

# 영어 학습을 위한 난이도 별 문제 생성 모델

2017010499 채선율  
2017021203 성유연  
2018020517 이창현

## 1. 목표

## 2. 문제 생성 개요

## 3. 토익 문제 경향 확인

## 4. 난이도 별 토익 문제 생성

- Word2Vec
- Doc2Vec
- Bi-directional LSTM

## 5. 결론 및 한계점

## 영어 학습을 위한 난이도 별 문제 생성

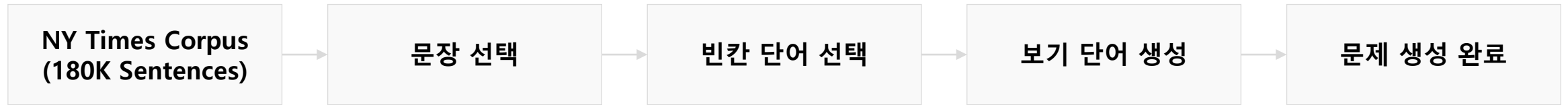
- 학습자(사용자)가 원하는 난이도에 따라 문제의 보기를 생성

### 난이도 : 상

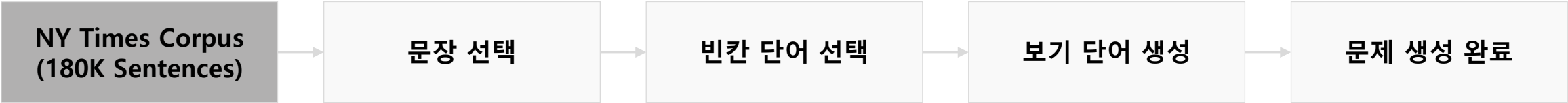
1. The indoor garden on the top floor of the new office building will be decorated with plants that need \_\_\_\_\_ care.

- (A) little
- (B) hardly
- (C) very
- (D) few

## 난이도별 영어 학습 문제 생성 flowchart



난이도별 영어 학습 문제 생성 flowchart



- 2011.01 – 2018.04 구간, 기사 5701건, 문장 184,032개
- 사용 method: beautifulsoup

*sections*

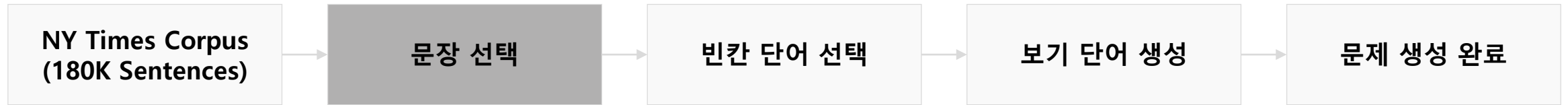
*sub\_url*

*title*

*sentence*

	sentence	sections	title	sub_url
1	Penny Vincenzi a British novelist whose sprawling	obituaries	Penny Vincenzi	<a href="https://www.ny">https://www.ny</a>
2	Her death was announced on her website	obituaries	Penny Vincenzi	<a href="https://www.ny">https://www.ny</a>
	The cause and location were not given Ms Vinc	obituaries	Penny Vincenzi	<a href="https://www.ny">https://www.ny</a>
	Vincenzi was a journalist when in 1989 her f	obituaries	Penny Vincenzi	<a href="https://www.ny">https://www.ny</a>
	more novels followed most recently A Q	obituaries	Penny Vincenzi	<a href="https://www.ny">https://www.ny</a>
	copies of her books are in print	obituaries	Penny Vincenzi	<a href="https://www.ny">https://www.ny</a>
	Most of her novels have lots of characters and l	obituaries	Penny Vincenzi	<a href="https://www.ny">https://www.ny</a>
8	Vincenzi writes long elaborate mannerly books	obituaries	Penny Vincenzi	<a href="https://www.ny">https://www.ny</a>
9	With strong female characters the stories have t	obituaries	Penny Vincenzi	<a href="https://www.ny">https://www.ny</a>
10	I know I must be a disappointment to people sh	obituaries	Penny Vincenzi	<a href="https://www.ny">https://www.ny</a>
11	The rest of my life doesnt match up with the so	obituaries	Penny Vincenzi	<a href="https://www.ny">https://www.ny</a>
12	Penelope Hannaford was born on April 10 1939	obituaries	Penny Vincenzi	<a href="https://www.ny">https://www.ny</a>
13	She was the only child of Stanley Hannaford a	obituaries	Penny Vincenzi	<a href="https://www.ny">https://www.ny</a>
14	She grew up in Devon and took to writing early	obituaries	Penny Vincenzi	<a href="https://www.ny">https://www.ny</a>
15	She was working as a secretary at Vogue when	obituaries	Penny Vincenzi	<a href="https://www.ny">https://www.ny</a>
16	Ms Vincenzi was a journalist for much of her w	obituaries	Penny Vincenzi	<a href="https://www.ny">https://www.ny</a>
17	It was the money that made me first have a crac	obituaries	Penny Vincenzi	<a href="https://www.ny">https://www.ny</a>
18	It was the late 1980s and quite a few women jo	obituaries	Penny Vincenzi	<a href="https://www.ny">https://www.ny</a>
19	So I decided to have a go myself	obituaries	Penny Vincenzi	<a href="https://www.ny">https://www.ny</a>
20	She asked Ms Cooper whom she was interviewi	obituaries	Penny Vincenzi	<a href="https://www.ny">https://www.ny</a>
21	I mentioned to Jilly that I had a story in mind	obituaries	Penny Vincenzi	<a href="https://www.ny">https://www.ny</a>
22	When I got back to the office she had got him	obituaries	Penny Vincenzi	<a href="https://www.ny">https://www.ny</a>
23	I mean you dont share your agent its like shari	obituaries	Penny Vincenzi	<a href="https://www.ny">https://www.ny</a>
24	Among her most successful novels were An Abs	obituaries	Penny Vincenzi	<a href="https://www.ny">https://www.ny</a>
25	She also wrote two books of short stories	obituaries	Penny Vincenzi	<a href="https://www.ny">https://www.ny</a>
26	Ms Vincenzi said that she never knew how her	obituaries	Penny Vincenzi	<a href="https://www.ny">https://www.ny</a>

## 난이도별 영어 학습 문제 생성 flowchart

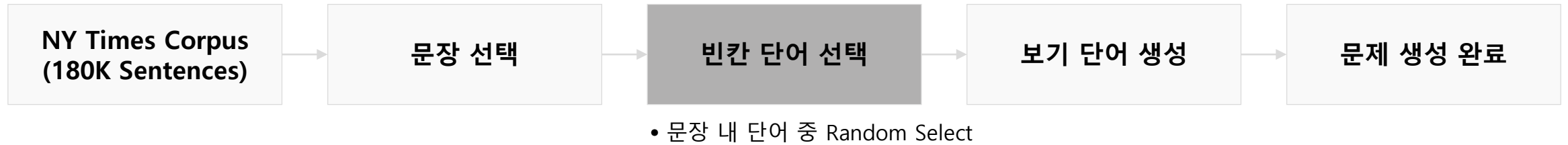


- Corpus 내 문장 중 Random Select

1. The indoor garden on the top floor of the new office building will be decorated with plants that need little care.

- (A)
- (B)
- (C)
- (D)

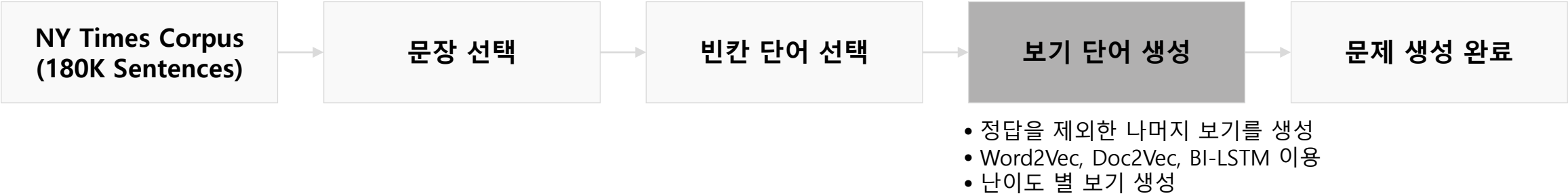
## 난이도별 영어 학습 문제 생성 flowchart



1. The indoor garden on the top floor of the new office building will be decorated with plants that need \_\_\_\_\_ care.

