EASE: Entity-Aware Contrastive Learning of Sentence Embedding

Sosuke Nishikawa[1,2*] **Ryokan Ri**[1,2] **Ikuya Yamada**[2,4]
sosuke-nishikawa@alumni.u-tokyo.ac.jp ryo0123@ousia.jp ikuya@ousia.jp

Yoshimasa Tsuruoka[1] **Isao Echizen**[1,3]
tsuruoka@logos.t.u-tokyo.ac.jp iechizen@nii.ac.jp

[1]The University of Tokyo [2]Studio Ousia
[3]National Institute of Informatics [4]RIKEN AIP

Abstract

We present EASE, a novel method for learning sentence embeddings via contrastive learning between sentences and their related entities. The advantage of using entity supervision is twofold: (1) entities have been shown to be a strong indicator of text semantics and thus should provide rich training signals for sentence embeddings; (2) entities are defined independently of languages and thus offer useful cross-lingual alignment supervision. We evaluate EASE against other unsupervised models both in monolingual and multilingual settings. We show that EASE exhibits competitive or better performance in English semantic textual similarity (STS) and short text clustering (STC) tasks and it significantly outperforms baseline methods in multilingual settings on a variety of tasks. Our source code, pretrained models, and newly constructed multilingual STC dataset are available at https://github.com/studio-ousia/ease.

1 Introduction

The current dominant approach to learning sentence embeddings is fine-tuning general-purpose pretrained language models, such as BERT (Devlin et al., 2019), with a particular training supervision. The type of supervision can be natural language inference data (Reimers and Gurevych, 2019), adjacent sentences (Yang et al., 2021), or a parallel corpus for multilingual models (Feng et al., 2020).

In this paper, we explore a type of supervision that has been under-explored in the literature: *entity hyperlink annotations* from Wikipedia. Their advantage is twofold: (1) entities have been shown to be a strong indicator of text semantics (Gabrilovich and Markovitch, 2007; Yamada et al., 2017, 2018; Ling et al., 2020) and thus should provide rich training signals for sentence embeddings; (2) entities are defined independently of languages and thus offer a useful cross-lingual alignment supervision (Calixto et al., 2021; Nishikawa et al., 2021; Jian et al., 2022; Ri et al., 2022). The extensive multilingual support of Wikipedia alleviates the need for a parallel resource to train well-aligned multilingual sentence embeddings, especially for low-resource languages. To demonstrate the effectiveness of entity-based supervision, we present **EASE** (Entity-Aware contrastive learning of Sentence Embeddings), which produces high-quality sentence embeddings in both monolingual and multilingual settings.

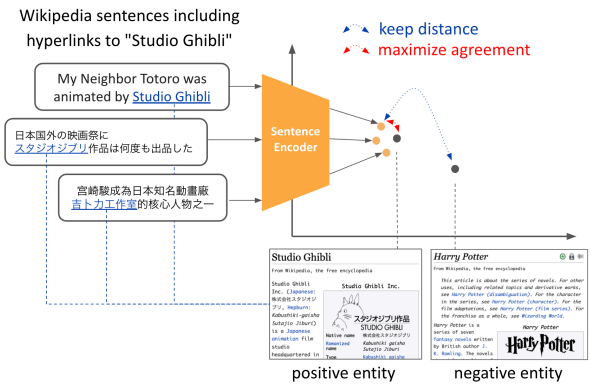

Figure 1: Illustration of the main concept behind EASE. Using a contrastive framework, sentences are embedded in the neighborhood of their hyperlink entity embeddings and kept apart from irrelevant entities. Here, we share the entity embeddings across languages for multilingual models to facilitate cross-lingual alignment of the representation.

EASE learns sentence embeddings with two types of objectives: (1) our novel entity contrastive learning (CL) loss between sentences and their related entities (Figure 1); (2) the self-supervised CL loss with dropout noise. The entity CL objective pulls the embeddings of sentences and their related entities close while keeping unrelated entities apart. The objective is expected to arrange the sentence embeddings in accordance with semantics captured

* Work done as an intern at Studio Ousia.

by the entities. To further exploit the knowledge in Wikipedia and improve the learned embeddings, we also introduce a method for mining hard negatives based on the entity type. The second objective, the self-supervised CL objective with dropout noise (Gao et al., 2021; Liu et al., 2021), is combined with the first one to enable sentence embeddings to capture fine-grained text semantics. We evaluate our model against other state-of-the-art unsupervised sentence embedding models, and show that EASE exhibits competitive or better performance on semantic textual similarity (STS) and short text clustering (STC) tasks.

We also apply EASE to multilingual settings. To facilitate the evaluation of the high-level semantics of multilingual sentence embeddings, we construct a multilingual text clustering dataset, MewsC-16 (Multilingual Short Text Clustering Dataset for News in 16 languages). Multilingual EASE is trained using the entity embeddings shared across languages. We show that, given the cross-lingual alignment supervision from the shared entities, multilingual EASE significantly outperforms the baselines in multilingual STS, STC, parallel sentence matching, and cross-lingual document classification tasks.

We further demonstrate the effectiveness of the multilingual entity CL in a more realistic scenario for low-resource languages. Using multilingual entity CL, we fine-tune a competitive multilingual sentence embedding model, LaBSE (Feng et al., 2020), and show that the tuning improves the performance of parallel sentence matching for low-resource languages under-supported by the model.

Finally, we analyze the EASE model by studying ablated models and the multilingual properties of the sentence embeddings to shed light on the source of the improvement in the model.

2 Related Work

2.1 Sentence Embeddings

Sentence embeddings, which represent the meaning of sentences in the form of a dense vector, have been actively studied. One of the earliest methods is Paragraph Vector (Le and Mikolov, 2014) in which sentence embeddings are trained to predict words within the text. Subsequently, various kinds of training tasks have been explored including reconstructing or predicting adjacent sentences (Kiros et al., 2015; Logeswaran and Lee, 2018) and solving a natural language inference (NLI) task (Conneau et al., 2017).

Recently, with the advent of general-purpose pretrained language models such as BERT (Devlin et al., 2019), it has become increasingly common to fine-tune pretrained models to produce high-quality sentence embeddings, revisiting the aforementioned supervision signals (Reimers and Gurevych, 2019; Yang et al., 2021), and using self-supervised objectives based on contrastive learning (CL). In this paper, we present a CL objective with entity-based supervision. We train our EASE model with entity CL together with self-supervised CL with dropout noise and show that the entity CL improves the quality of sentence embeddings.

