

ICT이노베이션스퀘어 AI복합교육 고급 언어과정

자연어처리를 위한 Text Similarity

현청천

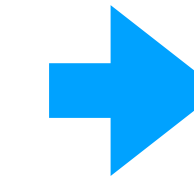
2021.04.19

What is Text Similarity

- 두 쌍의 text의 유사성을 예측하는 Task

한 남자가 음식을 먹고 있다.

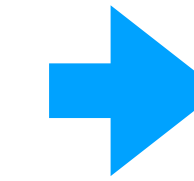
한 남자가 뭔가를 먹고 있다.



같은 의미

한 비행기가 착륙하고 있다.

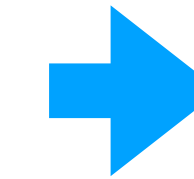
애니메이션화된 비행기 하나가 착륙하고 있다.



같은 의미

한 여성이 고기를 요리하고 있다.

한 남자가 말하고 있다.



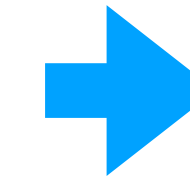
다른 의미

What is Text Similarity

- 두 쌍의 text의 유사성을 예측하는 Task

한 남자가 음식을 먹고 있다.

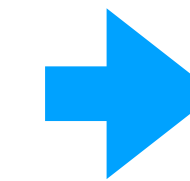
한 남자가 뭔가를 먹고 있다.



4.2

한 비행기가 착륙하고 있다.

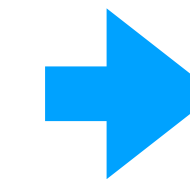
애니메이션화된 비행기 하나가 착륙하고 있다.



2.8

한 여성이 고기를 요리하고 있다.

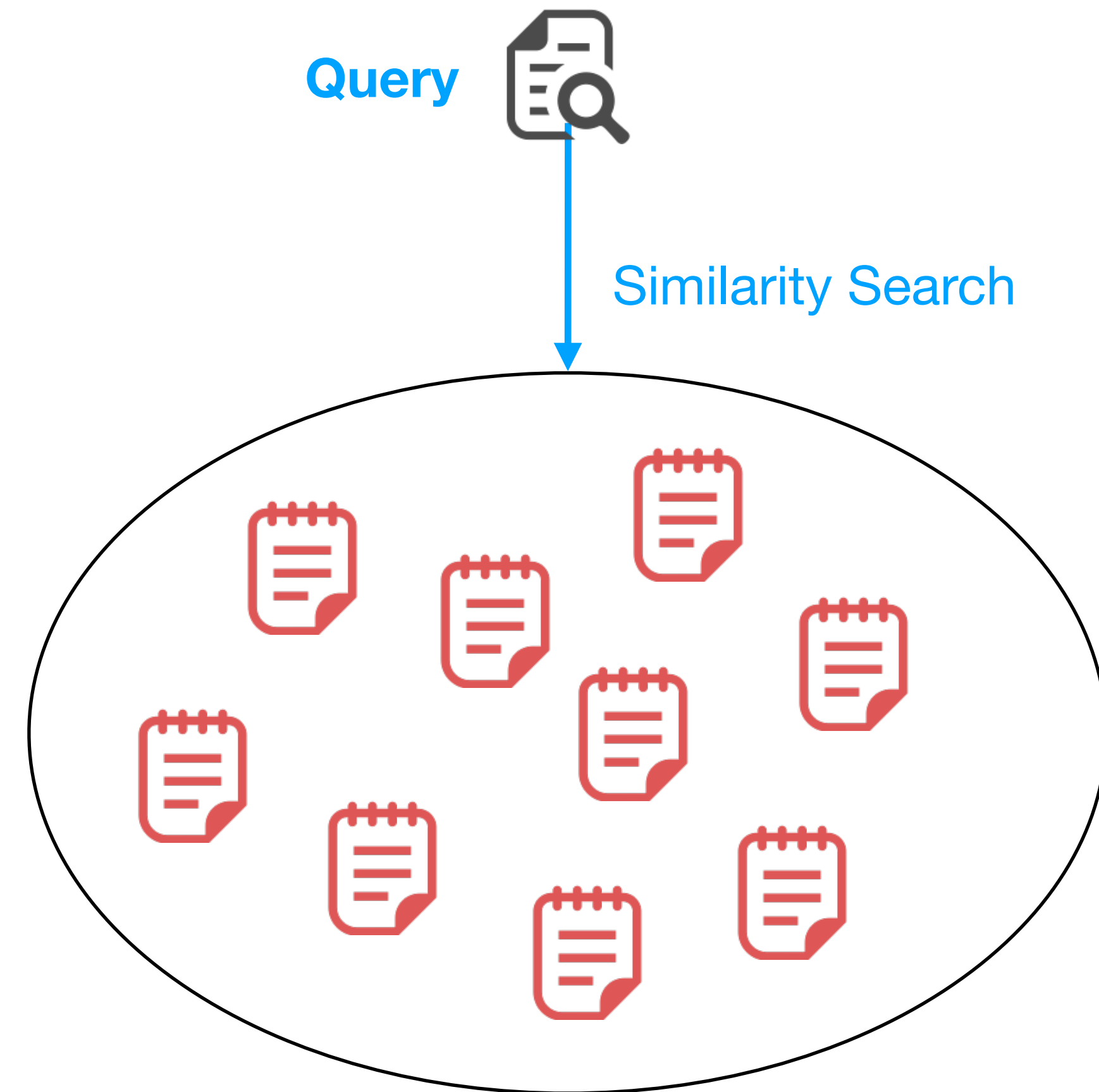
한 남자가 말하고 있다.



0.0

What is Text Similarity

- 두 쌍의 text의 유사성을 예측하는 Task
 - Information Retrieval (Search Engine)



What is Text Similarity

- 두 쌍의 text의 유사성을 예측하는 Task
 - Information Retrieval (Search Engine)
 - Data Matching

산업분류

특허

유(乳)제품; 분유 또는 분유제품;
보관; 초콜렛 우유; 아이스크림
또는 아이스크림을 조제하기 위
한 혼합물; 푸딩 또는 분말푸딩

말 및 양 사육업

각종 용도의 말, 양 및 염
소를 사육하는 산업활동
을 말한다.

젖소 사육업

우유를 생산하기 위하여
젖소를 사육하는 산업활
동을 말한다.

양돈업

각종 용도의 돼지, 멧돼지
를 번식 및 사육하는 산업
활동을 말한다.

내수면 어업

강, 호수, 하천 등 내수면
에서 어류 등의 각종 수
산 동.식물을 채취 또는
포획하는 산업활동을 말
한다.

해수면 양식 어업

해수면 또는 육상에서 해
수를 이용하여 각종 수산
동.식물을 증식 또는 양식
하는 산업활동을 말한다.

철 광업

철 성분을 주로 함유하고
있는 각종 철광석을 채굴
하는 산업활동으로서 철
광석의 정광 및 기타 부수
적인 처리활동과 철광석
의 응집처리 활동도 포함
한다.

