., 2021; Bai et al., 2022). Because AMR is agnostic to syntactic and wording variations, recent work has suggested the potential of AMR to work as an *interlingua* (Xue et al., 2014; Hajič et al., 2014; Damonte and Cohen, 2018). That is, we can represent the semantics in other languages using the corresponding AMR graph of the semantic equivalent in English. A number of *cross-lingual* AMR parsers (Damonte and Cohen, 2018; Blloshmi et al., 2020; Sheth et al., 2021; Procopio et al., 2021; Cai et al., 2021) have been developed to transform non-English texts into AMR graphs. Most of them rely on pre-trained multilingual language models and synthetic parallel data. In particular, Cai et al. (2021) proposed to learn a multilingual AMR parser from an English AMR parser via knowledge distillation. Their single parser is trained for five different languages (German, Spanish, Italian, Chinese, and English) and achieves state-of-the-art parsing accuracies. In addition, the

one-for-all design maintains parsing efficiency and reduces prediction inconsistency across different languages. Thus, we adopt the multilingual AMR parser of Cai et al. (2021) in our experiments.[1]

It is worth noting that the multilingual parser is capable of parsing many other languages, including those it has not been explicitly trained for, thanks to the generalization power inherited from pre-trained multilingual language models (Tang et al., 2020; Liu et al., 2020). In Section 4.2, we further extend the training of the multilingual parser to French, another major language, for improved performance.

## 4 Proposed Method

We first introduce how we learn AMR embeddings and then describe the whole pipeline for enhancing existing sentence embeddings.

### 4.1 Learning AMR Embeddings

**Linearization & Modeling** Given AMR is graph-structured, a variety of graph neural networks (Song et al., 2018; Beck et al., 2018; Ribeiro et al., 2019; Guo et al., 2019; Cai and Lam, 2020b; Ribeiro et al., 2019) have been proposed for the representation learning of AMR. However, recent work (Zhang et al., 2019a; Mager et al., 2020; Bevilacqua et al., 2021) has demonstrated that the power of existing pre-trained language models based on the Transformer architecture (Vaswani et al., 2017), such as BERT (Devlin et al., 2019), GPT2 (Radford et al., 2019) and BART (Lewis et al., 2020), can be leveraged for achieving better performance. Following them, we also take BERT as the backbone model.

Since Transformer-based language models are designed for sequential data, to encode graphical AMR, we resort to the linearization techniques in (Bevilacqua et al., 2021). Figure 1 illustrates the linearization of AMR graphs. For each AMR graph, a DFS traversal is performed starting from the root node of the graph, and the trajectory is recorded. We use parentheses to mark the hierarchy of node depths. Bevilacqua et al. (2021) also proposed to use special tokens for indicating variables in the linearized graph and for handling reentrancies (i.e., a node plays multiple roles in the graph). However, the introduction of special tokens significantly increases the length of the output sequence (almost 50% increase). We remove

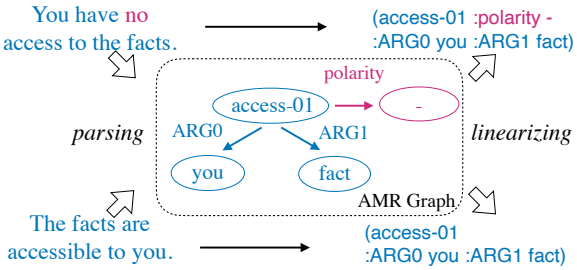


Figure 1: The parsing and linearization pipeline.

this feature and simply repeat the nodes when revisiting happens. This significantly reduces the length of the output sequence and allows more efficient modeling with Transformer-based language models. The downside is that reentrancy information becomes unrecoverable. However, we empirically found that the shortened sequences lead to better performance. The linearizations of AMR graphs are then treated as plain token sequences when being fed into Transformer-based language models. Note that AMR linearization introduces additional tokens that are rarely shown in English (e.g., "ARG2" and "belong-01"). These tokens may not be included in the original vocabulary of existing language models and could be segmented into sub-tokens (e.g., "belong-01" ⇒ "belong", "-", "01"), which are less meaningful and increase the sequence length. To deal with this problem, we extend the original vocabulary of existing language models to include all the relation and frame names occurring at least 5 times in the AMR sembank (LDC2017T10).

**Positive & Negative Examples** Contrastive learning aims to learn effective representations by pulling semantically similar examples together and pushing apart dissimilar examples. Following the discussion in Section 3.1, the most critical question in contrastive learning is how to obtain positive and negative examples. In language representations, positive examples $x_i^+$ are often constructed by applying minimal distortions (e.g., word deletion, reordering, and substitution) on $x_i$ (Wu et al., 2020; Meng et al., 2021) or introducing some random noise (e.g., dropout (Srivastava et al., 2014)) to the modeling function $f_\theta$ (Gao et al., 2021). On the other hand, negative examples $x_i^-$ are usually sampled from other sentences. However, prior work (Conneau et al., 2017; Gao et al., 2021) has demonstrated that entailment/contradiction sentence pairs in supervised natural language inference (NLI) datasets (Bowman et al., 2015; Williams et al.,

[1] https://github.com/jcyk/XAMR

2018) are better positive/negative pairs for learning sentence embeddings. Following (Gao et al., 2021), we borrow the supervisions from two NLI datasets, namely SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018). In the NLI datasets, given one premise, there are one entailment hypothesis and another contradiction hypothesis accompanying. Therefore, in each training example $(x_i, x_i^+, x_i^-)$, $x_i$ is the premise, $x_i^+$ is the entailment hypothesis, and $x_i^-$ is the contradiction hypothesis.

Specifically, we use the multilingual AMR parser described in Section 3.2 to parse sentences into AMR graphs. Because the sentences in the NLI datasets are in English, the resultant AMR graphs are all derived from English. This is in contrast to downstream applications where an AMR graph may be derived from a foreign language. To reduce the discrepancy between training and testing, we use OPUS-MT (Tiedemann and Thottingal, 2020) [2], an off-the-shelf translation system, to translate English sentences in the NLI datasets to other languages. The translations in other languages are then parsed by our multilingual AMR parser. In this way, we extend the training of AMR embeddings to multilingual scenarios as well.

**Mixed Training** To better cover both the monolingual and cross-lingual settings in downstream applications, the training aims to capture the interactions between AMR graphs derived from the same language as well as those derived from different languages. To this end, we mix up AMR graphs from different languages during training. Moreover, to alleviate the drawback of imperfect parsing and avoid catastrophic forgetting of pre-trained language models, we also mix up AMR graphs and original English sentences during training. The details are shown in Algorithm 1.

