xSIM++: An Improved Proxy to Bitext Mining Performance for Low-Resource Languages

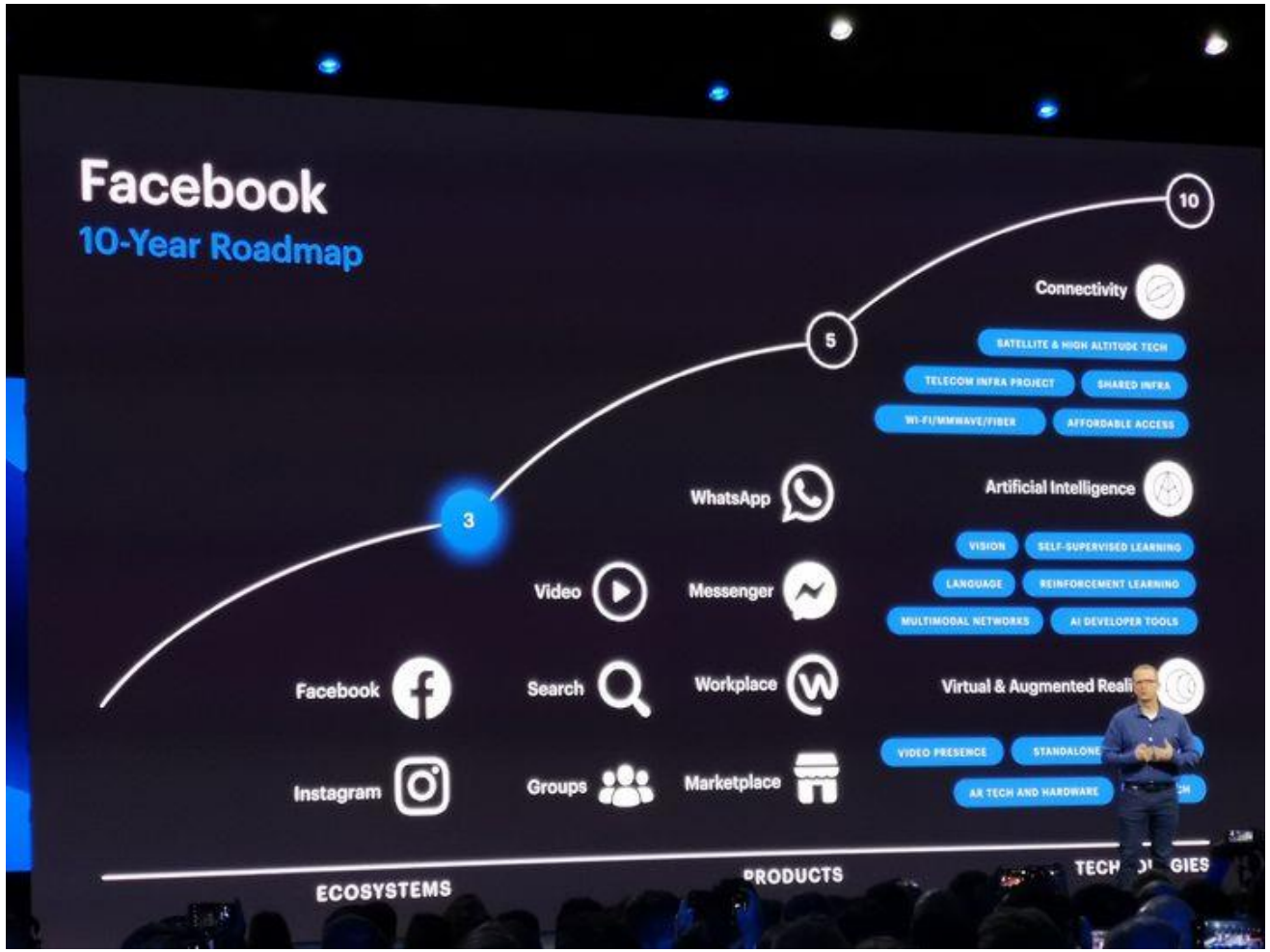
Meta AI (ACL 2023)

1. 동기

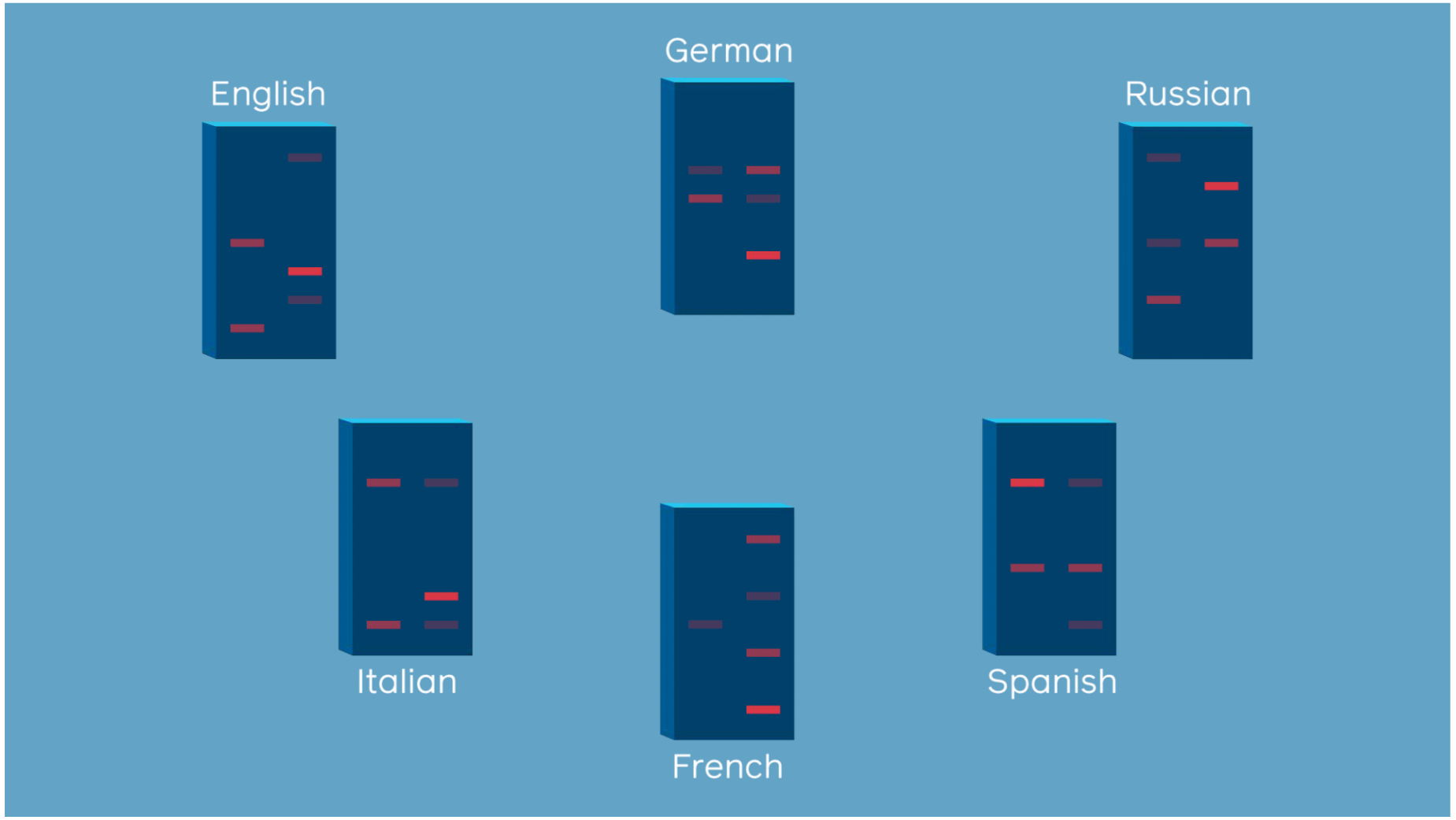
INDUSTRIES IN DEPTH

Mark Zuckerberg: Building a global community that works for everyone

Feb 17, 2017



2. Bitext mining이란



Our method automates and parallelizes this bitext mining process,
processing multiple batches of 50 million examples at a time