Text Similarity Dataset

- Text Similarity
 - SentEval

Task	Type	#train	#test	needs_train	set_classifier
MR	movie review	11k	11k	1	1
CR	product review	4k	4k	1	1
SUBJ	subjectivity status	10k	10k	1	1
MPQA	opinion-polarity	11k	11k	1	1
SST	binary sentiment analysis	67k	1.8k	1	1
SST	fine-grained sentiment analysis	8.5k	2.2k	1	1
TREC	question-type classification	6k	0.5k	1	1
SICK-E	natural language inference	4.5k	4.9k	1	1
SNLI	natural language inference	550k	9.8k	1	1
MRPC	paraphrase detection	4.1k	1.7k	1	1
STS 2012	semantic textual similarity	N/A	3.1k	0	0
STS 2013	semantic textual similarity	N/A	1.5k	0	0
STS 2014	semantic textual similarity	N/A	3.7k	0	0
STS 2015	semantic textual similarity	N/A	8.5k	0	0
STS 2016	semantic textual similarity	N/A	9.2k	0	0
STS B	semantic textual similarity	5.7k	1.4k	1	0
SICK-R	semantic textual similarity	4.5k	4.9k	1	0
COCO	image-caption retrieval	567k	5*1k	1	0

- SentEval은 Sentence Embedding 품질을 평가하기위한 라이브러리
- 18개 중 STS12 ~ STS16, STSB, SICK-R
- 두 문장 간에 유사도 측정
- <https://github.com/facebookresearch/SentEval>

Text Similarity Dataset

- Text Similarity

- SentEval
- Quora Question Pairs
- Paired Question
- KorSTS

- 질문 사이트인 Quora의 질문데이터
- 두 질문이 중복되는지 여부를 판단
- <https://www.kaggle.com/c/quora-question-pairs>

Text Similarity Dataset

- Text Similarity

- SentEval
- Quora Question Pairs
- Paired Question
- KorSTS

- Quora Question Pairs의 한국어 버전
- https://github.com/songys/Question_pair

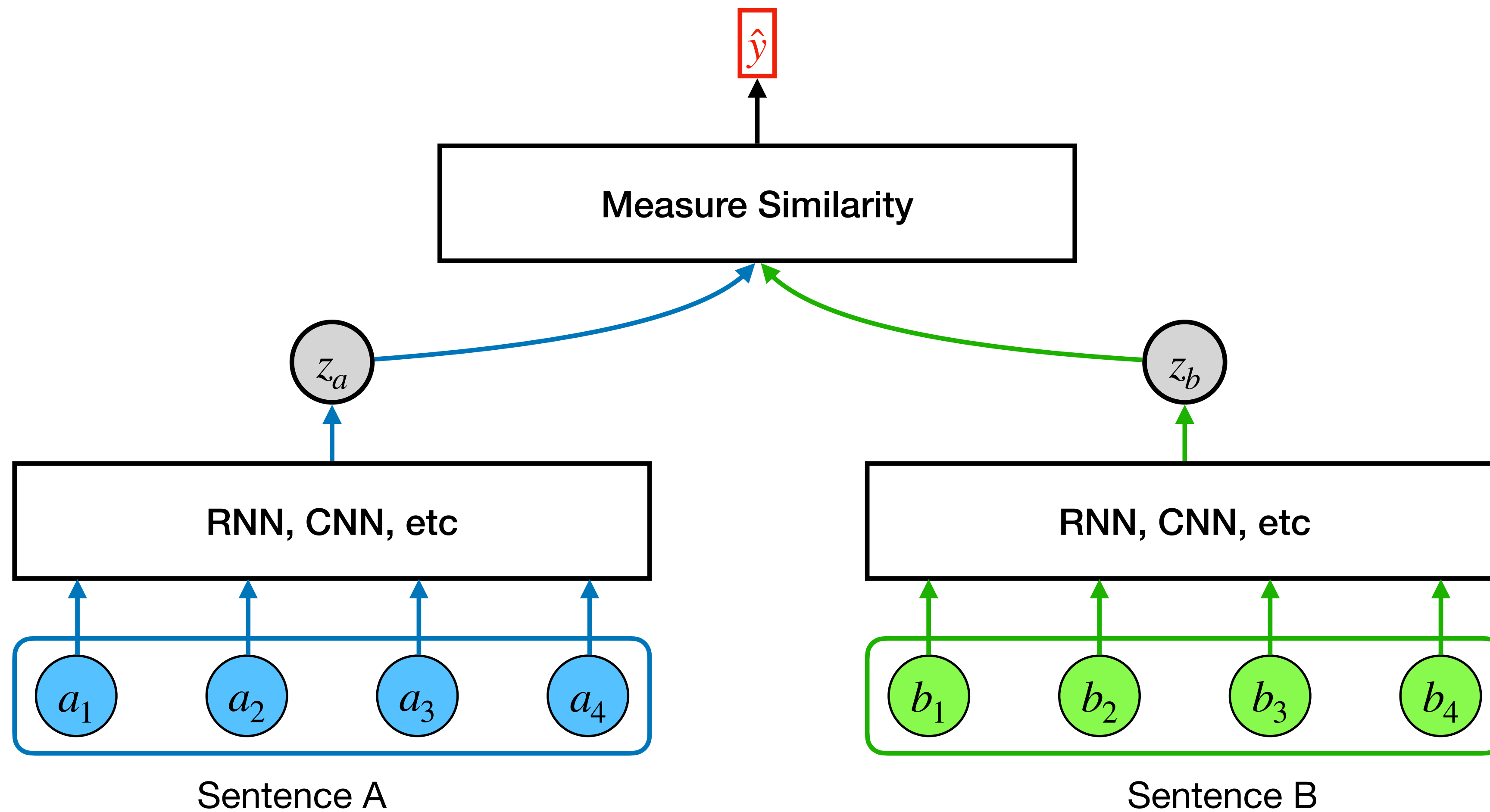
Text Similarity Dataset

- Text Similarity

- SentEval
- Quora Question Pairs
- Paired Question
- KorSTS

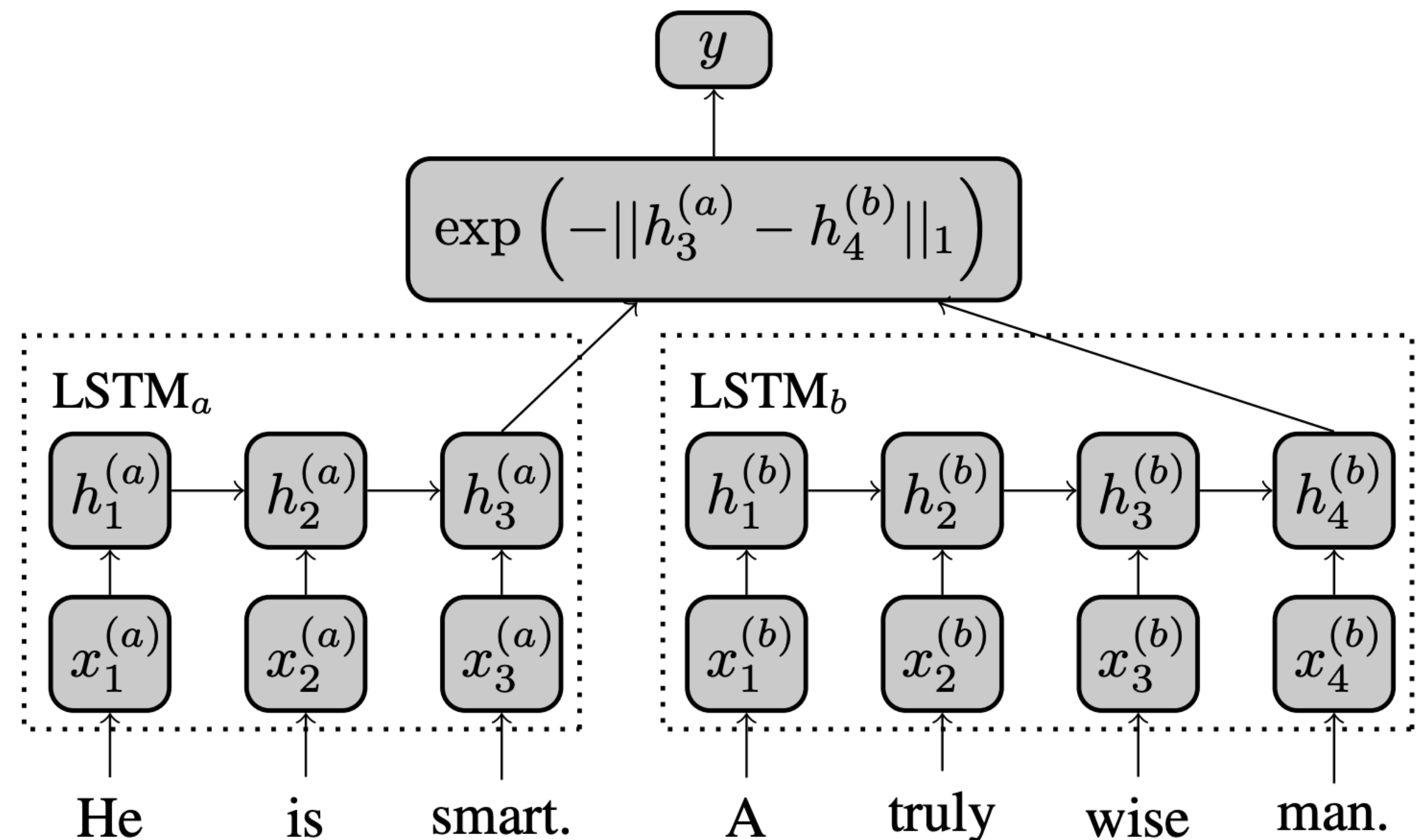
- STS B의 한국어 버전
- <https://github.com/kakaobrain/KorNLUDatasets>

