Cross-Lingual Semantic Search

Nils Reimers
HuggingFace
Creator of Sentence-Transformers (www.SBERT.net)

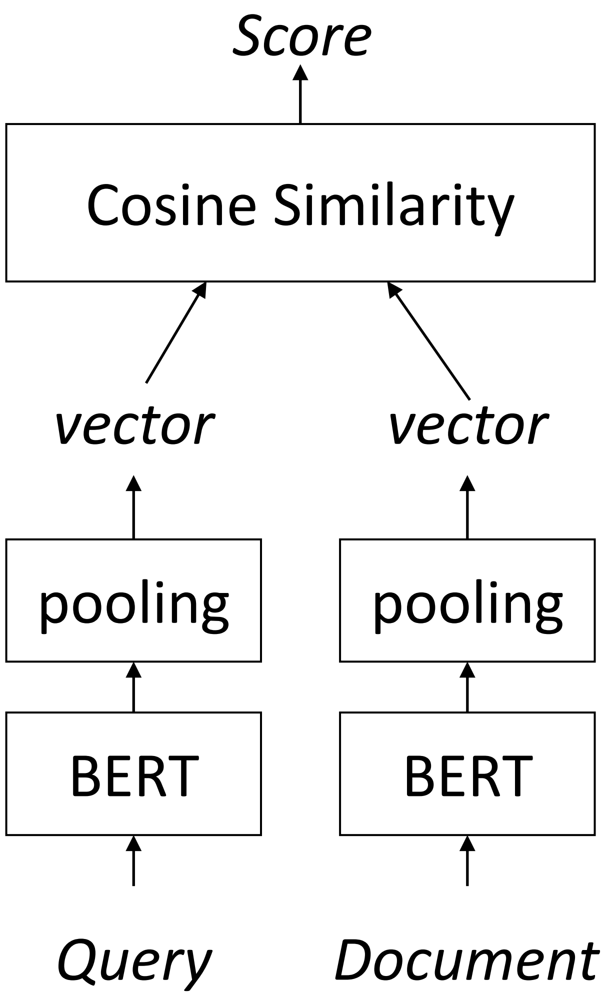
European Language Resource Coordination

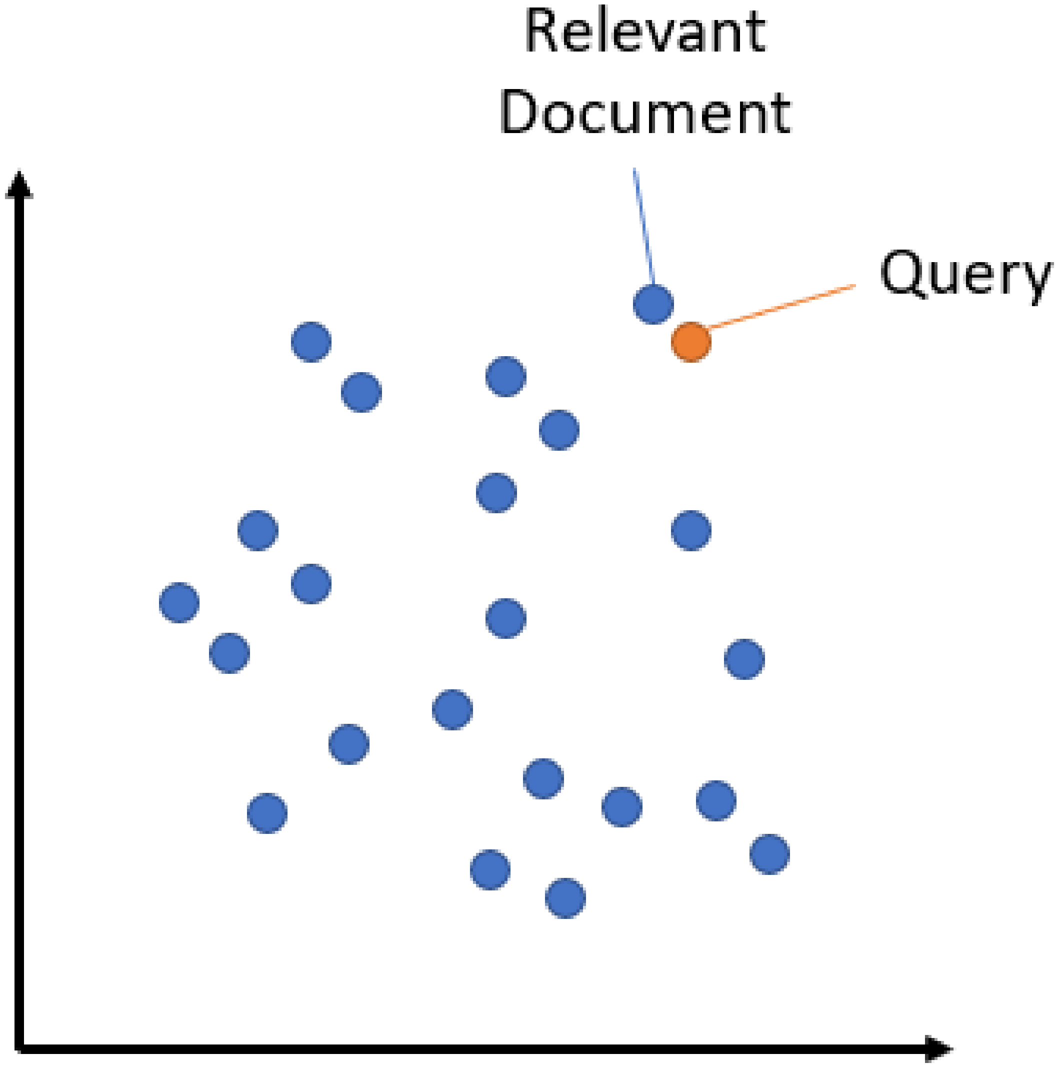
Neural Search – Why all the Hype?