- (A) little
- (B)
- (C)
- (D)

난이도별 영어 학습 문제 생성 flowchart



난이도 : 상?

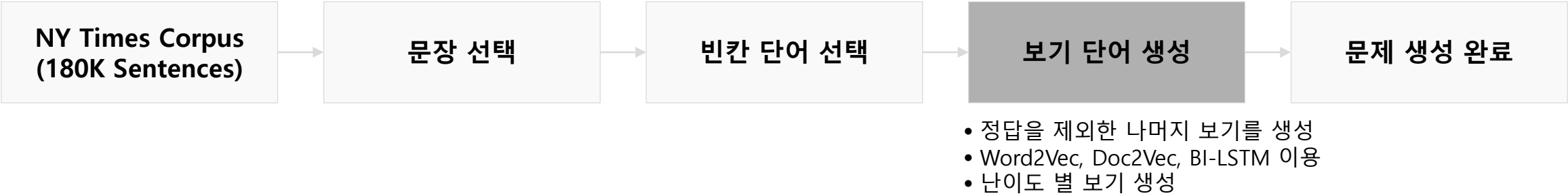
1. The indoor garden on the top floor of the new office building will be decorated with plants that need \_\_\_\_\_ care.

- (A) little
- (B) ✓
- (C) ✓
- (D) ✓

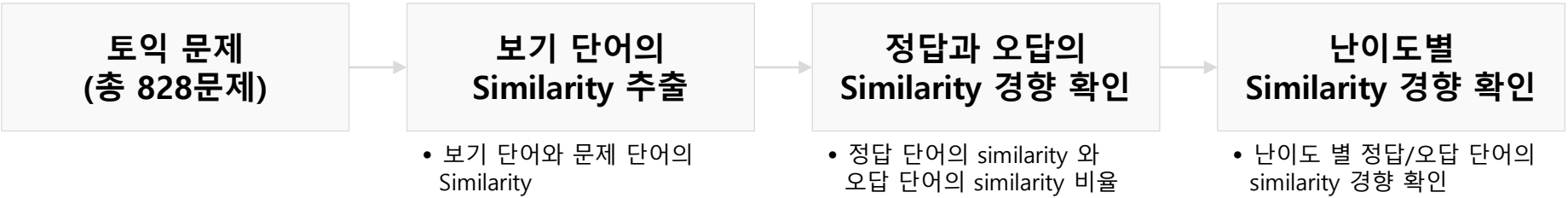
(B),(C),(D)에 따라 문제의 난이도가 결정



## 난이도별 영어 학습 문제 생성 flowchart

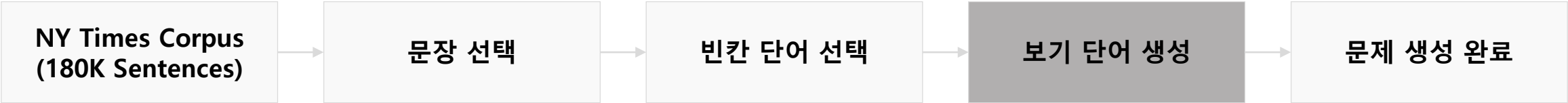


가정 : 보기 단어 생성에 따라 문제 난이도가 결정됨  
→ 실제 **토익 문제 보기 단어간의 관계 파악**을 통하여 난이도별 보기 단어의 경향 확인



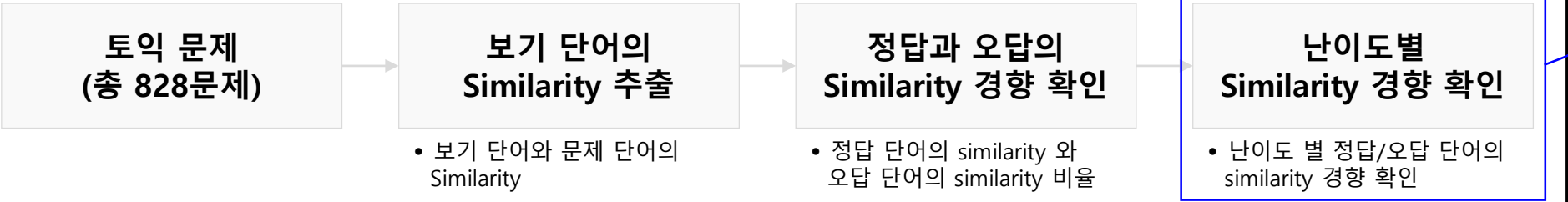
# 문제 생성 개요

## 난이도별 영어 학습 문제 생성 flowchart



- 정답을 제외한 나머지 보기를 생성
- Word2Vec, Doc2Vec, BI-LSTM 이용
- 난이도 별 보기 생성

가정 : 보기 단어 생성에 따라 문제 난이도가 결정됨  
→ 실제 **토익 문제 보기 단어간의 관계 파악**을 통하여 난이도별 보기 단어의 경향 확인



추출된 난이도별 보기 단어의 Similarity 값 비율을 보기 단어 생성에 이용

## 보기 단어와 주변 단어 간 Similarity 추출

1. The participants interrupted the speaker with questions ----- his presentation,  
so he requested the to query him after his delivery.

- (A) within
- (B) despite
- (C) during
- (D) following



#1

predict\_output\_word 이용



#2

word\_similarity 이용

## 보기 단어와 주변 단어 간 Similarity 추출

1. The participants interrupted the speaker with questions ----- his presentation, so he requested the to query him after his delivery.

- |               |                      |
|---------------|----------------------|
| (A) within    | $A_{sim} = 0.000235$ |
| (B) despite   | $B_{sim} = 0.000287$ |
| (C) during    | $C_{sim} = 0.000488$ |
| (D) following | $D_{sim} = 0.000296$ |

### #1 predict\_output\_word 이용

주변 단어를 predict\_output\_word 에 입력하여 각  
보기의 단어가 나올 확률 값 확인

### #2

word\_similarity 이용

## 보기 단어와 주변 단어 간 Similarity 추출

1. The participants interrupted the speaker with questions ----- his presentation,  
so he requested the to query him after his delivery.

- (A) within
- (B) despite
- (C) during
- (D) following

#1

predict\_output\_word 이용

#2 word\_similarity 이용

보기 단어와 주변 단어간의 유사도를 계산하여 그  
합을 보기 단어가 선택될 확률로 정의

보기 단어와 주변 단어 간 Similarity 추출

1. The participants interrupted the speaker with questions ----- his presentation,  
so he requested the to query him after his delivery.