Text Similarity Model (Type 1)

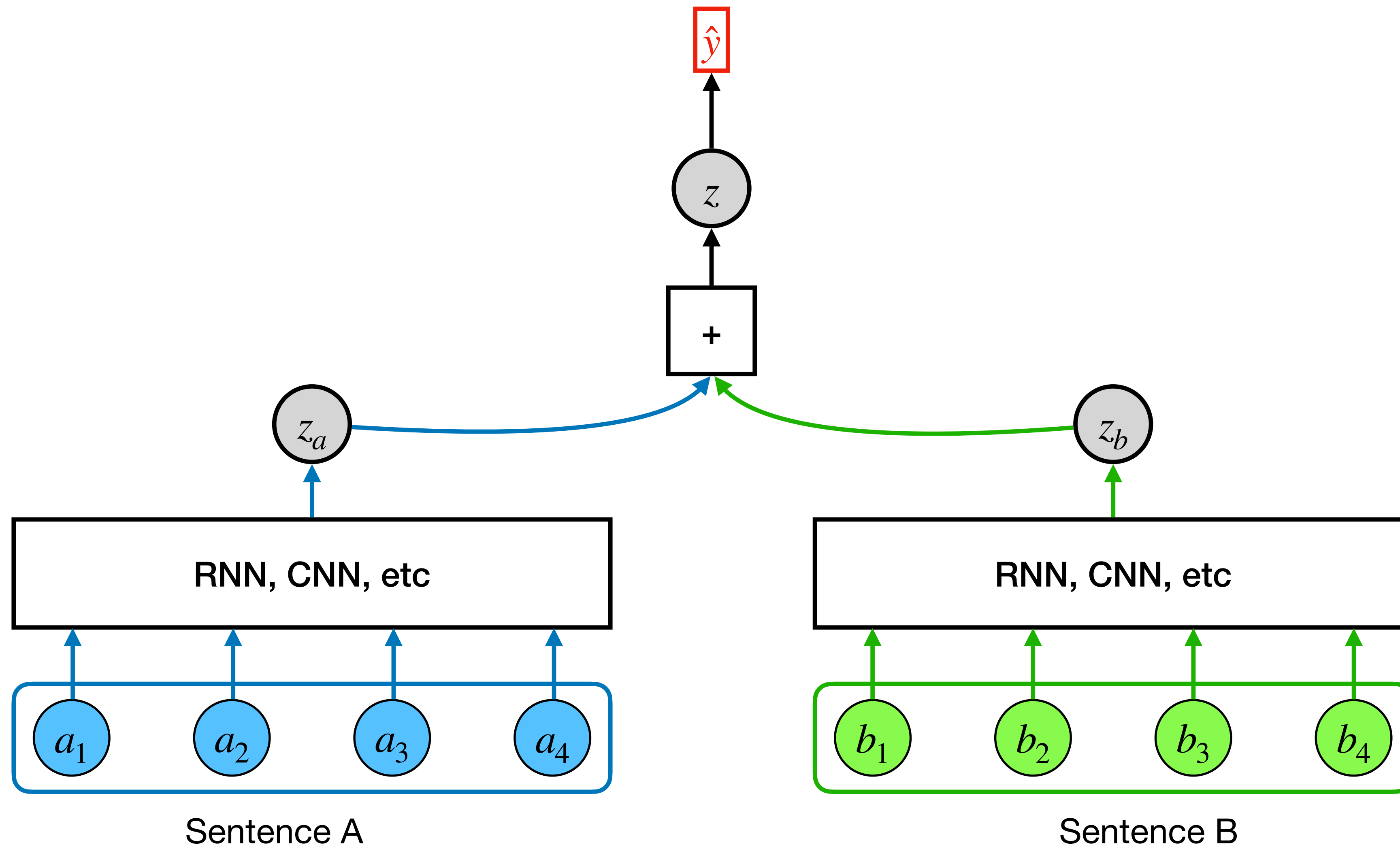


Text Similarity Model (Type 1)

- Siamese Recurrent Architectures for Learning Sentence Similarity
- Jonas Mueller et al. 2016
- 맨하탄 거리(Manhattan distance)를 사용하여 유사도 비교
- 모델이 단순하면서도 실용적임

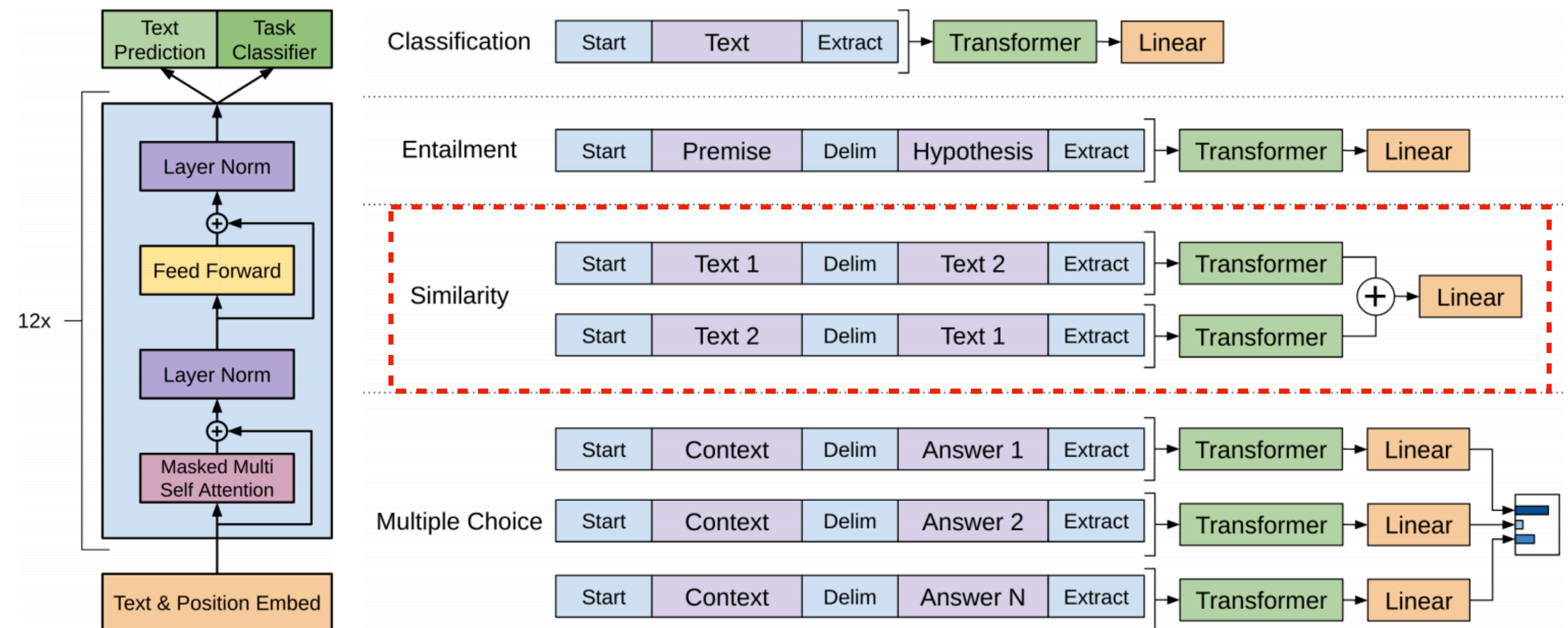


Text Similarity Model (Type 2)

