

# Generative Modeling For N-Acrostics

[최종발표]

Unstructured Data Analysis 6조

백인성 신욱수 김은비 강현규

# Table of Contents

## **01. Review**

- 1.1 Interim Presentation Summary
- 1.2 Feedback & Reflection
- 1.3 Project Flow

## **02. Process**

- 2.1 Topic Modeling(LDA)
- 2.2 Sentence Generation
- 2.3 Result

## **03. Conclusion**

- 3.1 Limitation
- 3.2 Wrap-up

## **\*Appendix**

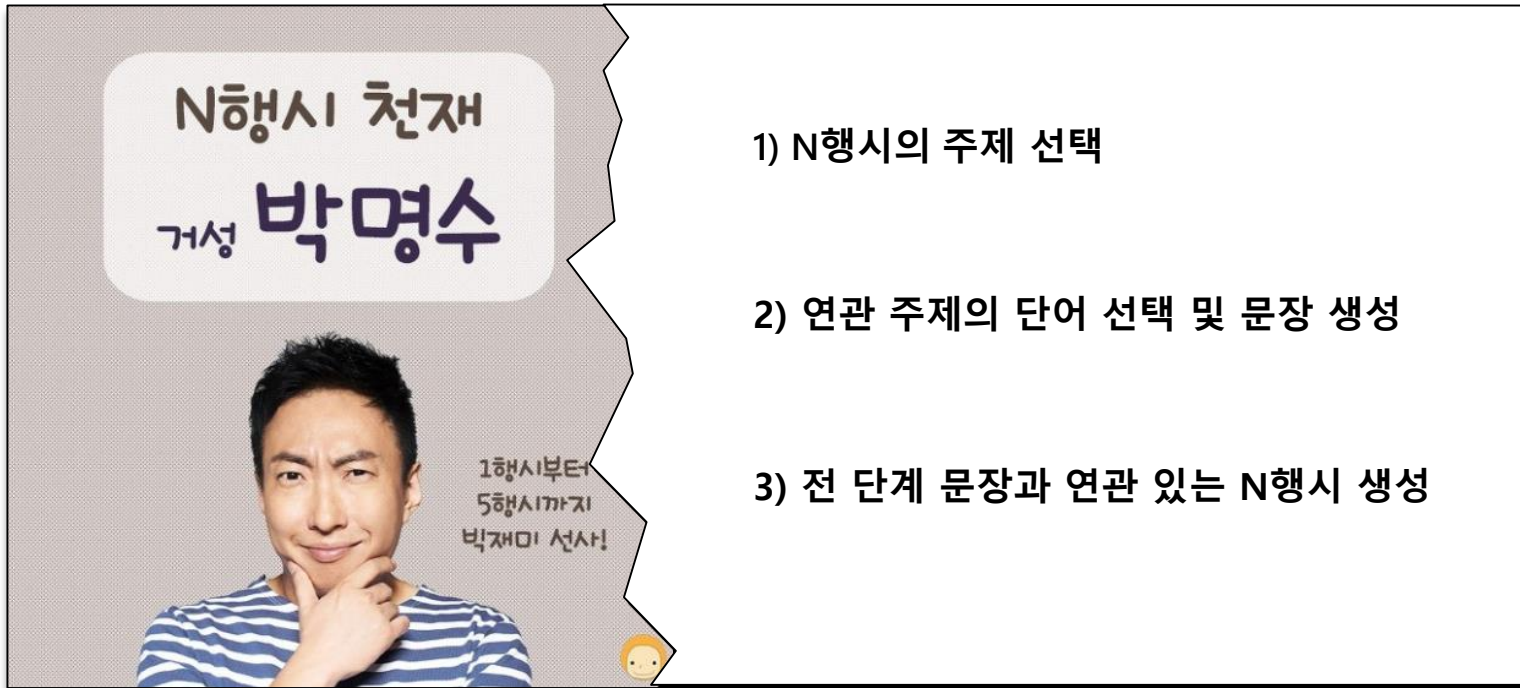
# 01. Review

# 01 | Review

## 1.1 Interim Presentation Summary

- Subject : Generative Modeling For N-Acrostics

“N행시 박명수Bot 만들기”



N행시 천재  
거성 박명수

1행시부터  
5행시까지  
박재미 선사!

- 1) N행시의 주제 선택
- 2) 연관 주제의 단어 선택 및 문장 생성
- 3) 전 단계 문장과 연관 있는 N행시 생성

# 01 | Review

## 1.1 Interim Presentation Summary

Data Crawling

Data preprocessing

Generative Modeling

### Completed works

---

- Bugs가사 data 수집 및 필터링 후 62,305건 가사 확보  
(중간 발표때는 이 중 3000건만 사용)
- 문장 단위 100만건 이상 확보 및 활용
- Morphological Analysis, Tokenize
- Frequency Analysis
- POS Tagging
- Frequency, Similarity 기반 Input Keyword 생성
- GRU 기반 문장 생성 모델 구현

## 1.1 Interim Presentation Summary

### ➤ Future works for final presentation

---

추가 전처리  
작업

- 추가 전처리 작업을 통해 현재 N행시 모델에서 더욱 자연스러운 문장이 도출되도록 함
  - 두음 법칙이 반영된 input keyword 생성
  - Hook Song 같은 반복 문장, 단어 제거

문장 간 유사도  
반영

- Topic modeling 을 활용하여 주제 카테고리 별 시작 단어 선정 및 연관 문장 생성되도록 구현
  - Latent Dirichlet Allocation 활용

추가 Data  
Crawling

- 재미를 반영하기위한 자료 추가 수집
  - Data crawling -> 자막, 유머 사이트 text data 수집
  - 반전 재미는 고민 중

비교 모델 구축

- 모델 구조 변경, 다른 모델 사용
  - Bidirectional GRU Model(문장의 양방향 순서 학습)

## 1.1 Interim Presentation Summary

### ➤ Future works for final presentation

---

추가 전처리  
작업

- 추가 전처리 작업을 통해 현재 N행시 모델에서 더욱 자연스러운 문장이 도출되도록 함
  - ✓ 두음 법칙이 반영된 input keyword 생성
  - ✓ Hook Song 같은 반복 문장, 단어 제거

문장 간 유사도  
반영

- Topic modeling 을 활용하여 주제 카테고리 별 시작 단어 선정 및 연관 문장 생성되도록 구현
  - ✓ Latent Dirichlet Allocation 활용

추가 Data  
Crawling

- 재미를 반영하기위한 자료 추가 수집  
~~Data crawling~~ → ~~자막, 유머 사이트 text data~~ 수집  
~~반전 재미는 고민 중~~

비교 모델 구축

- 모델 구조 변경, 다른 모델 사용
  - ✓ Bidirectional GRU Model(문장의 양방향 순서 학습)

