EASE: Entity-Aware Contrastive Learning of Sentence Embedding

Sosuke Nishikawa[1,2], Ryokan Ri[1,2], Ikuya Yamada[1,4], Yoshimasa Tsuruoka[2] and Isao Echizen[2,3]
[1]Studio Ousia, Japan
[2]The University of Tokyo, Japan
[3]National Institute of Informatics, Japan
[4]RIKEN, Japan

Outline

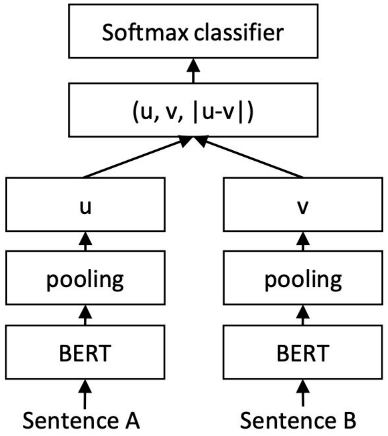
- Backgound
- Proposed Method
- Experiment
- Analysis
- Conclusion

Outline

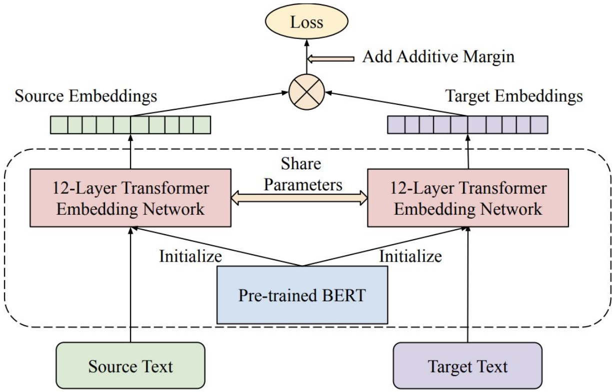
- **Backgound**
- Proposed Method
- Experiment
- Analysis
- Conclusion

Sentence embedding

Learning universal sentence embeddings is a fundamental problem



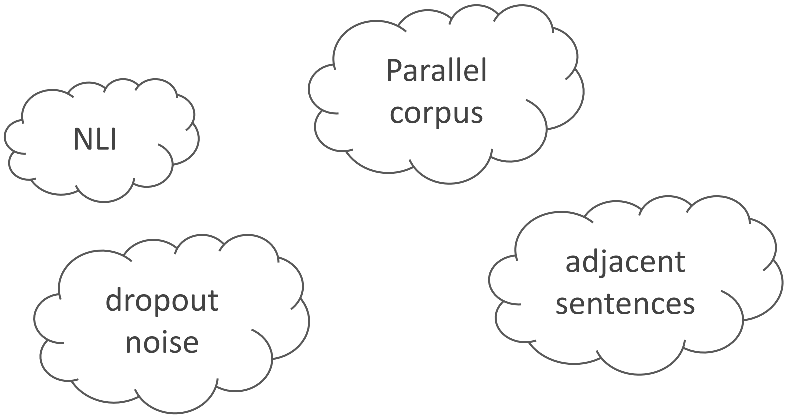
S-BERT
[Reimers and Gurevych, 2019]



LaBSE
[Feng et al., 2022]

How to learn Sentence embedding?

Sentence embedding model is trained with various training supervision



How to learn Sentence embedding?

Sentence embedding model is trained with various training supervision

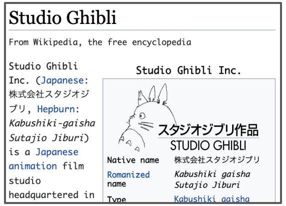
NLI

Parallel corpus

dropout noise

adjacent sentences

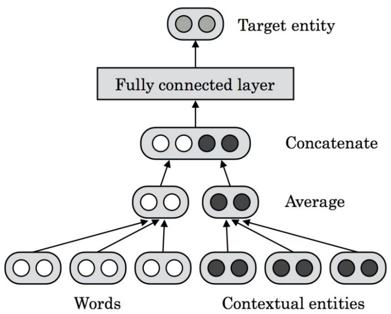
→ We utilize **entity hyperlink annotations from Wikipedia** as a training resource for sentence embeddings