Contrastive learning The basic idea of contrastive representation learning is to pull semantically similar samples close and keep dissimilar samples apart (Hadsell et al., 2006). CL for sentence embeddings can be classified by the type of positive pairs used. As representative examples, several methods use entailment pairs as positive pairs in NLI datasets (Gao et al., 2021; Zhang et al., 2021). To alleviate the need for an annotated dataset, self-supervised approaches are also being actively studied. Typical self-supervised methods involve generating positive pairs using data augmentation techniques, including discrete operations such as word deletion and shuffling (Yan et al., 2021; Meng et al., 2021), back-translation (Fang et al., 2020), intermediate BERT hidden representations (Kim et al., 2021), and dropout noise within transformer layers (Gao et al., 2021; Liu et al., 2021). Contrastive tension (CT)-BERT (Carlsson et al., 2021) regards as positive pairs the outputs of the same sentence from two individual encoders. DeCLUTR (Giorgi et al., 2021) uses different spans of the same document. In contrast to these methods that perform CL between sentences, our method performs CL between sentences and their associated entities.

Multilingual sentence embeddings Another yet closely related line of research is focused on learning multilingual sentence embeddings, which capture semantics across multiple languages. Early competitive methods typically utilize the sequence-to-sequence objective with parallel corpora to learn multilingual sentence embeddings (Schwenk and Douze, 2017; Artetxe and Schwenk, 2019); recently fine-tuned multilingual pretrained models have achieved state-of-the-art performance (Feng et al., 2020; Goswami et al., 2021). However, one

drawback of such approaches is that, to achieve strong results for a particular language pair, they need rich parallel or semantically related sentence pairs, which are not necessarily easy to obtain. In this work, we explore the utility of Wikipedia entity annotations, which are aligned across languages and already available in over 300 languages. We also show that the entity CL in a multilingual scenario effectively improves the alignment of sentence embeddings between English and low-resource languages not well supported in an existing multilingual model.

2.2 Learning Representations Using Entity-based Supervision

Entities have been conventionally used to model text semantics (Gabrilovich and Markovitch, 2007, 2006). Several recently proposed methods learn text representations based on entity-based supervision by predicting entities from their relevant text (Yamada et al., 2017) or entity-masked sentences (Ling et al., 2020). In the proposed EASE model, the existing self-supervised CL method based on BERT (Gao et al., 2021) is extended using entity-based supervision with carefully designed hard negatives. Moreover, it is applied to the multilingual setting by leveraging the language-agnostic nature of entities.

3 Model and Training Data

In this section, we describe the components of our learning method for sentence embeddings, EASE, which is trained using entity hyperlink annotations available in Wikipedia.

3.1 Contrastive Learning with Entities

Given pairs of a sentence and a semantically related entity (positive entity) $\mathcal{D} = \{(s_i, e_i)\}_{i=1}^{m}$, we train our model to predict the entity embedding $\mathbf{e_i} \in \mathbb{R}^{\mathrm{d_e}}$ from the sentence embedding $\mathbf{s_i} \in \mathbb{R}^{\mathrm{d_s}}$. Following the contrastive framework in Chen et al. (2020), the training loss for (s_i, e_i) with a minibatch of N pairs is:

$$l_i^e = -\log \frac{e^{\mathrm{sim}(\mathbf{s_i}, \mathbf{We_i})/\tau}}{\sum_{j=1}^{N} e^{\mathrm{sim}(\mathbf{s_i}, \mathbf{We_j})/\tau}}, \quad (1)$$

where $\mathbf{W} \in \mathbb{R}^{\mathrm{d_e} \times \mathrm{d_s}}$ is a learnable matrix weight, τ is a temperature hyperparameter, and $\mathrm{sim}(\cdot)$ is the cosine similarity $\frac{\mathbf{s_1}^\top \mathbf{s_2}}{\|\mathbf{s_1}\| \cdot \|\mathbf{s_2}\|}$.

Data We construct the sentence-entity paired datasets from the January 2019 version of Wikipedia dump. We split text in the articles into sentences using `polyglot`.[1] For each sentence, we extract the hyperlink entities as semantically related entities.[2] Each entity forms a training instance (s_i, e_i) for the sentence. We restrict the entities to those that appear more than ten times as hyperlinks in the training corpus. They are converted into Wikidata entities, which are shared across languages, using inter-language links obtained from the March 2020 version of the Wikidata dump.[3]

3.2 Hard Negative Entities

The introduction of hard negatives (data that are difficult to distinguish from an anchor point) has been reported to be effective in improving CL models (Gao et al., 2021; Robinson et al., 2021). We introduce a hard negative mining technique that finds negative entities similar to the positive entity but yet unrelated to the sentence.

Specifically, for each positive entity, we collect hard negative entity candidates that satisfy the following two conditions: (1) entities with the same type as the positive entity. Entity types are defined as the entities in the "instance of" relation on Wikidata, following the work of Xiong et al. (2020). If there are more than one appropriate type, we randomly choose one; (2) entities that do not appear on the same Wikipedia page. Our assumption here is that entities on the same page are topically related to the positive entity and thus are not appropriate for negative data. Finally, we randomly choose one of the candidates to construct hard negative training data. For example, the "Studio Ghibli" entity has the type "animation studio" and one of the hard negative entity candidates is "Walt Disney Animation Studios".

Given datasets with hard negative entities $\mathcal{D} = \{(s_i, e_i, e_i^-)\}_{i=1}^{m}$, the loss function is

$$l_i^e = -\log \frac{e^{\mathrm{sim}(\mathbf{h_i}, \mathbf{We_i})/\tau}}{\sum_{j=1}^{N} (e^{\mathrm{sim}(\mathbf{h_i}, \mathbf{We_j})/\tau} + e^{\mathrm{sim}(\mathbf{h_i}, \mathbf{We_j^-})/\tau})}. \quad (2)$$

[1] https://polyglot.readthedocs.io/en/latest/Tokenization.html

[2] In a preliminary experiment, we also tried constructing entity-sentence paired data from entities and the first sentence on their page, and found that the current approach performs better.

[3] https://en.wikipedia.org/wiki/Help:Interlanguage_links

3.3 Pretrained Entity Embeddings

We initialize entity embeddings using English entity embeddings pretrained on Wikipedia. These embeddings are trained using the open-source `Wikipedia2Vec` tool (Yamada et al., 2020) and the January 2019 English Wikipedia dump. The vector dimension is set to 768, which is the same as those of the hidden representations of the base pretrained models, and the other hyperparameters to their default values. The parameters of the entity embedding matrix are updated during the training process.

3.4 Self-supervised Contrastive Learning with Dropout Noise

Self-supervised CL with dropout noise, which inputs a sentence and predicts itself using dropout as noise, is an effective method for learning sentence embeddings in an unsupervised way (Liu et al., 2021; Gao et al., 2021). We combine this method with our entity CL.

Given two embeddings with different dropout masks $\mathbf{s_i}, \mathbf{s_i^+}$, the training loss of self-supervised CL l_i^s is defined by

$$l_i^s = -\log \frac{e^{\mathrm{sim}(\mathbf{s_i}, \mathbf{s_i^+})/\tau}}{\sum_{j=1}^{N} e^{\mathrm{sim}(\mathbf{s_i}, \mathbf{s_j^+})/\tau}}. \quad (3)$$

In summary, our total loss is

$$l_i^{ease} = \lambda l_i^e + l_i^s, \quad (4)$$

where l^e and l^s are defined in Equations (2) and (3) respectively, and λ denotes a hyperparameter that defines the balance between the entity CL and self-supervised CL with dropout noise. The details on the hyperparameters of the models can be found in Appendix A.