# 01 | Review

## 1.2 Feedback & Reflection

Generative model 이 생성한 문장의 마지막 단어가 다음 문장의 키워드와 관련성이 높도록 설계

마지막 문장은 종결 어미가 되도록 제약을 걸어서 자연스러운 문장이 되도록 할 것

GPU 이슈가 있으면 google colab 사용을 시도해볼 것

- 이전 문장의 마지막 단어를 다음 문장의 Input 으로 활용하여 문장간 유사도 확보

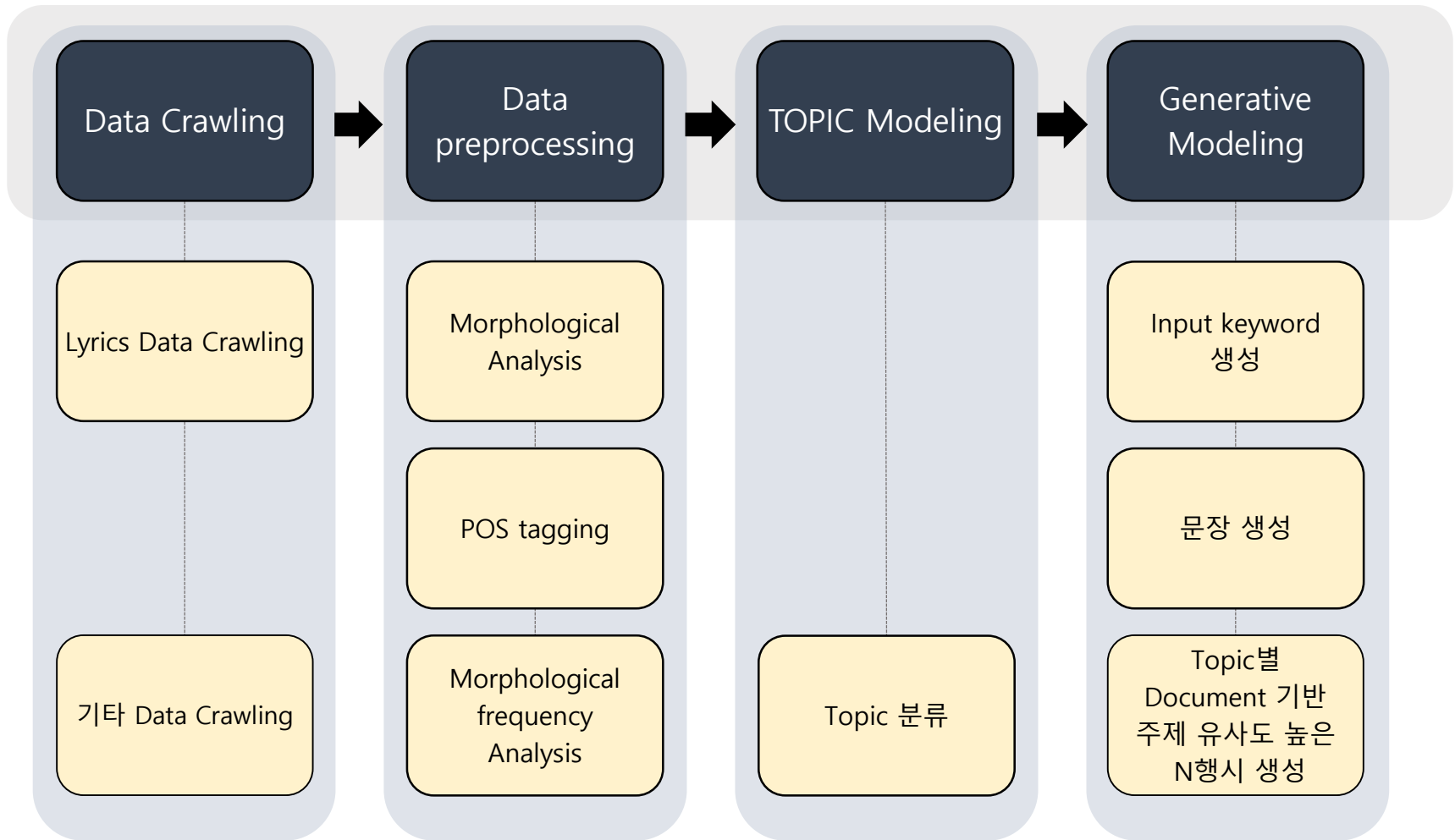
- 문장의 마지막에 <EOS> Token을 추가하여 해당 토큰 출력 시 문장 생성 종료

- 서버 컴퓨터를 활용하여 메모리 이슈 해결



# 01 | Review

## 1.3 Project Flow

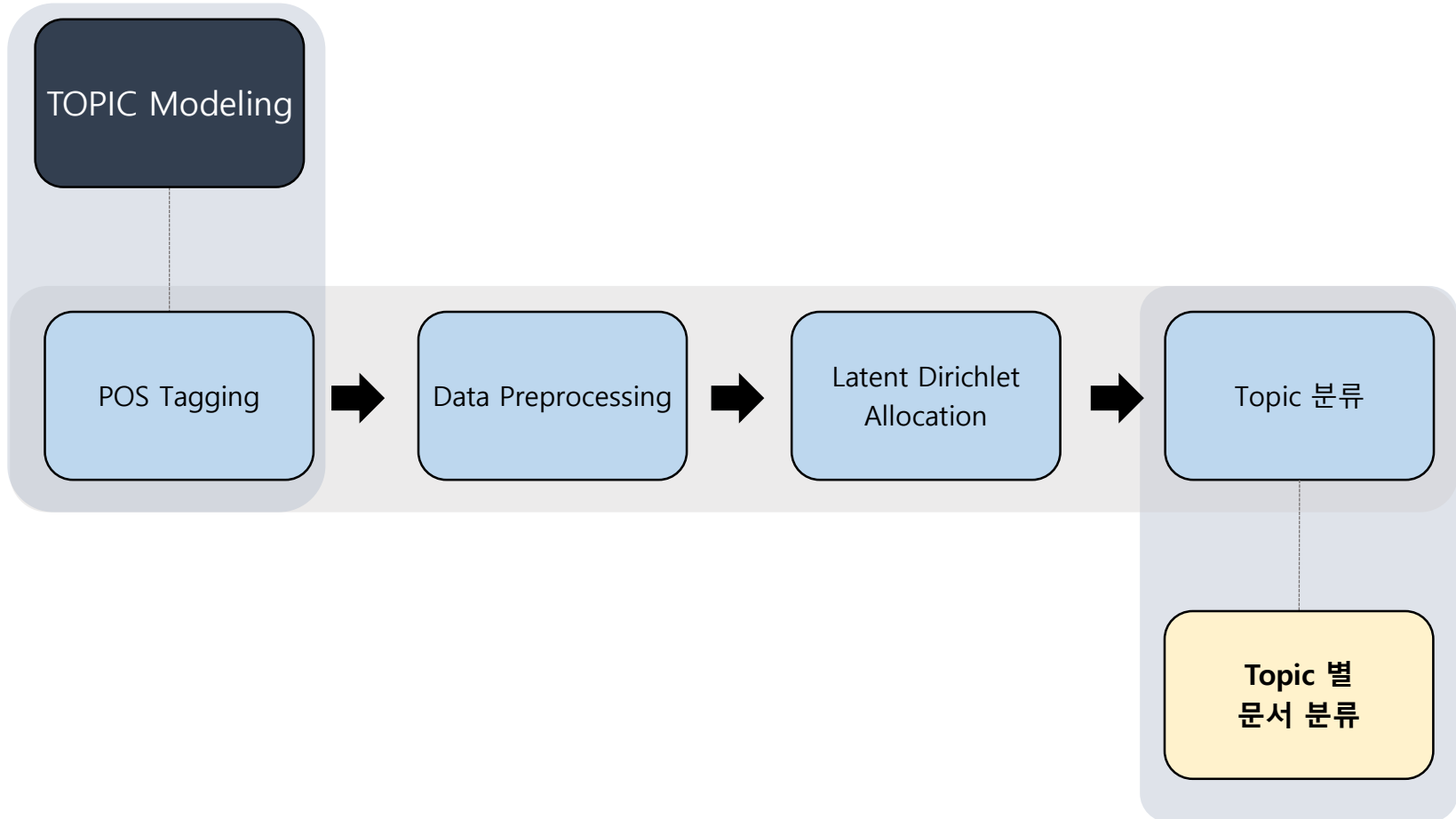


## 02. Process

# 02 | Process

## 2.1 Topic Modeling

### ➤ Topic Modeling Process



# 02 | Process

## 2.1 Topic Modeling

### ➤ POS\_tagging

- 의미 있는 Topic 추출을 위해 단어별 POS\_tagging 진행 후 명사, 동사, 형용사 단어만 추출

<Konlpy 內 Okt(Twitter)>

```
def pos_tagging(docs):  
  
    title = [doc.split('\n')[0] for doc in docs]  
    artist = [doc.split('\n')[1] for doc in docs]  
    docs = [doc.split('\n')[2] for doc in docs]  
  
    #POS tagging  
  
    pos_tag = []  
    for doc in docs:  
        pos = Okt.pos(doc)  
        pos_tag.append(pos)
```