We hypothesize that the noise introduced by automatic translation could negatively affect the performance but a suitable amount of noise might also serve as a helpful regularizer. Unfortunately, due to the lack of gold translations, we could not perform a rigorous quantitative comparison. In our preliminary experiments, we also tried another automatic translation system, mBART-mmt (Tang et al., 2020) [3], other than OPUS-MT. We found that mBART-mmt leads to worse performance in general, likely due to its lower translation quality.

**Algorithm 1:** Learning AMR Embeddings.

**Input:** Dataset: $\mathcal{D} = \{(x_i, x_i^+, x_i^-)\}_{i=1}^N$, Systems: AMR parser $\texttt{parse}(\cdot)$ and English-to-$l$ translator $\texttt{translate}(\cdot, l)$, Maximum training steps: $T$, Batch size $M$, Language set: $\mathcal{L}$.

**for** $t \leftarrow 1$ *to* $T$ **do**
  Draw a mini-batch $\mathcal{B} = \{(x_i, x_i^+, x_i^-)\}_{i=1}^M$ from $\mathcal{D}$
  **foreach** sentence $x$ **in** $\mathcal{B}$ **do**
    Draw a language $l \sim \mathcal{L}$
    **if** $l$ **is not** *en* **then**
      $x \leftarrow \texttt{parse}(\texttt{translate}(x, l))$
    **else**
      Draw a text/graph factor $q \sim U(0, 1)$
      **if** $q >$ *0.5* **then**
        $x \leftarrow \texttt{parse}(x)$
  Optimize the model $f_\theta$ with Eq. (1) on the updated $\mathcal{B}$

**Output:** Optimized Model $f_\theta$

## 4.2 Incorporating AMR Embeddings

The learned AMR embeddings can be used to augment any existing sentence embedding model. For any input sentence $x$, it is processed through two channels: (1) the sentence is first parsed into an AMR graph $y = \texttt{parse}(x)$. The graph is then fed into our AMR encoder: $\mathbf{h} = f_\theta(y)$. (2) the sentence is directly encoded by an off-the-shelf sentence embedding model $g(\cdot)$: $\mathbf{s} = g(x)$. Lastly, we combine the text and graph embeddings ($\mathbf{s}$ and $\mathbf{h}$) to produce the final sentence representation.

**Parsing** Theoretically, the multilingual AMR parser introduced in Cai et al. (2021) can parse 50 different languages as it inherits the multilingual encoder pre-trained on these languages from Tang et al. (2020). However, the original parser has only been explicitly trained for German (de), Spanish (es), Italian (it), Chinese (zh), and English (en). We hypothesize that including more languages in training can help improve the overall parsing accuracy. Therefore, we add French (fr), another major language, to the training of the parser.[4]

**Integration** We explore four different choices for the integration of the text embedding $\mathbf{s}$ and the AMR embedding $\mathbf{h}$: $\mathbf{s} \oplus \mathbf{h}$, $\mathbf{s} + \mathbf{h}$, $\frac{\mathbf{s}}{\|\mathbf{s}\|} \oplus \frac{\mathbf{h}}{\|\mathbf{h}\|}$, $\frac{\mathbf{s}}{\|\mathbf{s}\|} + \frac{\mathbf{h}}{\|\mathbf{h}\|}$, where $\oplus$ denotes the concatenation of

[2] https://huggingface.co/docs/transformers/model_doc/marian

[3] https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt

[4] The extension only requires an English-to-French translation system, which is the OPUS-MT system in our implementation. We refer readers to Cai et al. (2021) for more details.

two vectors. Empirically, we find that $\frac{\mathbf{s}}{\|\mathbf{s}\|} \oplus \frac{\mathbf{h}}{\|\mathbf{h}\|}$ generally works best.

## 5 Evaluation Benchmark

To provide a more comprehensive evaluation of multilingual sentence representations, In addition to traditional semantic textual similarity tasks, we also introduce a set of downstream transfer tasks.

### 5.1 Semantic Textual Similarity

**Multilingual STS** The goal of semantic textual similarity (STS) is to assign for a pair of sentences a score indicating their semantic similarity. For example, a score of 0 indicates not related and 5 indicates semantically equivalent. We use the datasets in Reimers and Gurevych (2020), which is an extended version of the multilingual STS 2017 dataset (Cer et al., 2017). The evaluation is done by comparing the distance in the embedding space and the human-annotated scores in the dataset.

### 5.2 Transfer Tasks

We evaluate the quality of the multilingual sentence embeddings on the following cross-lingual sentence/sentence-pair classification benchmarks:

**XNLI** The Cross-lingual Natural Language Inference benchmark (Conneau et al., 2018) is used to estimate the capability of cross-lingual / multilingual models in recognizing textual entailment. The evaluation sets of XNLI are created by manually translating the development corpus and the testing corpus of MultiNLI (Williams et al., 2018) to 15 other languages.

**PAWS-X** The Cross-lingual Paraphrase Adversaries from Word Scrambling benchmark (Yang et al., 2019a) consists of golden English paraphrase identification pairs from PAWS (Zhang et al., 2019b) and around 24k human translations of PAWS evaluation sets (i.e., development set and testing set) in English, French, Spanish, German, Chinese, Japanese (ja), and Korean (ko).

**QAM** The Question-Answer Matching task aims to predict if the given (question, passage) pair is a QA pair. We use the multilingual QAM dataset from XGLUE (Liang et al., 2020), which provides the labeled instance (question, passage, label) in English, French, and German, to evaluate the effectiveness of multilingual sentence embeddings.

| Task | Languages |
|---|---|
| XNLI | en, fr, de ,es, zh, el, bg, ru, tr, ar, vi, th, hi, sw, ur |
| PAWS-X | en, fr, de, es, ja, ko |
| QAM | en, fr, de |
| MLDoc | en, fr, de, es, zh, ru, it, ja |
| MARC | en, fr, de, es, zh, ja |

Table 1: Test languages in different transfer tasks.

**MLDoc** The Multilingual Document Classification benchmark (Schwenk and Li, 2018) is a multilingual corpus with a collection of news documents written in English, German, Spanish, French, Italian, Chinese, Japanese, and Russian (ru). The entire corpus is manually classified into four groups according to the topic of the document.

**MARC** The Multilingual Amazon Review Corpus (Keung et al., 2020) is a large-scale collection of Amazon user reviews for multilingual rating classification. The corpus covers 6 languages, including English, German, French, Spanish, and Chinese, Japanese.