4 Experiment: Monolingual

We first evaluate EASE in monolingual settings. We fine-tune monolingual pre-trained language models using only English Wikipedia data.

4.1 Setup

We use one million pairs sampled from the English entity-sentence pairs described in Section 3 as training data. In this setting, we train sentence embedding models from pre-trained checkpoints of BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019) and take the [CLS] representation as the sentence embedding. We add a linear layer after the output sentence embeddings only during training, as in Gao et al. (2021).

Model	7 STS avg.	8 STC avg.
GloVe embedding (avg.)	61.3†	56.4
BERT (avg.)	52.6	50.9
CT-BERT$_{\text{base}}$	72.1	61.6
SimCSE-BERT$_{\text{base}}$	76.3	57.1
EASE-BERT$_{\text{base}}$	**77.0**	**63.1**
RoBERTa (avg.)	53.5	40.9
DeCLUTR-RoBERTa$_{\text{base}}$	70.0	**60.0**
SimCSE-RoBERTa$_{\text{base}}$	76.6	57.4
EASE-RoBERTa$_{\text{base}}$	**76.8**	58.6

Table 1: Sentence embedding performance on seven monolingual STS tasks (Spearman's correlation) and eight monolingual STC tasks (clustering accuracy). The highest values among the models with the same pre-trained encoder are in bold. †: results from Reimers and Gurevych (2019); all other results are reproduced or reevaluated by us using published checkpoints. The complete results are available in Appendix G.

We compare our method with unsupervised sentence embedding methods including average GloVe embeddings (Pennington et al., 2014), average embeddings of vanilla BERT or RoBERTa, and previous state-of-the-art approaches such as SimCSE (Gao et al., 2021), CT (Carlsson et al., 2021), and DeCLUTR (Giorgi et al., 2021).

We evaluate sentence embeddings using two tasks: STS and STC. These tasks are supposed to measure the degree of sentence embeddings capturing fine-grained and broad semantic structures.

Semantic textual similarity STS is a measure of the capability of capturing graded similarity of sentences. We use seven monolingual STS tasks: STS 2012-2016 (Agirre et al., 2012, 2013, 2014, 2015, 2016), STS Benchmark (Cer et al., 2017), and SICK-Relatedness (Marelli et al., 2014). Following the settings of Reimers and Gurevych (2019), we calculate Spearman's rank correlation coefficient between the cosine similarity of the sentence embeddings and the ground truth similarity scores.

Short text clustering Another important aspect of sentence embeddings is the ability to capture categorical semantic structure, i.e., to map sentences from the same categories close together and those from different categories far apart (Zhang et al., 2021). We also evaluate sentence embeddings using eight benchmark datasets for STC (Zhang et al., 2021) to investigate how well our method can encode high-level categorical structures into

Model	EN-EN	AR-AR	ES-ES	EN-AR	EN-DE	EN-TR	EN-ES	EN-FR	EN-IT	EN-NL	Avg.
$\text{mBERT}_{\text{base}}$ (avg.)	54.4	50.9	56.7	18.7	33.9	16.0	21.5	33.0	34.0	35.3	35.4
$\text{SimCSE-mBERT}_{\text{base}}$	78.3	62.5	76.7	26.2	55.6	23.8	37.9	48.1	49.6	50.3	50.9
$\text{EASE-mBERT}_{\text{base}}$	**79.3**	**62.8**	**79.4**	**31.6**	**59.8**	**26.4**	**53.7**	**59.2**	**59.4**	**60.7**	**57.2**
$\text{XLM-R}_{\text{base}}$ (avg.)	52.2	25.5	49.6	15.7	21.3	12.1	10.6	16.6	22.9	23.9	25.0
$\text{SimCSE-XLM-R}_{\text{base}}$	77.9	63.4	**80.6**	**36.3**	56.2	28.9	38.9	**51.8**	52.6	54.2	54.1
$\text{EASE-XLM-R}_{\text{base}}$	**80.6**	**65.3**	80.4	34.2	**59.1**	**37.6**	**46.5**	51.2	**56.6**	**59.5**	**57.1**

Table 2: Spearman's correlation for multilingual semantic textual similarity on extended version of STS 2017 dataset.

Model	ar	ca	cs	de	en	eo	es	fa	fr	ja	ko	pl	pt	ru	sv	tr	Avg.
$\text{mBERT}_{\text{base}}$ (avg.)	27.0	27.2	**44.3**	36.2	37.9	25.6	**41.1**	35.0	25.9	44.2	31.0	35.0	30.1	23.4	28.9	34.9	33.0
$\text{SimCSE-mBERT}_{\text{base}}$	30.1	26.9	41.3	32.5	37.3	27.2	36.2	**36.9**	29.0	48.9	33.9	37.6	37.9	**27.1**	26.9	35.3	34.1
$\text{EASE-mBERT}_{\text{base}}$	**31.9**	**29.6**	38.8	**38.5**	**30.2**	**34.5**	37.2	36.7	**30.4**	**49.3**	**36.2**	**40.0**	**41.0**	27.0	**30.5**	**44.7**	**36.0**
$\text{XLM-R}_{\text{base}}$ (avg.)	**26.0**	24.7	28.2	29.4	23.0	23.5	22.1	36.6	23.6	38.8	22.0	24.2	32.8	18.0	**33.2**	26.0	27.0
$\text{SimCSE-XLM-R}_{\text{base}}$	24.6	26.3	34.6	28.6	33.4	31.7	32.9	35.9	29.1	41.1	31.1	33.1	30.0	26.0	32.9	37.2	31.8
$\text{EASE-XLM-R}_{\text{base}}$	25.3	**26.7**	**43.2**	**37.0**	**34.9**	**34.2**	**37.2**	**42.4**	**32.0**	**46.0**	**32.8**	**41.6**	**33.4**	**31.3**	27.2	**41.8**	**35.4**

Table 3: Clustering accuracy for multilingual short text clustering on MewsC-16 dataset.

sentence embeddings. These datasets contain short sentences, ranging from 6 to 28 average words in length, from a variety of domains such as news, biomedical, and social network service (Twitter). We cluster the sentence embeddings using K-Means (MacQueen, 1967) and compute the clustering accuracy using the Hungarian algorithm (Munkres, 1957) averaged over three independent runs.

4.2 Results

Table 1 shows the evaluation results for the seven STS and eight STC tasks. Overall, our EASE methods significantly improve the performance of the base models (i.e., BERT and RoBERTa), and on average outperform the previous state-of-the-art methods on all tasks except STC with the RoBERTa backbone. The most significant improvement is observed for EASE-BERT, with an average improvement of 61.6% to 63.1% over the previous best result for STC tasks. These results suggest that EASE is able to measure the semantic similarity between sentences, and simultaneously excel at capturing high-level categorical semantic structure.