(‘출발’, ‘Noun’)  
(‘있을지’, ‘Adjective’)  
(‘가보고’, ‘Verb’)  
(‘까지’, ‘Josa’)  
(‘.', ‘Punctuation’)  
(‘그’, ‘Determiner’)

<명사, 동사, 형용사 단어 추출>

```
#extract noun, adjective, adverb, verb  
pos_doc = []  
for i in range(len(pos_tag)):  
    temp = []  
    pos_doc.append(temp)  
  
    for i in range(len(pos_tag)):  
        for word, tag in pos_tag[i]:  
            if tag in ['Noun', 'Adjective', 'Verb']:  
                pos_doc[i].append(word)  
  
    return pos_doc, title, artist
```

(‘출발’, ‘Noun’)  
(‘있을지’, ‘Adjective’)  
(‘가보고’, ‘Verb’)

## 2.1 Topic Modeling

### ➤ Data\_preprocessing

- 의미 있는 Topic 추출에 도움이 되지 못하는 단어 길이가 '1'인 글자 삭제

[한 글자 단어 제외 전]

['출발', '김동률', '아주', '멀리', '가보고', '싶어', '곳', '누구', '만  
날', '수가', ...]  
['오래된', '노래', '김동률', '찾아낸', '낯은', '테입', '속', '노텔',  
'들었어', '서투른', ...]  
['아이', '김동률', '사랑', '한다', '말', '날', '받아줄', '때', '더',  
'이상', ...]  
['로', '쿨', '어', '두운', '불빛', '아래', '촛불', '하나', '와인',  
'잔', ...]  
['잘가요', '정재욱', '미안', '해마', '이제야', '난', '깨', '달아요',  
'내', '절대', ...]  
['김현성', '왜', '이제', '왔나요', '더', '아원', '그대', '나', '힘들었  
나요', '두', ...]  
['여자', '키스', '도대체', '알', '수가', '없어', '남자', '마음', '원  
할', '땀', ...]  
['산책', '박기영', '별일', '없니', '햇살', '좋은', '날', '돌아서', '견  
던', '이길', ...]  
['지오디', '내', '가는', '이', '길이', '어디', '가는지', '어디', '날',  
'데려가는지', ...]  
['아름다운', '날', '장혜진', '미안한', '맘', '들곤', '했', '지', '날',  
'다그쳐', ...]

'어', '깨', '지' 등 Topic 추출 방해 단어 多數

[한 글자 단어 제외 후]

['출발', '김동률', '아주', '멀리', '가보고', '싶어', '누구', '만날',  
'수가', '있을지', ...]  
['오래된', '노래', '김동률', '찾아낸', '낯은', '테입', '노텔', '들었어',  
'서투른', '피아노', ...]  
['아이', '김동률', '사랑', '한다', '받아줄', '이상', '바랄게', '없다고',  
'자신', '있게', ...]  
['두운', '불빛', '아래', '촛불', '하나', '와인', '담긴', '약속', '하  
나', '할상', ...]  
['잘가요', '정재욱', '미안', '해마', '이제야', '달아요', '절대', '그대',  
'아름', '편찮을게요', ...]  
['김현성', '이제', '왔나요', '아원', '그대', '힘들었나요', '하네요', '모  
든', '버리려', '했죠', ...]  
['여자', '키스', '도대체', '수가', '없어', '남자', '마음', '원할', '연  
제', '주니', ...]  
['산책', '박기영', '별일', '없니', '햇살', '좋은', '돌아서', '견던',  
'이길', '걸곤', ...]  
['지오디', '가는', '길이', '어디', '가는지', '어디', '데려가는지', '어  
딘', '없지만', '없지만', ...]  
['아름다운', '장혜진', '미안한', '들곤', '다그쳐', '원한', '가졌을', '그  
땀', '그게', '사랑', ...]

Topic 추출에 유의미한 두 자 이상 단어만 남김

# 02 | Process

## 2.1 Topic Modeling

### ➤ Data\_preprocessing

- 개수가 매우 많아, 편향 된 결과를 만드는 '그대', '사랑', '나를' 단어 제거

[ '그대', '사랑', '나를' 단어 제외 전 ]

	A	B	C
1	Topic	Word	문서 개수
2	31	사랑,그대,우리,나를,다시...	59953
3	27	그대,안녕,말아요,당신,사랑...	430
4	26	보여,엄마,하게,예뻐,아래...	315
5	8	주님,야야,없도록,꾸네,멀게만..	137
6	33	원해,문제,가자,새끼,위험해...	126

특정 단어가 속한 Topic  
-> 대부분 문서에 할당

[ '그대', '사랑', '나를' 단어 제외 후 ]

	A	B	C
1	Topic	Word	문서 개수
2	0	오늘,우리,좋아,사람,있어...	36022
3	12	기억,다시,우리,시간,눈물...	17638
4	17	조금,혹시,누가,사람,없어...	5331
5	18	당신,가득,태양,멋진,사라져	1813
6	9	주님,라라라라,하나님,예수	587

특정 단어 제거 후,  
Topic 별 문서가 고르게 분포

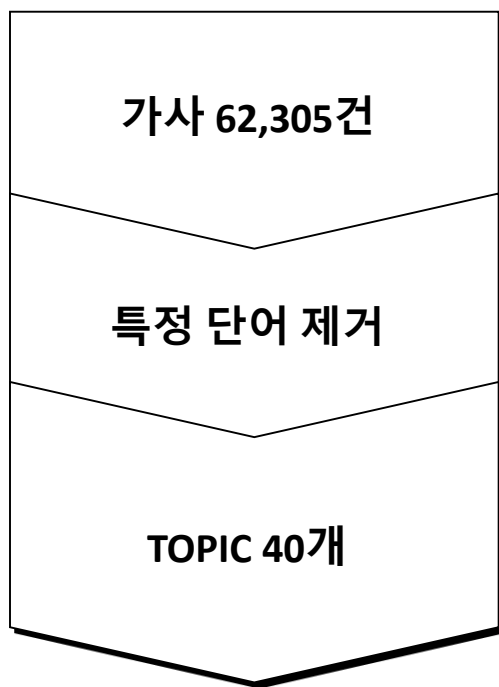
# 02 | Process

## 2.1 Topic Modeling

### ➤ LDA Model

- Topic 개수를 여러 개 시도해 본 결과, Topic 개수 = 40개 일 때 가장 좋은 결과를 보임

#### [Topic Modeling Process]



#### [결과]

Document per topic matrix 구축  
(Document 각각에 확률이 가장 높은 Topic을 그 문서의 대표 토픽으로 선정)

Probability	Topic	Title	Artist
0.077428468	1	워키토키	힌트
0.832233012	6	가시나	선미
0.077664524	1	폰서트	십센치
0.078086756	1	사르르	케이윌 씨스타 보이프렌드 매드 클
0.001015014	0	난 늘	태윤
0.10795407	0	고개 숙인 봄꽃들	하정완
0.001003662	0	숨 그 끝의	송지혜
0.001001218	0	봄날의 꽃잎과 하늘의 구름처럼	예동할창단
0.045320898	1	나의 잘못	지루박 이성우
0.039328363	1	나는 왜 이럴까	폴립
0.183877364	1	이런 세상에	찬영
0.092498943	1	구십구세까지 팔팔하게	이백길
0.076688424	1	한강의 밤	주노가
0.115405343	1	설탕분수	수민
0.051304325	1	불타올라	큐바니즘
0.153626278	1	지향	라데아토
0.001614313	0	봄노래	무적기타
0.049266055	1	좋아해 소녀	형섭의웅
0.019073647	3	제가 씹니다	최완수