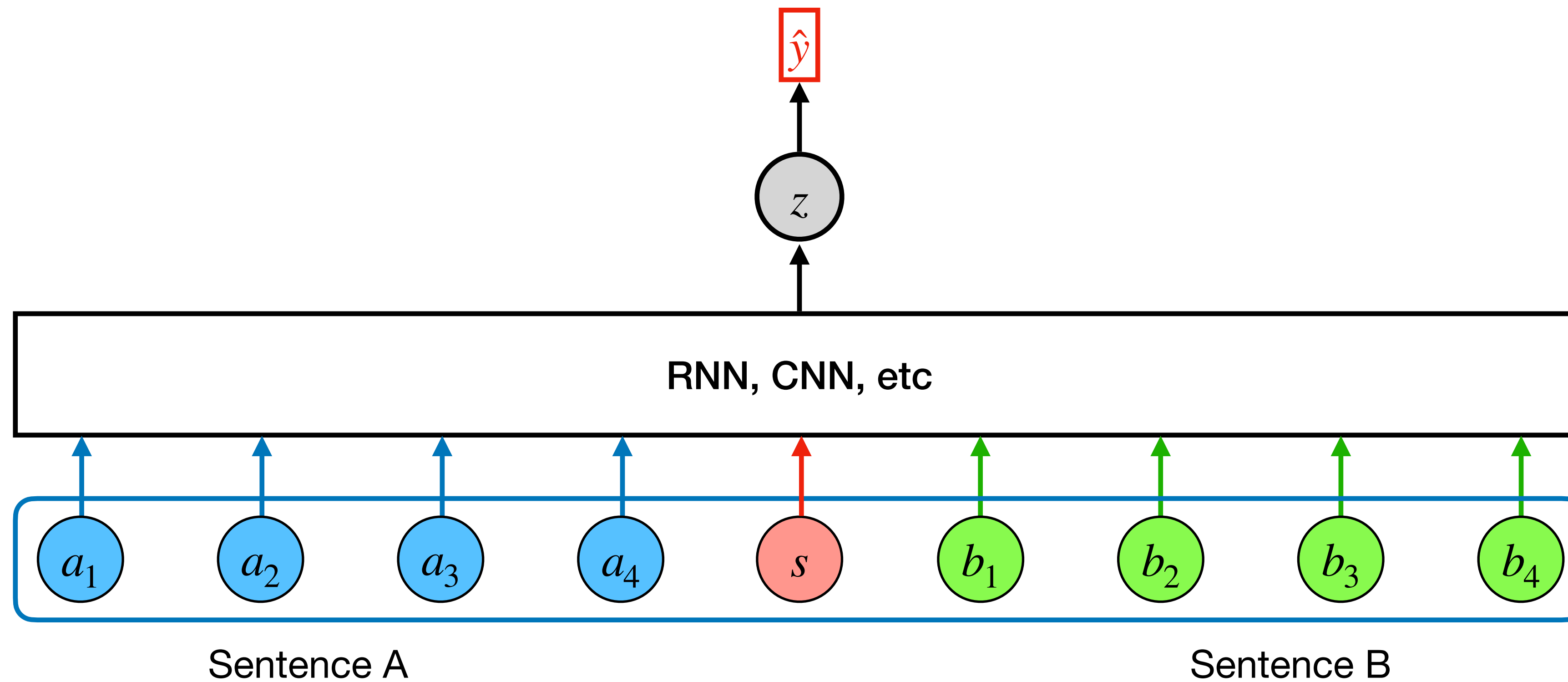


Text Similarity Model (Type2)

- Improving Language Understanding by Generative Pre-Training
- Alec Radford et al. 2018
- 하나의 모델로 다양한 Task에서 사용 가능
- Transformer 기반의 단 방향 Pretrained Language Model