5 Experiment: Multilingual

To further explore the advantage of entity annotations as cross-lingual alignment supervision, we test EASE in multilingual settings: we fine-tune multilingual pre-trained language models using Wikipedia data in multiple languages.

5.1 Setup

We sample 50,000 pairs for each language and use them together as training data from the entity-sentence paired data in 18 languages.[4] As our primary baseline model, we use a SimCSE model trained using the same multilingual data as EASE (i.e., sentences in entity-sentence paired data).[5] In this setting, we start fine-tuning from pre-trained checkpoints of mBERT or XLM-R (Conneau et al., 2020) and take mean pooling to obtain sentence embeddings for both training and evaluation on both EASE and SimCSE. We also tested other pooling methods, but mean pooling was the best in this experiment for both models (Appendix B).

5.2 Multilingual STS and STC

We evaluate our method using the extended version of the STS 2017 dataset (Reimers and Gurevych, 2020), which contains annotated sentences for ten language pairs: EN-EN, AR-AR, ES-ES, EN-AR, EN-DE, EN-TR, EN-ES, EN-FR, EN-IT, and EN-NL. We compute Spearman's rank correlation as in Section 4.1. We also conduct experiments on our newly introduced multilingual STC dataset described as follows:

[4] We chose 18 languages (ar, ca, cs, de, en, eo, es, fa, fr, it, ja, ko, nl, pl, pt, ru, sv, tr) present in both the MewsC-16 dataset (see Section 5.2) and the extended version of STS 2017.

[5] We use the same entity embeddings trained on English Wikipedia as those of the monolingual settings (Section 3.3). Note that entities in all languages are converted to Wikidata entities that are shared across languages using inter-language links as described in Section 3.1.

Model	ar	ca	cs	de	eo	es	fr	it	ja	ko	nl	pl	pt	ru	sv	tr	Avg.
mBERT$_{base}$ (avg.)	20.6	49.2	32.8	62.8	12.2	57.7	55.6	50.8	38.6	33.1	54.8	40.2	58.5	51.4	45.8	30.1	43.4
SimCSE-mBERT$_{base}$	16.4	51.5	30.7	57.0	18.2	54.8	54.5	49.9	39.6	28.1	52.7	37.9	53.6	46.8	45.5	25.0	41.4
EASE-mBERT$_{base}$	**32.1**	**66.5**	**47.7**	**74.2**	**26.1**	**70.1**	**66.7**	**65.3**	**59.2**	**46.8**	**69.2**	**55.4**	**69.1**	**64.4**	**59.4**	**38.1**	**56.9**
XLM-R$_{base}$ (avg.)	10.3	15.3	16.5	49.6	7.5	36.4	30.8	25.6	15.0	19.3	45.2	24.1	42.0	37.4	42.8	17.9	27.2
SimCSE-XLM-R$_{base}$	38.4	57.6	55.7	80.6	46.0	68.9	70.4	66.4	60.0	54.1	73.1	65.3	75.1	71.1	76.7	56.4	63.5
EASE-XLM-R$_{base}$	**42.6**	**65.1**	**63.8**	**87.2**	**56.1**	**75.9**	**74.1**	**70.8**	**68.2**	**60.5**	**77.9**	**71.9**	**80.6**	**76.5**	**79.2**	**60.9**	**69.4**

Table 4: Accuracy on Tatoeba dataset averaged over forward and backward directions (en to target language and vice-versa).

Model	Avg.
mBERT$_{base}$ (avg.)	17.3
SimCSE-mBERT$_{base}$	16.8
EASE-mBERT$_{base}$	**25.4**
XLM-R$_{base}$ (avg.)	9.4
SimCSE-XLM-R$_{base}$	28.5
EASE-XLM-R$_{base}$	**32.1**

Table 5: Average accuracy for 94 languages not included in EASE training on Tatoeba.

Model	en (dev)	de	es	fr	it	ja	ru	zh	Avg.
mBERT$_{base}$ (avg.)	**89.5**	68.0	68.1	70.6	62.7	61.2	61.5	69.6	65.9
SimCSE-mBERT$_{base}$	88.4	62.3	**73.2**	78.2	64.3	**63.7**	61.3	**75.0**	68.3
EASE-mBERT$_{base}$	89.0	**69.9**	69.2	**80.1**	**66.8**	62.8	**64.4**	73.2	**69.5**
XLM-R$_{base}$ (avg.)	**90.9**	**82.7**	**79.8**	72.1	72.5	71.1	69.6	71.4	74.2
SimCSE-XLM-R$_{base}$	90.7	74.9	74.1	81.5	70.3	71.7	70.1	76.6	74.2
EASE-XLM-R$_{base}$	90.6	77.9	75.6	**83.9**	**72.6**	**72.8**	**71.1**	**81.6**	**76.5**

Table 6: Classification accuracy for zero-shot cross-lingual text classification on MLDoc dataset.

MewsC-16 To evaluate the ability of sentence embeddings to encode high-level categorical concepts in a multilingual setting, we constructed MewsC-16 (**M**ultilingual Short Text **C**lustering Dataset for **News** in **16** languages) from Wikinews.[6] MewsC-16 contains topic sentences from Wikinews articles in 13 categories and 16 languages. More detailed information is available in Appendix E. We perform clustering and compute accuracy for each language as in Section 4.1.

Tables 2 and 3 show the results of our multilingual STS and STC experiments. Overall, EASE substantially outperforms the corresponding base models (i.e., mBERT and XLM-R) on both tasks. Similar to the results for the monolingual setting, the average performance of EASE exceeds that of SimCSE for multilingual STC tasks with an improvement of 34.1% to 36.0% for mBERT and 31.8% to 35.4% for XLM-R. This result suggests that even in a multilingual setting, EASE can encode high-level categorical semantic structures into sentence embeddings. Moreover, EASE performs better than SimCSE on multilingual STS tasks, which is a slightly different result than for the monolingual setting. More specifically, the performance of EASE-mBERT is better than that of SimCSE-mBERT (50.9 vs 57.2), and that of EASE-XLM-R is better than that of SimCSE-XLM-R (54.1 vs 57.1). This indicates that cross-lingual alignment supervision from language-independent entities is beneficial in learning multilingual sentence embeddings.

[6] https://en.wikinews.org/wiki/Main_Page

5.3 Cross-lingual Parallel Matching

We evaluate EASE on the Tatoeba dataset (Artetxe and Schwenk, 2019) to assess more directly its ability to capture cross-lingual semantics. This task is to retrieve the correct target sentence for each query sentence, given a set of parallel sentences. We perform the retrieval using the cosine similarity scores of the sentence embeddings. For each language-pair dataset, we compute the retrieval accuracy averaged over the forward and backward directions (English to the target language and vice-versa).