## 6 Experiments

### 6.1 Experimental Setup

For STS tasks, following previous work (Gao et al., 2021), we define the similarity score as the cosine similarity of sentence embeddings and compute the Spearman's rank correlation between the computed score and the gold score.

For downstream transfer tasks, we follow the conventional zero-shot cross-lingual transfer setting (Liang et al., 2020; Hu et al., 2020), where annotated training data is provided in English but none is provided in other languages. We fit a logistic regression classifier on top of fixed sentence representations and follow default configurations in Conneau and Kiela (2018); Gao et al. (2021). To faithfully reflect the multilinguality of multilingual sentence embeddings, we train exactly one model for each task. The union of the development sets in different languages is adopted for model selection.

### 6.2 Implementation Details

We initialize our AMR encoder with BERT (Devlin et al., 2019) (uncased) and take the `[CLS]` representation as the sentence embedding. By default, the AMR encoder is trained on English, German, Spanish, Italian, Chinese, French, and Arabic (ar). Each model is trained for a maximum of 9 epochs

with a learning rate of $5e-5$ and a batch size of 512. The temperature in Eq. (1) is set to be 0.05. For model selection, we use the STS-B development (Cer et al., 2017). We train a multilingual AMR parser on English, German, Spanish, Italian, Chinese, and French using the same recipe in Cai et al. (2021). We release our code at https://github.com/jcyk/MSE-AMR.

### 6.3 Baseline Systems

We evaluate the following systems:

**mBERT / XLM-R** We use the mean pooling of the outputs from the pre-trained mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020), which are pre-trained on multilingual data. However, no parallel or labeled data was used.

**mUSE** Multilingual Universal Sentence Encoder (Chidambaram et al., 2019) uses a dual-encoder transformer architecture and adopts contrastive objectives. It was trained on mined question-answer pairs, SNLI data, translated SNLI data, and parallel corpora over 16 languages.

**LASER** LASER (Artetxe and Schwenk, 2019) trains a sequence-to-sequence encoder-decoder architecture on parallel corpora for machine translation. The sentence representation is obtained via max-pooling over the output of the encoder. LASER was trained over 93 languages.

**LaBSE** Language-agnostic BERT Sentence Embedding (LaBSE) (Feng et al., 2020) was trained similar to mUSE with a translation ranking loss. It fine-tunes a dual-BERT architecture with 6 Billion translation pairs for 109 languages.

**Xpara** Reimers and Gurevych (2020) fine-tunes XLM-R to imitate SBERT-paraphrases (Reimers and Gurevych, 2019), a RoBERTa model trained on more than 50 Million English paraphrase pairs, with massive bilingual sentence pairs over 50 languages.

### 6.4 Model Variants

To study the effect of each modeling choice, we implement a series of model variants.

• **#1**: To show if learning from English data suffices, we train the AMR encoder with only English sentences and the AMR graphs derived from them.

• **#2**: To study the effect of extending the training of the multilingual AMR parser to French, we use the original parser in Cai et al. (2021), which does not include French.

• **#3**: To measure the help of involving more languages when training the AMR encoder, we train the AMR encoder without the AMR graphs derived from French and Arabic.

• **#4**: To validate the usefulness of adding the English sentences to the training of the AMR encoder, we train the AMR encoder without English sentences.

• **#5**: The standard model as described in Section 6.2.

For each model variant, we report the average performance over five different runs (different random seeds) throughout this paper.

| Model | # | EN-EN | ES-ES | AR-AR | Avg. (Δ) |
|---|---|---|---|---|---|
| mBERT | | 54.36 | 56.69 | 50.86 | 53.97 |
| XLM-R | | 52.18 | 49.58 | 25.50 | 42.42 |
| mUSE | | 86.39 | 86.86 | 76.41 | 83.22 |
| LASER | | 77.62 | 79.69 | 68.84 | 75.38 |
| LaBSE | | 79.45 | 80.83 | 69.07 | 76.45 |
| Xpara | | 88.10 | 85.71 | 79.10 | 84.30 |
| mUSE++ | 1 | 88.51 | 86.53 | 80.12 | 85.05 (+1.83) |
| | 2 | 88.57 | **87.57** | 80.45 | 85.53 (+2.31) |
| | 3 | 88.30 | 87.07 | 80.32 | 85.23 (+2.01) |
| | 4 | 88.38 | 86.95 | 80.56 | 85.30 (+2.08) |
| | 5 | 88.74 | 87.14 | 80.67 | 85.52 (+2.30) |
| Xpara++ | 1 | 89.31 | 85.89 | 80.62 | 85.28 (+0.98) |
| | 2 | 89.19 | 86.60 | **81.85** | **85.88 (+1.58)** |
| | 3 | 89.06 | 86.40 | 80.78 | 85.42 (+1.12) |
| | 4 | 89.27 | 86.34 | 80.74 | 85.45 (+1.15) |
| | 5 | **89.45** | 86.52 | 81.04 | 85.66 (+1.36) |

Table 2: Performance (Spearman's correlation) on STS tasks (monolingual setup). Δ indicates the improvements from our methods.

### 6.5 Results

**Multilingual STS** Table 2 and Table 3 show the evaluation results on 3 monolingual STS tasks and 7 cross-lingual STS tasks respectively. As seen, the best-performing models in the literature are mUSE and Xpara. Thus, we present the results of augmenting mUSE and Xpara with our AMR embeddings, denoted by mUSE++ and Xpara++ respectively. Using AMR embeddings substantially improves both two models across the monolingual (up to +2.31 on avg.) and cross-lingual settings (up to +2.22 on avg.), greatly advancing the state-of-the-art performance. The average scores of monolingual and cross-lingual settings are pushed to 85.88 and 84.25 respectively. The improvements for mUSE are generally greater than those for Xpara, even though the training data of mUSE overlaps