- Real example on (Simple) Wikipedia (170k documents)
- Query: `What is the capital of the United States?`
- Top-3 Hits

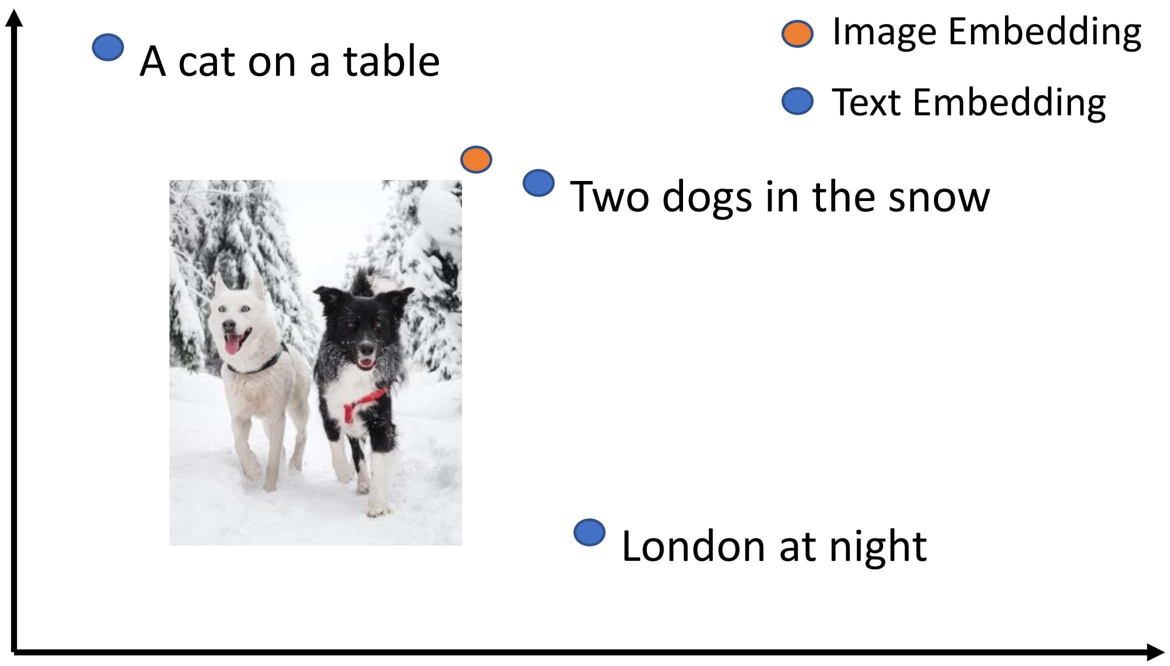
Lexical Search (BM25)	**Neural Search**
▪ **Capital** punishment (the death penalty) has existed in the **United States** […] ▪ Ohio is one of the 50 **states** in the **United States**. Its **capital** is Columbus. […] ▪ Nevada is one of the **United States'** **states**. Its **capital** […]	▪ Washington, D.C. […] is the **capital of the United States**. […] ▪ A capital city (or capital town or just capital) is a city or town, […] ▪ The United States **Capitol** is the building where the United States Congress meets […]