Table 4 shows the evaluation results for the languages in the CL training data. EASE significantly outperforms the corresponding base models and SimCSE for all languages. Notably, the mean performance of EASE-mBERT is better than that of vanilla mBERT by 13.5 percentage points. This indicates that EASE can capture cross-lingual semantics owing to the cross-lingual supervision of entity annotations, which aligns semantically similar sentences across languages. One interesting observation is that the performance of SimCSE-mBERT is worse than that of vanilla mBERT. We conjecture that this is because the SimCSE model is trained using only the positive sentence pairs within the same language, which sometimes leads to less language-neutral representations.

To further explore the cross-lingual ability of

EASE, we evaluate it on languages not included in the EASE training set (Table 5). The results show that EASE performs robustly on these languages as well, which suggests that, in EASE, the cross-lingual alignment effect propagates to other languages not used in additional training with EASE (Kvapilíková et al., 2020).

5.4 Cross-lingual Zero-shot Transfer

We further evaluate our sentence embeddings on a downstream task in which sentence embeddings are used as input features, especially in the cross-lingual zero-shot transfer setting. For evaluation in this setting, we use MLDoc (Schwenk and Li, 2018), a cross-lingual document classification dataset that classifies news articles in eight languages into four categories. We train a linear classifier using sentence embeddings as input features on the English training data, and evaluate the resulting classifier in the remaining languages. To directly evaluate the ability of the resulting sentence embeddings, we do not update the parameters of the sentence encoder but only train the linear classifier in this setting. The detailed settings are shown in Appendix D.

As shown in Table 6, our EASE models achieve the best average performance on both back-bones, suggesting that multilingual embeddings learned with the CL are also effective in the cross-lingual transfer setting.

6 Case Study: Fine-tuning Supervised Model with EASE

Existing multilingual sentence representation models trained on a large parallel corpus do not always perform well, especially for languages that are not included in the training data. In contrast, EASE requires only the Wikipedia text corpus, which is available in more than 300 languages.[7] Thus, one possible use case for EASE would be to complement the performance of existing models in low-resource languages by exploiting the Wikipedia data in those languages.

To test this possibility, we fine-tune LaBSE (Feng et al., 2020), which is trained on both monolingual and bilingual data in 109 languages, with our EASE framework in five low-resource languages (kab, pam, cor, tr, mhr). These languages are not present in the original training corpus, so the model performed particularly poorly on these languages. We fine-tune the model using 5,000 pairs each from English and the corresponding language data.

[7] https://meta.wikimedia.org/wiki/List_of_Wikipedias

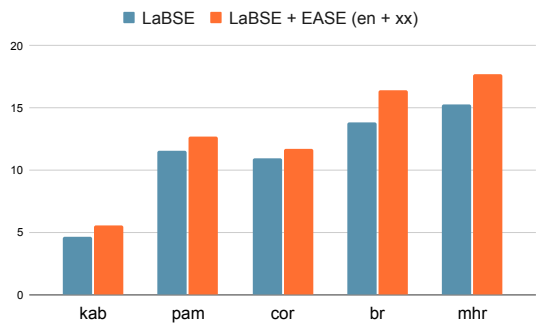

Figure 2: Results of fine-tuning LaBSE with EASE framework on Tatoeba dataset.

As shown in Figure 2, EASE improves the performance of LaBSE across all target languages, which is an intriguing result considering that LaBSE has already been trained on about six billion parallel corpora. These results suggest the potential benefit of combining EASE with other models using parallel corpora, especially for languages without or with a few parallel corpora.

7 Analysis

7.1 Ablation Study

We conduct ablations to better understand how each component of EASE contributes to its performance. We measure the performance of the models using monolingual STS in the monolingual setting and multilingual STS in the multilingual setting, without one of the following components: the self-supervised CL loss, hard negatives, and Wikipedia2Vec initialization (Table 7). As a result, we find all of the components to make an important contribution to the performance.

It is worth mentioning that entity CL alone (i.e., w/o self-supervised CL) also improves the baseline performance significantly. The performance contributions in the multilingual setting are particularly significant (53.1 for XLM-R and 49.3 for mBERT) and comparable to those for the SimCSE models. These results suggest that CL with entities by itself is effective in learning multilingual sentence embeddings.

Setting	EASE-BERT$_{\text{base}}$ STS avg.	EASE-RoBERTa$_{\text{base}}$ STS avg.	EASE-mBERT$_{\text{base}}$ mSTS avg.	EASE-XLM-R$_{\text{base}}$ mSTS avg.
Full model	**76.9**	**76.8**	**57.2**	**57.1**
w/o self-supervised CL	65.3	66.1	49.3	53.1
w/o hard negative	75.3	76.1	53.8	52.7
w/o Wikipedia2Vec	73.8	76.3	52.1	54.3
w/o all (vanilla model)	31.4	43.6	35.4	25.0

Table 7: Results of ablation study.

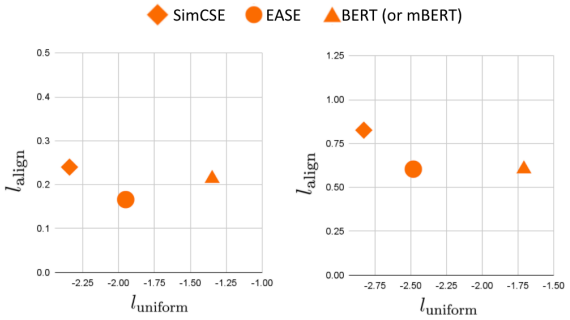

Figure 3: $l_{\text{align}} - l_{\text{uniform}}$ plot of BERT-based (or mBERT-based) models in monolingual (left) and multilingual (right) settings.

7.2 Alignment and Uniformity

To further understand the source of the performance improvement with EASE, we evaluate two key properties to measure the quality of the representations obtained from contrastive learning (Wang and Isola, 2020): *alignment* measures the closeness of representations between positive pairs; *uniformity* measures how well the representations are uniformly distributed. We let $f(x)$ denote the normalized representation of x, and compute the two measures using

$$l_{\text{align}} \triangleq \mathop{\mathbb{E}}_{(x,x^+)\sim p_{\text{pos}}} \|f(x) - f(x^+)\|^2, \quad (5)$$

$$l_{\text{uniform}} \triangleq \log \mathop{\mathbb{E}}_{x,y \overset{i.i.d.}{\sim} p_{\text{data}}} e^{-2\|f(x)-f(y)\|^2}, \quad (6)$$

where p_{pos} denotes positive pairs and p_{data} denotes the entire data distribution. We compute these metrics using BERT-based models on the STS-B development set data. For investigation in the multilingual setting, we compute them using mBERT-based models on the multilingual STS data used in the experiment Section 5. We compute the averages of alignment and uniformity for each language pair. For each setting, we take STS pairs with a score higher than 4 in the 0-to-5 scale as p_{pos} and all STS sentences as p_{data}.

As shown in Figure 3, the trends are similar in both settings: (1) both EASE and SimCSE significantly improve uniformity compared with that for the vanilla model; (2) EASE is inferior to SimCSE in terms of uniformity and superior in terms of alignment. This result suggests that entity CL does not have the effect of biasing embeddings towards a more uniform distribution. Instead, it has the effect of aligning semantically similar samples, which leads to the improved performance of the resultant sentence embeddings.