| Model | # | EN-AR | EN-DE | EN-TR | EN-ES | EN-FR | EN-IT | EN-NL | Avg. (Δ) |
|---|---|---|---|---|---|---|---|---|---|
| mBERT | | 18.67 | 33.86 | 16.02 | 21.47 | 32.98 | 34.02 | 35.30 | 27.47 |
| XLM-R | | 15.71 | 21.30 | 12.07 | 10.60 | 16.64 | 22.88 | 23.95 | 17.59 |
| mUSE | | 79.27 | 82.13 | 75.47 | 79.62 | 82.64 | 84.55 | 84.07 | 81.11 |
| LASER | | 66.53 | 64.20 | 71.99 | 57.93 | 69.06 | 70.83 | 68.67 | 67.03 |
| LaBSE | | 74.51 | 73.85 | 72.07 | 65.71 | 76.98 | 76.99 | 75.22 | 73.62 |
| Xpara | | 81.81 | 83.66 | 80.16 | 84.05 | 83.16 | 85.66 | 83.67 | 83.17 |
| mUSE++ | 1 | 79.99 | 83.81 | 75.08 | 81.74 | 83.39 | 86.84 | 86.61 | 82.49 (+1.38) |
| | 2 | 81.85 | 85.01 | 75.12 | 83.25 | 83.68 | 85.62 | 84.78 | 82.76 (+1.65) |
| | 3 | 80.22 | 84.18 | 76.53 | 82.79 | 84.60 | 86.75 | 86.28 | 83.05 (+1.94) |
| | 4 | 80.43 | 84.41 | 76.58 | 82.59 | 84.52 | 86.72 | 86.38 | 83.09 (+1.98) |
| | 5 | 80.52 | 84.87 | 76.57 | 83.01 | 84.91 | 86.71 | **86.71** | 83.33 (+2.22) |
| Xpara++ | 1 | 81.16 | 84.86 | 77.75 | 83.71 | 83.61 | 87.32 | 85.42 | 83.40 (+0.23) |
| | 2 | **82.89** | 85.56 | 77.66 | **85.14** | 84.44 | 86.35 | 84.08 | 83.73 (+0.56) |
| | 3 | 81.47 | 85.28 | 79.21 | 84.55 | 84.77 | 87.02 | 85.15 | 83.92 (+0.75) |
| | 4 | 81.45 | 85.58 | 79.20 | 84.47 | 84.84 | 87.13 | 85.34 | 84.00 (+0.83) |
| | 5 | 81.73 | **85.62** | **79.50** | 84.76 | **85.22** | **87.33** | 85.58 | **84.25 (+1.08)** |

Table 3: Performance (Spearman's correlation) on STS tasks (cross-lingual setup).

| Model | # | MLDoc seen/all | XNLI seen/all | PAWS-X seen/all | MARC seen/all | QAM | Avg. (Δ) seen/all |
|---|---|---|---|---|---|---|---|
| mBERT | | 80.17/77.90 | 47.23/44.41 | 57.30/57.05 | 38.66/38.43 | 55.25 | 55.72/54.61 |
| XLM-R | | 79.99/77.86 | 48.57/46.83 | 56.10/56.06 | 50.63/50.03 | 55.58 | 58.17/57.27 |
| mUSE | | 79.79/77.20 | 55.60/48.63 | 57.68/57.34 | 47.28/46.37 | 60.82 | 60.23/58.07 |
| LASER | | 77.42/74.63 | 60.36/58.87 | 73.89/70.81 | 49.08/47.97 | 58.28 | 63.81/62.11 |
| LaBSE | | 84.93/82.29 | 58.24/56.65 | 58.75/58.31 | 49.95/48.85 | 59.34 | 62.24/61.09 |
| Xpara | | 65.68/62.42 | 55.81/53.33 | 58.50/58.06 | 49.92/48.79 | 56.25 | 57.23/55.77 |
| LASER++ | 1 | 81.67/78.67 | 63.48/57.53 | 73.64/70.49 | 49.42/48.32 | 59.16 | 65.47/62.83 (+1.66/+0.72) |
| | 2 | 81.71/78.71 | 63.64/57.33 | 73.51/70.36 | 49.46/48.18 | 59.08 | 65.48/62.73 (+1.67/+0.62) |
| | 3 | 81.91/79.03 | 63.65/57.86 | 73.68/70.50 | 49.62/48.51 | 59.37 | 65.65/63.05 (+1.84/+0.94) |
| | 4 | 82.74/79.74 | 63.45/57.66 | 73.72/70.52 | 49.32/48.27 | 59.42 | 65.73/63.12 (+1.92/+1.01) |
| | 5 | 82.65/79.80 | **63.88/57.98** | **73.79/70.61** | 49.44/48.31 | 59.41 | **65.83/63.22 (+2.02/+1.11)** |
| LaBSE++ | 1 | 85.59/82.86 | 59.24/53.06 | 59.92/59.13 | 51.08/50.07 | 59.73 | 63.11/60.97 (+0.87/-0.12) |
| | 2 | 85.68/82.77 | 59.66/53.07 | 59.44/58.80 | 50.87/49.84 | 59.85 | 63.10/60.87 (+0.86/-0.22) |
| | 3 | 85.56/82.82 | 59.69/53.54 | 59.66/58.95 | 51.15/50.08 | 59.99 | 63.21/61.08 (+0.97/-0.01) |
| | 4 | **85.89/83.10** | 59.44/53.31 | 59.68/58.98 | 51.13/50.11 | **60.21** | 63.27/61.14 (+1.03/+0.05) |
| | 5 | 85.70/83.02 | 59.66/53.55 | 59.81/59.07 | **51.20/50.21** | 59.99 | 63.27/61.17 (+1.03/+0.08) |

Table 4: Performance (accuracy) on transfer tasks. Δ indicates the improvements from our methods.

with our AMR encoders. We hypothesize that it is because Xpara is trained on paraphrase corpus, which diminishes the ability of AMR to group different expressions of the same semantics.

One interesting finding is that model variant #2 performs best on monolingual settings while model variant #5 attains the best results on cross-lingual settings. We believe that adding more languages to the training of the AMR parser helps the generalization to other languages and reduces the parsing inconsistency across different languages. Thus, the AMR graphs from different languages are better aligned, leading to a better-aligned vector space. On the other hand, adding more language may decrease the parsing accuracies on individual languages due to the fixed model capacity. Note that all other model variants except #2 underperform #5, confirming the effectiveness of the proposed mixed training strategy.

**Transfer Tasks** Table 4 shows the evaluation results on transfer tasks. For each task, we report the macro-average scores across different languages. The results for each language can be found in Appendix. Different to previous work, our AMR encoders are only trained with a few languages (en, de, es, it, zh, fr, and ar) at most. To isolate the effect on unseen languages, we separate the results on those seen languages from all languages (seen/all). First of all, we find that the rankings of existing models are quite different to the results on STS tasks. LASER and LaBSE achieve the best results on most transfer tasks except for QAM, and outperforms mUSE and Xpara by large margins in most cases. The results demonstrate the limitation of solely testing on sentence similarity measurement.