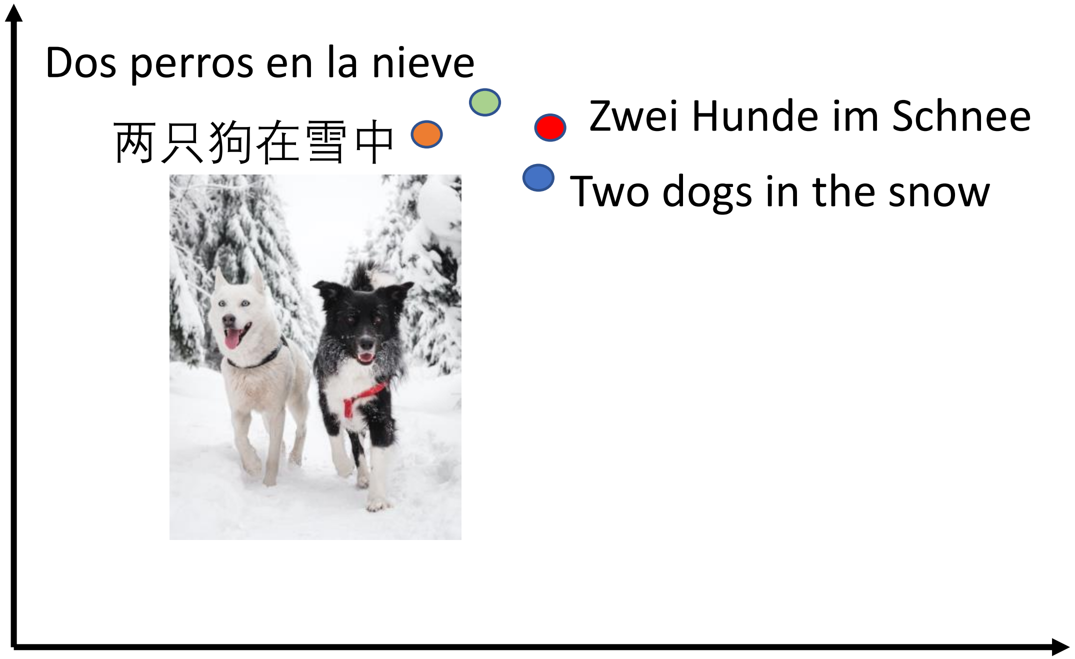
Bi-Encoders


Score
Cosine Similarity
vector
vector
pooling
pooling
BERT
BERT
Query
Document


Relevant
Document
Query

Multi-Modal & Multi-Lingual Search





Multiple Negative Ranking Loss

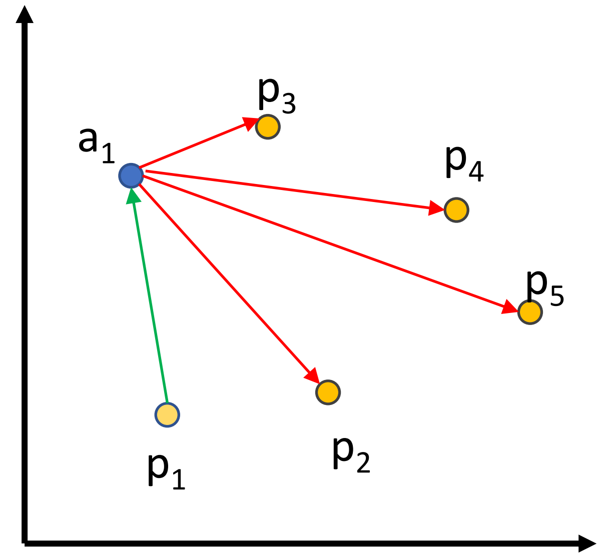
- Have positive pairs:
  (a_1, p_1)
  (a_2, p_2)
  (a_3, p_3)

- Examples:
  - (query, answer-passage)
  - (English sentence, French Sentence)