7.3 Qualitative analysis

To investigate how EASE improves the quality of sentence embeddings, we conduct qualitative analysis by comparing the results of EASE and SimCSE on STS Benchmark. Table 8a shows typical cases of how EASE improves sentence embeddings. We find that EASE embeddings are more robust to synonyms and grammatical differences since they are more aware of the topic similarity between sentences, resulting in more accurate score inference. On the other hand, as shown in the deterioration cases in Table 8b, EASE embeddings are sometimes overly sensitive to topical similarity, making it difficult to capture the correct meaning of the whole sentence.

8 Discussion and Conclusion

Our experiments have demonstrated that entity supervision in EASE improves the quality of sentence embeddings both in the monolingual setting and, in particular, the multilingual setting. As recent studies have shown, entity annotations can be used as *anchors* to learn quality cross-lingual representations (Calixto et al., 2021; Nishikawa et al., 2021; Jian et al., 2022; Ri et al., 2022), and our work is another demonstration of their utility, particularly in sentence embeddings. One promising future direction is exploring how to better exploit the cross-lingual nature of entities.

Sentence1	Sentence2	Gold	EASE	SimCSE
i think you 're looking for mikey (1992) .	i think you 're looking for the movie	3.00	2.32	1.62
the new york senator 's new book , " living history , " appears a certain bestseller .	hillary clinton , the new york senator and former first lady , has a book out monday titled living history .	3.20	3.57	1.94
he was referring to john s. reed , the former citicorp chief executive who became interim chairman and chief executive of the exchange last sunday .	next week , john s. reed , the former citicorp chief executive who sunday became interim chairman and chief executive of the exchange , will take up his position .	4.00	3.52	2.73

(a) Improvement cases

Sentence1	Sentence2	Gold	EASE	SimCSE
it 's not a good idea .	it 's a good question .	0.00	2.88	1.33
suicide attack kills eight in baghdad	suicide attacks kill 24 people in baghdad	2.40	3.92	2.43
the nasdaq composite index rose 19.67 , or 1.3 percent , to 1523.71 , its highest since june 18 .	the s and p 500 had climbed 16 percent since its march low and yesterday closed at its highest since dec. 2 .	0.80	3.25	2.04

(b) Deterioration cases

Table 8: Comparison of STS Benchmark results by monolingual EASE and SimCSE.

Our experiments also demonstrate the utility of Wikipedia as a multilingual database. As described in Section 6, Wikipedia entity annotations can compensate for the lack of parallel resources in learning cross-lingual representations. Wikipedia currently supports more than 300 languages, and around half of them have over 10,000 articles.[8] Moreover, Wikipedia is ever growing; it is expected to include more and more languages.[9] This will motivate researchers to develop methods for multilingual models including low-resource languages in the aid of entity annotations in Wikipedia.

However, the reliance on Wikipedia for training data may limit the application of the models to specific domains (e.g., general or encyclopedia domains). To apply EASE to other domains, one may need to annotate text from the domain either manually or automatically. Future work can investigate the effectiveness of the entity CL in other domains and possibly its the combination with an entity linking system.

Finally, we note that the supervision signal in EASE is inherently noisy. Different entities have different characteristics as a topic indicator, and sentences that contain the same entity do not necessarily share meaning. Future work can address this by considering how an entity is used in a sentence to obtain more reliable supervision.

[8] https://meta.wikimedia.org/wiki/List_of_Wikipedias

[9] https://incubator.wikimedia.org/wiki/Incubator:Main_Page

Acknowledgements

This work was partially supported by JSPS KAKENHI Grants JP16H06302, JP18H04120, JP21H04907, and JP21K18023, and by JST CREST Grants JPMJCR18A6 and JPMJCR20D3, Japan. JPMJCR18A6: JST CREST (FY2018-FY2023) PI Prof. Yamagishi JPMJCR20D3: JST CREST (FY2020-FY2025) PI Prof. Echizen JP16H06302: KAKENHI, Scientific Research (S) (FY2016-FY2020) PI Prof. Babaguchi, Co-PI Prof. Echizen JP18H04120: KAKENHI, Scientific Research (A) (FY2018-FY2020) PI Prof. Echizen, Co-PI Prof. Yamagishi

<Follows are newly accepted.> JP21H04907: KAKENHI, Scientific Research (A) (FY2021-FY2023) PI Prof. Echizen, Co-PI Prof. Yamagishi JP21K18023: KAKENHI, Early-Career Scientists (FY2021-FY2023) PI Dr. Le Trung-Nghia

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)*.

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)*.

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)*.

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)*.

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*.

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

P. B. Baxendale. 1958. Machine-made index for technical literature: An experiment. *IBM J. Res. Dev.*, 2(4):354–361.

Iacer Calixto, Alessandro Raganato, and Tommaso Pasini. 2021. Wikipedia entities as rendezvous across languages: Grounding multilingual language models by predicting Wikipedia hyperlinks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Fredrik Carlsson, Amaru Cuba Gyllensten, Evangelia Gogoulou, Erik Ylipää Hellqvist, and Magnus Sahlgren. 2021. Semantic re-tuning with contrastive tension. In *9th International Conference on Learning Representations*.

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)*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119.

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

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.

H. P. Edmundson. 1969. New methods in automatic extracting. *J. ACM*, 16(2):264–285.

Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. 2020. CERT: Contrastive self-supervised learning for language understanding. *ArXiv*, arXiv/2005.12766.

Fangxiaoyu Feng, Yinfei Yang, Daniel Matthew Cer, N. Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *ArXiv*, abs/2007.01852.

Evgeniy Gabrilovich and Shaul Markovitch. 2006. Overcoming the brittleness bottleneck using wikipedia: Enhancing text categorization with encyclopedic knowledge. In *AAAI*.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on*

Empirical Methods in Natural Language Processing, Online and Punta Cana, Dominican Republic.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2021. DeCLUTR: Deep contrastive learning for unsupervised textual representations. 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)*.

Koustava Goswami, Sourav Dutta, Haytham Assem, Theodorus Fransen, and John P. McCrae. 2021. Cross-lingual sentence embedding using multi-task learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

R. Hadsell, S. Chopra, and Y. LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2.

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 *Proceedings of the 37th International Conference on Machine Learning*, volume 119.

Xiaoze Jian, Yaobo Liang, Weizhu Chen, and Nan Duan. 2022. XLM-K: Improving cross-lingual language model pre-training with multilingual knowledge. In *AAAI*.

Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. 2021. Self-guided contrastive learning for BERT sentence representations. 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)*.

Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in Neural Information Processing Systems*, volume 28.

Ivana Kvapilíková, Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Ondřej Bojar. 2020. Unsupervised multilingual sentence embeddings for parallel corpus mining. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14.