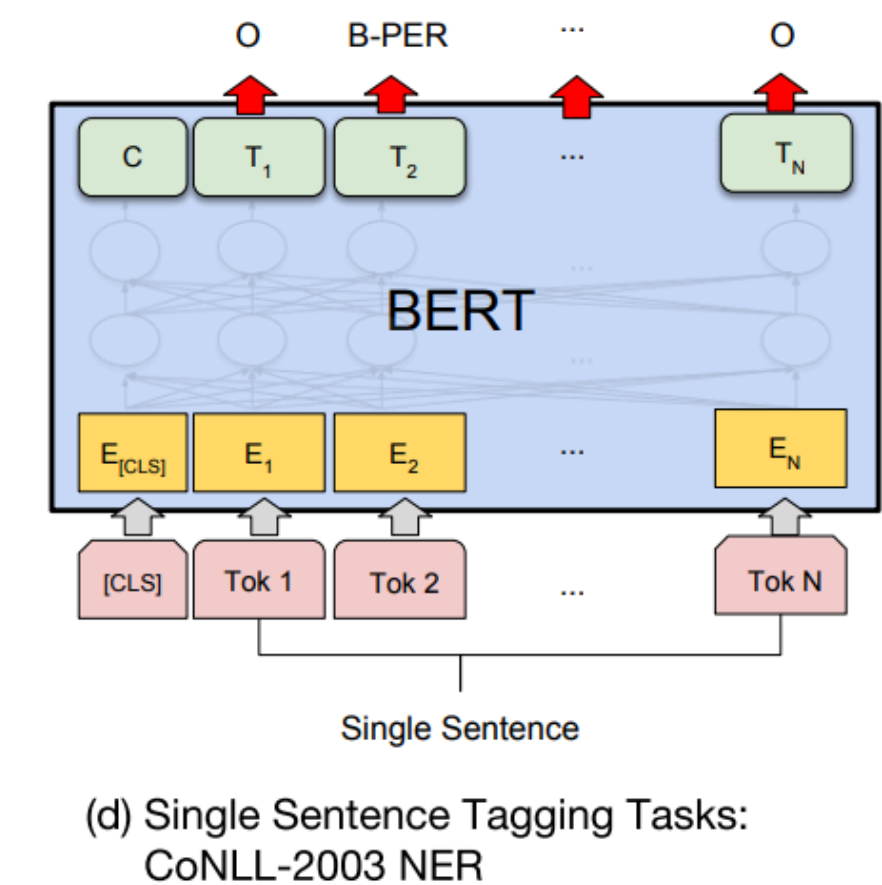
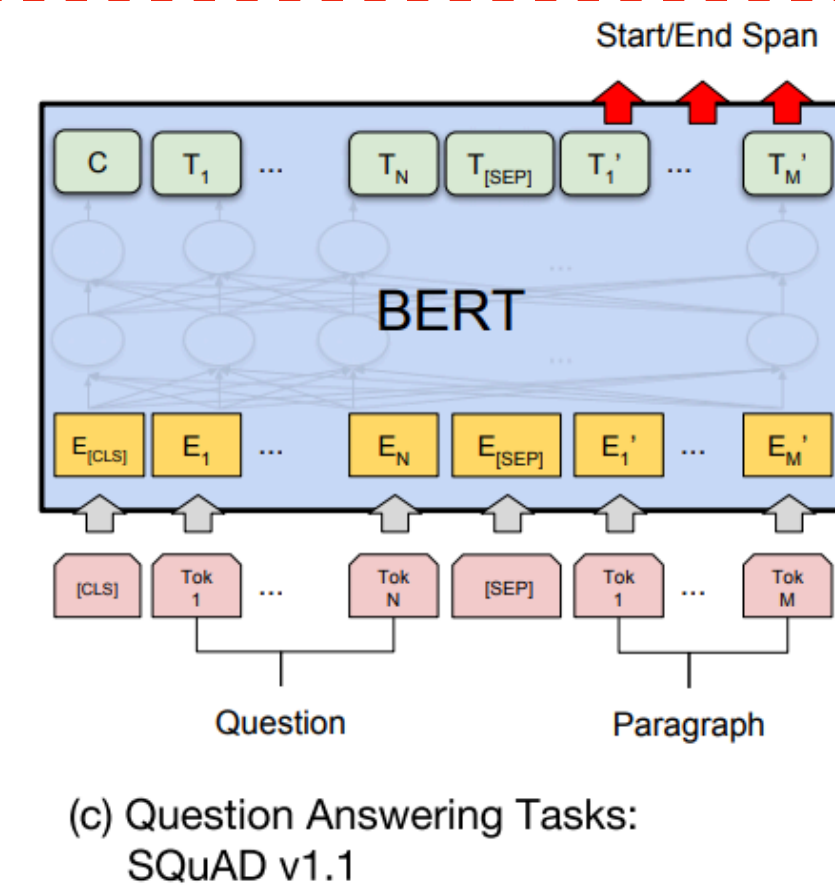
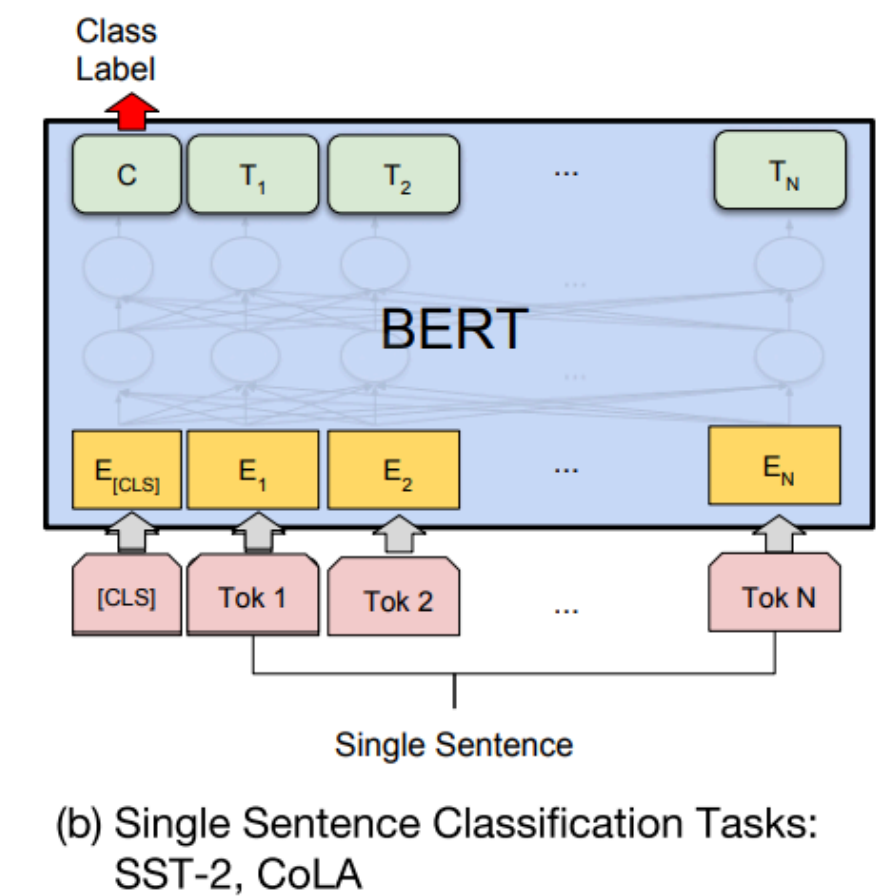
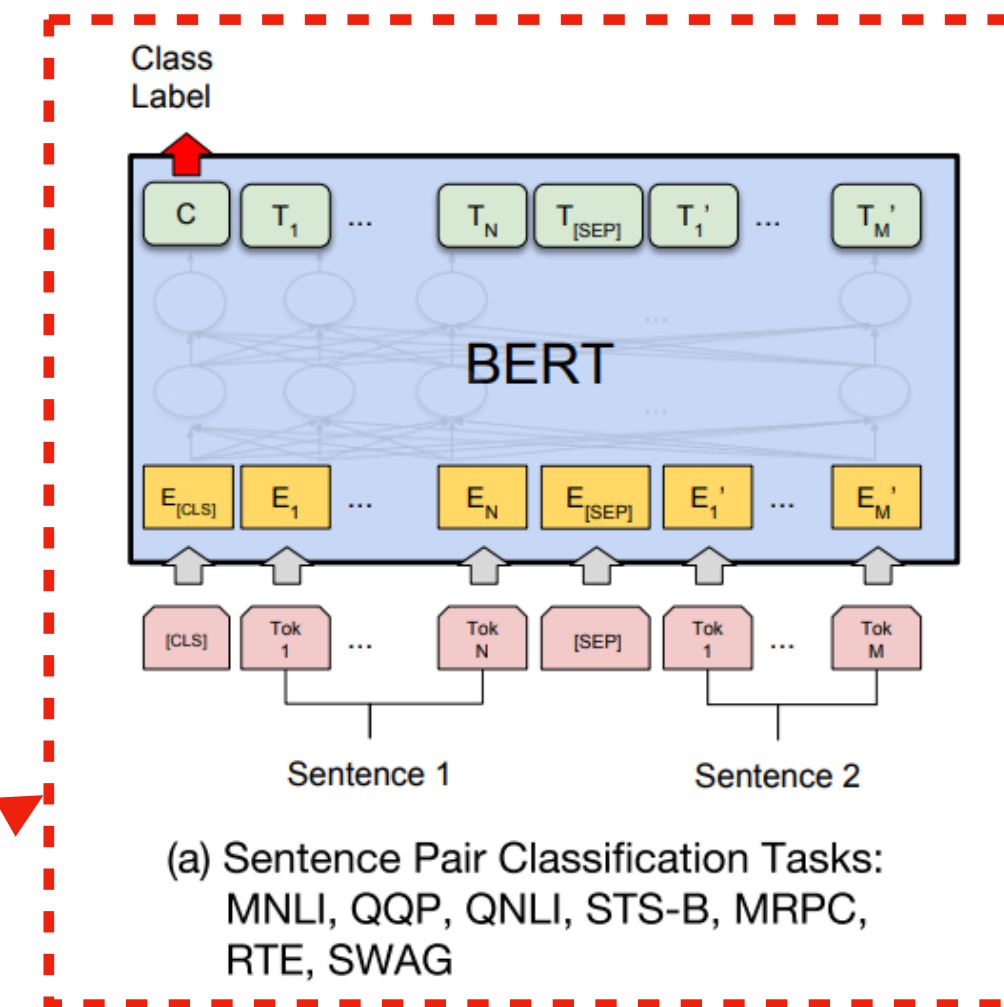
Text Similarity Model (Type 3)



Text Similarity Model (Type3)

- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
- Jacob Devlin et al. 2018
- 현재 자연어처리에서 가장 중요한 모델
- Transformer 기반의 양방향 Pretrained Language Model

[CLS] SentenceA [SEP] SentenceB [SEP]



Text Similarity Model (SOTA)