3 4

(A) within  
(B) despite  
(C) during  
(D) following

$A_p = \text{Front similarity} + \text{Back similarity}$

$\text{Front similarity} = \text{Similarity}(\text{within}, \text{questions}) + \text{Similarity}(\text{within}, \text{with})$   
 $\quad + \text{Similarity}(\text{within}, \text{speaker}) + \text{Similarity}(\text{within}, \text{the})$

$\text{Back similarity} = \text{Similarity}(\text{within}, \text{his}) + \text{Similarity}(\text{within}, \text{presentation})$   
 $\quad + \text{Similarity}(\text{within}, \text{so}) + \text{Similarity}(\text{within}, \text{he})$

$$= \sum_{i=1}^n \text{sim}(A, X_i) + \sum_{j=1}^n \text{sim}(A, X_j)$$

#2 word\_similarity 이용

보기 단어와 주변 단어간의 유사도를 계산하여 그  
합을 보기 단어가 선택될 확률로 정의

보기 단어와 주변 단어 간 Similarity 추출

1. The participants interrupted the speaker with questions ----- his presentation,  
so he requested the to query him after his delivery.

- (A) within  $A_{sim} = 0.157$
- (B) despite  $B_{sim} = 0.179$
- (C) during  $C_{sim} = 0.240$
- (D) following  $D_{sim} = 0.163$

$A_p = \text{Front similarity} + \text{Back similarity}$

$\text{Front similarity} = \text{Similarity}(\text{within}, \text{questions}) + \text{Similarity}(\text{within}, \text{with})$   
 $\quad + \text{Similarity}(\text{within}, \text{speaker}) + \text{Similarity}(\text{within}, \text{the})$

$\text{Back similarity} = \text{Similarity}(\text{within}, \text{his}) + \text{Similarity}(\text{within}, \text{presentation})$   
 $\quad + \text{Similarity}(\text{within}, \text{so}) + \text{Similarity}(\text{within}, \text{he})$

$$= \sum_{i=1}^n w_i \text{sim}(A, X_i) + \sum_{j=1}^n w_j \text{sim}(A, X_j) \quad w_i = \frac{1}{i+a}, w_j = \frac{1}{j+a}$$

$w_i$  : 단어 인접도 가중치  
 $a$  : 가중치 조절 parameter

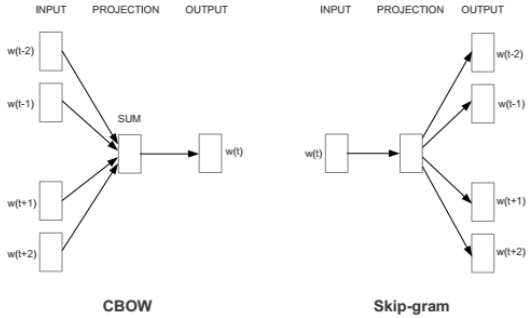
#2 word\_similarity 이용

보기 단어와 주변 단어간의 유사도를 계산하여 그  
합을 보기 단어가 선택될 확률로 정의

Word2vec model 별 문제 풀이 결과

- Predict\_output\_word를 사용한 #1에서는 42.8%, Word\_similarity를 사용한 #2에서는 42.4%의 정답률
- #1은 CBOW를 이용한 모델에서 정답률이 높은 경향, #2는 skip-gram을 이용한 모델에서 정답률이 높은 경향을 보임

Size	Window	Min_count	Training algorithm	Stop-words elimination	Words count	#1	#2
128	10	5	CBOW		32573	42.8%	33.9%
128	10	5	skip-gram		32573	36.6%	39.6%
128	10	5	skip-gram	O	32573	30.4%	36.0%
128	10	10	skip-gram		20921	37.2%	38.3%
256	10	5	CBOW		32573	42.8%	30.7%
256	10	5	CBOW	O	32573	27.8%	25.7%
256	10	5	skip-gram		32573	36.0%	42.4%
256	10	5	skip-gram	O	32573	29.7%	36.1%
256	10	10	skip-gram		20921	36.2%	38.0%
384	10	2	skip-gram		59682	37.3%	38.2%
384	10	5	skip-gram		32573	36.4%	39.4%
384	10	10	skip-gram		20921	36.5%	38.5%





가중치를 적용한 문제 풀이 결과

- 가중치를 주었을 때 정답률이 상승하는 경향을 보임
- 해당 방법론이 우연히 정답을 유추하는 것이 아니라 인접 단어와의 의미적인 관계를 고려함을 알 수 있음

Size	Window	Min_count	Training algorithm	Stop-words elimination	Words count	#1	#2	#2 +가중치(w)
128	10	5	CBOW		32573	42.8%	33.9%	31.4%
128	10	5	skip-gram		32573	36.6%	39.6%	41.5%
128	10	5	skip-gram	O	32573	30.4%	36.0%	37.0%
128	10	10	skip-gram		20921	37.2%	38.3%	40.3%
256	10	5	CBOW		32573	42.8%	30.7%	29.5%
256	10	5	CBOW	O	32573	27.8%	25.7%	25.8%
256	10	5	skip-gram		32573	38.0%	42.4%	43.5%
256	10	5	skip-gram	O	32573	29.7%	36.1%	36.4%
256	10	10	skip-gram		20921	36.2%	38.0%	38.5%
384	10	2	skip-gram		59682	37.3%	38.2%	39.9%
384	10	5	skip-gram		32573	36.4%	39.4%	39.3%
384	10	10	skip-gram		20921	36.5%	38.5%	38.4%

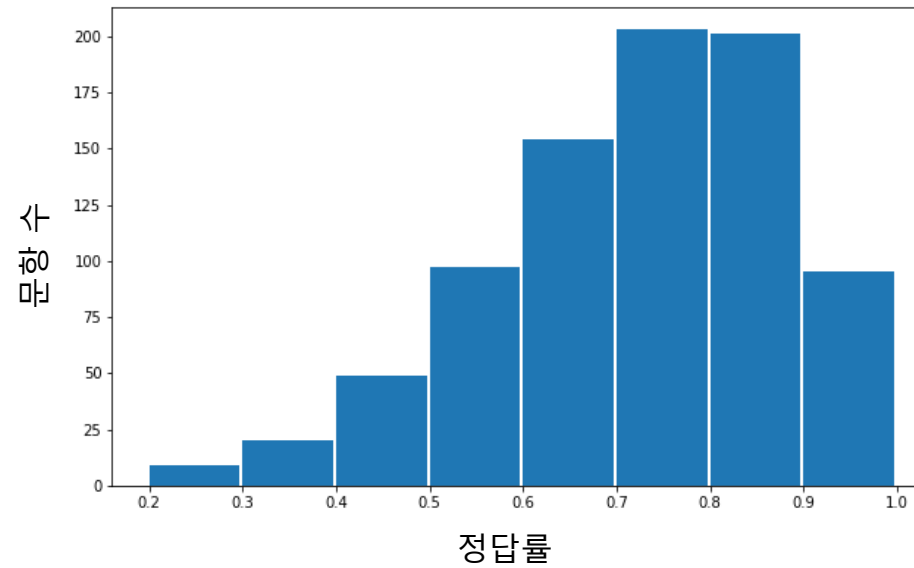
문제 유형별 문제 풀이 결과

- Vocabulary 문제 유형의 정답률 상승 확인
- 단어의 의미상의 관계를 학습한 모델이기 때문에 전치사나 시제를 묻는 문제보다는 어휘 문제를 더 잘 해결하는 것으로 보임