- (a_i, p_i) should be close in vector space and (a_i, p_j) should be distant in vector space ($i != j$)
  - Unlikely that e.g. two randomly selected questions are similar

- Computed as ranking loss with Cross-Entropy:
  - Given a_1, which is the right answer out of [p_1, p_2, p_3]?
  - Compute scores: [$s(a_1, p_1)$, $s(a_1, p_2)$, $s(a_1, p_3)$]
  - Cross-Entropy loss with gold label: [1, 0, 0]

- Also called “training with in-batch negatives”, InfoNCE or NTXentLoss



Multiple Negative Ranking Loss Intuitive Explanation

- a_1: How do you feel today?
  - p_1: Wie fühlst du dich heute? *(How do you feel today?)*
  - p_2: Vielen Dank für die Frage *(Thank you for the question)*
  - p_3: Gibt es weitere Fragen *(Are there further questions?)*

- Compute text embeddings & compute similarities:
  - $sim(a_1, p_1) = 0.5$
  - $sim(a_1, p_2) = 0.3$
  - $sim(a_1, p_3) = 0.1$

- See it as classification task and use Cross-Entropy Loss:
  - Prediction: [0.5, 0.3, 0.1]
  - Gold: [1, 0, 0]

Multiple Negative Ranking Loss
Hard Negatives

- Larger batch size => task more difficult => better results
  - Given query, which of the 10 passages provide the answer?
  - Given query, which of the 1k passages provide the answer?