My Neighbor Totoro was animated by [Studio Ghibli]

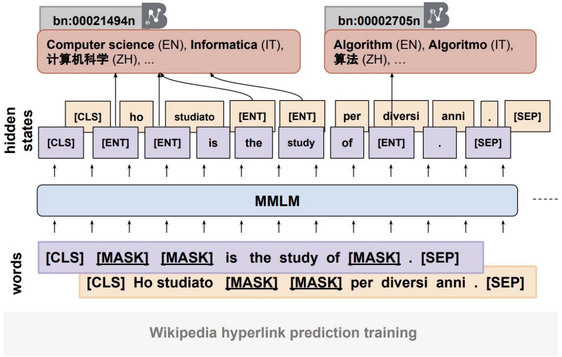
Entity as training resource

- ➢ Entities have been shown to be a strong indicator of text semantics
- ➢ Entities are defined independently of languages



TextEnt

[Yamada et al., 2018]



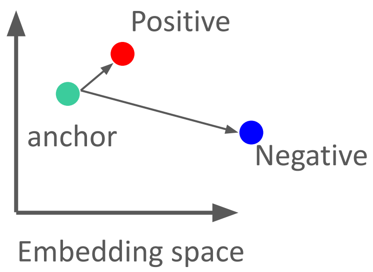
Multilingual Wikipedia

hyperlink prediction [Calixto et al., 2021]

→ We train sentence embeddings exploiting these properties of entity

Contrastive learning

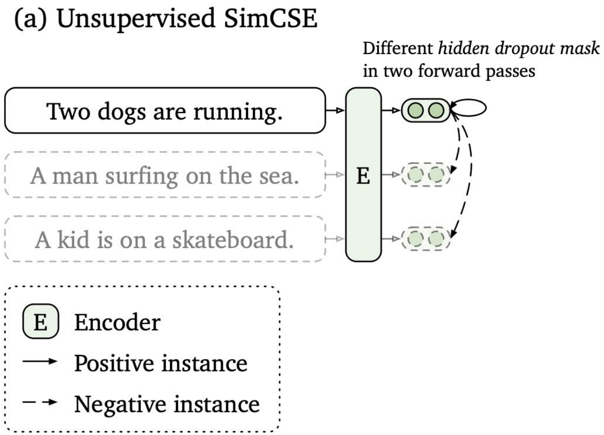
- ➢ Contrastive learning (CL) puts semantically similar samples close and keeps dissimilar samples apart [Hadsell et al., 2006]
- ➢ It is a popular and effective way for representation learning



-> Our approach:
contrastive learning between **sentences** and their associated **entities**

SimCSE

Unsup. SimCSE predicts input sentence itself with dropout noise
[Gao et al., 2021]



→ We extend Unsup. SimCSE model with entity-based contrastive learning (Entity CL)

Outline

- Backgound
- **Proposed Method**
- Experiment
- Analysis
- Conclusion

Entity CL

My Neighbor Totoro is animated by Studio Ghibli

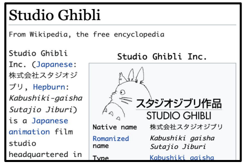
Proposed method

Entity CL

My Neighbor Totoro is animated by Studio Ghibli

My Neighbor Totoro is animated by [Studio Ghibli]

[Studio Ghibli]