Size	Window	Min_count	Training algorithm	Stop-words elimination	Words count	#1	#1 (Voca)	#2 +가중치	#2 +가중치 (Voca)
128	10	5	CBOW		32573	42.8%	47.8%	31.4%	40.3%
128	10	5	skip-gram		32573	36.6%	44.7%	41.5%	49.6%
128	10	5	skip-gram	○	32573	30.4%	46.9%	37.0%	52.2%
128	10	10	skip-gram		20921	37.2%	46.5%	40.3%	47.3%
256	10	5	CBOW		32573	42.8%	52.2%	29.5%	31.4%
256	10	5	CBOW	○	32573	27.8%	41.6%	25.8%	33.2%
256	10	5	skip-gram		32573	38.0%	42.5%	43.5%	52.2%
256	10	5	skip-gram	○	32573	29.7%	45.6%	36.4%	51.3%
256	10	10	skip-gram		20921	36.2%	42.0%	38.5%	49.1%
384	10	2	skip-gram		59682	37.3%	38.9%	39.9%	44.2%
384	10	5	skip-gram		32573	36.4%	39.8%	39.3%	46.9%
384	10	10	skip-gram		20921	36.5%	41.6%	38.4%	44.2%

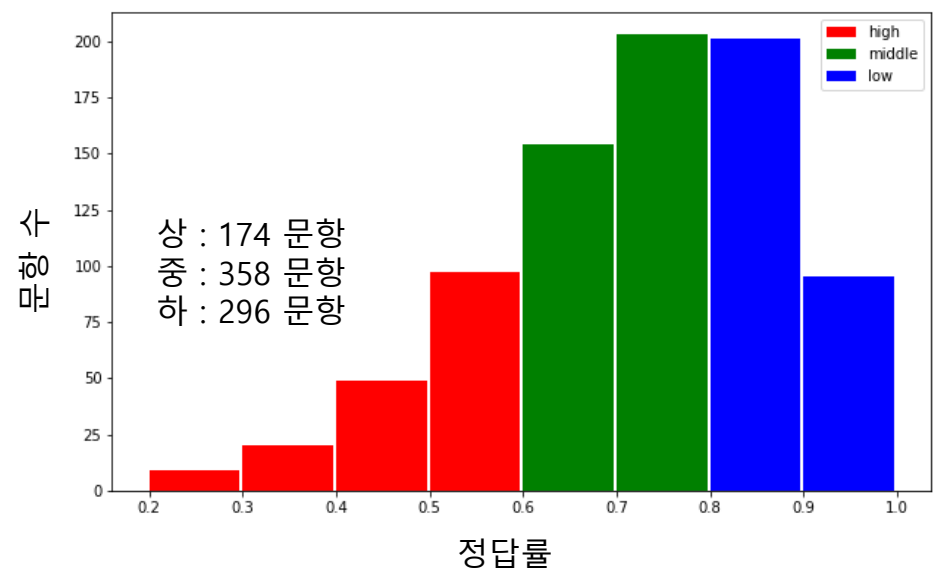
## 문제 난이도 별 분포 확인

- 난이도 별 문제 경향을 확인 하기 위하여 전체 문항에 대한 난이도 분포 확인



## 문제 난이도 별 분포 확인

- 세가지 난이도(상, 중, 하)로 구분하여 분석 진행
- 총 828 문항(상: 174문항, 중: 358문항, 하: 296문항)

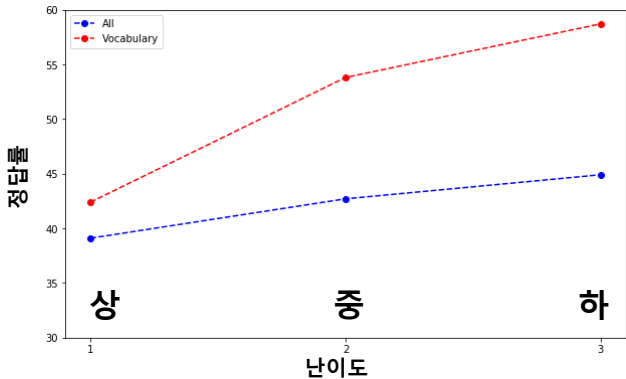


## 문제 난이도 별 정답률 확인

- 난이도 별 정답률 확인 결과 난이도가 어려운 문제보다 쉬운 문제에서 정답률 상승 경향 확인
- 해당 모델이 문제를 해결하는 과정이 난이도를 고려한다고 판단 할 수 있음.

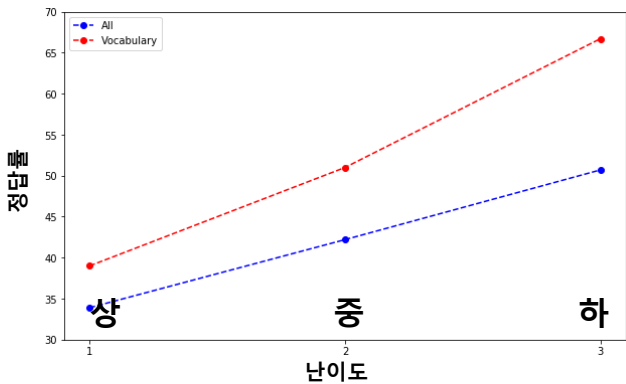
Size 256, Win 10, min 5, CBOW, #1

난이도	총 문항수	정답수	정답률	어휘 문항수	정답수	정답률
상	174	68	39.1%	59	25	<b>42.4%</b>
중	358	153	42.7%	104	56	<b>53.8%</b>
하	296	133	44.9%	63	37	<b>58.7%</b>
SUM	828	354	42.8%	226	118	52.2%



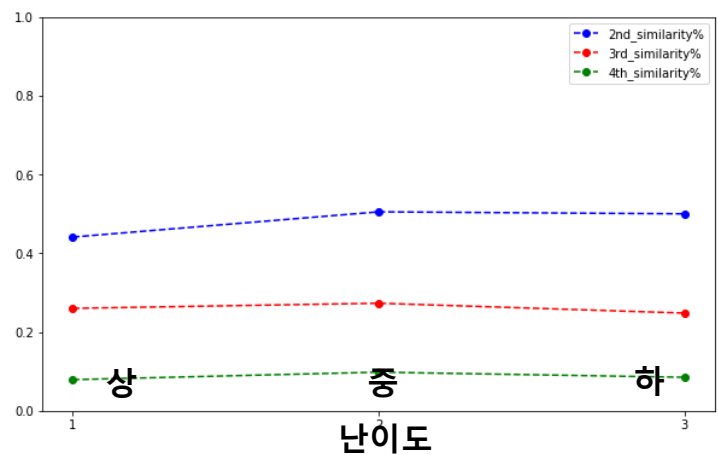
Size 256, Win 10, min 5, Skip-gram, #2

난이도	총 문항수	정답수	정답률	어휘 문항수	정답수	정답률
상	174	59	33.9%	59	23	<b>39.0%</b>
중	358	151	42.2%	104	53	<b>51.0%</b>
하	296	150	50.7%	63	42	<b>66.7%</b>
SUM	828	360	43.5%	226	118	52.2%