State fo Art
Text Similarity

<https://paperswithcode.com/task/semantic-textual-similarity>

Similarity Metric (Jaccard Similarity)

- 두 집합의 유사도를 측정하는 방법

$$J(A, B) = \frac{A \cap B}{A \cup B}$$

$$A = \{\text{머, 신, 러, 닝}\}$$

$$B = \{\text{딕, 러, 닝}\}$$

$$A \cup B = \{\text{머, 신, 러, 닝, 딕}\}$$

$$A \cap B = \{\text{러, 닝}\}$$

$$\frac{A \cap B}{A \cup B} = \frac{2}{5} = 0.4$$

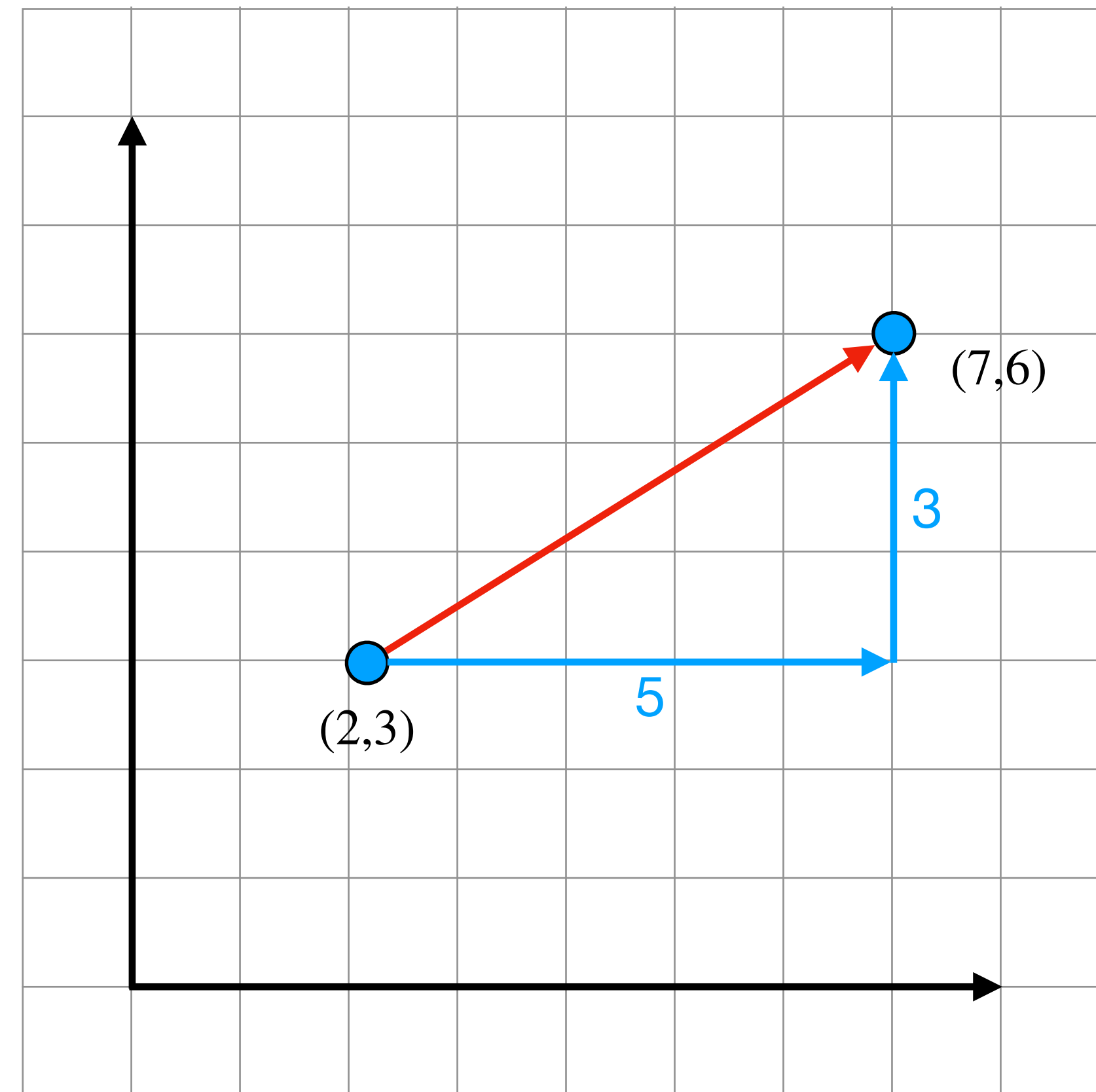
Similarity Metric (Distance)

- distance의 특징
 - $d(a, b) \geq 0$
 - $d(a, b) = 0 \Leftrightarrow a = b$
 - $d(a, b) = d(b, a)$
 - $d(a, c) \leq d(a, b) + d(b, c)$

Similarity Metric (Euclidean Distance)

유클리드 공간에서의 직선거리

$$\begin{aligned} d(a, b) &= \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \dots + (a_n - b_n)^2} \\ &= \sqrt{\sum_{i=1}^n (a_i - b_i)^2} \end{aligned}$$