Studio Ghibli

From Wikipedia, the free encyclopedia

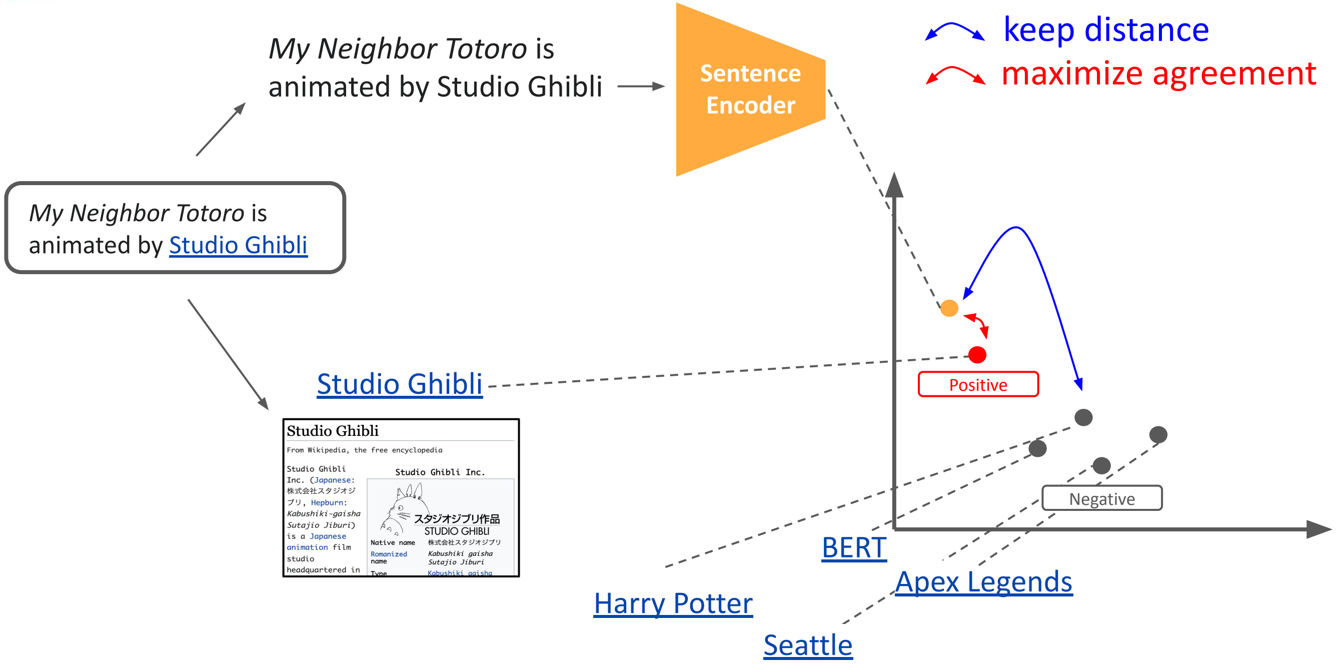
Studio Ghibli Inc. (Japanese: 株式会社スタジオジブリ, Hepburn: *Kabushiki-gaisha Sutajio Jiburi*) is a Japanese animation film studio headquartered in

Studio Ghibli Inc.