# 02 | Process

## 2.1 Topic Modeling

### ➤ LDA Model

- Topic 개수를 여러 개 시도해 본 결과, Topic 개수 = 40개 일 때 가장 좋은 결과를 보임

Topic	W1	W2	W3	W4	W5	...	Naming	# of Docs
0	오늘	우리	좋아	마음	지금	...	사랑	36022
1	날아가	나비	만나게	좋아서	알겠어	...	사랑	70
...	..	...	...	...	...	...	...	...
9	주님	라라라	하나님	예수	찬양	...	찬송가	587
10	슬프게	기다리지	기다랗게	행복해야	울면	...	이별	11
12	기억	다시	우리	시간	눈물	...	이별	129
...	..	...	...	...	...	...	...	...
29	하겠어	축제	향기로운	부족해	거부	...	놀이	38
...	..	...	...	...	...	...	...	...
38	돼요	지워지지	없네요	믿어요	날아올라	...	-	19
39	봅니다	데리러	실랑	추고	아픈데	...	-	8



# 02 | Process

## 2.1 Topic Modeling

### ➤ LDA Model

- Topic 개수를 여러 개 시도해 본 결과, Topic 개수 = 40개 일 때 가장 좋은 결과를 보임

Topic	W1	W2	W3	W4	W5	...	Naming	# of Docs
0	오늘	우리	좋아	마음	지금	...	사랑	36022
1	날아가	나비	만나게	좋아서	알겠어	...	사랑	70
...	..	...	...	...	...	...	...	...
9	주님	라라라	하나님	예수	찬양	...	찬송가	587
10	슬프게	기다리지	기다랗게	행복해야	울면	...	이별	11
12	기억	다시	우리	시간	눈물	...	이별	129
...	..	...	...	...	...	...	...	...
29	하겠어	축제	향기로운	부족해	거부	...	놀이	38
...	..	...	...	...	...	...	...	...
38	돼요	지워지지	없네요	믿어요	날아올라	...	-	19
39	봅니다	데리러	실랑	추고	아픈데	...	-	8

# 02 | Process

## 2.1 Topic Modeling

### ➤ LDA Model

- Topic 개수를 여러 개 시도해 본 결과, Topic 개수 = 40개 일 때 가장 좋은 결과를 보임

Topic	W1	W2	W3	W4	W5	...	Naming	# of Docs
0	오늘	우리	좋아	마음	지금	...	사랑	36022
1	날아가	나비	만나게	좋아서	알겠어	...	사랑	70
...	..	...	...	...	...	...	...	...
9	주님	라라라	하나님	예수	찬양	...	찬송가	587
10	슬프게	기다리지	기다랗게	행복해야	울면	...	이별	11
12	기억	다시	우리	시간	눈물	...	이별	129
...	..	...	...	...	...	...	...	...
29	하겠어	축제	향기로운	부족해	거부	...	놀이	38
...	..	...	...	...	...	...	...	...
38	돼요	지워지지	없네요	믿어요	날아올라	...	-	19
39	봅니다	데리러	실랑	추고	아픈데	...	-	8

# 02 | Process

## 2.1 Topic Modeling

### ➤ Topic Modeling 최종 활용 방안

- 가장 구분이 잘 되고, 대조적인 주제를 갖는 '사랑', '이별' Topic 2개 내의 데이터를 활용 (총 61,476건)

#### 사랑 TOPIC

# of documents = 38,223  
# of sentences = 182,272  
# of words = 54,103

[예시]

Topic	제목	아티스트
0	벌써 12시	청하
1	활활	워너원
4	진진자라	태진아
...	...	...
32	예뻐 예뻐	레이디스코드

#### 이별 TOPIC

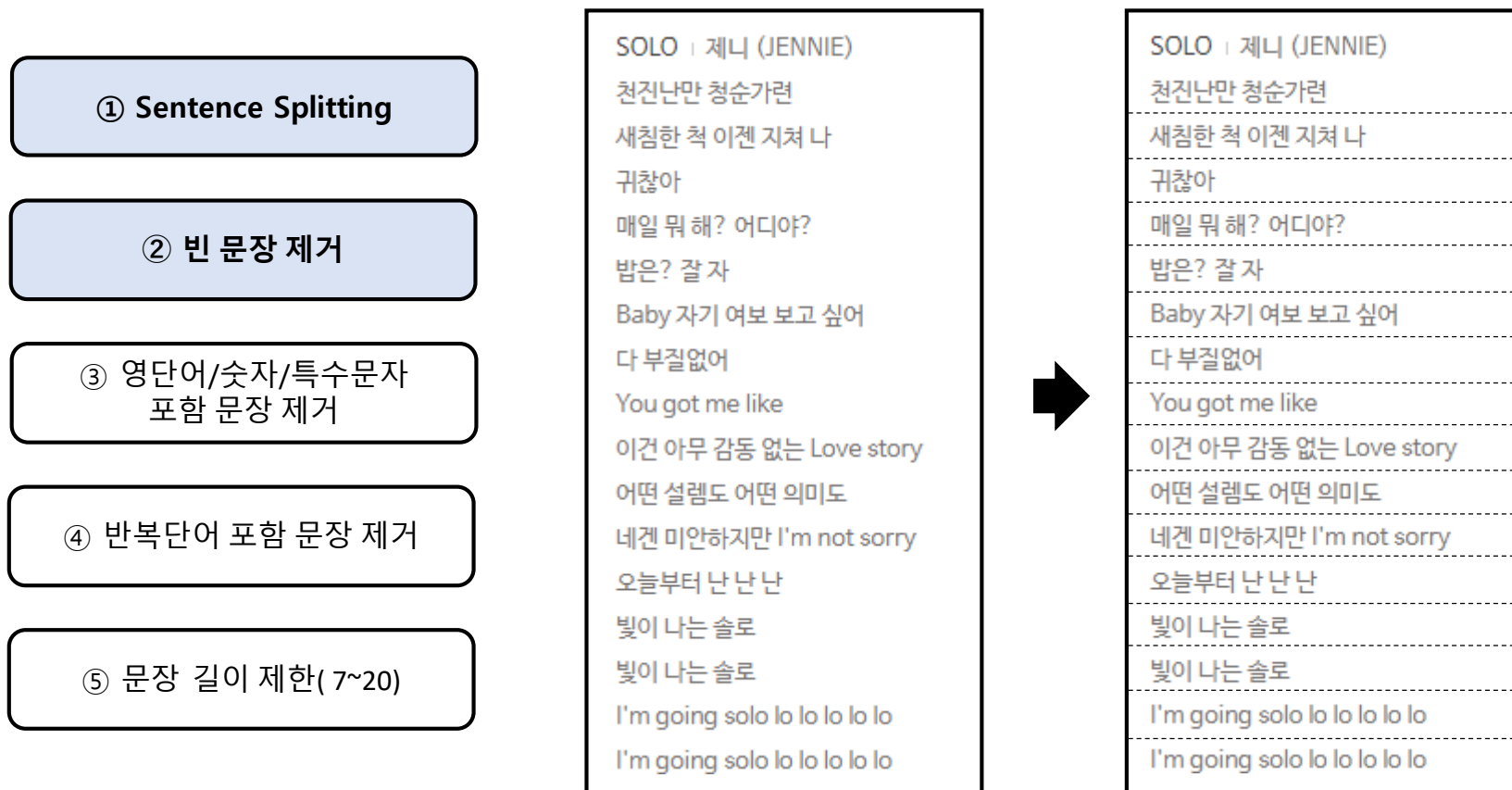
# of documents = 23,253  
# of sentences = 80,636  
# of words = 28,300