3. Bitext mining pipeline

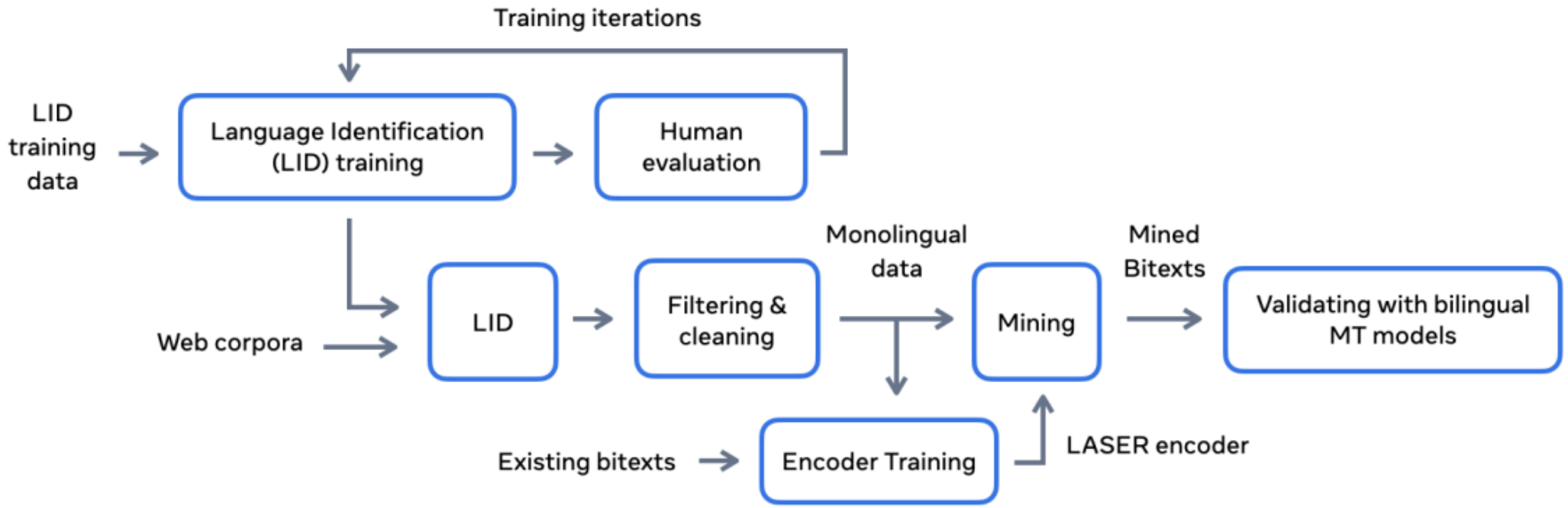


Figure 7: **Overview of our Bitext Mining Pipeline.** Language identification is applied on web corpora to extract monolingual sentences. Aligned pairs are later identified with LASER3.

4. 선행연구: Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings

	농산물은 차, 쌀, 설탕, 담배, 장뇌, 과일, 그리고 비단으로 구성되어 있습니다.
(A)	*Les produits agricoles sont constitués de thé, de riz, de sucre, de tabac, de camphre, de fruits et de soie.*
0.818	Main crops include wheat, sugar beets, potatoes, cotton, tobacco, vegetables, and fruit.
0.817	The fertile soil supports wheat, corn, barley, tobacco, sugar beet, and soybeans.
0.814	Main agricultural products include grains, cotton, oil, pigs, poultry, fruits, vegetables, and edible fungus.
0.808	The important crops grown are cotton, jowar, groundnut, rice, sunflower and cereals.
	하지만 현재의 상황에서는 위험 없이 그것들을 무시할 수 있을 것입니다.
(B)	*Mais dans le contexte actuel, nous pourrons les ignorer sans risque.*
0.737	But, in view of the current situation, we can safely ignore these.
0.499	But without the living language, it risks becoming an empty shell.
0.498	While the risk to those working in ceramics is now much reduced, it can still not be ignored.
0.488	But now they have discovered they are not free to speak their minds.

Table 1: Motivating example of the proposed method. We show the nearest neighbors of two French sentences on the BUCC training set along with their cosine similarities. Only the nearest neighbor of B is a correct translation, yet that of A has a higher cosine similarity. We argue that this is caused by the cosine similarity of different sentences being in different scales, making it a poor indicator of the confidence of the prediction. Our method tackles this issue by considering the margin between a given candidate and the rest of the k nearest neighbors.

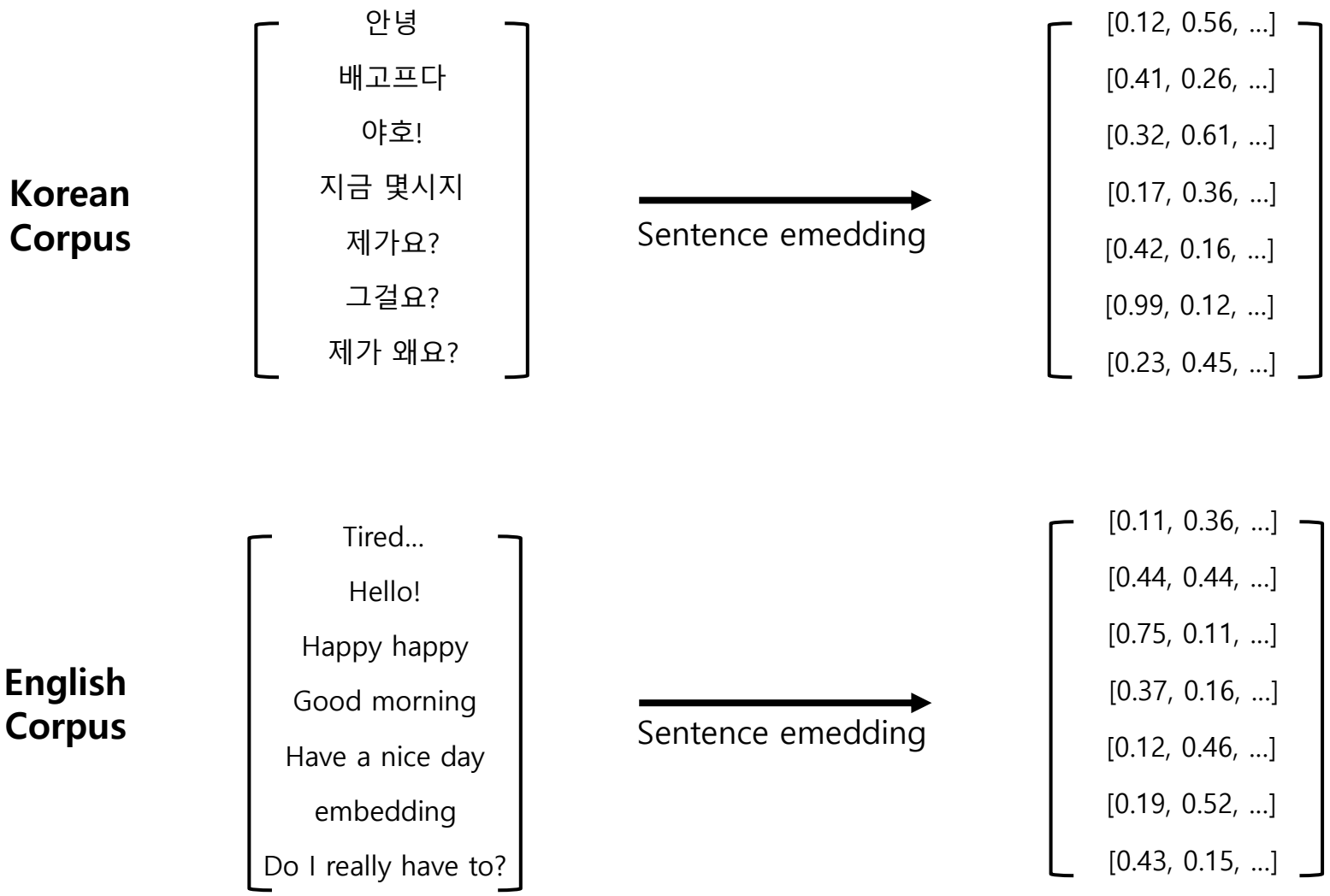
4. 선행연구: Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings

$$\text{score}(x, y) = \text{margin}\left(cos(x, y), \sum_{z \in NN_k(x)} \frac{cos(x, z)}{2k} + \sum_{v \in NN_k(y)} \frac{cos(y, v)}{2k}\right)$$

where NN*k*(*x*) denotes the *k* nearest neighbors of *x* in the other language excluding duplicates