Jeffrey Ling, Nicholas FitzGerald, Zifei Shan, Livio Baldini Soares, Thibault Févry, David Weiss, and Tom Kwiatkowski. 2020. Learning cross-context entity representations from text. *ArXiv*, arXiv/2001.03765.

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021. Fast, effective, and self-supervised: Transforming masked language models into universal lexical and sentence encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. *ArXiv*, arXiv/1907.11692.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In *6th International Conference on Learning Representations*.

J. B. MacQueen. 1967. Some methods for classification and analysis of multivariate observations. In *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1.

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 *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*.

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. In *Conference on Neural Information Processing Systems*.

James R. Munkres. 1957. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, 5(1):32–38.

Sosuke Nishikawa, Ikuya Yamada, Yoshimasa Tsuruoka, and Isao Echizen. 2021. A multilingual bag-of-entities model for zero-shot cross-lingual text classification. *ArXiv*, arXiv/2110.07792.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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)*.

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)*.

Ryokan Ri, Ikuya Yamada, and Yoshimasa Tsuruoka. 2022. mLUKE: The power of entity representations in multilingual pretrained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2021. Contrastive learning with hard negative samples. In *9th International Conference on Learning Representations*.

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*.

Holger Schwenk and Xian Li. 2018. A corpus for multilingual document classification in eight languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*.

Wenhan Xiong, Jingfei Du, William Yang Wang, and Veselin Stoyanov. 2020. Pretrained encyclopedia: Weakly supervised knowledge-pretrained language model. In *8th International Conference on Learning Representations*.

Ikuya Yamada, Akari Asai, Jin Sakuma, Hiroyuki Shindo, Hideaki Takeda, Yoshiyasu Takefuji, and Yuji Matsumoto. 2020. Wikipedia2Vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from Wikipedia. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. 2017. Learning distributed representations of texts and entities from knowledge base. *Transactions of the Association for Computational Linguistics*, 5:397–411.

Ikuya Yamada, Hiroyuki Shindo, and Yoshiyasu Takefuji. 2018. Representation learning of entities and documents from knowledge base descriptions. In *Proceedings of the 27th International Conference on Computational Linguistics*.

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 *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)*.

Ziyi Yang, Yinfei Yang, Daniel Cer, Jax Law, and Eric Darve. 2021. Universal sentence representation learning with conditional masked language model. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Dejiao Zhang, Shang-Wen Li, Wei Xiao, Henghui Zhu, Ramesh Nallapati, Andrew O. Arnold, and Bing Xiang. 2021. Pairwise supervised contrastive learning of sentence representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

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*.

A Training Details

We implement our EASE model using `transformers`[10] libraries. For the monolingual settings, we use the STS-B development set as in (Gao et al., 2021). For multilingual settings, we use the STS-B and SICK-R development set. In this setting, we simply concatenate the entity-sentence paired data for all 18 languages and randomly sample from the concatenated data to construct batches.[11] In both settings, we train our model for one epoch, compute evaluation scores every 250 training steps on the development data, and keep the best model. We conduct a grid-search for batch size $\in \{64, 128, 256, 512\}$ and learning rate $\in \{3e-05, 5e-05\}$. The chosen hyperparameters for each model is shown in Table 9.

Model	Batch size	Learning Rate
SimCSE-mBERT$_{\text{base}}$	128	3e-05
SimCSE-XLM-R$_{\text{base}}$	128	3e-05
EASE-BERT$_{\text{base}}$	64	3e-05
EASE-RoBERTa$_{\text{base}}$	128	5e-05
EASE-mBERT$_{\text{base}}$	256	5e-05
EASE-XLM-R$_{\text{base}}$	64	3e-05

Table 9: Hyperparameters for experiment.

For the loss balancing term λ and softmax temperature τ in the EASE models (section 3), we empirically find that $\lambda = 0.01$, $\tau = 100$ for the monolingual setting and $\tau = 10$ for the multilingual setting work well.

Computing Infrastructure We run the experiments on a server with AMD EPYC 7302 16-Core CPU and a NVIDIA A100-PCIE-40GB GPU. The training of EASE takes approximately 1 hour.

B Pooling Methods for SimCSE and EASE

We compare several pooling methods on both SimCSE and EASE in the multilingual setting: [CLS] with MLP; [CLS] with MLP during training only; [CLS] without MLP; mean pooling. Table 10 shows the evaluation results based on the STS-B and SICK-R development set.

Pooler	SimCSE	EASE
[CLS] pooling		
w/ MLP	63.0	65.0
w/ MLP (train)	72.0	73.3
w/o MLP	72.0	73.4
mean pooling	**72.1**	**73.8**

Table 10: Average Spearman's correlation for different pooling methods for SimCSE and EASE in multilingual setting on STS-B and SICK-R development set.

The mean pooling representation performs best on both models. We thus use mean pooling on both models in Section 5.

[10] https://huggingface.co/docs/transformers/index

[11] In our preliminary experiments, we also tested a setting in which data in the same language were used within the same batch; we did not observe a consistent improvement in the performance of either the SimCSE or EASE models.

C Parallel Sentence Mining

We evaluate the multilingual sentence embeddings with the parallel sentence mining task using the BUCC 2018 shared task dataset (Zweigenbaum et al., 2018). The task is to find the parallel pairs given monolingual sentence pools in two languages, with 2–3% of the sentences being parallel, to find the parallel pairs.

Each model uses the raw embedding output and performance is evaluated without fine-tuning. We first encode all sentences into embeddings and compute the cosine similarity scores between all possible sentence pairs. We then retrieve the sentence pairs with above a fixed threshold and compute the F1 score using the ground-truth parallel pairs.

As the test set is not publicly available, we use the sample set to tune the threshold of the parallel sentence mining and the training set for evaluation, which is a common practice in similar studies (Hu et al., 2020; Feng et al., 2020).

The results are summarized in Table 11. Our EASE models outperform the SimCSE baselines across the languages, demonstrating that the entity contrastive objective improves the alignment of the multilingual sentence embeddings without a parallel corpora. However, performance is significantly poor than that of LaBSE, which is trained using massive amounts of parallel corpora, suggesting that we still need parallel resources to be competitive on this task.

	en-de	en-fr	en-ru	en-zh
SimCSE-mBERT$_{\text{base}}$	13.2	19.2	7.9	11.5
EASE-mBERT$_{\text{base}}$	26.9	33.8	24.2	32.9
SimCSE-XLM-R$_{\text{base}}$	31.8	32.3	28.9	19.9
EASE-XLM-R$_{\text{base}}$	33.3	33.2	33.6	23.4
LaBSE	89.0	88.2	84.7	74.2

Table 11: The F1 scores on BUCC 2018 the training set. Retrieval is performed in forward search, i.e., English sentences as the targets and the other language as the queries.

D Detailed Settings for MLDoc Experiment

We use the english.train.1000 and english.dev datasets for the training and validation data, respectively. We conduct a grid-search for batch size $\in \{32, 64, 128\}$ and learning rate $\in \{0.1, 0.01, 0.001\}$ using validation data 12. We run the experiment three times with different random seeds and record the average scores.