[예시]

Topic	제목	아티스트
10	울면 안돼	캐롤
12	기억을 걷는 시간	넬
17	이러지마 제발	케이월
...	...	...
23	이젠 잊기로 해요	멜로디데이

## 2.2 Sentence Generation

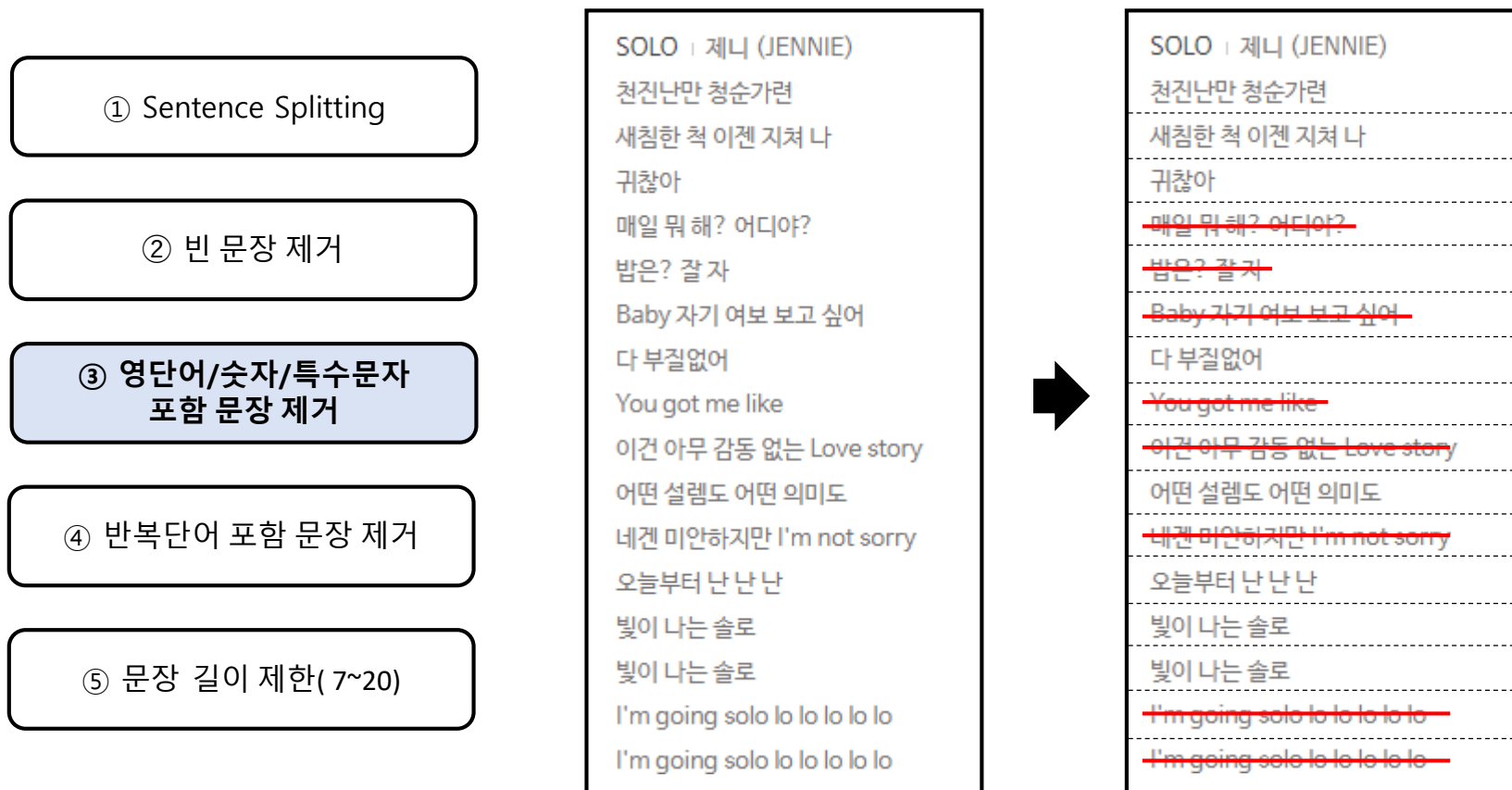
### ➤ 의미 있는 문장 추출을 위한 전처리



# 02 | Process

## 2.2 Sentence Generation

### ➤ 의미 있는 문장 추출을 위한 전처리



## 2.2 Sentence Generation

### ➤ 의미 있는 문장 추출을 위한 전처리

- ① Sentence Splitting
- ② 빈 문장 제거
- ③ 영단어/숫자/특수문자 포함 문장 제거
- ④ 반복단어 포함 문장 제거
- ⑤ 문장 길이 제한( 7~20)

SOLO | 제니 (JENNIE)  
천진난만 청순가련  
새침한 척 이젠 지쳐 나  
귀찮아  
매일 뭐 해? 어디야?  
밤은? 잘 자  
Baby 자기 여보 보고 싶어  
다 부질없어  
You got me like  
이건 아무 감동 없는 Love story  
어떤 설렘도 어떤 의미도  
네겐 미안하지만 I'm not sorry  
오늘부터 난 난 난  
빛이 나는 솔로  
빛이 나는 솔로  
I'm going solo lo lo lo lo lo  
I'm going solo lo lo lo lo lo



SOLO | 제니 (JENNIE)  
천진난만 청순가련  
새침한 척 이젠 지쳐 나  
귀찮아  
~~매일 뭐 해? 어디야?~~  
~~밤은? 잘 자~~  
~~Baby 자기 여보 보고 싶어~~  
다 부질없어  
~~You got me like~~  
~~이건 아무 감동 없는 Love story~~  
~~어떤 설렘도 어떤 의미도~~  
~~네겐 미안하지만 I'm not sorry~~  
~~오늘부터 난 난 난~~  
빛이 나는 솔로  
빛이 나는 솔로  
~~I'm going solo lo lo lo lo lo~~  
~~I'm going solo lo lo lo lo lo~~

# 02 | Process

## 2.2 Sentence Generation

### ➤ 의미 있는 문장 추출을 위한 전처리

- ① Sentence Splitting
- ② 빈 문장 제거
- ③ 영단어/숫자/특수문자 포함 문장 제거
- ④ 반복단어 포함 문장 제거
- ⑤ 문장 길이 제한( 7~20)

SOLO | 제니 (JENNIE)  
천진난만 청순가련  
새침한 척 이젠 지쳐 나  
귀찮아  
매일 뭐 해? 어디야?  
밤은? 잘 자  
Baby 자기 여보 보고 싶어  
다 부질없어  
You got me like  
이건 아무 감동 없는 Love story  
어떤 설렘도 어떤 의미도  
네겐 미안하지만 I'm not sorry  
오늘부터 난 난 난  
빛이 나는 솔로  
빛이 나는 솔로  
I'm going solo lo lo lo lo lo  
I'm going solo lo lo lo lo lo