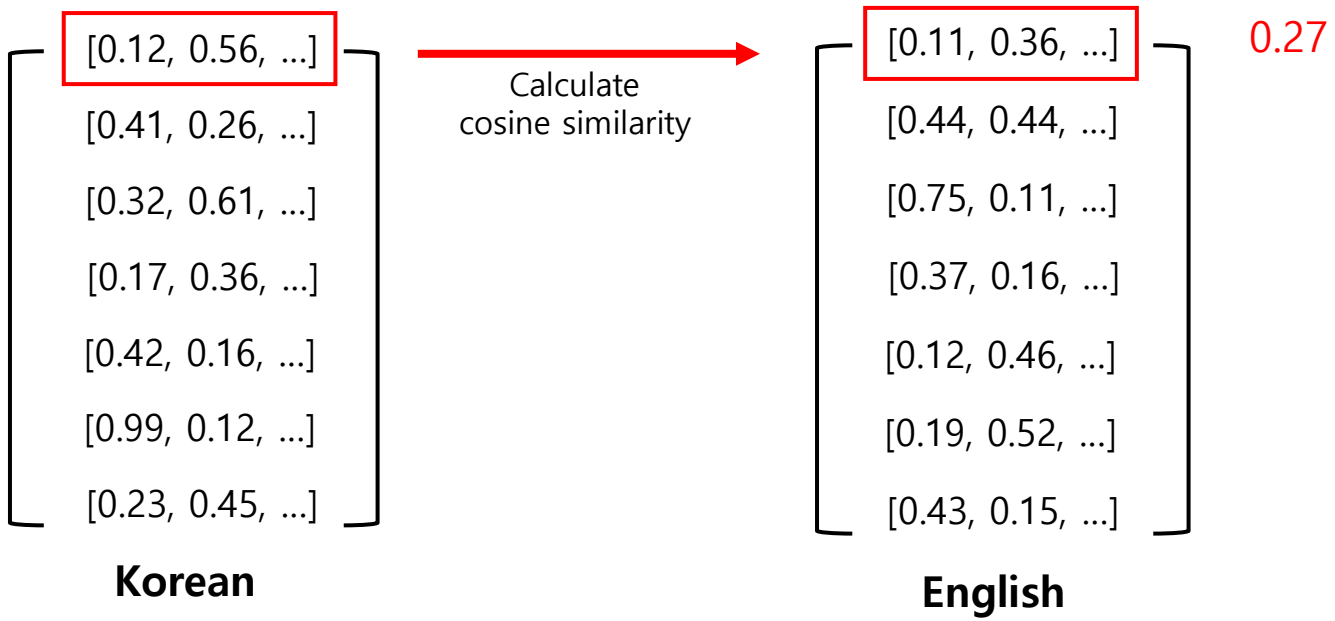
- **Absolute** (margin(*a*, *b*) = *a*): Ignoring the average. This is equivalent to **cosine similarity** and thus our baseline.
- **Distance** (margin(*a*, *b*) = *a* − *b*): Subtracting the average cosine similarity from that of the given candidate.
- **Ratio** (margin(*a*, *b*) = *a* / *b*): The ratio between the candidate and the average cosine of its nearest neighbors in both directions. (**best)**