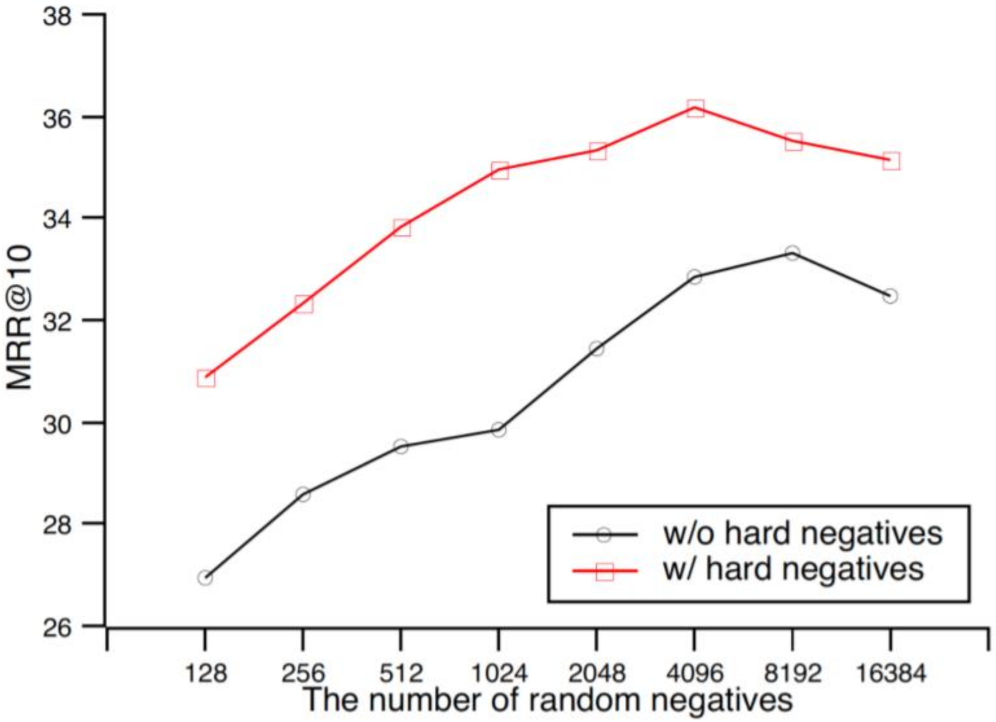


Image: https://arxiv.org/pdf/2010.08191.pdf

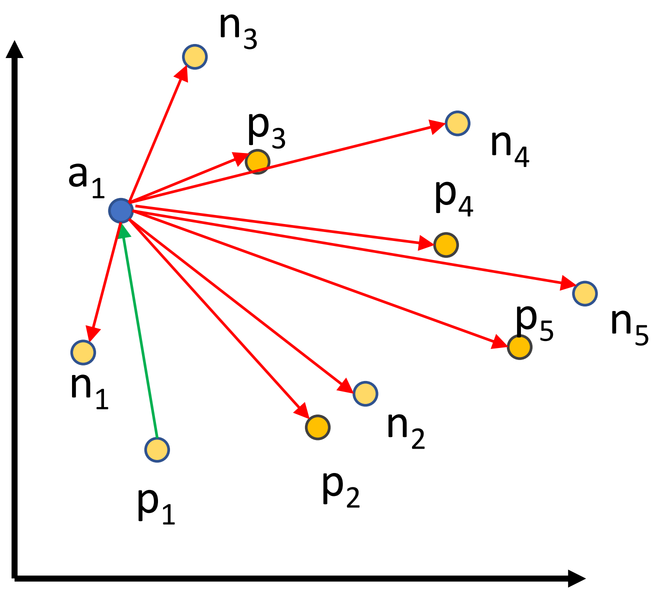
Multiple Negative Ranking Loss
Hard Negatives

- Train with tuples:
  (a_1, p_1, n_1)
  (a_2, p_2, n_2)

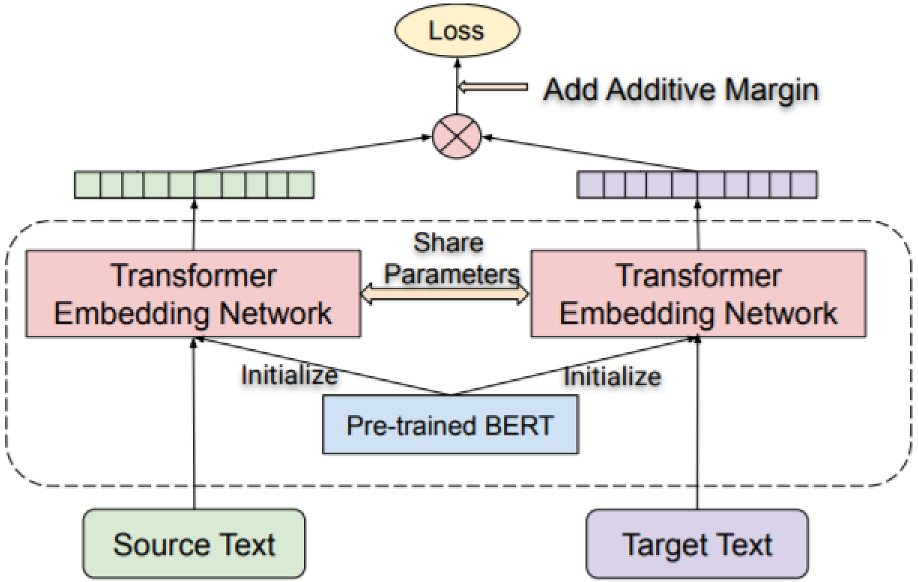
- n_i should be similar to p_i but not match with a_i

- Bad example:
  a: How many people live in London?
  p: Around 9 million people live in London
  n: London has a population of 9 million people.

- Good example:
  a: How many people live in London?
  p: Around 9 million people live in London
  n: Around 1 million people live in Birmingham, second to London.