Next, we augment the best-performing models,

LASER and LaBSE, with our AMR embeddings (LASER++ and LaBSE++). For seen languages, our methods substantially boost the performance of these two models across different tasks (up to +2.02 on avg.). The performance gains over LASER are greater than those over LaBSE. Note that LASER is trained with an encoder-decoder architecture and both LaBSE and our AMR encoders are trained with a Siamese network. Therefore, we believe the AMR embeddings are more complementary to LASER.

When considering all languages, the improvements over LASER are also considerable (up to +1.11 on avg.). However, according to the average scores over different tasks, the AMR embeddings seem to fail to improve LaBSE; We even observe a performance drop for model variants #1-#3. Nevertheless, the performance drop largely comes from XNLI while the scores on other tasks are instead boosted. This is because the test sets of XNLI include some distant languages (e.g., Swahili and Urdu) that our multilingual AMR parser cannot handle well (see the results on individual languages in Table 6 in Appendix). We conjecture that further extending the multilingual AMR parser to more languages can alleviate this problem. The comparison among different model variants provides a basis for the above speculation. As we can see, model variant #2 (exclude French from the training of the multilingual AMR parser) performs worst. Also, model variants #1 (drop all non-English AMR graphs for training) and #2 (drop the AMR graphs derived from French and Arabic) are the other two variants that negatively impact the average performance. Another interesting observation is that model variant #4 performs best on MLDoc and QAM, suggesting English sentences might not be necessary.

## 7 Conclusion

This paper presented a thorough evaluation of existing multilingual sentence embeddings, ranging from traditional text similarity measurement to a new variety of transfer tasks. We found that different methods excel at different tasks. We then proposed to improve existing methods with universal AMR embeddings, which leads to better performance on all tasks.

## Limitations

Although our work provides an effective solution for improving multilingual sentence embeddings with AMR, we acknowledge some limitations of this study and further discuss them in the following: (1) Our framework treats the text encoder as a black box and does not care too much about its implementation. Although it is flexible and straightforward to apply our framework to any multilingual sentence embedding model, designing more specific interaction mechanisms for different text encoders is supposed to be better and we leave it as future work. (2) The improvement from our framework is higher in seen languages than unseen languages. Further extending the language coverage in the training phases of both the multilingual AMR parser and the AMR encoder is presumably beneficial to the cross-lingual generalization capability of the AMR embeddings. However, due to the limit of computational resources, we only consider a few languages in the experiments.

## References


Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. SemEval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. SemEval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 81–91.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511. Association for Computational Linguistics.

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop*

*on Semantic Evaluation (SemEval 2012)*, pages 385–393.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *SEM 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association of Computational Linguistics (TACL)*, 7:597–610.

Xuefeng Bai, Yulong Chen, and Yue Zhang. 2022. Graph pre-training for AMR parsing and generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6001–6015, Dublin, Ireland. Association for Computational Linguistics.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Daniel Beck, Gholamreza Haffari, and Trevor Cohn. 2018. Graph-to-sequence learning using gated graph neural networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 273–283, Melbourne, Australia. Association for Computational Linguistics.

Michele Bevilacqua, Rexhina Blloshmi, and Roberto Navigli. 2021. One spring to rule them both: Symmetric amr semantic parsing and generation without a complex pipeline. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12564–12573.

Rexhina Blloshmi, Rocco Tripodi, and Roberto Navigli. 2020. XL-AMR: Enabling cross-lingual AMR parsing with transfer learning techniques. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2487–2500, Online. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 632–642.

Jane Bromley, James W Bentz, Léon Bottou, Isabelle Guyon, Yann LeCun, Cliff Moore, Eduard Säckinger, and Roopak Shah. 1993. Signature verification using a "siamese" time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, 7(04):669–688.

Deng Cai and Wai Lam. 2019. Core semantic first: A top-down approach for AMR parsing. 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 3799–3809, Hong Kong, China. Association for Computational Linguistics.

Deng Cai and Wai Lam. 2020a. AMR parsing via graph-sequence iterative inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1290–1301, Online. Association for Computational Linguistics.

Deng Cai and Wai Lam. 2020b. Graph transformer for graph-to-sequence learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7464–7471.

Deng Cai, Xin Li, Jackie Chun-Sing Ho, Lidong Bing, and Wai Lam. 2021. Multilingual amr parsing with noisy knowledge distillation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2778–2789.

Fredrik Carlsson, Amaru Cuba Gyllensten, Evangelia Gogoulou, Erik Ylipää Hellqvist, and Magnus Sahlgren. 2021. Semantic re-tuning with contrastive tension. In *International Conference on Learning Representations (ICLR)*.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations*, pages 169–174.

Minmin Chen. 2017. Efficient vector representation for documents through corruption. In *International Conference on Learning Representations (ICLR)*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, pages 1597–1607.

Ting Chen, Yizhou Sun, Yue Shi, and Liangjie Hong. 2017. On sampling strategies for neural network-based collaborative filtering. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 767–776.

Xinlei Chen and Kaiming He. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758.

Muthu Chidambaram, Yinfei Yang, Daniel Cer, Steve Yuan, Yunhsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Learning cross-lingual sentence representations via a multi-task dual-encoder model. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 250–259, Florence, Italy. Association for Computational Linguistics.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. 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 4069–4082, Hong Kong, China. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard 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.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In *International Conference on Language Resources and Evaluation (LREC)*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 670–680.

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

Marco Damonte and Shay B. Cohen. 2018. Cross-lingual Abstract Meaning Representation parsing. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1146–1155, New Orleans, Louisiana. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186.

Akiko Eriguchi, Melvin Johnson, Orhan Firat, Hideto Kazawa, and Wolfgang Macherey. 2018. Zero-shot cross-lingual classification using multilingual neural machine translation. *arXiv preprint arXiv:1809.04686*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Zhe Gan, Yunchen Pu, Ricardo Henao, Chunyuan Li, Xiaodong He, and Lawrence Carin. 2017. Learning generic sentence representations using convolutional neural networks. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 2390–2400, Copenhagen, Denmark. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2021. DeCLUTR: Deep contrastive learning for unsupervised textual representations. In *Association for Computational Linguistics and International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 879–895.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. 2020. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*.

Mandy Guo, Qinlan Shen, Yinfei Yang, Heming Ge, Daniel Cer, Gustavo Hernandez Abrego, Keith Stevens, Noah Constant, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Effective parallel corpus mining using bilingual sentence embeddings. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 165–176, Brussels, Belgium. Association for Computational Linguistics.