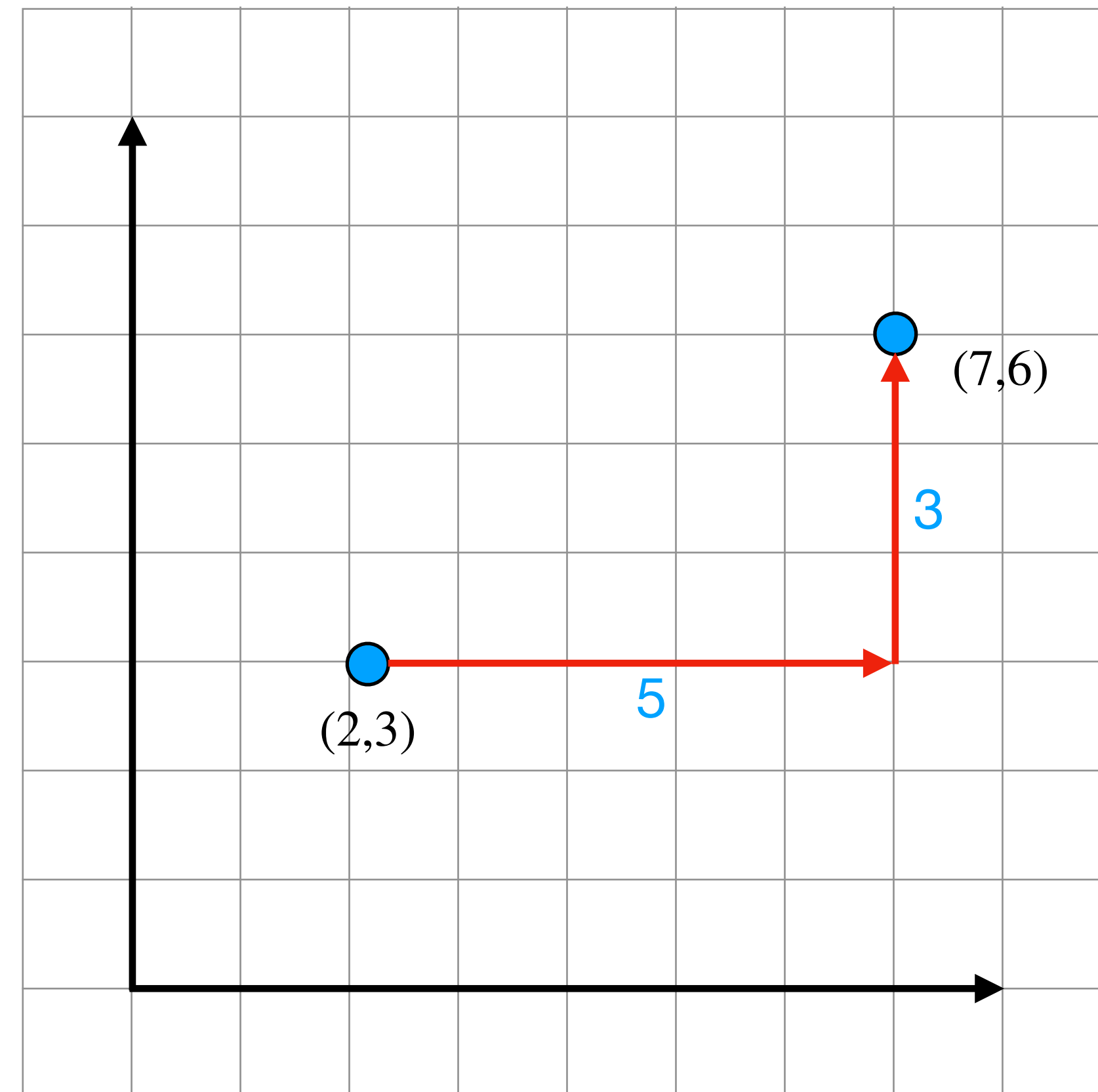


$$\sqrt{5^2 + 3^2} = \sqrt{34}$$

Similarity Metric (Manhattan Distance)

두 점의 좌표의 거리의 차이

$$d(a, b) = |a_1 - b_1| + |a_2 - b_2| + \dots + |a_n - b_n|$$
$$= \sqrt{\sum_{i=1}^n |a_i - b_i|}$$



$$5 + 3 = 8$$

What is Natural Language Inference

- 두 문장(premise, hypothesis)이 의미적으로 수반(entail)하는지 예측하는 Task

Entailment

같은 의미

Neutral

관계 없음

Contradiction

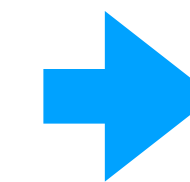
반대 의미

What is Natural Language Inference

- 두 문장(premise, hypothesis)이 의미적으로 수반(entail)하는지 예측하는 Task

Premise: 저는, 그냥 알아내려고 거기 있었어요.

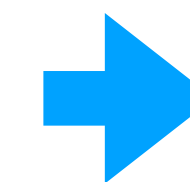
Hypothesis: 이해하려고 노력하고 있었어요.



Entailment

Premise: 저는, 그냥 알아내려고 거기 있었어요.

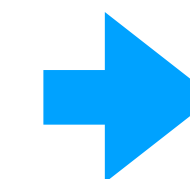
Hypothesis: 나는 돈이 어디로 갔는지 이해하려고 했어요.



Neutral

Premise: 저는, 그냥 알아내려고 거기 있었어요.

Hypothesis: 나는 처음부터 그것을 잘 이해했다.



Contradiction

Natural Language Inference Dataset

- Natural Language Inference
 - SNLI
 - MNLI
 - KorNLI

- Stanford에서 공개한 데이터
- Image caption data
- 약 50만개의 큰 데이터셋
- <https://nlp.stanford.edu/projects/snli/>

Natural Language Inference Dataset

- Natural Language Inference
 - SNLI
 - MNLI
 - KorNLI

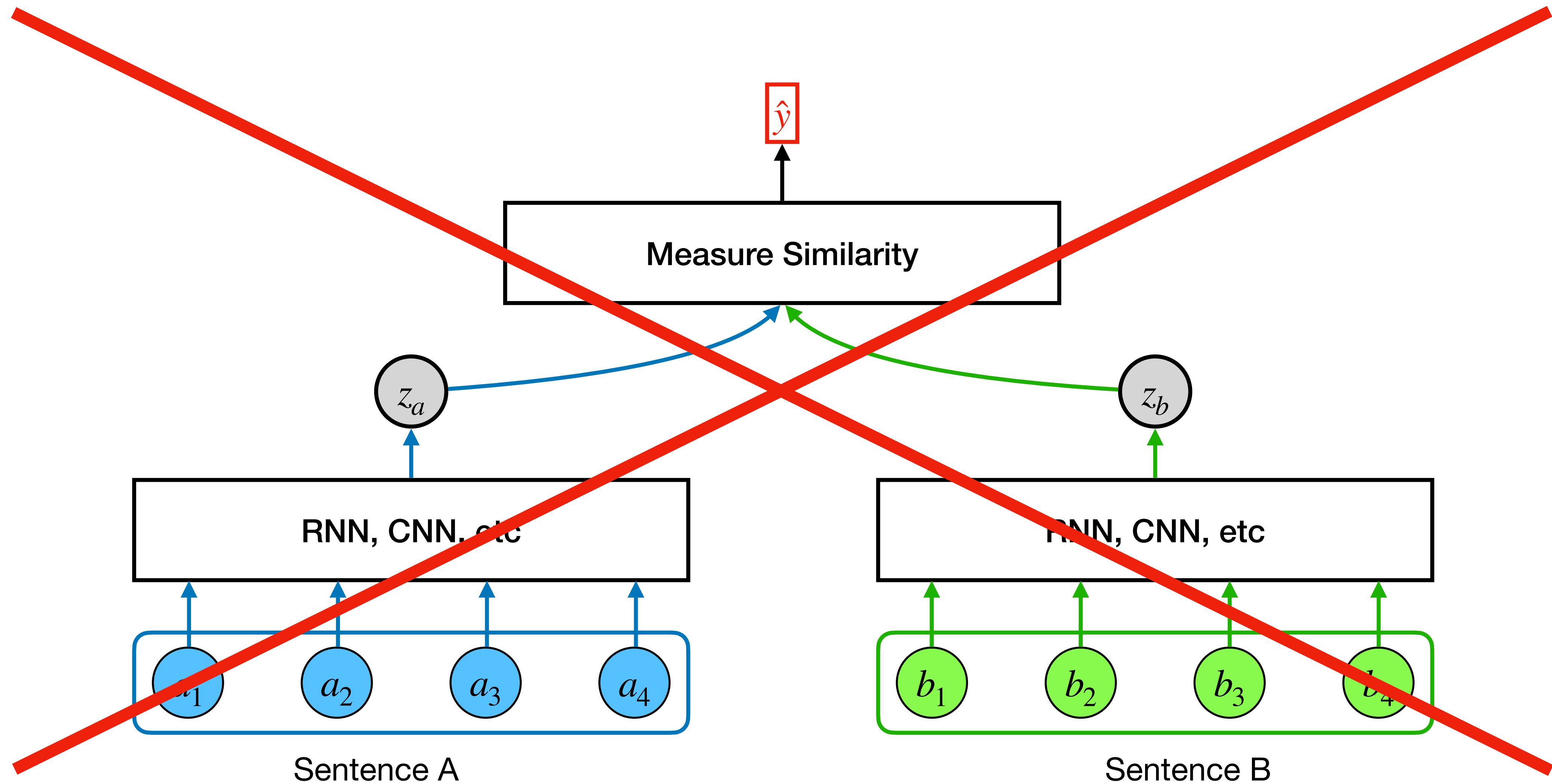
- SNLI의 단점을 보완
- 다양한 도메인 data
- <https://cims.nyu.edu/~sbowman/multinli/>

Natural Language Inference Dataset

- Natural Language Inference
 - SNLI
 - MNLI
 - KorNLI

- SNLI, SNLI의 한국어 버전
- 학습 데이터는 기계번역 평가 데이터를 사람이 번역
- <https://github.com/kakaobrain/KorNLUDatasets>


Natural Language Inference Model (Type 1)