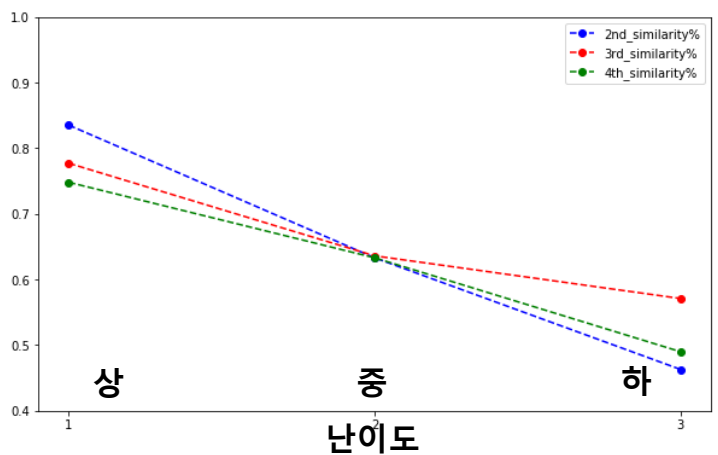
Algorithm 에 따른 난이도별 Similarity 경향 확인

- #1의 경우 보기 별 확률값의 비율이 의미없게 나타남
- #2의 경우 정답 단어와 두번째 단어가 예측될 확률값의 비율이 난이도에 따라 상승하는 경향을 보임



Size 256, Win 10, min 5, COW, #1

난이도	비율1%	비율2%	비율3%	비율4%
상	1.000	0.441	0.260	0.079
중	1.000	0.505	0.273	0.098
하	1.000	0.500	0.248	0.085

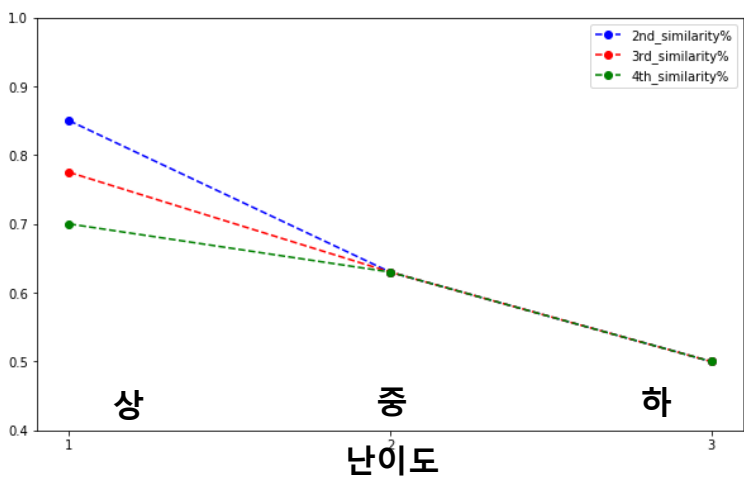


Size 256, Win 10, min 5, Skip-gram, #2

난이도	비율1	비율2%	비율3%	비율4%
상	1.000	0.835	0.633	0.463
중	1.000	0.777	0.636	0.571
하	1.000	0.748	0.633	0.490

최종 제안하는 문제 난이도별 보기의 Similarity 구성

- 문제의 난이도를 결정하는 기준으로 정답 단어와 두번째 보기의 확률값의 비율을 다음과 같이 설정
- 상: 0.85, 중: 0.075, 하: 0.70