Zhijiang Guo, Yan Zhang, Zhiyang Teng, and Wei Lu. 2019. Densely connected graph convolutional networks for graph-to-sequence learning. *Transactions of the Association for Computational Linguistics*, 7:297–312.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 1735–1742. IEEE.

Jan Hajič, Ondřej Bojar, and Zdeňka Urešová. 2014. Comparing Czech and English AMRs. In *Proceedings of Workshop on Lexical and Grammatical Resources for Language Processing*, pages 55–64, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv:1705.00652*.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual models for compositional distributed semantics. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 58–68, Baltimore, Maryland. Association for Computational Linguistics.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning distributed representations of sentences from unlabelled data. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1367–1377.

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 generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Junjie Huang, Duyu Tang, Wanjun Zhong, Shuai Lu, Linjun Shou, Ming Gong, Daxin Jiang, and Nan Duan. 2021. Whiteningbert: An easy unsupervised sentence embedding approach. *arXiv preprint arXiv:2104.01767*.

Yacine Jernite, Samuel R Bowman, and David Sontag. 2017. Discourse-based objectives for fast unsupervised sentence representation learning. *arXiv preprint arXiv:1705.00557*.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The multilingual Amazon reviews corpus. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 4563–4568, Online. Association for Computational Linguistics.

Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. 2021. Self-guided contrastive learning for BERT sentence representations. In *Association for Computational Linguistics and International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 2528–2540.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3294–3302.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International Conference on Machine Learning (ICML)*, pages 1188–1196. PMLR.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Association for Computational Linguistics (ACL)*, pages 7871–7880.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In *International Conference on Learning Representations (ICLR)*.

Manuel Mager, Ramón Fernandez Astudillo, Tahira Naseem, Md Arafat Sultan, Young-Suk Lee, Radu Florian, and Salim Roukos. 2020. GPT-too: A language-model-first approach for AMR-to-text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1846–1852, Online. Association for Computational Linguistics.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *International Conference on Language Resources and Evaluation (LREC)*, pages 216–223.

Yu Meng, Chenyan Xiong, Payal Bajaj, Saurabh Tiwary, Paul Bennett, Jiawei Han, and Xia Song. 2021. COCO-LM: Correcting and contrasting text sequences for language model pretraining. *arXiv preprint arXiv:2102.08473*.

Juri Opitz and Anette Frank. 2022. Sbert studies meaning representations: Decomposing sentence embeddings into explainable amr meaning features. *arXiv preprint arXiv:2206.07023*.

Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2018. Unsupervised learning of sentence embeddings using compositional n-gram features. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 528–540.

Hieu Pham, Thang Luong, and Christopher Manning. 2015. Learning distributed representations for multilingual text sequences. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 88–94, Denver, Colorado. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Luigi Procopio, Rocco Tripodi, and Roberto Navigli. 2021. Sgl: Speaking the graph languages of semantic parsing via multilingual translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 325–337.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Leonardo F. R. Ribeiro, Claire Gardent, and Iryna Gurevych. 2019. Enhancing AMR-to-text generation with dual graph 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 3183–3194, Hong Kong, China. Association for Computational Linguistics.

Holger Schwenk. 2018. Filtering and mining parallel data in a joint multilingual space. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 228–234, Melbourne, Australia. Association for Computational Linguistics.

Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 157–167, Vancouver, Canada. Association for Computational Linguistics.

Holger Schwenk and Xian Li. 2018. A corpus for multilingual document classification in eight languages. In *International Conference on Language Resources and Evaluation (LREC)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Janaki Sheth, Young-Suk Lee, Ramón Fernandez Astudillo, Tahira Naseem, Radu Florian, Salim Roukos, and Todd Ward. 2021. Bootstrapping multilingual amr with contextual word alignments. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 394–404.

Karan Singla, Dogan Can, and Shrikanth Narayanan. 2018. A multi-task approach to learning multilingual representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 214–220, Melbourne, Australia. Association for Computational Linguistics.

Linfeng Song, Yue Zhang, Zhiguo Wang, and Daniel Gildea. 2018. A graph-to-sequence model for AMR-to-text generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1616–1626, Melbourne, Australia. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research (JMLR)*, 15(1):1929–1958.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv:2103.15316*.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*, pages 6000–6010.

John Wieting and Kevin Gimpel. 2018. ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Association for Computational Linguistics (ACL)*, pages 451–462.

John Wieting, Kevin Gimpel, Graham Neubig, and Taylor Berg-Kirkpatrick. 2019. Simple and effective paraphrastic similarity from parallel translations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4602–4608, Florence, Italy. Association for Computational Linguistics.

John Wieting, Graham Neubig, and Taylor Berg-Kirkpatrick. 2020. A bilingual generative transformer for semantic sentence embedding. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1581–1594.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1112–1122.

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020. Clear: Contrastive learning for sentence representation. *arXiv preprint arXiv:2012.15466*.

Nianwen Xue, Ondřej Bojar, Jan Hajič, Martha Palmer, Zdeňka Urešová, and Xiuhong Zhang. 2014. Not an interlingua, but close: Comparison of English AMRs to Chinese and Czech. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1765–1772, Reykjavik, Iceland. European Language Resources Association (ELRA).

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. ConSERT: A contrastive framework for self-supervised sentence representation transfer. In *Association for Computational Linguistics and International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 5065–5075.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94, Online. Association for Computational Linguistics.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019a. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Ziyi Yang, Chenguang Zhu, and Weizhu Chen. 2019b. Parameter-free sentence embedding via orthogonal basis. In *Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 638–648, Hong Kong, China. Association for Computational Linguistics.

Katherine Yu, Haoran Li, and Barlas Oguz. 2018. Multilingual seq2seq training with similarity loss for cross-lingual document classification. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 175–179, Melbourne, Australia. Association for Computational Linguistics.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. 2021. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv preprint arXiv:2103.03230*.

Sheng Zhang, Xutai Ma, Kevin Duh, and Benjamin Van Durme. 2019a. AMR parsing as sequence-to-graph transduction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 80–94, Florence, Italy. Association for Computational Linguistics.

Yan Zhang, Ruidan He, Zuozhu Liu, Lidong Bing, and Haizhou Li. 2021. Bootstrapped unsupervised sentence representation learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5168–5180.

Yan Zhang, Ruidan He, Zuozhu Liu, Kwan Hui Lim, and Lidong Bing. 2020. An unsupervised sentence embedding method by mutual information maximization. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1601–1610.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019b. PAWS: Paraphrase adversaries from word scrambling. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1298–1308.