スタジオジブリ作品
STUDIO GHIBLI

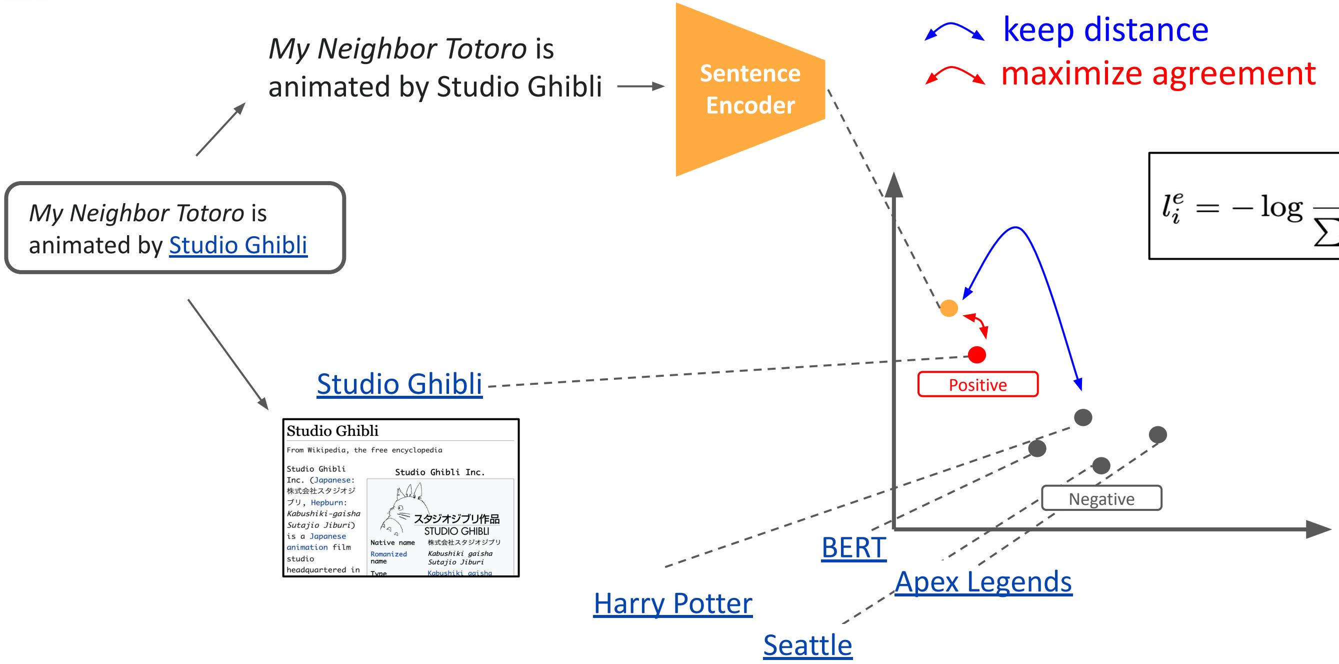
Native name	株式会社スタジオジブリ
Romanized name	*Kabushiki gaisha Sutajio Jiburi*
Type	Kabushiki gaisha

Entity CL



Pull the sentence embeddings and **their related entities (hyperlinks)** closer while pushing random entities apart