난이도	비율1	비율2%	비율3%	비율4%
상	1	0.85	0.63	0.50
중	1	0.775	0.63	0.50
하	1	0.70	0.63	0.50

## 문제 생성 모델

1. Word2Vec 모델
2. Doc2Vec 모델
3. Bidirectional LSTM 모델

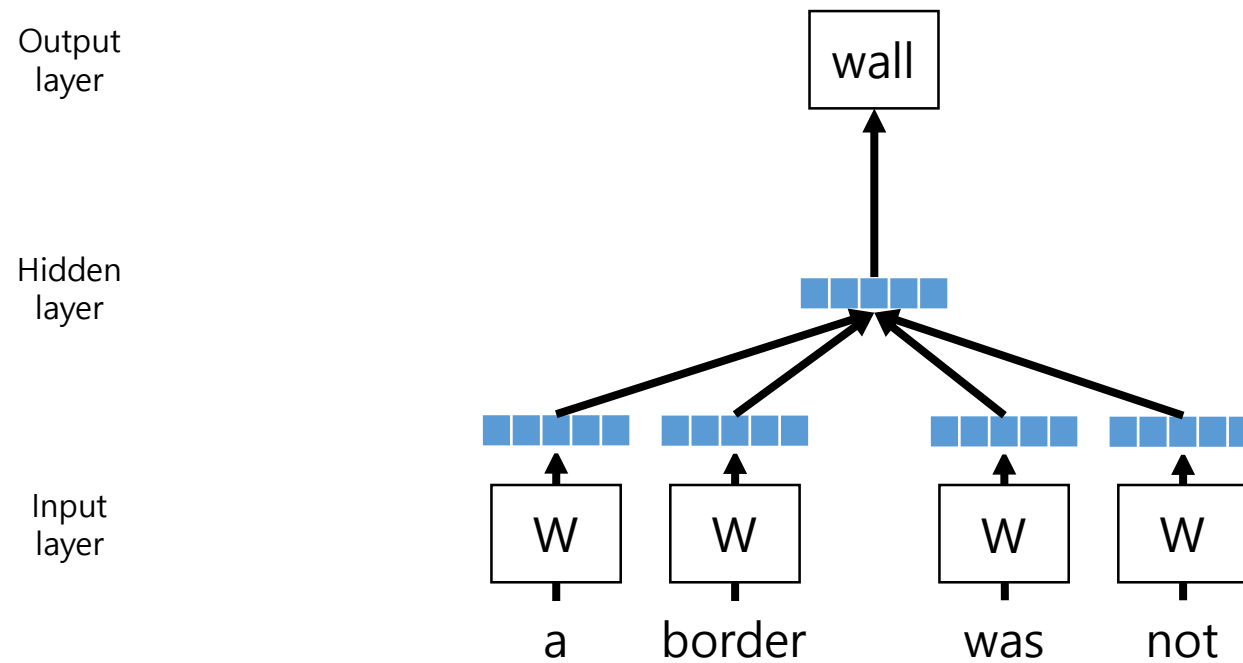


## 문제 생성 모델

1. Word2Vec 모델
2. Doc2Vec 모델
3. Bidirectional LSTM 모델

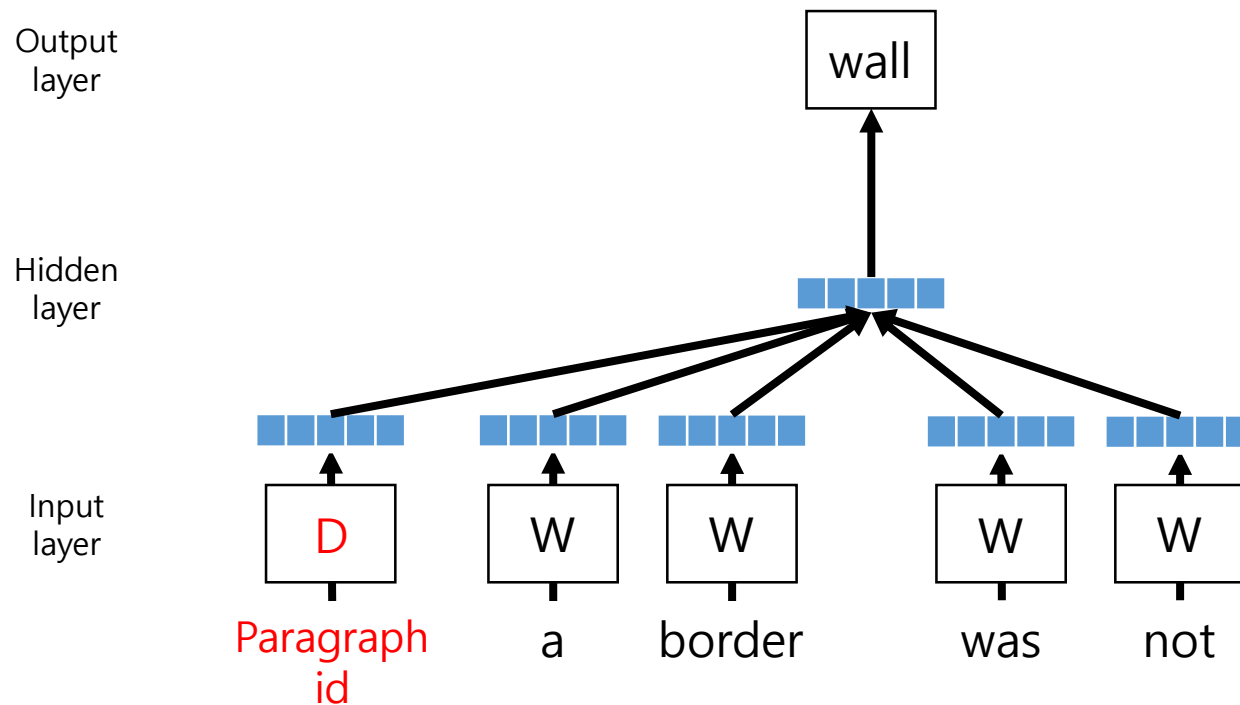
## 텍스트 embedding 방법론 – Word2Vec

- 각 단어는 하나의 벡터로 표현됨
- 같은 문맥 상의 다른 단어들과 함께 target 단어를 예측하는 데에 사용됨



## 텍스트 embedding 방법론 – Doc2Vec

- Word2Vec에서 sentence level로 확장
- Paragraph vector도 같은 문맥 상의 다른 단어들과 함께 target 단어를 예측하는 데에 사용됨



## NY Times Corpus 내에서 문제를 랜덤 선택

- NY Times Corpus에서 기사를 수집
- 랜덤으로 문장을 선택


2. President Trump [contradicted his chief of staff, John Kelly](#), above, who said on Wednesday that the president's campaign promise to build a border wall was not "fully informed."

You have 4 free articles remaining.  
[Subscribe to The Times](#)

"The Wall is the Wall, it has never changed or evolved from the first day I conceived of it," Mr. Trump wrote on Twitter.

And there was confusion over Mr. Trump's tweets on another subject: negotiations to avoid a government shutdown. [The House passed a stopgap bill](#) to keep the government open late Thursday, but Senate Democrats appear ready to block it.

<https://www.nytimes.com/2018/01/18/briefing/congress-california-amazon.html>



President Trump contradicted his chief of staff, John Kelly, above, who said on Wednesday that the president's campaign promise to build a border **wall** was not "fully informed".

## 어휘 문제를 만들기 위하여 POS-tagging 진행 후 명사 선택

- 빈칸을 어휘 추론하는 문제를 생성하기 위해 pos tagging 진행
- 명사 중 하나를 랜덤하게 선택

President Trump contradicted his chief of staff, John Kelly, above, who  
said on Wednesday that the president's campaign promise to build a  
border **wall** was not "fully informed".

- (A)
- (B)
- (C)
- (D)

### 선택된 단어 빈칸 처리 및 보기 생성

- 선택된 단어를 빈칸 처리 후 해당 단어를 보기 중 하나에 위치

President Trump contradicted his chief of staff, John Kelly, above, who said on Wednesday that the president's campaign promise to build a border ---- was not "fully informed".

- (A) wall
- (B)
- (C)
- (D)

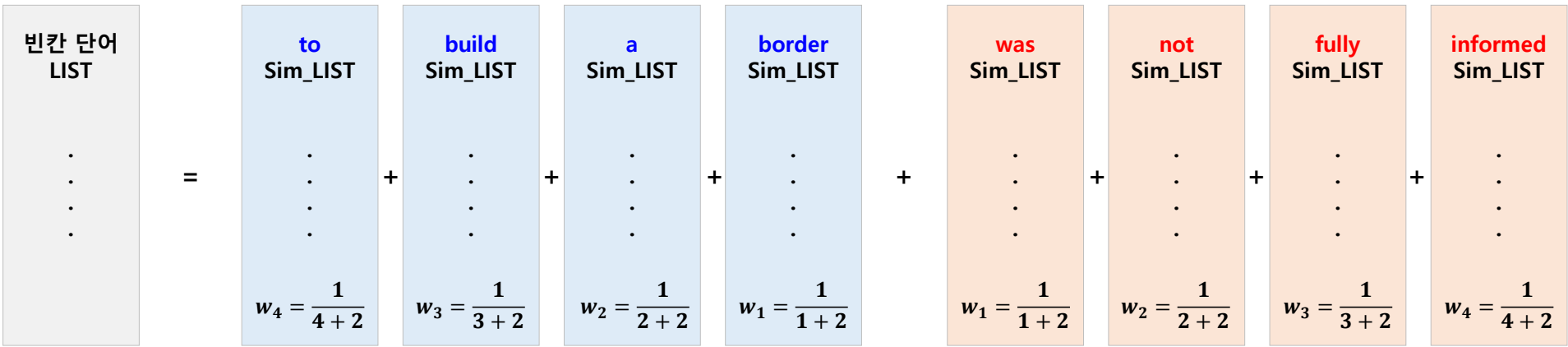
빈칸에 들어갈 수 있는 단어 LIST 생성

- 빈 칸 앞 뒤 4 단어를 이용하여 빈 칸에 들어갈 수 있는 단어 List 생성

President Trump contradicted his chief of staff, John Kelly, above, who  
said on Wednesday that the president's campaign promise to build a  
border ----- was not "fully informed".

1 1 2 3 4 4 3 2

(A) wall



## 정답 단어의 Similarity 계산

- 빈 칸 단어 list에서 정답인 단어에 대한 similarity 추출

President Trump contradicted his chief of staff, John Kelly, above, who said on Wednesday that the president's campaign promise to build a border ---- was not "fully informed".

(A) wall

(B)

(C)

(D)

빈칸 단어 LIST	
	.
(wall, 0.5039)	.
	.

$$A_{sim} = 0.5039$$



난이도에 맞는 각 보기의 Similarity 계산

- 난이도 별 문제 생성을 위하여 나머지 보기가 가질 수 있는 Similarity 계산

President Trump contradicted his chief of staff, John Kelly, above, who said on Wednesday that the president's campaign promise to build a border ---- was not "fully informed".

- (A) wall
- (B)
- (C)
- (D)

$A_{sim} = 0.5039$   
 $B_{sim} = A_{sim} * 0.85 = 0.4283$   
 $C_{sim} = A_{sim} * 0.63 = 0.3174$   
 $D_{sim} = A_{sim} * 0.50 = 0.2519$

난이도	sim1	sim2%	sim3%	sim4%
상	1	0.85	0.63	0.50
중	1	0.775	0.63	0.50
하	1	0.70	0.63	0.50

## 난이도에 맞는 Similarity 를 가지는 단어 선택

- 계산된 similarity를 가지는 단어 선택

President Trump contradicted his chief of staff, John Kelly, above, who said on Wednesday that the president's campaign promise to build a border ----- was not "fully informed".