4. 선행연구: Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings

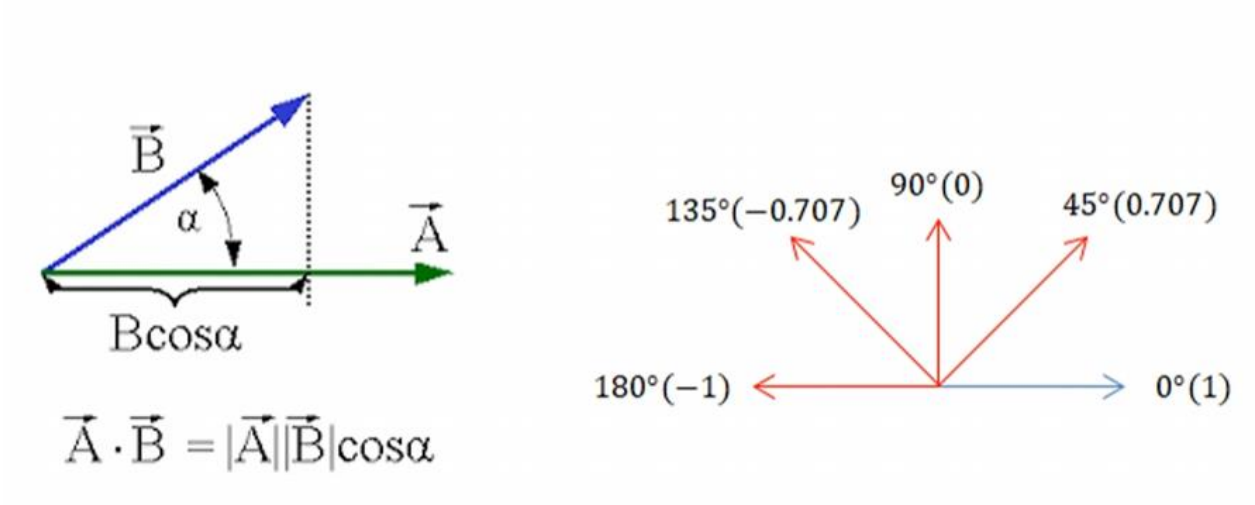


여기서 embedding 모델로 LASER encoder 를 사용, 후속연구에서 LASER3 encoder와 LaBSE를 같이 사용

4. 선행연구: Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings

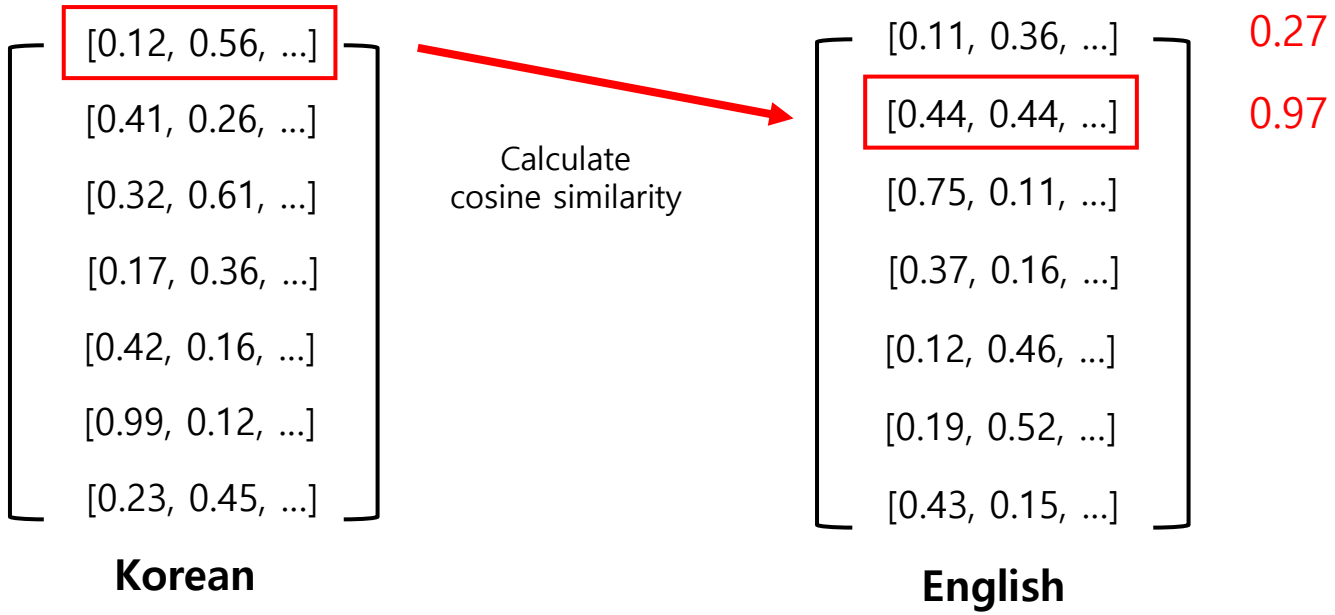


4. 선행연구: Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings

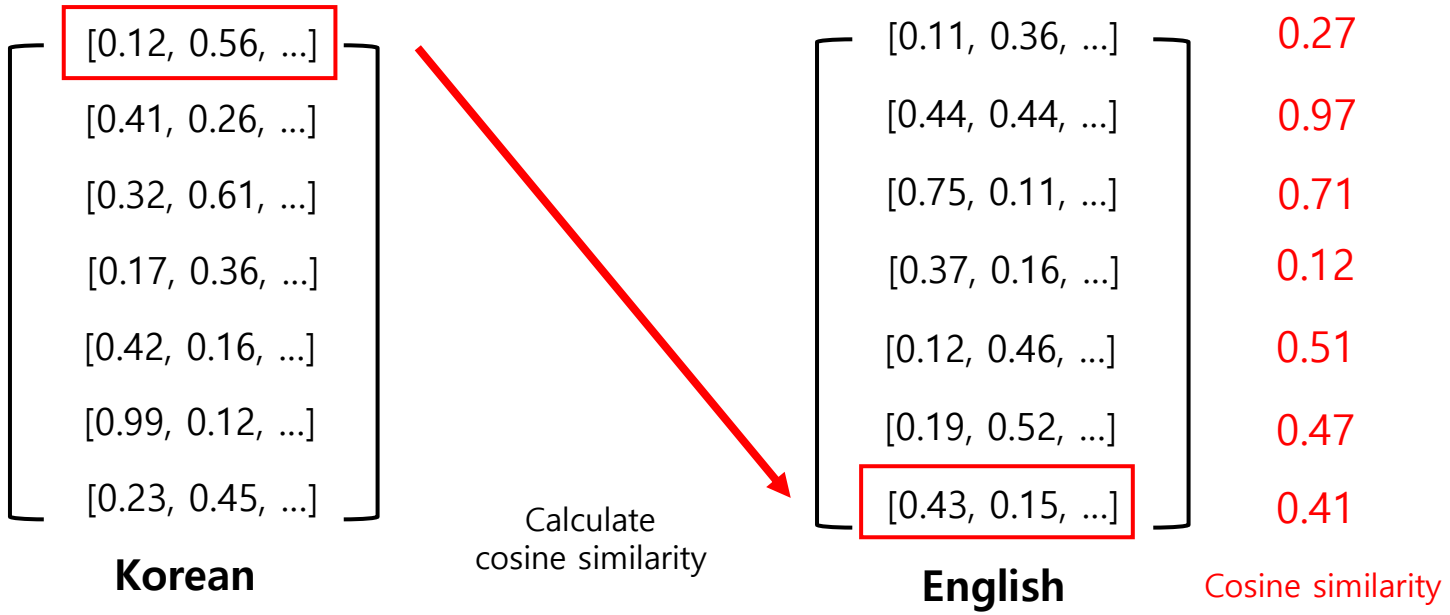


Cosine similarity

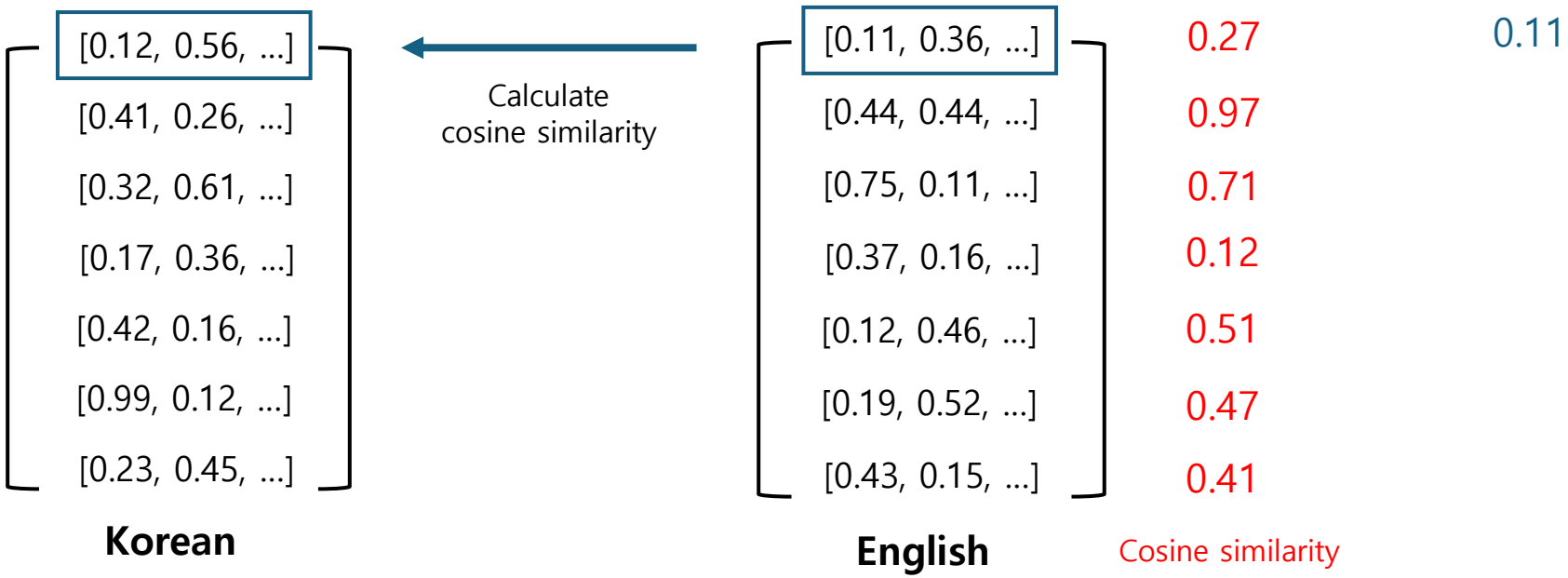
4. 선행연구: Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings



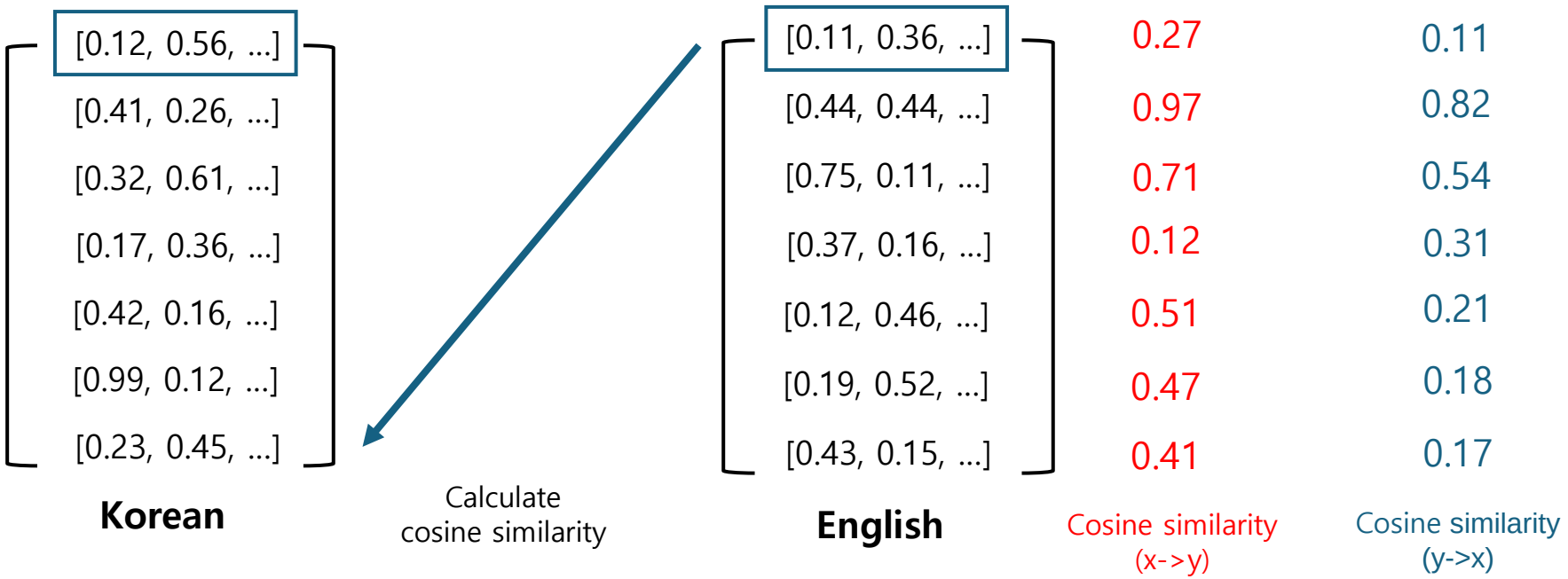
4. 선행연구: Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings



4. 선행연구: Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings



4. 선행연구: Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings



4. 선행연구: Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings

$$\text{score}(x, y) = \text{margin}\left(cos(x, y), \sum_{z \in NN_k(x)} \frac{cos(x, z)}{2k} + \sum_{v \in NN_k(y)} \frac{cos(y, v)}{2k}\right)$$

Unless otherwise indicated, we use $k = 4$.

4. 선행연구: Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings

$$\text{score}(x, y) = \text{margin}\left(cos(x, y), \sum_{z \in NN_k(x)} \frac{cos(x, z)}{2k} + \sum_{v \in NN_k(y)} \frac{cos(y, v)}{2k}\right)$$

0.27

Unless otherwise indicated, we use $k = 4$.

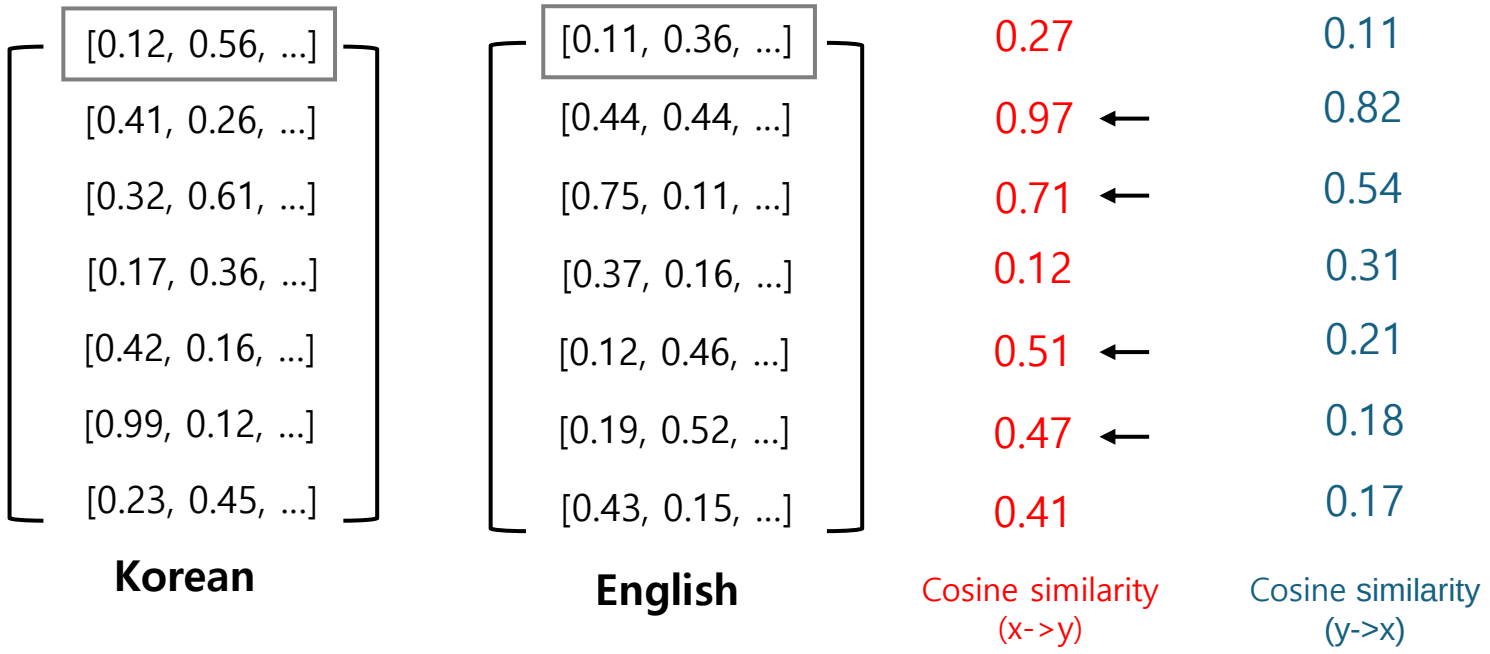
Korean	English	Cosine similarity (x->y)	Cosine similarity (y->x)
[0.12, 0.56, ...]	[0.11, 0.36, ...]	0.27	0.11
[0.41, 0.26, ...]	[0.44, 0.44, ...]	0.97	0.82
[0.32, 0.61, ...]	[0.75, 0.11, ...]	0.71	0.54
[0.17, 0.36, ...]	[0.37, 0.16, ...]	0.12	0.31
[0.42, 0.16, ...]	[0.12, 0.46, ...]	0.51	0.21
[0.99, 0.12, ...]	[0.19, 0.52, ...]	0.47	0.18
[0.23, 0.45, ...]	[0.43, 0.15, ...]	0.41	0.17

4. 선행연구: Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings

$$\text{score}(x, y) = \text{margin}\left(cos(x, y), \sum_{z \in NN_k(x)} \frac{cos(x, z)}{2k} + \sum_{v \in NN_k(y)} \frac{cos(y, v)}{2k}\right)$$

0.27 (cos(x, y)) 0.332 (sum terms)

Unless otherwise indicated, we use $k = 4$.