Entity CL

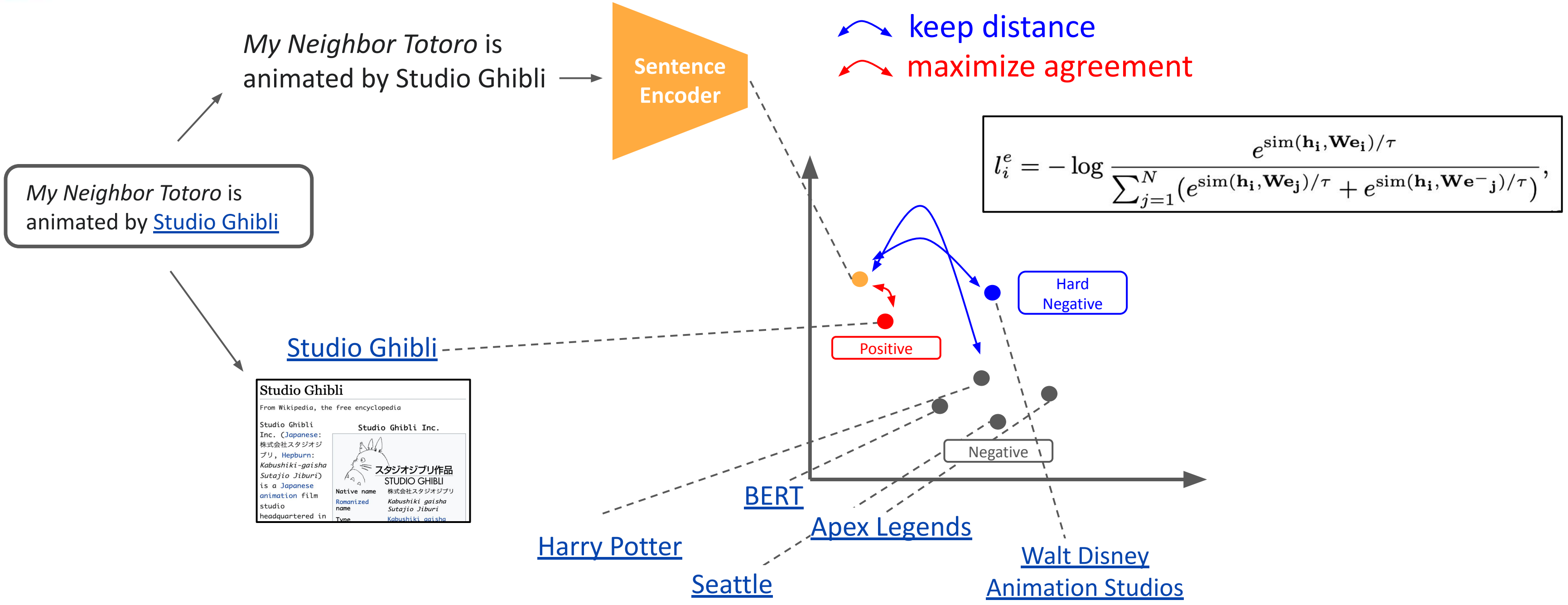


$$l_i^e = -\log \frac{e^{\mathrm{sim}(\mathbf{h_i}, \mathbf{We_i})/\tau}}{\sum_{j=1}^{N} e^{\mathrm{sim}(\mathbf{h_i}, \mathbf{We_j})/\tau}},$$

Pull the sentence embeddings and **their related entities (hyperlinks)** closer while pushing random entities apart

Proposed method

Entity CL

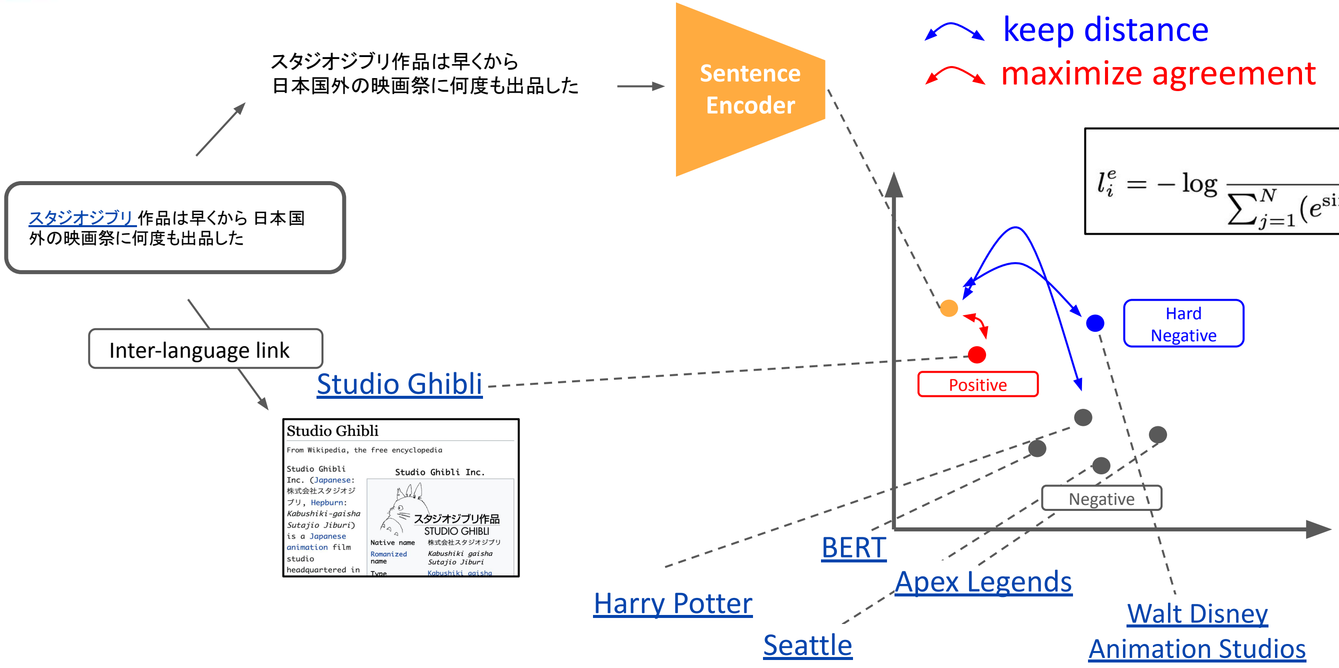


Introduce **hard negative entities** that satisfy the following two conditions:

- entities with the same type as the positive entity
- entities that do not appear on the same Wikipedia page

Proposed method

Entity CL

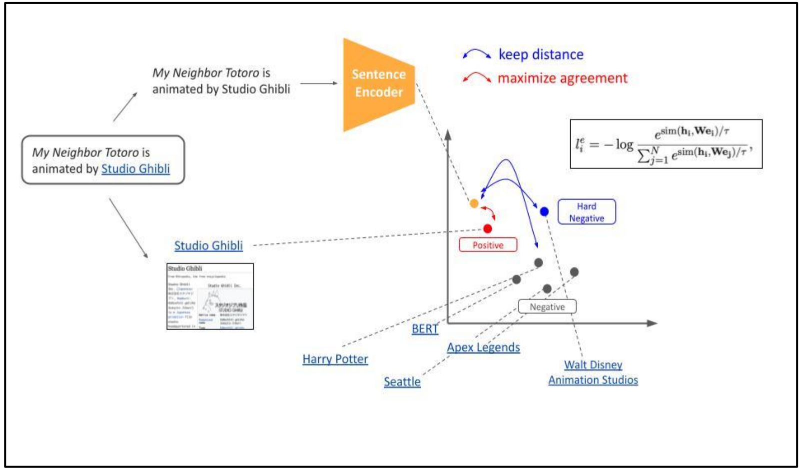


$$l_i^e = -\log \frac{e^{\mathrm{sim}(\mathbf{h_i}, \mathbf{We_i})/\tau}}{\sum_{j=1}^{N} \left(e^{\mathrm{sim}(\mathbf{h_i}, \mathbf{We_j})/\tau} + e^{\mathrm{sim}(\mathbf{h_i}, \mathbf{We^-_j})/\tau}\right)},$$

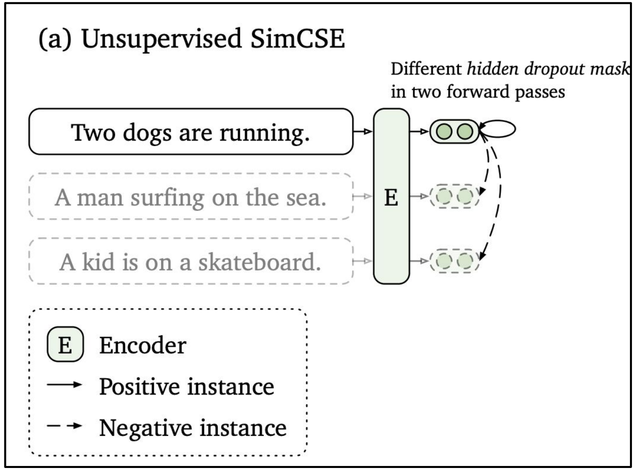
Language-independent entities allow us to use **multilingual sentences** during EASE training

EASE

Combined entity CL with self-supervised CL



+



Entity CL

SimCSE

Overall loss: $l_i^{ease} = \lambda l_i^e + l_i^s,$

Outline

- Backgound
- Proposed Method
- **Experiment**
- Analysis
- Conclusion

Experiment: Overview

- **Monolingual Setting**
  - Training data: **English** Wikipedia
  - Evaluation: **monolingual** tasks
- **Multilingual Setting**
  - Training data: Wikipedia in **multiple languages**
  - Evaluation: **multilingual** tasks
- **Case Study**
  - Fine-tune LaBSE with EASE framework

Experiment: Monolingual

- Data

  One million entity-sentence pairs sampled from **English** Wikipedia

- Baselines

  SOTA unsupervised sentence embedding methods

- Entity embedding

  Wikipedia2Vec [Yamada et al., 2020] embeddings trained from English Wikipedia

- Tasks

  Semantic textual similarity (STS), Short text clustering (STC)

Results

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).