(A) wall

(B)

(C)

(D)

$A_{sim} = 0.5039$

$B_{sim} = 0.4283$

$C_{sim} = 0.3174$

$D_{sim} = 0.2519$

빈칸 단어 LIST	
.	.
(wall,	0.5039)
.	.
(gate,	0.4283)
.	.
(rebel,	0.3174)
.	.
(año,	0.2519)
.	.

## 난이도에 맞는 Similarity 를 가지는 단어 선택

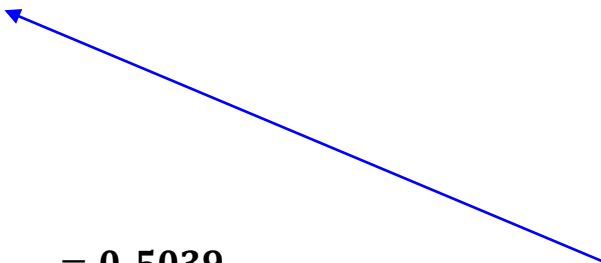
- 각 난이도 기준 확률에 맞는 similarity를 가지는 단어 선택

President Trump contradicted his chief of staff, John Kelly, above, who said on Wednesday that the president's campaign promise to build a border ---- was not "fully informed".

- (A) wall
- (B) gate
- (C) rebel
- (D) año

$$\begin{aligned}A_{sim} &= 0.5039 \\B_{sim} &= 0.4283 \\C_{sim} &= 0.3174 \\D_{sim} &= 0.2519\end{aligned}$$

빈칸 단어 LIST	
.	
(wall,	0.5039)
.	
(gate,	0.4283)
.	
(rebel,	0.3174)
.	
(año,	0.2519)
.	



Word2Vec

Size 256, Win 10, min 5, Skip-gram

President Trump contradicted his chief of staff, John Kelly, above, who said on Wednesday that the president’s campaign promise to build a border ---- was not “fully informed”.

난이도 상

- (A) wall 0.5039
- (B) gate 0.4283
- (C) rebel 0.3174
- (D) año 0.2519

난이도 중

- (A) wall 0.5039
- (B) vehicle 0.3905
- (C) rebel 0.3174
- (D) año 0.2519

난이도 하

- (A) wall 0.5039
- (B) strengthen 0.3527
- (C) rebel 0.3174
- (D) año 0.2519

난이도	sim1	sim2%	sim3%	sim4%
상	1	0.85	0.63	0.50
중	1	0.775	0.63	0.50
하	1	0.70	0.63	0.50

Doc2Vec

Size 256, Win 10, min 5, PV-DM

President Trump contradicted his chief of staff, John Kelly, above, who said on Wednesday that the president’s campaign promise to build a border ---- was not “fully informed”.

난이도 상

- (A) wall 0.1730
- (B) security 0.1471
- (C) crowds 0.1090
- (D) fiat 0.0865

난이도 중

- (A) wall 0.1730
- (B) provocateur 0.1341
- (C) crowds 0.1090
- (D) fiat 0.0865

난이도 하

- (A) wall 0.1730
- (B) contraception 0.1211
- (C) crowds 0.1090
- (D) fiat 0.0865

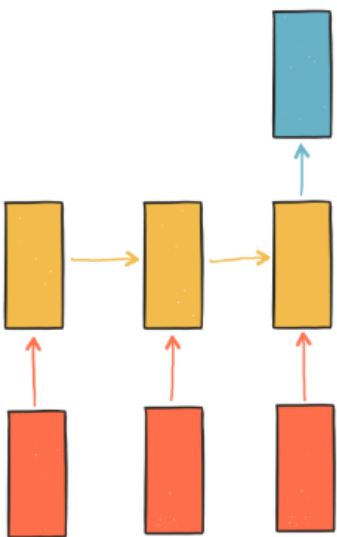
난이도	sim1	sim2%	sim3%	sim4%
상	1	0.85	0.63	0.50
중	1	0.775	0.63	0.50
하	1	0.70	0.63	0.50

## 문제 생성 모델

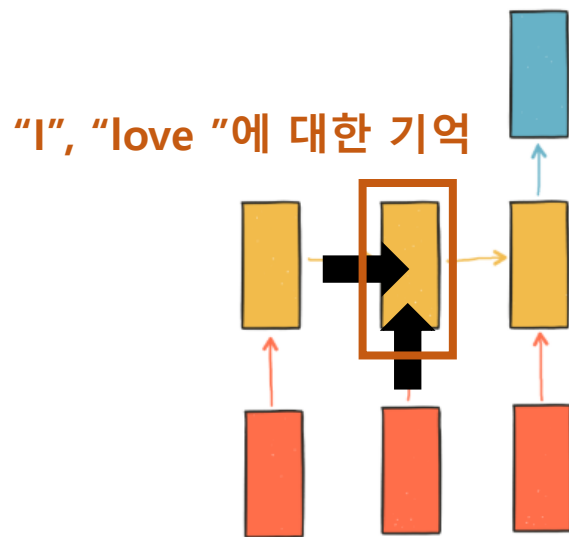
1. Word2Vec 모델
2. Doc2Vec 모델
3. Bidirectional LSTM 모델

## Bidirectional LSTM

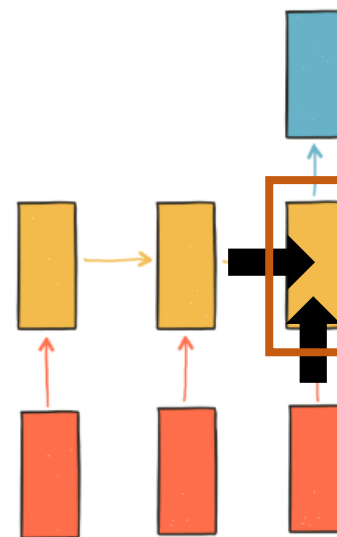
- 순차적으로 등장하는 데이터를 학습시키는 데에 적합한 모델
- Hidden state로 인해 현재까지의 입력 데이터를 요약한 정보를 '기억'
- 모든 입력을 처리하고 난 후에 마지막 기억은 시퀀스 전체를 요약하며, 이를 역 sequence로 진행