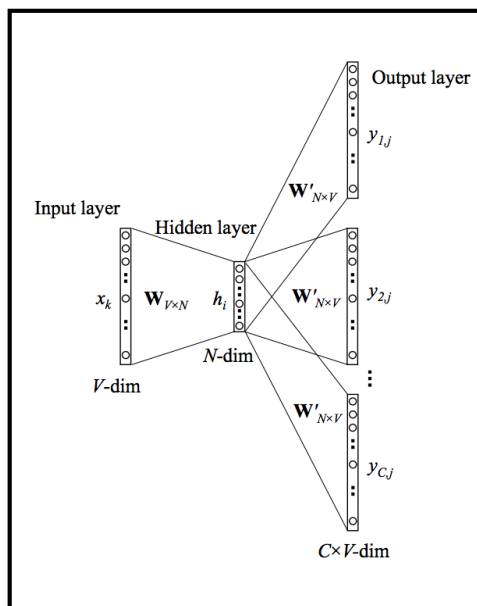
SOLO | 제니 (JENNIE)  
~~천진난만 청순가련~~  
~~새침한 척 이젠 지쳐 나~~  
~~귀찮아~~  
~~매일 뭐 해? 어디야?~~  
~~밤은? 잘 자~~  
~~Baby 자기 여보 보고 싶어~~  
~~다 부질없어~~  
~~You got me like~~  
~~이건 아무 감동 없는 Love story~~  
~~어떤 설렘도 어떤 의미도~~  
~~네겐 미안하지만 I'm not sorry~~  
~~오늘부터 난 난 난~~  
빛이 나는 솔로  
빛이 나는 솔로  
~~I'm going solo lo lo lo lo lo~~  
~~I'm going solo lo lo lo lo lo~~

# 02 | Process

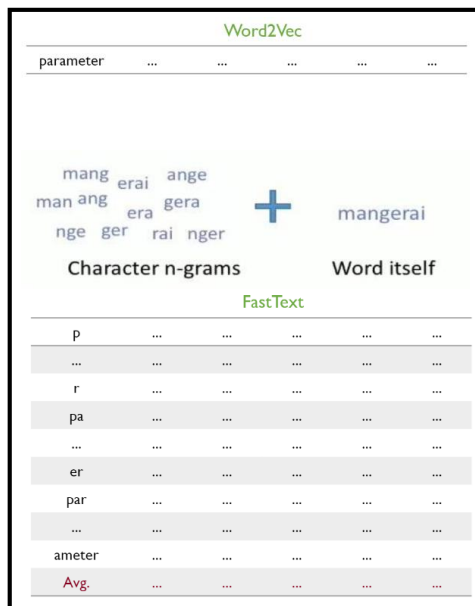
## 2.2 Sentence Generation

- 시작단어 생성을 위한 토픽별 Word Embedding Model 학습

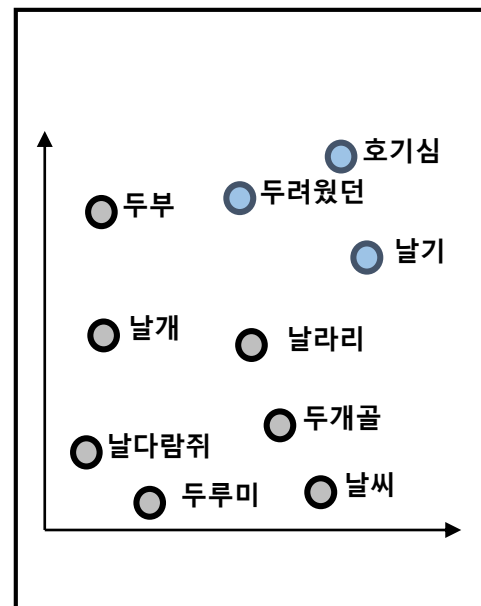
Word2Vec



FastText



Word Embedding



### <Option>

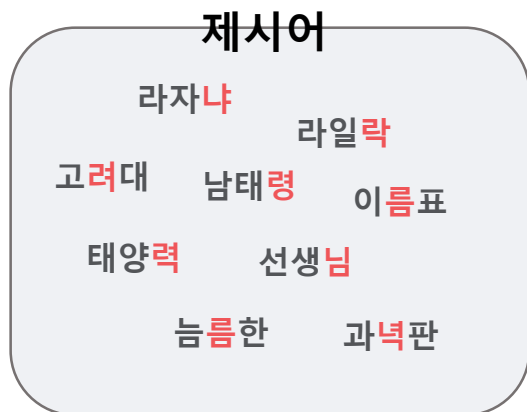
- Embedding Dimension = 128
- Training Algorithm = skip-gram
- Window = 5
- Minimum count = 1
- Iteration = 200



# 02 | Process

## 2.2 Sentence Generation

- 한글 발음에 특화된 첫소리 법칙 적용



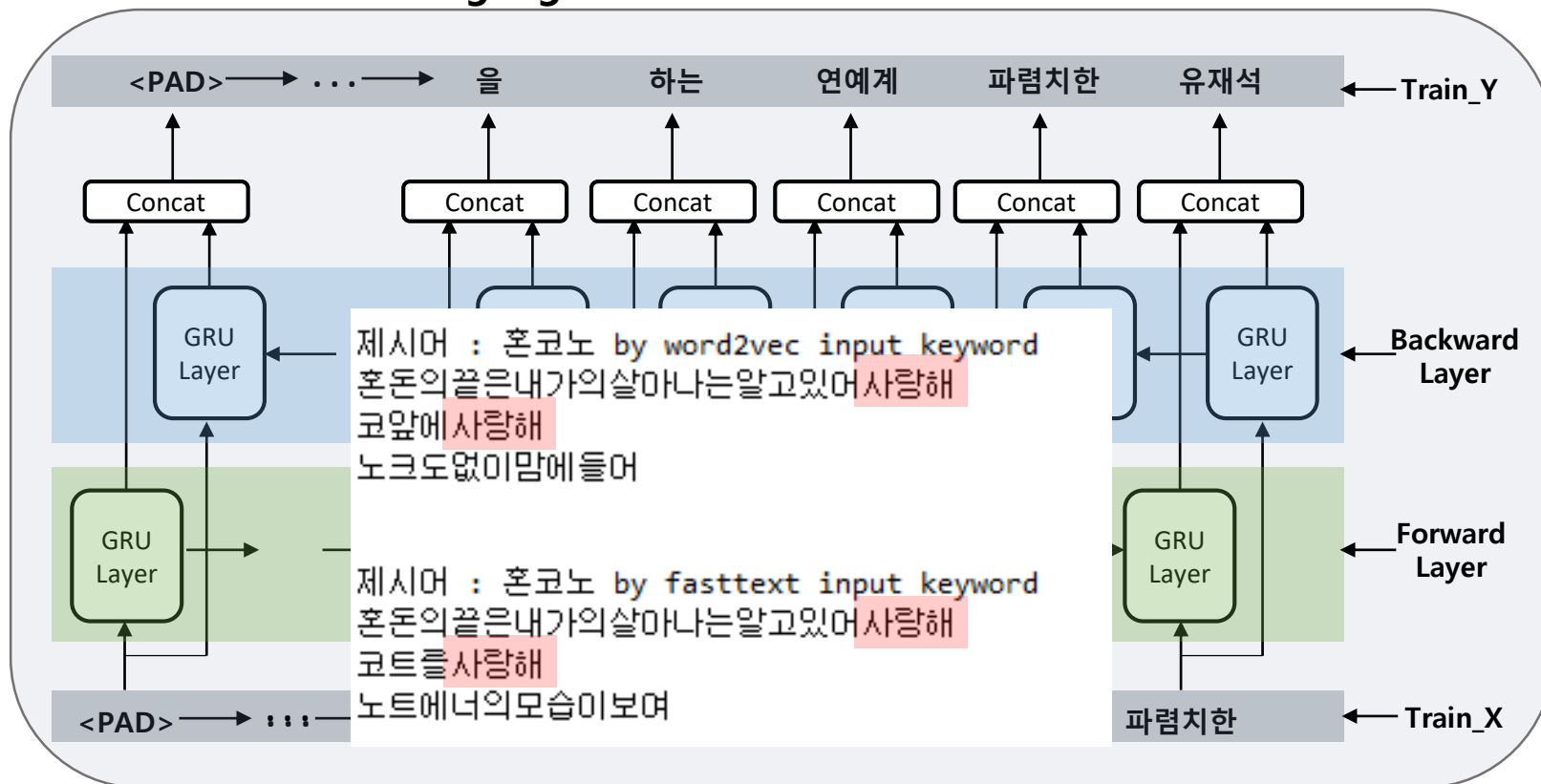
문장 시작 단어 생성 시  
두음법칙 적용

```
def dueum_keyword(keyword):  
    if keyword == "리" or keyword == "니":  
        dueum_keyword = "이"  
    elif keyword == "릴" or keyword == "님":  
        dueum_keyword = "임"  
    elif keyword == "링" or keyword == "녕":  
        dueum_keyword = "업"  
    elif keyword == "린" or keyword == "닌":  
        dueum_keyword = "인"  
    elif keyword == "련" or keyword == "년":  
        dueum_keyword = "연"  
    elif keyword == "려" or keyword == "녀":  
        dueum_keyword = "여"  
    elif keyword == "령" or keyword == "녕":  
        dueum_keyword = "영"  
    elif keyword == "랑" or keyword == "냥":  
        dueum_keyword = "알"  
    elif keyword == "름" or keyword == "름":  
        dueum_keyword = "을"  
    elif keyword == "라" or keyword == "나":  
        dueum_keyword = "아"  
    elif keyword == "렉" or keyword == "낙":  
        dueum_keyword = "역"  
    elif keyword == "료" or keyword == "뇨":  
        dueum_keyword = "요"  
    elif keyword == "류" or keyword == "뉴":  
        dueum_keyword = "유"  
    elif keyword == "룻":  
        dueum_keyword = "룻"  
    elif keyword == "로":  
        dueum_keyword = "노"  
    elif keyword == "룬":  
        dueum_keyword = "논"  
    elif keyword == "뢰":  
        dueum_keyword = "뇌"
```