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 Transfer task

We provide the detailed results on each language for each transfer task in Table 5-9.

| Model | # | DE | EN | ES | FR | IT | JA | RU | ZH | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|
| mBERT | | 83.73 | 89.88 | 75.75 | 83.73 | 68.25 | 71.12 | 71.08 | 79.65 | 77.90 |
| XLM-R | | 84.60 | 88.78 | 77.98 | 80.20 | 74.00 | 73.25 | 69.70 | 74.38 | 77.86 |
| mUSE | | 85.80 | 87.95 | 77.00 | 84.45 | 68.45 | 69.35 | 69.50 | 75.08 | 77.20 |
| LASER | | 84.28 | 84.95 | 72.85 | 79.25 | 69.67 | 67.10 | 65.42 | 73.50 | 74.63 |
| LaBSE | | 88.42 | 90.88 | 81.03 | 87.90 | 76.40 | 73.00 | 75.70 | 84.97 | 82.29 |
| Xpara | | 69.23 | 88.35 | 67.88 | 65.15 | 61.62 | 52.45 | 52.83 | 41.85 | 62.42 |
| | 1 | 87.99 | 89.22 | 79.41 | 82.75 | 71.33 | 72.72 | 66.62 | 79.35 | 78.67 |
| | 2 | 87.56 | 89.80 | 79.35 | 82.65 | 71.77 | 71.64 | 67.78 | 79.16 | 78.71 |
| LASER++ | 3 | 87.62 | 89.49 | 79.65 | 83.02 | 71.29 | 73.69 | 67.08 | 80.39 | 79.03 |
| | 4 | 88.22 | 90.27 | 80.39 | 83.93 | 72.17 | 73.99 | 67.53 | 81.45 | 79.74 |
| | 5 | 88.02 | 90.10 | 80.17 | 83.59 | 72.54 | 74.38 | 68.12 | 81.50 | 79.80 |
| | 1 | 89.38 | 91.70 | 83.28 | 88.07 | 76.22 | 75.45 | 73.87 | 84.87 | 82.86 |
| | 2 | 89.72 | 91.64 | 83.72 | 87.84 | 76.41 | 74.73 | 73.32 | 84.76 | 82.77 |
| LaBSE++ | 3 | 89.08 | 91.46 | 83.59 | 88.07 | 76.08 | 75.53 | 73.67 | 85.09 | 82.82 |
| | 4 | 89.41 | 91.45 | 84.51 | 88.30 | 76.37 | 75.75 | 73.76 | 85.27 | 83.11 |
| | 5 | 89.39 | 91.41 | 84.11 | 88.00 | 76.28 | 75.67 | 74.30 | 85.02 | 83.02 |

Table 5: MLDoc results of different sentence embedding models.

| Model | # | AR | BG | DE | EL | EN | ES | FR | HI | RU | SW | TH | TR | UR | VI | ZH | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| mBERT | | 42.57 | 45.35 | 46.75 | 43.99 | 53.53 | 47.64 | 47.60 | 41.54 | 46.07 | 37.49 | 36.75 | 43.17 | 40.46 | 47.96 | 45.31 | 44.41 |
| XLM-R | | 45.15 | 48.90 | 45.85 | 49.28 | 54.53 | 49.56 | 49.96 | 45.23 | 47.35 | 38.64 | 42.28 | 47.43 | 41.92 | 49.94 | 46.37 | 46.83 |
| mUSE | | 53.09 | 48.76 | 55.05 | 35.39 | 59.02 | 56.25 | 55.81 | 35.91 | 54.93 | 39.20 | 54.23 | 53.59 | 37.86 | 35.97 | 54.35 | 48.63 |
| LASER | | 58.80 | 60.26 | 60.96 | 60.68 | 61.54 | 60.60 | 60.52 | 56.19 | 59.50 | 53.13 | 59.26 | 59.46 | 52.46 | 59.92 | 59.76 | 58.87 |
| LaBSE | | 56.75 | 57.56 | 57.07 | 57.88 | 60.58 | 58.74 | 58.08 | 55.57 | 56.41 | 54.77 | 53.79 | 56.43 | 52.08 | 55.89 | 58.22 | 56.65 |
| Xpara | | 54.33 | 56.15 | 56.45 | 55.97 | 57.07 | 55.95 | 56.31 | 52.18 | 55.31 | 34.17 | 53.67 | 54.73 | 48.20 | 54.75 | 54.75 | 53.33 |
| | 1 | 58.52 | 51.44 | 63.94 | 45.89 | 67.37 | 63.83 | 65.02 | 57.12 | 61.74 | 43.72 | 49.46 | 59.67 | 51.40 | 61.58 | 62.20 | 57.53 |
| | 2 | 58.10 | 48.34 | 64.45 | 45.63 | 67.59 | 65.03 | 64.17 | 57.32 | 61.58 | 44.29 | 49.98 | 58.57 | 51.34 | 60.98 | 62.52 | 57.33 |
| LASER++ | 3 | 58.76 | 51.51 | 64.45 | 46.45 | 67.26 | 64.46 | 64.96 | 57.16 | 62.44 | 43.94 | 50.04 | 60.29 | 51.84 | 62.36 | 62.01 | 57.86 |
| | 4 | 58.67 | 51.59 | 64.02 | 46.39 | 66.76 | 64.24 | 64.71 | 57.09 | 62.09 | 43.88 | 49.73 | 60.08 | 51.51 | 61.90 | 62.30 | 57.67 |
| | 5 | 59.10 | 51.66 | 64.84 | 46.25 | 67.02 | 64.61 | 65.19 | 57.66 | 62.80 | 44.00 | 49.89 | 60.23 | 51.78 | 62.07 | 62.55 | 57.98 |
| | 1 | 54.90 | 45.94 | 58.92 | 39.92 | 63.98 | 59.72 | 60.04 | 54.54 | 56.88 | 39.56 | 42.32 | 55.40 | 49.04 | 56.87 | 57.85 | 53.06 |
| | 2 | 55.38 | 43.67 | 59.54 | 40.08 | 64.28 | 60.88 | 59.24 | 54.47 | 56.81 | 40.22 | 41.74 | 54.76 | 49.47 | 56.95 | 58.62 | 53.07 |
| LaBSE++ | 3 | 55.98 | 46.60 | 59.68 | 40.21 | 63.93 | 60.70 | 59.99 | 54.79 | 57.50 | 39.64 | 42.42 | 56.24 | 49.90 | 57.61 | 57.85 | 53.54 |
| | 4 | 55.91 | 46.34 | 59.30 | 40.71 | 63.49 | 60.52 | 59.87 | 54.55 | 57.18 | 39.38 | 42.09 | 56.09 | 49.47 | 57.24 | 57.56 | 53.31 |
| | 5 | 56.41 | 46.39 | 59.30 | 40.67 | 63.63 | 60.51 | 60.16 | 55.12 | 57.31 | 39.73 | 42.26 | 56.21 | 49.80 | 57.77 | 57.93 | 53.55 |