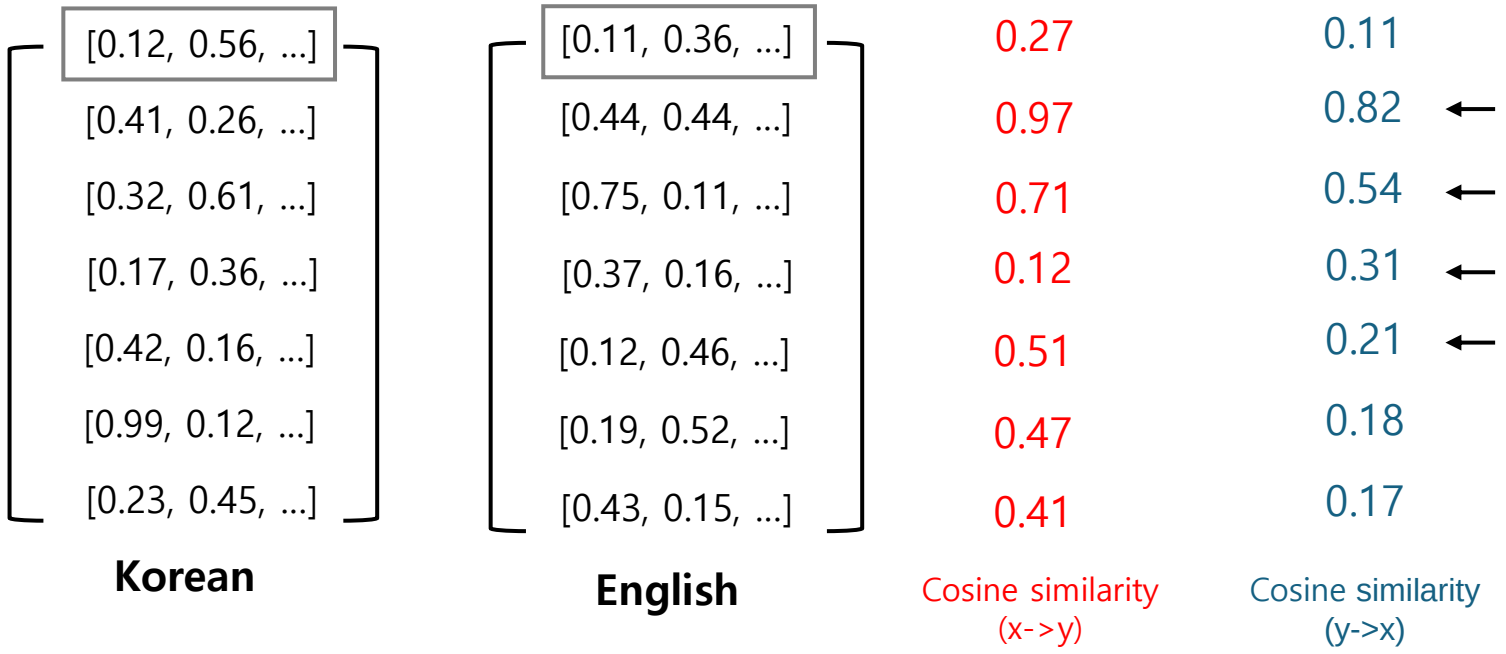


4. 선행연구: Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings

$$\text{score}(x, y) = \text{margin}\left(cos(x, y), \sum_{z \in NN_k(x)} \frac{cos(x, z)}{2k} + \sum_{v \in NN_k(y)} \frac{cos(y, v)}{2k}\right)$$

0.27 0.332 0.235

Unless otherwise indicated, we use $k = 4$.



4. 선행연구: Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings

$$\text{score}(x, y) = \text{margin}\left(cos(x, y), \sum_{z \in NN_k(x)} \frac{cos(x, z)}{2k} + \sum_{v \in NN_k(y)} \frac{cos(y, v)}{2k}\right)$$

Score(x, y) = margin(*0.27*, *0.567*)

- **Absolute** (margin(a, b) = a): Ignoring the average. This is equivalent to **cosine similarity** and thus our baseline.
- **Distance** (margin(a, b) = $a - b$): Subtracting the average cosine similarity from that of the given candidate.
- **Ratio** (margin(a, b) = a / b): The ratio between the candidate and the average cosine of its nearest neighbors in both directions. (**best)**

4. 선행연구: Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings

Func.	Retrieval	EN-DE			EN-FR		
		P	R	F1	P	R	F1
Abs. (cos)	Forward	78.9	75.1	77.0	82.1	74.2	77.9
	Backward	79.0	73.1	75.9	77.2	72.2	74.7
	Intersection	84.9	80.8	82.8	83.6	78.3	80.9
	Max. score	83.1	77.2	80.1	80.9	77.5	79.2
Dist.	Forward	94.8	94.1	94.4	91.1	**91.8**	91.4
	Backward	94.8	94.1	94.4	91.5	91.4	91.4
	Intersection	94.9	94.1	94.5	91.2	**91.8**	91.5
	Max. score	94.9	94.1	94.5	91.2	**91.8**	91.5
Ratio	Forward	95.2	**94.4**	**94.8**	**92.4**	91.3	91.8
	Backward	95.2	**94.4**	**94.8**	92.3	91.3	91.8
	Intersection	**95.3**	**94.4**	**94.8**	**92.4**	91.3	**91.9**
	Max. score	**95.3**	**94.4**	**94.8**	**92.4**	91.3	**91.9**

Table 2: BUCC results (precision, recall and F1) on the training set, used to optimize the filtering threshold.

	en-de	en-fr	en-ru	en-zh
Azpeitia et al. (2017)	83.7	79.5	-	-
Azpeitia et al. (2018)	85.5	81.5	81.3	77.5
Bouamor and Sajjad (2018)	-	76.0	-	-
Schwenk (2018)	76.9	75.8	73.8	71.6
Proposed method (Europarl)	**95.6**	**92.9**	-	-
Proposed method (UN)	-	-	**92.0**	**92.6**

Table 3: BUCC results (F1) on the test set. We use the *ratio* function with *maximum score* retrieval and the filtering threshold optimized on the training set.

5. 기존 연구의 한계

농산물은 차, 쌀, 설탕, 담배, 장뇌, 과일, 그리고 비단으로 구성되어 있습니다.

(A)	*Les produits agricoles sont constitués de thé, de riz, de sucre, de tabac, de camphre, de fruits et de soie.*
0.818	Main crops include wheat, sugar beets, potatoes, cotton, tobacco, vegetables, and fruit.
0.817	The fertile soil supports wheat, corn, barley, tobacco, sugar beet, and soybeans.
0.814	Main agricultural products include grains, cotton, oil, pigs, poultry, fruits, vegetables, and edible fungus.
0.808	The important crops grown are cotton, jowar, groundnut, rice, sunflower and cereals.

하지만 현재의 상황에서는 위험 없이 그것들을 무시할 수 있을 것입니다.

(B)	*Mais dans le contexte actuel, nous pourrons les ignorer sans risque.*
0.737	But, in view of the current situation, we can safely ignore these.
0.499	But without the living language, it risks becoming an empty shell.
0.498	While the risk to those working in ceramics is now much reduced, it can still not be ignored.
0.488	But now they have discovered they are not free to speak their minds.

Table 1: Motivating example of the proposed method. We show the nearest neighbors of two French sentences on the BUCC training set along with their cosine similarities. Only the nearest neighbor of B is a correct translation, yet that of A has a higher cosine similarity. We argue that this is caused by the cosine similarity of different sentences being in different scales, making it a poor indicator of the confidence of the prediction. Our method tackles this issue by considering the margin between a given candidate and the rest of the k nearest neighbors.

6. xSIM++: An Improved Proxy to Bitext Mining Performance for Low-Resource Languages

Transformation Category	Original Sentence	Transformed Sentence
Causality Alternation	Apart from the fever and a sore throat, I feel well and in good shape to carry out my work by telecommuting.	Apart from the fever and a sore throat, I feel well and in bad shape to carry out my work by telecommuting
Entity Replacement	Charles was the first member of the British Royal Family to be awarded a degree.	M. Smith was the first member of The University to be awarded a degree.
Number Replacement	Nadal bagged 88% net points in the match winning 76 points in the first serve.	Nadal bagged 98% net points in the match winning 71 points in the sixth serve.

Table 1: Examples of the transformations applied to the English sentences from FLORES200 dev set. The red texts indicate the places of alternations.

- **Causality Alternation** To alter causality in a sentence
  (1) replace adjectives with their antonyms
  (2) negate the meaning of sentences by adding or removing negation function words to the sentences
  (3) leverage the negation strengthening approach which changes the causal relationships through more assertive function words

- **Entity Replacement** replace entities in sentences with the ones randomly sampled from the candidate set

- **Number Replacement** We use spaCy to detect dates, ordinals, cardinals, times, numbers, and percentages and then randomly replace their values.

7. Experiment

- **Data**
  - Source: CommonCrawl, ParaCrawl
  - approximately 3.7 billion sentences of English.
  - Others: 140K ~ 120M
    - Faroese (fao), Kabuverdianu (kea), Tok Pisin (tpi), Kikuyu (kik), Friulian (fur), Igbo (ibo), Luxembourgish (ltz), Swahili (swh), Zulu (zul), Bemba (bem).