# 02 | Process

## 2.2 Sentence Generation

### ➤ Bidirectional GRU Language Model Structure

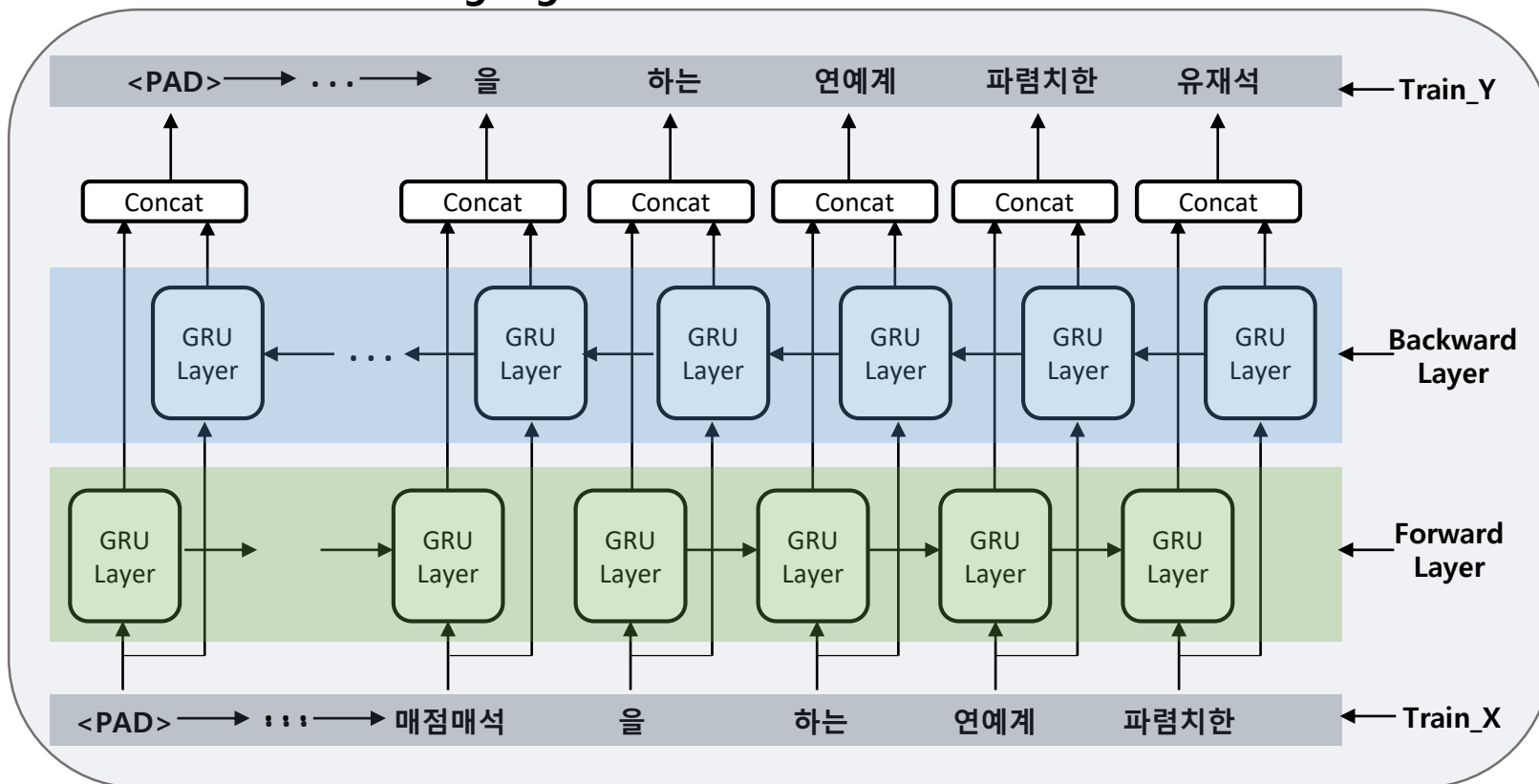


- 문장 학습 시 양방향 순서를 고려하여 Language model의 성능 향상 기대
- 특정 명사와 조사 or 동사가 항상 Set으로 등장하여 다양한 문장을 생성하지 못하는 기존 GRU의 문제점을 개선할 수 있을 것으로 판단(Ex : "그대/Noun + 를/Josa", "사랑/Noun + 해/Verb" 등)

# 02 | Process

## 2.2 Sentence Generation

### ➤ Bidirectional GRU Language Model Structure



#### <Option>

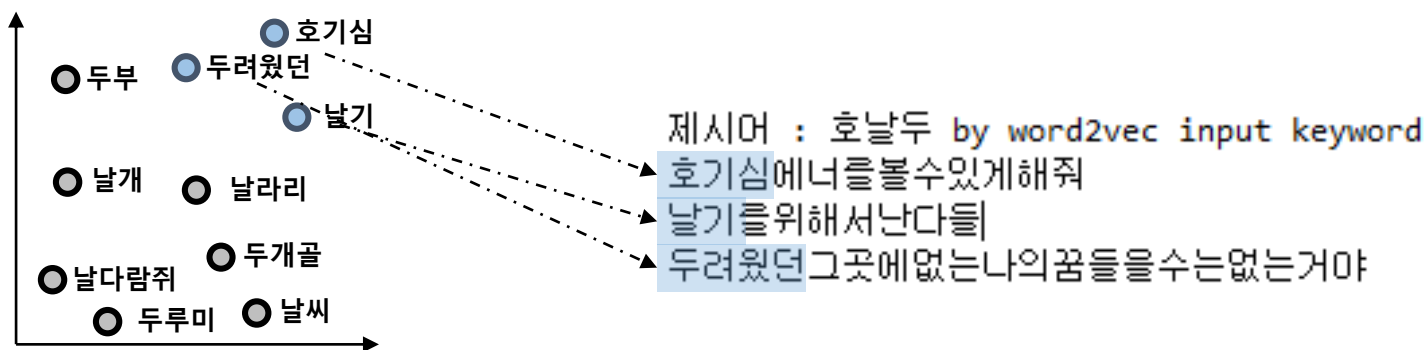
- Embedding Dimension = 128
- Batch Size = 128
- Optimizer = rmsprop
- Validation\_Split = 0.2
- Hidden Layer Size = 256
- Epoch = 100
- Loss = sparse\_categorical\_crossentropy