- ➢ EASE exhibits competitive or better performance in monolingual STS and STC
- ➢ EASE excel at capturing high-level categorical semantic structure

Experiment: Multilingual

- Data
  - Aggregated Wikipedia data of 50,000 pairs for each language (18 language)
- Baseline
  - SimCSE trained using the same multilingual data as EASE
- Entity embedding
  - Wikipedia2Vec embeddings trained from **English** Wikipedia
- Tasks
  - Multilingual STS, Multilingual STC (**MewsC-16**)
  - Cross-lingual Parallel Matching (Tatoeba), Cross-lingual Text Classification (MLDoc)

MewsC-16

➢ Our novel dataset: MewsC-16 (**M**ultilingual Short Text **C**lustering Dataset for N**ews** in **16** languages)

➢ MewsC-16 contains topic sentences from Wikinews articles in 13 categories

Language	Sentence	Category
En	December 22, 2004 The controversial European Union Directive on the Patentability of Computer Implemented Inventions, also called the “software patent directive” has been put to rest for 2004.	Science and technology
Ja	7月14日、第133回の芥川賞、直木賞（日本文学振興会）の選考会が東京の築地・新喜楽で行われた	Culture and entertainment

Examples of MewsC-16

Results

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

EASE significantly outperforms SimCSE in multilingual STS and STC

Results

Model	ar	ca	cs	de	eo	es	fr	it	ja	ko	nl	pl	pt	ru	sv	tr	Avg.
mBERT$_{\text{base}}$ (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$_{\text{base}}$	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$_{\text{base}}$	**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$_{\text{base}}$ (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$_{\text{base}}$	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$_{\text{base}}$	**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	en (dev)	de	es	fr	it	ja	ru	zh	Avg.
mBERT$_{\text{base}}$ (avg.)	**89.5**	68.0	68.1	70.6	62.7	61.2	61.5	69.6	65.9
SimCSE-mBERT$_{\text{base}}$	88.4	62.3	**73.2**	78.2	64.3	**63.7**	61.3	**75.0**	68.3
EASE-mBERT$_{\text{base}}$	89.0	**69.9**	69.2	**80.1**	**66.8**	62.8	**64.4**	73.2	**69.5**
XLM-R$_{\text{base}}$ (avg.)	**90.9**	**82.7**	**79.8**	72.1	72.5	71.1	69.6	71.4	74.2
SimCSE-XLM-R$_{\text{base}}$	90.7	74.9	74.1	81.5	70.3	71.7	70.1	76.6	74.2
EASE-XLM-R$_{\text{base}}$	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.

EASE significantly outperforms SimCSE in multilingual CLPM and CLTC

Case Study

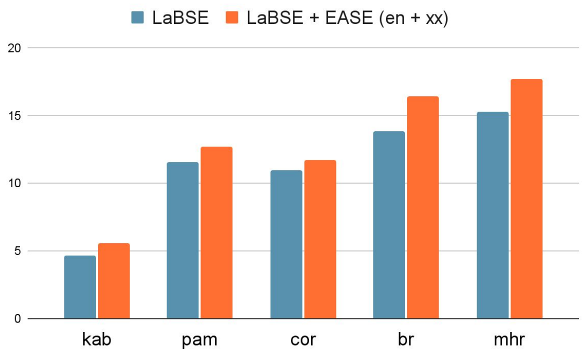
Can we leverage Wikipedia to complement the performance of existing models?

-> We **fine-tuned** LaBSE with our EASE framework in low-resource languages

Case Study

Can we leverage Wikipedia to complement the performance of existing models?

-> We **fine-tuned** LaBSE with our EASE framework in low-resource languages



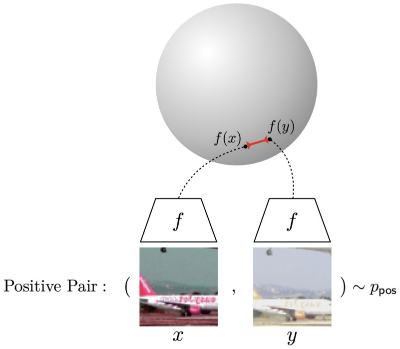
EASE successfully complement the performance of LaBSE!