- **Training Data**
  - LASER3 에 태워서 xSIM으로 pairing한 bitext
  - LaBSE 에 태워서 xSIM으로 pairing한 bitext

  - LASER3 에 태워서 xSIM++으로 pairing한 bitext
  - LaBSE 에 태워서 xSIM++으로 pairing한 bitext (개수모름;;)

  - 기존에 갖고있었던 mined bitext 데이터 (개수모름;;)

- **Test Data**
  - Flores-200

- **Model**
  - Plain Transformers
  - 72M

B Sizes of Monolingual data for Low-Resource Languages

Language	Size
kik	147,902
kea	226,507
fur	737,178
fao	1,179,475
tpi	1,661,743
bem	2,302,805
ibo	8,124,418
zul	20,477,331
swh	55,399,821
ltz	123,944,670

Table 6: Number of monolingual sentences for each language.

C Hyperparameters for NMT systems

encoder layers	6
encoder attention heads	8
encoder embed dim	512
encoder FFNN embed dim	4096
decoder layers	6
decoder attention heads	8
decoder embed dim	512
decoder FFNN embed dim	4096
optimiser	Adam
adam betas	(0.9, 0.98)
learning rate	0.001
dropout	0.3
spm vocab size	7000

Table 7: Hyperparameters for NMT systems.

7-1. pairwise ranking accuracy

$$\Delta = \text{score}(\text{System A}) - \text{score}(\text{System B})$$

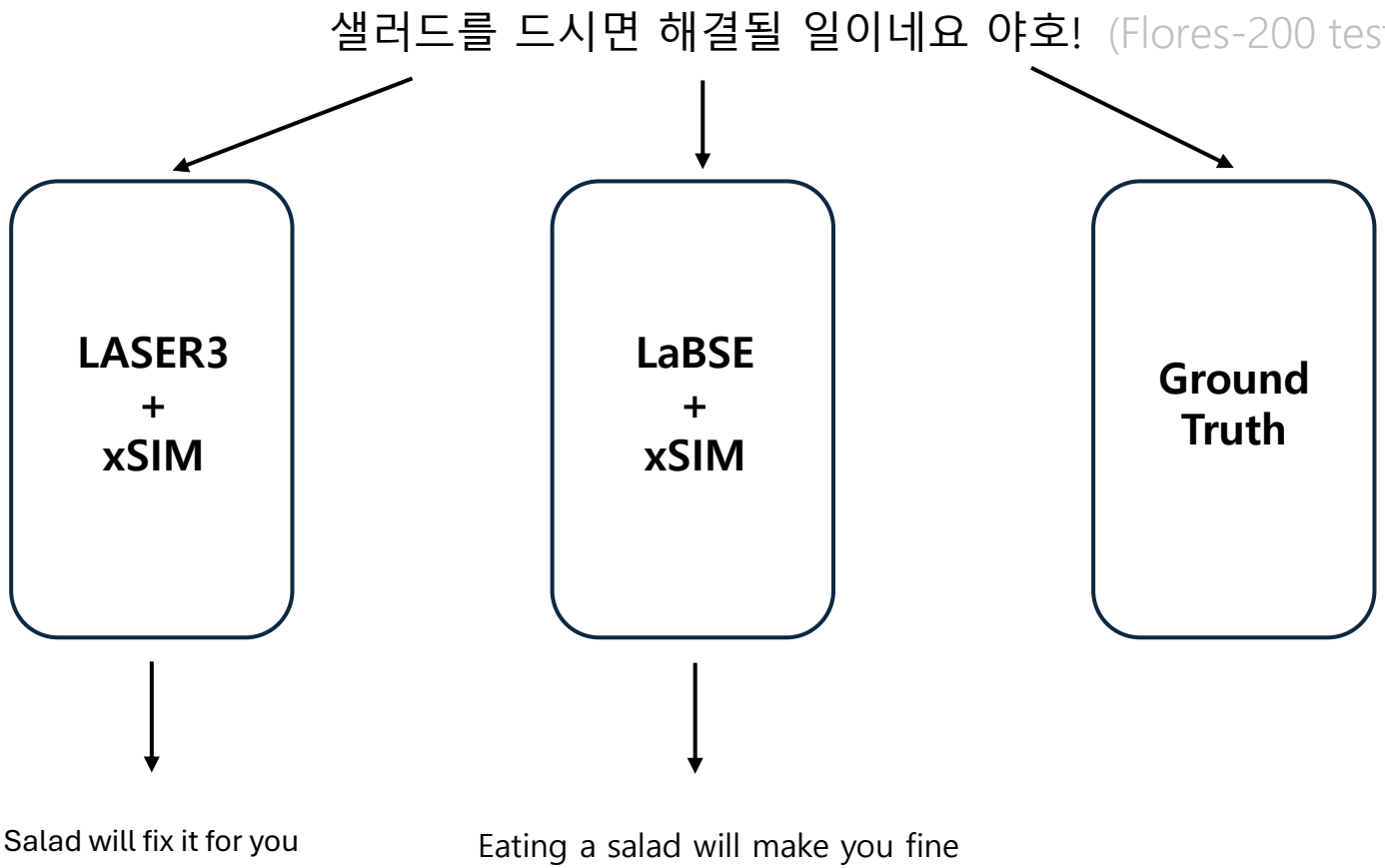
$$\text{Accuracy} = \frac{|\text{s}(\text{proxy}\Delta) = \text{s}(\text{mining}\Delta) \text{ for all system pairs}|}{|\text{all system pairs}|}$$

Pairwise ranking Accuracy

- where, proxyΔ is the difference of the xsim or xsim++ scores,
- miningΔ is the difference of the BLEU scores,
- s(·) is the sign function,
- and | · | returns the cardinal number of the input.

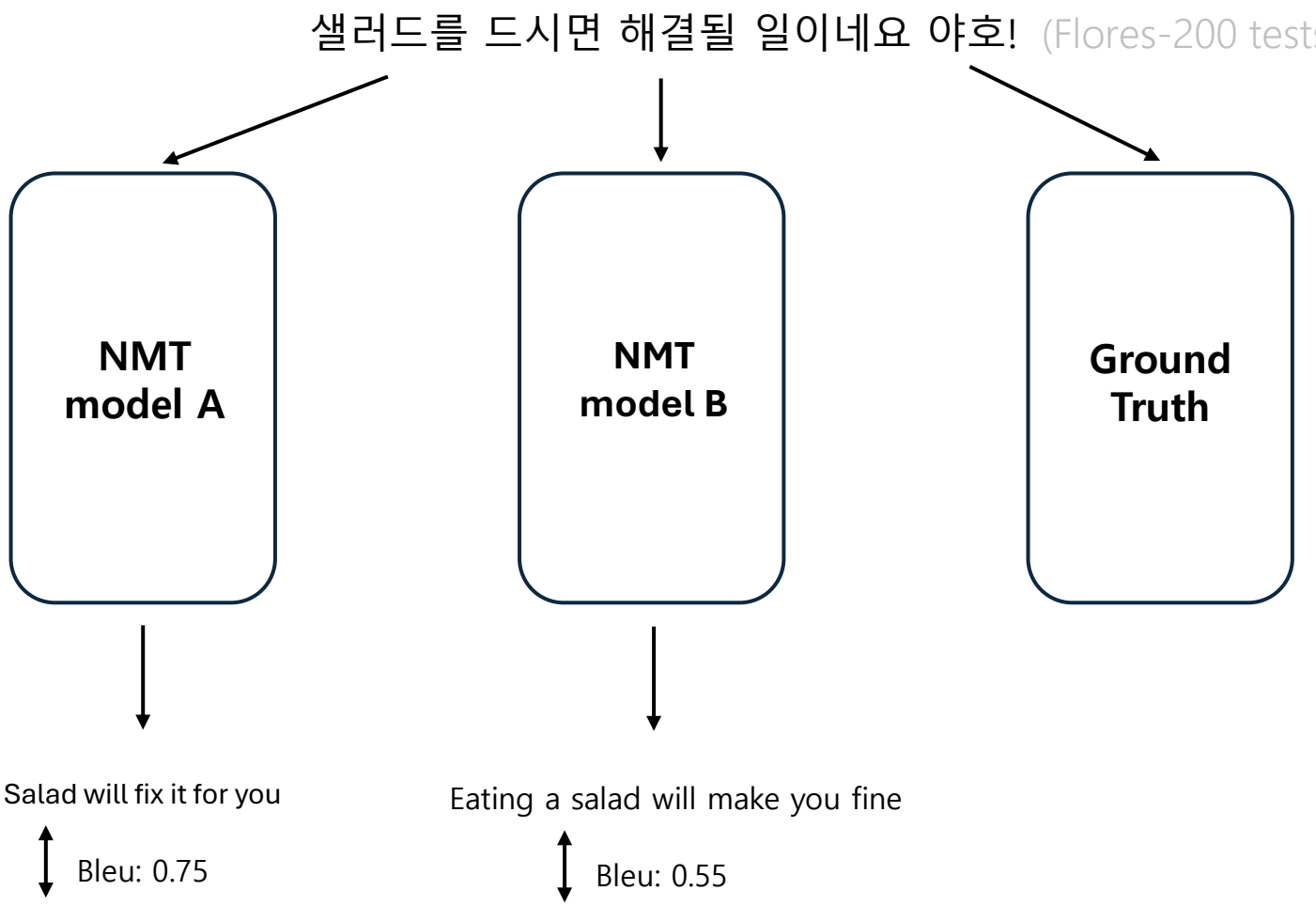
$$Sgn(x) = \begin{cases} -1, & if\ x < 0 \\ 0, & if, x = 0 \\ 1, & if, x > 0 \end{cases}$$