Table 6: XNLI results of different sentence embedding models.

| Model | # | DE | EN | ES | FR | JA | KO | ZH | Avg. |
|---|---|---|---|---|---|---|---|---|---|
| mBERT | | 57.00 | 57.30 | 57.45 | 57.40 | 56.85 | 56.00 | 57.35 | 57.05 |
| XLM-R | | 55.70 | 55.70 | 55.65 | 55.25 | 56.05 | 55.85 | 58.20 | 56.06 |
| mUSE | | 57.70 | 58.10 | 56.45 | 57.35 | 56.70 | 56.25 | 58.80 | 57.34 |
| LASER | | 72.20 | 79.80 | 75.00 | 74.80 | 65.40 | 60.85 | 67.65 | 70.81 |
| LaBSE | | 58.80 | 58.90 | 57.55 | 59.50 | 57.10 | 57.30 | 59.00 | 58.31 |
| Xpara | | 59.00 | 57.35 | 58.35 | 59.30 | 57.45 | 56.50 | 58.50 | 58.06 |
| | 1 | 72.13 | 80.06 | 74.62 | 74.34 | 64.41 | 60.79 | 67.05 | 70.48 |
| | 2 | 72.51 | 79.57 | 74.67 | 74.12 | 64.49 | 60.46 | 66.69 | 70.36 |
| LASER++ | 3 | 72.15 | 80.34 | 74.59 | 74.26 | 64.37 | 60.75 | 67.06 | 70.50 |
| | 4 | 72.09 | 80.18 | 74.92 | 74.52 | 64.26 | 60.79 | 66.87 | 70.52 |
| | 5 | 72.23 | 80.01 | 74.82 | 74.74 | 64.37 | 60.91 | 67.16 | 70.61 |
| | 1 | 59.50 | 61.61 | 58.74 | 60.28 | 57.25 | 57.06 | 59.49 | 59.13 |
| | 2 | 59.17 | 60.87 | 57.95 | 59.16 | 57.51 | 56.88 | 60.04 | 58.79 |
| LaBSE++ | 3 | 59.83 | 60.69 | 58.43 | 59.70 | 57.30 | 57.01 | 59.66 | 58.95 |
| | 4 | 59.42 | 60.63 | 58.77 | 60.01 | 57.40 | 57.03 | 59.58 | 58.98 |
| | 5 | 59.55 | 61.25 | 58.61 | 60.05 | 57.29 | 57.13 | 59.61 | 59.07 |

Table 7: PAWS-X results of different sentence embedding models.

| Model | # | DE | EN | ES | FR | JA | ZH | Avg. |
|---|---|---|---|---|---|---|---|---|
| mBERT | | 38.28 | 45.54 | 38.32 | 38.40 | 32.78 | 37.28 | 38.43 |
| XLM-R | | 52.16 | 54.78 | 48.70 | 48.08 | 49.42 | 47.02 | 50.03 |
| mUSE | | 47.18 | 50.90 | 48.52 | 47.76 | 42.02 | 41.82 | 46.37 |
| LASER | | 51.44 | 52.68 | 49.26 | 50.00 | 42.02 | 42.42 | 47.97 |
| LaBSE | | 51.46 | 52.40 | 49.86 | 50.46 | 45.58 | 43.36 | 48.85 |
| Xpara | | 52.24 | 53.50 | 49.22 | 50.12 | 44.50 | 43.14 | 48.79 |
| LASER++ | 1 | 51.48 | 53.90 | 50.22 | 49.92 | 41.58 | 42.84 | 48.32 |
| | 2 | 51.76 | 54.02 | 50.25 | 49.90 | 41.39 | 41.76 | 48.18 |
| | 3 | 51.54 | 54.24 | 50.18 | 50.23 | 41.90 | 42.97 | 48.51 |
| | 4 | 51.74 | 54.13 | 49.64 | 50.04 | 41.04 | 43.06 | 48.28 |
| | 5 | 51.70 | 54.20 | 49.80 | 50.42 | 41.07 | 42.68 | 48.31 |
| LaBSE++ | 1 | 53.26 | 54.10 | 50.78 | 51.07 | 46.18 | 45.01 | 50.07 |
| | 2 | 52.80 | 54.24 | 50.78 | 50.45 | 46.06 | 44.69 | 49.84 |
| | 3 | 53.27 | 54.35 | 50.71 | 51.07 | 46.34 | 44.74 | 50.08 |
| | 4 | 53.33 | 54.02 | 50.92 | 51.16 | 46.23 | 45.03 | 50.12 |
| | 5 | 53.53 | 54.33 | 50.77 | 51.34 | 46.02 | 45.29 | 50.21 |

Table 8: MARC results of different sentence embedding models.

| Model | # | DE | EN | FR | Avg. |
|---|---|---|---|---|---|
| mBERT | | 54.21 | 56.60 | 54.94 | 55.25 |
| XLM-R | | 55.30 | 57.18 | 54.25 | 55.58 |
| mUSE | | 62.60 | 58.01 | 61.84 | 60.82 |
| LASER | | 57.95 | 58.63 | 58.25 | 58.28 |
| LaBSE | | 59.06 | 58.15 | 60.82 | 59.34 |
| Xpara | | 58.90 | 57.01 | 60.08 | 58.66 |
| LASER++ | 1 | 59.17 | 59.62 | 58.69 | 59.16 |
| | 2 | 58.81 | 59.52 | 58.91 | 59.08 |
| | 3 | 59.24 | 59.96 | 58.89 | 59.37 |
| | 4 | 59.46 | 59.88 | 58.91 | 59.42 |
| | 5 | 59.34 | 59.94 | 58.94 | 59.41 |
| LaBSE++ | 1 | 61.03 | 57.42 | 60.75 | 59.73 |
| | 2 | 60.81 | 57.26 | 61.48 | 59.85 |
| | 3 | 60.94 | 57.99 | 61.05 | 59.99 |
| | 4 | 61.23 | 58.48 | 60.91 | 60.21 |
| | 5 | 60.55 | 58.83 | 60.60 | 59.99 |

Table 9: QAM results of different sentence embedding models.