Outline

Alignment and Uniformity

Two key properties for the CL-based representations [Wang and Isola, 2020]

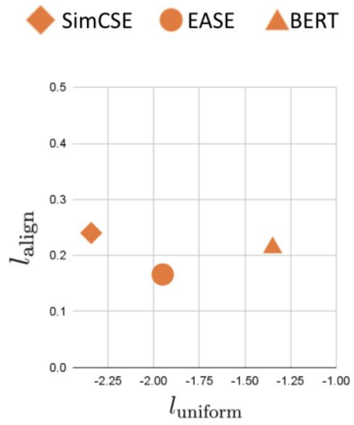
- Alignment: the closeness of representations between positive pairs
- Uniformity: how well the representations are uniformly distributed



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

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

Alignment and Uniformity



Alignment and uniformity

Alignment: **EASE** < BERT < SimCSE

Uniformity: SimCSE < **EASE** < BERT

(Lower numbers are Better)

Entity CL has the effect of **aligning semantically similar examples**

Outline

- ➢ Backgound
- ➢ Proposed Method
- ➢ Experiment
- ➢ Analysis
- ➢ **Conclusion**

Conclusion

- ➢ Proposed EASE, a novel method for learning sentence embeddings via contrastive learning between sentences and their associated entities

- ➢ EASE significantly outperforms baseline methods in both monolingual settings, and, especially multilingual settings

Published our source code, pre-trained models, and newly constructed multilingual STC (MewsC-16) dataset

https://github.com/studio-ousia/ease

Thank you for listening

Question?

Results for unseen languages

evaluate EASE on languages not included in the training set

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

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

EASE cross-lingual alignment effect propagates to other languages

Ablation Study

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.

- ➢ All components contribute to the performance!
- ➢ Entity CL alone also improves the baseline performance significantly

Quantitative analysis

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

- ➢ EASE embeddings are more robust to synonyms and grammatical differences
- ➢ EASE embeddings are sometimes overly sensitive to topical similarity

Limitation

	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.

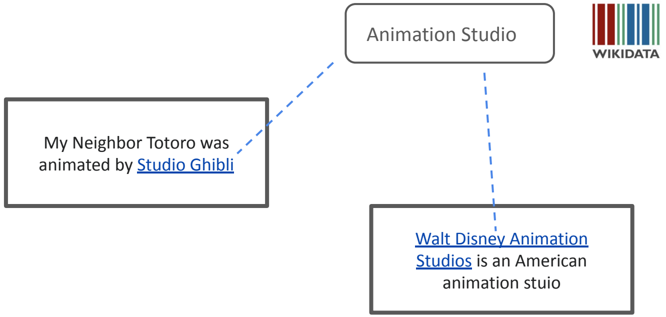
EASE performance is significantly poor
than that of LaBSE for the parallel sentence mining task

Hard negative entity

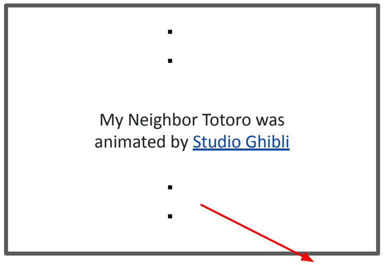
1. Similar to the positive entity
- The same type as the positive entity

2. Yet unrelated to the sentence
- Not on the same Wikipedia page as the positive entity

Entity type in Wikidata



Article A



Walt Disney Animation Studios
is not in this article!

Details of evaluation dataset

- **Monolingual Setting**
  - STS 2012-2016 (Agirre et al., 2012, 2013, 2014, 2015, 2016), STS Benchmark (Cer et al., 2017), and SICK-Relatedness (Marelli et al., 2014)
  - eight benchmark datasets for STC (Zhang et al., 2021)
- **Multilingual Setting**
  - STS 2017 dataset (Reimers and Gurevych, 2020)
  - MewsC-16 (created by us)
  - Tatoeba dataset (Artetxe and Schwenk, 2019)
  - MLDoc (Schwenk and Li, 2018)