RNN 다이어그램



RNN 다이어그램



RNN 다이어그램

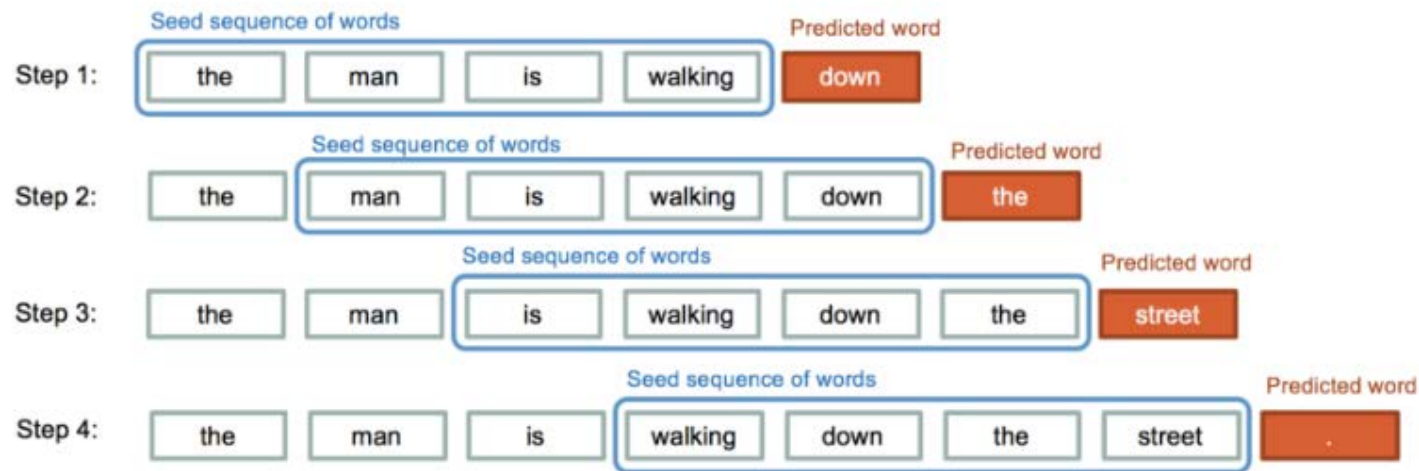
"I love you"에 대한 기억

<마지막 기억>  
모든 입력값에 대한 요약

## Bidirectional LSTM

- Bidirectional LSTM을 문제 생성 분야에 적용
- Train data: 모든 단어에 대한 seed sequence와 그 뒤의 단어 학습
- 선정된 문제에 대해 빈칸을 앞의 sequence들로 예측하고 예측 확률값을 통해 보기 생성

Sequence length = 4



steps to generate sentences




## Bidirectional LSTM

- Bidirectional LSTM을 문제 생성 분야에 적용
- Train data: 모든 단어에 대한 seed sequence와 그 뒤의 단어 학습
- 선정된 문제에 대해 빈칸을 앞의 sequence들로 예측하고 예측 확률값을 통해 보기 생성

Sequence length = 7

building owners in new york city are legally obligated to provide heat for their



Softmax Value	
tenant	
Disposition	
reimbursement	
monsoon	
dissect	
⋮	

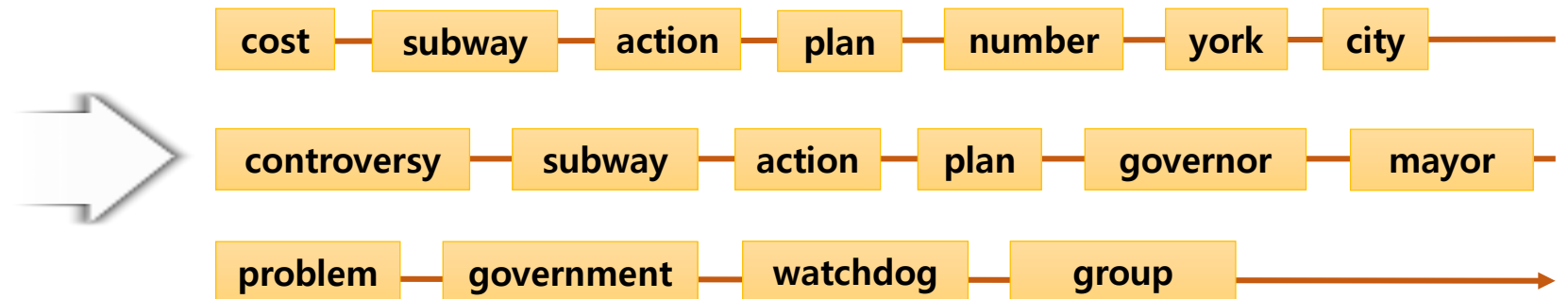
## Bidirectional LSTM 전처리

- Stop words를 제거하여 in, that, of, the 와 같은 단어들은 학습에서 제외
- Lemmatization을 수행하여 모델이 sequence 간의 패턴을 인식하도록 학습
- Vocabulary 문제는 한 sequence안의 단어들의 의미적 유사성을 학습해야 함
- 명사를 추출하여 명사 간의 관계를 학습하는 모델

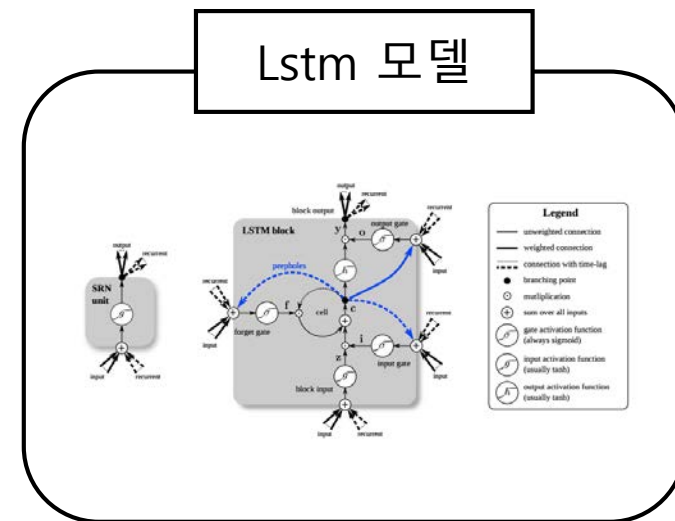
### Input data

The original **cost** of the **subway action plan** was 836 million but that **number** was based on New **York city** providing **half** the **cost**. Given the amount of **controversy** around the **subway action plan** and who pays **governor** or **mayor** match up with the **problems** causing the delays right now asked John Kaehny of Reinvent Albany a **government watchdog group**.

### Proprocessed data



1. One hot vector를 사용하여 Dictionary를 만들어 모든 단어에 대한 index 부여
2. 단어 sequence를 boolean matrix(1 or 0)로 변경



Bidirectional LSTM

President Trump contradicted his chief of staff, John Kelly, above, who said on Wednesday that the president’s campaign promise to build a border ---- was not “fully informed”.

빈칸 단어  
LIST

(wall, 0.7213)

(grinch, 0.6131)

(gory, 0.4544)

(tautou, 0.3606)

난이도 상

- (A) wall 0.7213
- (B) grinch 0.6131
- (C) gory 0.4544
- (D) tautou 0.3606

난이도 중

- (A) wall 0.7213
- (B) function 0.5590
- (C) gory 0.4544
- (D) tautou 0.3606

난이도 하

- (A) wall 0.7213
- (B) pray 0.5049
- (C) gory 0.4544
- (D) tautou 0.3606

난이도	sim1	sim2%	sim3%	sim4%
상	1	0.85	0.63	0.50
중	1	0.775	0.63	0.50
하	1	0.70	0.63	0.50

- Corpus 내 문장을 사용하여 빈칸 어휘 문제 생성
- 토익 문제를 통한 난이도 별 문제의 보기 구성 확인
  - 단어 간의 의미적 유사성을 반영하는 것에 대한 검증
  - 난이도 별 어떤 보기로 구성되어 있는지 검증
- 토익 문제의 난이도 별로 보기의 확률값을 반영하여 문제 생성하여 비교
  - Word2Vec Similarity
  - Doc2Vec Similarity
  - Bi-directional LSTM
- 한계점
  - 문제의 난이도가 단어의 의미적 유사성 뿐만 아니라 단어 자체의 난이도로 결정될 수 있음
  - 보기의 단어가 의미 있는 단어로 구성되지 않을 가능성이 큼
  - 생성된 문제에 대한 검증 구체화

**감사합니다.**