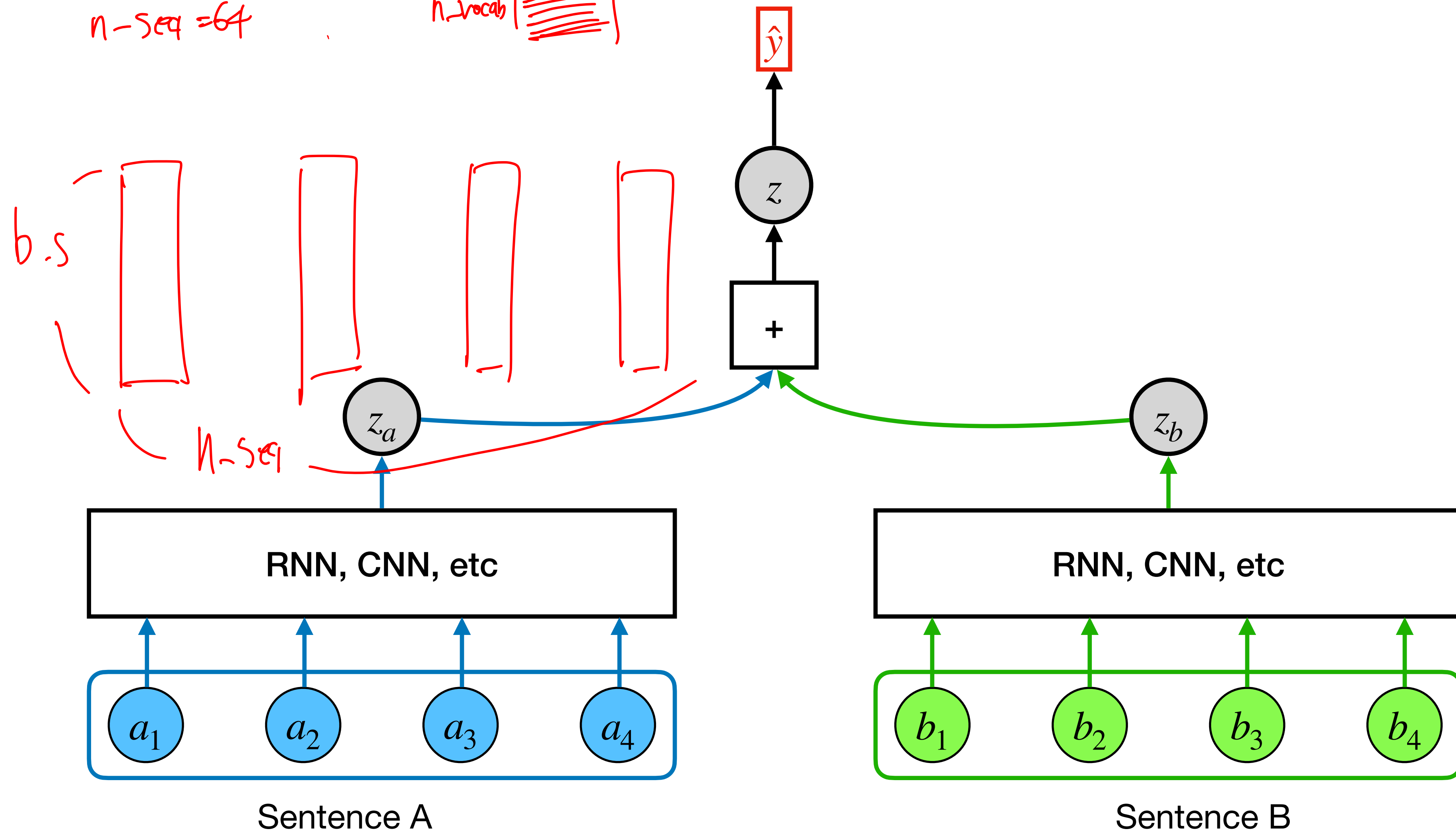


단어 ID, $[0, 1, 10, 20, 7] \rightarrow$ embedding
 을 통해 vector.

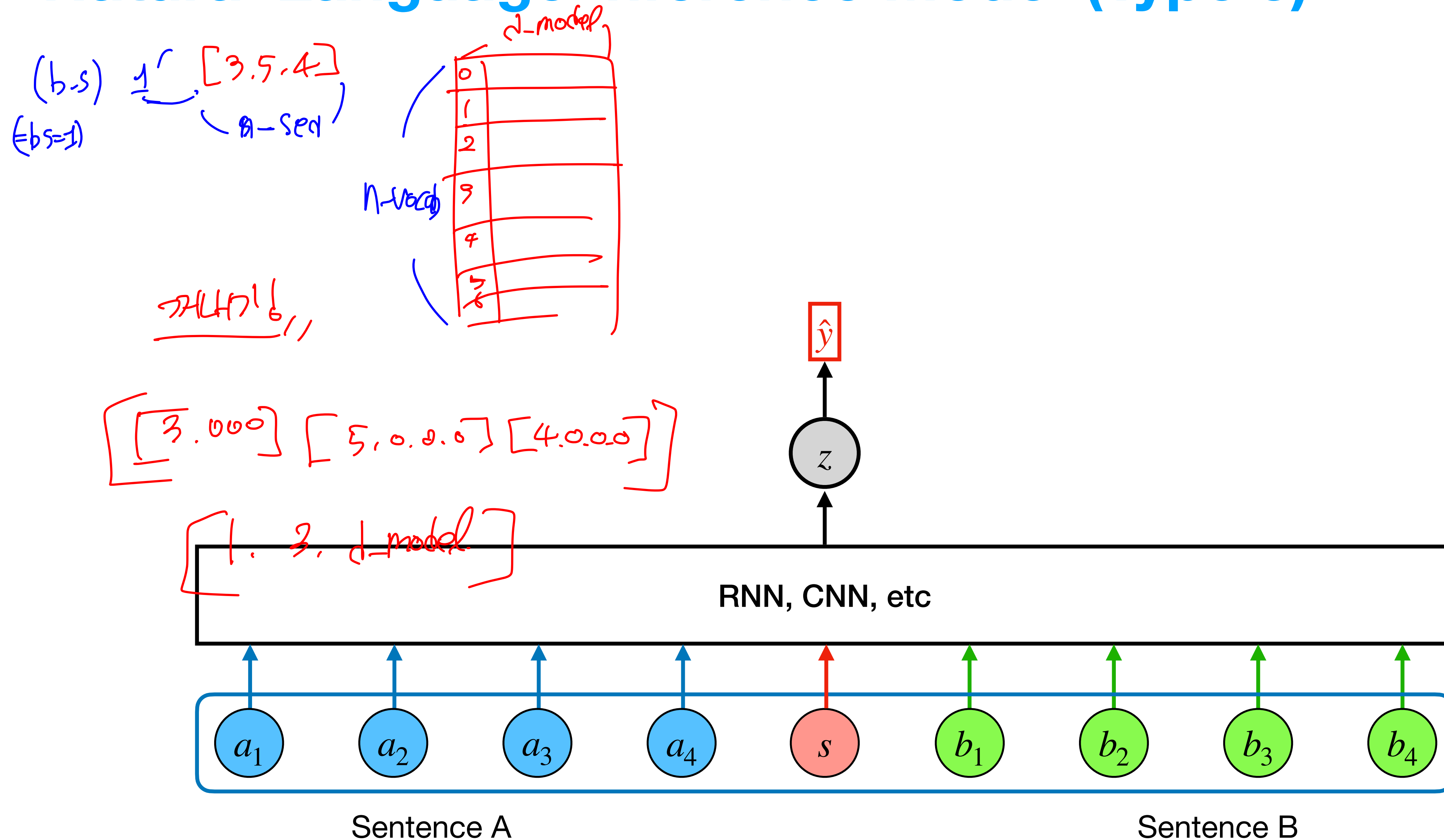
Natural Language Inference Model (Type 2)

$b.s = 10$
 $n.seq = 64$

$n.vocab$ 



Natural Language Inference Model (Type 3)



Natural Language Inference Model (SOTA)

State fo Art
Natural Language Inference

<https://paperswithcode.com/task/natural-language-inference>

감사합니다.