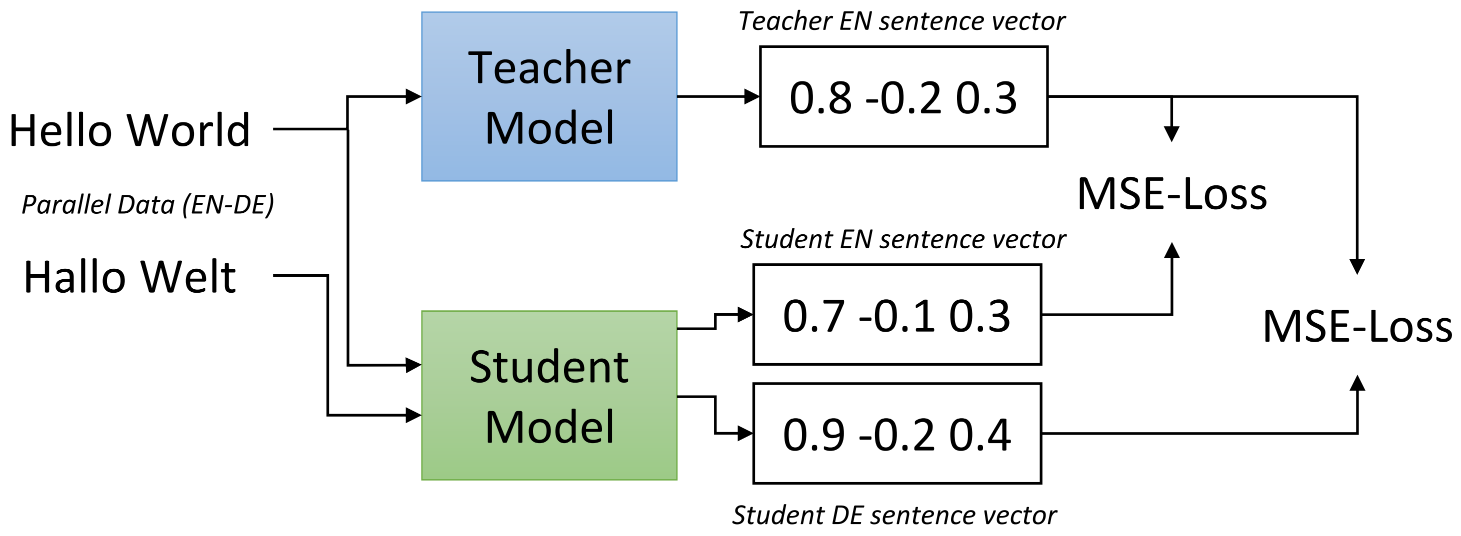
LaBSE



- Pre-Training
  - Trained on large mono-lingual dataset via MLM
  - Trained on translation pairs via TLM (Translation Lang. Model)
- Fine-tuned on translation pairs via MultipleNegativesRankingLoss

https://arxiv.org/abs/2007.01852

Multilingual Knowledge Distillation



- Given:
  - Teacher sentence embedding model T (e.g. SBERT trained on English STS)
  - Parallel sentence data $((s_1, t_1), \ldots, (s_n, t_n))$
  - Student model S with multilingual vocabulary (e.g. XLM-R + Mean Pooling)
- Train student S such that:

$$S(s_i) \approx T(s_i) \qquad S(t_i) \approx T(s_i)$$

https://arxiv.org/abs/2004.09813

Results – Semantic Similarity

- Given two sentences, predict semantic similarity (0…5)

Model	EN-AR	EN-DE	EN-TR	EN-ES	EN-FR	EN-IT	EN-NL	Avg.
mBERT mean	16.7	33.9	16.0	21.5	33.0	34.0	35.6	27.2
XLM-R mean	17.4	21.3	9.2	10.9	16.6	22.9	26.0	17.8
mBERT-nli-stsb	30.9	62.2	23.9	45.4	57.8	54.3	54.1	46.9
XLM-R-nli-stsb	44.0	59.5	42.4	54.7	63.4	59.4	66.0	55.6
Knowledge Distillation								
mBERT ← SBERT-nli-stsb	77.2	78.9	73.2	79.2	78.8	78.9	77.3	77.6
DistilmBERT ← SBERT-nli-stsb	76.1	77.7	71.8	77.6	77.4	76.5	74.7	76.0
XLM-R ← SBERT-nli-stsb	77.8	78.9	74.0	79.7	78.5	78.9	77.7	77.9
XLM-R ← SBERT-paraphrases	82.3	84.0	80.9	83.1	84.9	86.3	84.5	**83.7**
Other Systems								
LASER	66.5	64.2	72.0	57.9	69.1	70.8	68.5	67.0
mUSE	79.3	82.1	75.5	79.6	82.6	84.5	84.1	81.1
LaBSE	74.5	73.8	72.0	65.5	77.0	76.9	75.1	73.5

- mBERT / XLM-R perform badly when trained on English only
- Knowledge Distillation incorporates knowledge from teacher model
- LASER & LaBSE perform badly

Bitext Mining

- Given two corpora: Find parallel (translated) sentences

Model	DE-EN	FR-EN	RU-EN	ZH-EN	Avg.
mBERT mean	44.1	47.2	38.0	37.4	41.7
XLM-R mean	5.2	6.6	22.1	12.4	11.6
mBERT-nli-stsb	38.9	39.5	26.4	30.2	33.7
XLM-R-nli-stsb	44.0	51.0	51.5	44.0	47.6
Knowledge Distillation					
XLM-R $\leftarrow$ SBERT-nli-stsb	86.8	84.4	86.3	85.1	85.7
XLM-R $\leftarrow$ SBERT-paraphrase	90.8	87.1	88.6	87.8	88.6
Other systems					
mUSE	88.5	86.3	89.1	86.9	87.7
LASER	95.4	92.4	92.3	91.7	93.0
LaBSE	95.9	92.5	92.4	93.0	93.5

Table 3: F_1 score on the BUCC bitext mining task.