Model	Batch size	Learning Rate
mBERT$_{\text{base}}$(avg.)	32	0.1
XLM-R$_{\text{base}}$(avg.)	32	0.1
SimCSE-mBERT$_{\text{base}}$	32	0.1
SimCSE-XLM-R$_{\text{base}}$	32	0.01
EASE-mBERT$_{\text{base}}$	32	0.01
EASE-XLM-R$_{\text{base}}$	32	0.01

Table 12: Hyperparameters for MLDoc experiment

E Construction of MewsC-16 Dataset

To construct the MewsC-16 dataset, we collect sentences for each category in each language from the Wikinews dump.[12] We first select 13 topic categories in the English Wikinews [13] that are also defined in other languages (Science and technology, Politics and conflicts, Environment, Sports, Health, Crime and law, Obituaries, Disasters and accidents, Culture and entertainment, Economy and business, Weather, Education, Media). We then collect pages with topic categories for each language and remove the pages with two or more topic categories. We clean the text on each page with the `Wikiextractor` tool[14], and split it into sentences using the `polyglot` sentence tokenizer. Finally, we use the first sentence assuming that it well represents the topic of the entire article (Baxendale, 1958; Edmundson, 1969). The corpus statistics for each language are shown in Table 13.

Language	# of sentences	# of label types	Language	# of sentences	# of label types
ar	2,243	11	fr	10,697	13
ca	3,310	11	ja	1,984	12
cs	1,534	9	ko	344	10
de	6,398	8	pl	7,247	11
en	12,892	13	pt	8,921	11
eo	227	8	ru	1,406	12
es	6,415	11	sv	584	7
fa	773	9	tr	459	7
			total	65,425	13

Table 13: Corpus statistics for MewsC-16

[12] https://dumps.wikimedia.org/backup-index.html
[13] https://en.wikinews.org/wiki/Category:News_articles_by_section
[14] https://github.com/attardi/wikiextractor

F Baselines

For average GloVe embedding (Pennington et al., 2014), we use open-source GloVe vectors trained on Wikipedia and Gigaword with 300 dimensions.[15] We use the pretrained model from HuggingFace's `Transformers`[16] for vanilla pretrained language models, including BERT (bert-base-uncased) (Devlin et al., 2019), RoBERTa (roberta-base) (Liu et al., 2019), mBERT (bert-base-multilingual-cased) and XLM-R (xlm-roberta-base) (Conneau et al., 2020). We use the published checkpoints for unsupervised SimCSE (Gao et al., 2021)[17], CT (Carlsson et al., 2021)[18], and DeCLUTR (Giorgi et al., 2021).[19]

G Monolingual STS and STC

Table 14 and 15 show the complete results for seven STS tasks and eight STC tasks. For STS, the average EASE performance is slightly better than that of SimCSE, although the advantage is not consistent across tasks. For most of the STC tasks, EASE consistently outperforms SimCSE. These results indicate that EASE stands out at capturing high-level categorical semantic structures and that its ability to measure sentence semantic similarity is comparable to or better than that of SimCSE.

Model	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
GloVe embeddings (avg.)	55.1	70.7	59.7	68.3	63.7	58.0	53.8	61.3
$\text{BERT}_{\text{base}}$ (avg.)	30.9	59.9	47.7	60.3	63.7	47.3	58.2	52.6
$\text{BERT}_{\text{base}}$-flow	58.4	67.1	60.9	75.2	71.2	68.7	64.5	66.6
$\text{BERT}_{\text{base}}$-whitening	57.8	66.9	60.9	75.1	71.3	68.2	63.7	66.3
IS-$\text{BERT}_{\text{base}}^{\heartsuit}$	56.8	69.2	61.2	75.2	70.2	69.2	64.3	66.6
CT-$\text{BERT}_{\text{base}}$	61.6	76.8	68.5	77.5	76.5	74.3	69.2	72.1
SimCSE-$\text{BERT}_{\text{base}}$	68.4	**82.4**	**74.4**	80.9	78.6	76.9	**72.2**	76.3
EASE-BERT$_{\text{base}}$	**72.8**	81.8	73.7	**82.3**	**79.5**	**78.9**	69.7	**77.0**
$\text{RoBERTa}_{\text{base}}$ (avg.)	32.1	56.3	45.2	61.3	62.0	55.4	62.0	53.5
$\text{RoBERTa}_{\text{base}}$ (first-last avg.)	40.9	58.7	49.1	65.6	61.5	58.6	61.6	56.6
DeCLUTR-$\text{RoBERTa}_{\text{base}}$	52.4	75.2	65.5	77.1	78.6	72.4	**68.6**	70.0
SimCSE-$\text{RoBERTa}_{\text{base}}$	68.7	**82.6**	**73.6**	81.5	**80.8**	**80.5**	67.9	76.5
EASE-RoBERTa$_{\text{base}}$	**70.9**	81.5	73.5	**82.6**	80.5	80.0	68.4	**76.8**

Table 14: Spearman's correlation for monolingual semantic textual similarity tasks.

Model	AG	Bio	G-S	G-T	G-TS	SO	SS	Tweet	Avg.
GloVe embeddings (avg.)	83.2	30.7	59.0	58.3	67.4	29.9	70.4	52.1	56.4
$\text{BERT}_{\text{base}}$ (avg.)	79.8	32.5	55.0	47.0	62.4	21.7	64.0	44.6	50.9
CT-$\text{BERT}_{\text{base}}$	79.2	**38.7**	**65.5**	**60.7**	**69.8**	67.9	55.5	**55.2**	61.6
SimCSE-$\text{BERT}_{\text{base}}$	74.4	34.3	59.5	57.8	64.4	49.6	64.3	52.1	57.1
EASE-BERT$_{\text{base}}$	**85.8**	36.2	60.5	60.4	67.0	**68.1**	**71.7**	54.8	**63.1**
$\text{RoBERTa}_{\text{base}}$ (avg.)	66.5	26.6	47.9	42.8	58.3	16.7	30.0	38.6	40.9
DeCLUTR-$\text{RoBERTa}_{\text{base}}$	**80.7**	**41.0**	**65.2**	**60.5**	**69.6**	32.9	**73.6**	**56.8**	**60.0**
SimCSE-$\text{RoBERTa}_{\text{base}}$	69.8	37.3	60.0	58.0	66.6	69.3	48.3	50.0	57.4
EASE-RoBERTa$_{\text{base}}$	69.4	39.3	60.7	57.7	66.3	**73.9**	49.4	51.8	58.6

Table 15: Clustering accuracy for monolingual short text clustering tasks.

[15] https://nlp.stanford.edu/projects/glove/
[16] https://github.com/huggingface/transformers
[17] https://github.com/princeton-nlp/SimCSE
[18] https://github.com/FreddeFrallan/Contrastive-Tension
[19] https://github.com/JohnGiorgi/DeCLUTR