7-1. pairwise ranking accuracy



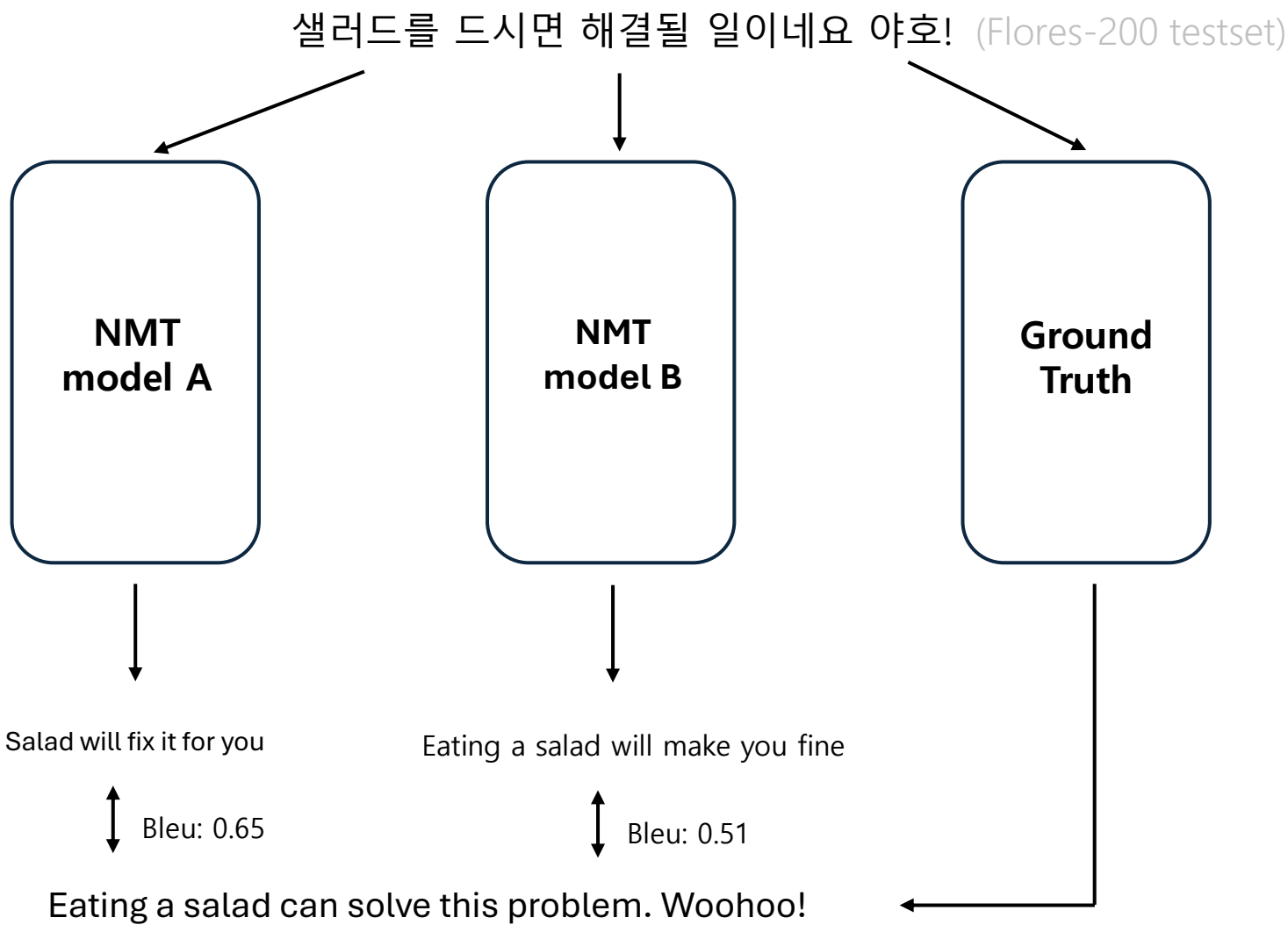
Having a salad will solve the problem. Yay! (Flores-200 testset)

7-1. pairwise ranking accuracy



	NMT model A	NMT model B
TestSet	O	
Mined Text model		

7-1. pairwise ranking accuracy



	NMT model A	NMT model B
TestSet	O	
Mined Text model	O	

8. Result

Metric	Accuracy	GPU hours
`xsim`	35.48	0.43
`xsim++`	72.00*	0.52
Mining BLEU (Oracle)	100	19569

Table 3: Pairwise ranking accuracy along with the total number of GPU hours. For all experiments, we used NVIDIA A100 GPUs. An $*$ indicates that the result passes the significance test proposed by Koehn (2004) with p-value < 0.05 when compared to `xsim`.

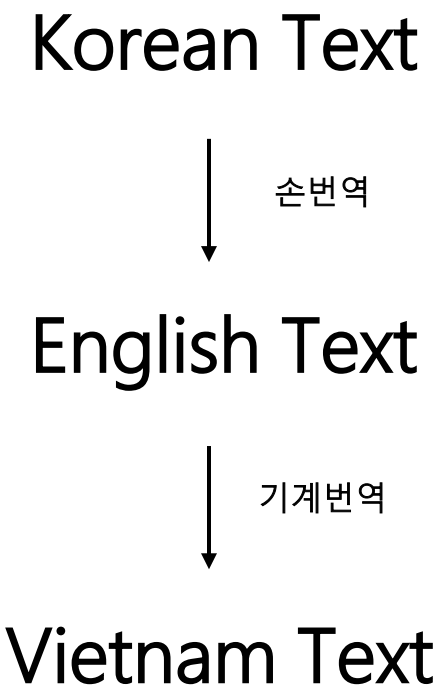
9. 한계

Transformation Category	Original Sentence	Transformed Sentence
Causality Alternation	Apart from the fever and a sore throat, I feel well and in good shape to carry out my work by telecommuting.	Apart from the fever and a sore throat, I feel well and in bad shape to carry out my work by telecommuting
Entity Replacement	Charles was the first member of the British Royal Family to be awarded a degree.	M. Smith was the first member of The University to be awarded a degree.
Number Replacement	Nadal bagged 88% net points in the match winning 76 points in the first serve.	Nadal bagged 98% net points in the match winning 71 points in the sixth serve.

Table 1: Examples of the transformations applied to the English sentences from FLORES200 dev set. The red texts indicate the places of alternations.

Entity Replacement. We collect candidate entities from large amounts of monolingual data. Then we replace entities in sentences with the ones randomly sampled from the candidate set. For both stages, we use the named entity recognizer from NLTK (Bird et al., 2009).

10. 활용 방안



10. 활용 방안

Korean Text

손번역

English Text

기계번역
+
Beam Search or Sampling

Vietnam Text 1
Vietnam Text 2
Vietnam Text 3
Vietnam Text 4