- LASER & LaBSE better than mUSE & Knowledge Distillation
- Issue with mUSE & KD: They find similar sentences, that are not perfect translations

Data Efficiency

Dataset	#DE	EN-DE	#AR	EN-AR
XLM-R mean	-	21.3	-	17.4
XLM-R-nli-stsb	-	59.5	-	44.0
MUSE Dict	101k	75.8	27k	68.8
Wikititles Dict	545k	71.4	748k	67.9
MUSE + Wikititles	646k	76.0	775k	69.1

Dataset size	EN-DE	EN-AR
XLM-R mean	21.3	17.4
XLM-R-nli-stsb	59.5	44.0
1k	71.5	48.4
5k	74.5	59.6
10k	77.0	69.5
25k	80.0	70.2
Full TED2020	80.4	78.0

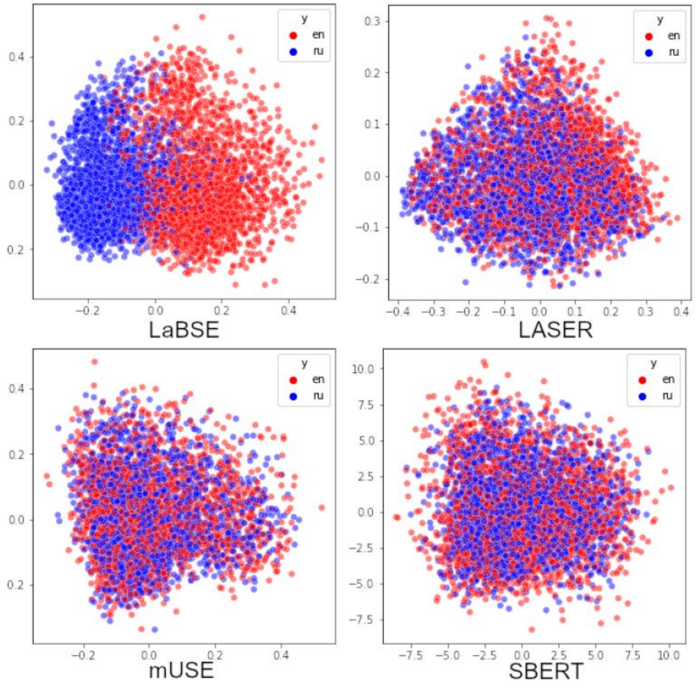
Table 6: Performance on STS 2017 dataset when trained with reduced TED2020 dataset sizes.

Knowledge Distillation vs. Training on Target Language

Model	KO-KO
LASER	68.44
mUSE	76.32
Trained on KorNLI & KorSTS	
Korean RoBERTa-base	80.29
Korean RoBERTa-large	80.49
XLM-R	79.19
XLM-R-large	81.84
Multiling. Knowledge Distillation	
XLM-R ← SBERT-nli-stsb	81.47
XLM-R-large ← SBERT-large-nli-stsb	83.00

Table 7: Spearman rank correlation on Korean STS-benchmark test-set (Ham et al., 2020).

Language Bias



- Preference of certain language combinations
- Language bias impacts performance negatively on multilingual pools
- LASER and LaBSE with strong language bias

Model	Expected Score	Actual Score	Difference
LASER	69.5	68.6	-0.92
mUSE	81.7	81.6	-0.19
LaBSE	74.4	73.1	-1.29
XLM-R $\leftarrow$ SBERT-paraphrases	84.0	83.9	-0.11

Language Bias – Good or Bad?

Side-effects **with** language bias:

- Same language results are ranked higher just because of language
- There might be better hits / answers in other languages

Side-Effects **without** Language Bias

wedding

शादी (hindi: wedding)

Who is the president?

A: Joe Biden is the current president

qui est le président?

A: Joe Biden is the current president

Conclusions

- Sematic search much better than keyword search
- Can work on many languages & modalities
- Usage of translation pairs to align vector spaces
- Training setup impacts if language bias exist or not
- How to get a language bias free model that still respects cultural / language specific differences
  - cat vs बिल्ली
  - wedding vs शादी