## 2.2 Sentence Generation

### ➤ Sentence Similarity : Interim

- Input Keyword의 Similarity에만 의존
- 각 단어로 시작하는 문장 생성시 문장간 유사도를 반영하지 않았음

Word2vec Embedding Space



$$\text{sim}(\text{호기심}, \text{날기}) > \text{sim}(\text{호기심}, \text{날씨}) > \text{sim}(\text{호기심}, \text{날라리})$$

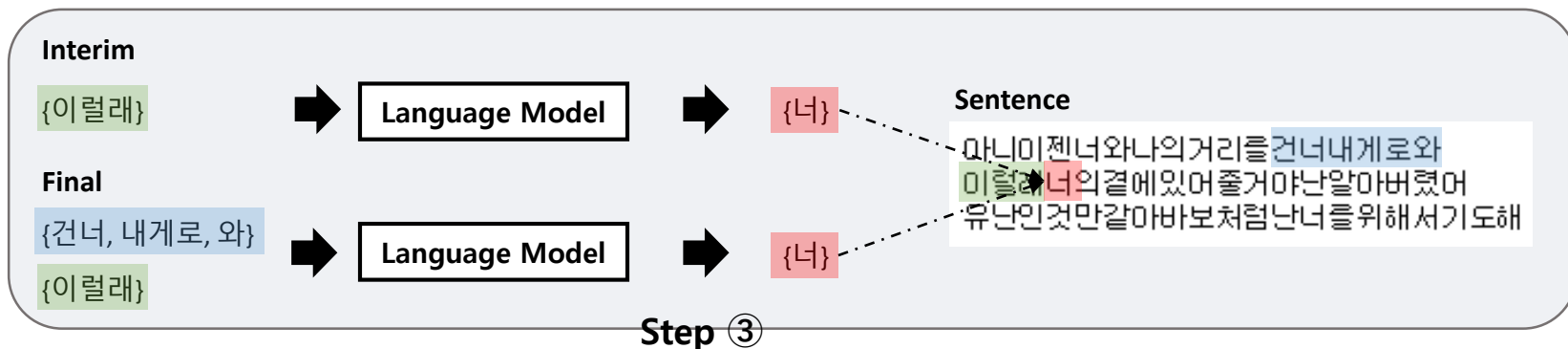
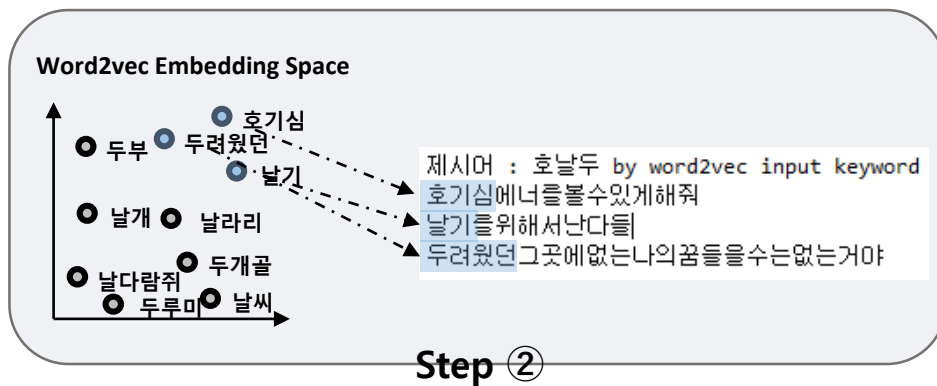
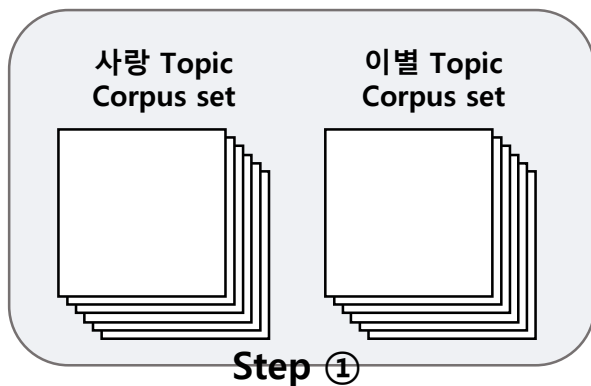
$$\text{sim}(\text{날기}, \text{두려웠던}) > \text{sim}(\text{날기}, \text{두부}) > \text{sim}(\text{호기심}, \text{두개골})$$

# 02 | Process

## 2.2 Sentence Generation

### ➤ Sentence Similarity : Final

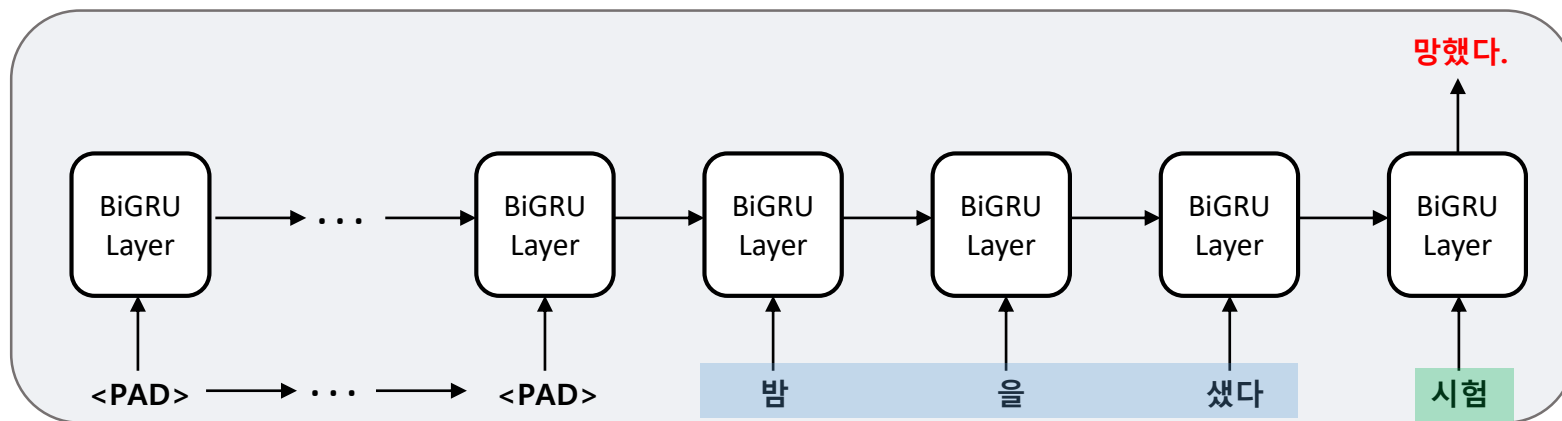
- Step ① : Topic Modeling으로 학습 Corpus의 유사성을 1차적으로 Filtering
- Step ② : Input Keyword의 Similarity 반영
- Step ③ : 문장 생성 시 앞 문장의 마지막 3개 단어 + 시작 단어를 Input으로 넣어 유사도 반영



## 2.2 Sentence Generation

### ➤ Sentence Similarity : Final – Step ③

- 문장 생성 시 앞 문장의 마지막 3개 단어 + 시작 단어를 Input으로 넣어 유사도 반영
- Exaple) 2<sup>nd</sup> Sentence 생성 Process
  - 1<sup>st</sup> Sentence로 '깜지 쓰느라 **밤을 샀다**' 라는 문장이 생성되었고
  - 2<sup>nd</sup> 시작 단어로 깜지와 유사한 '**시험**'이라는 단어가 선택된 경우



Output = 'Verb' 인 경우 문장 생성 종료 → 문장 출력

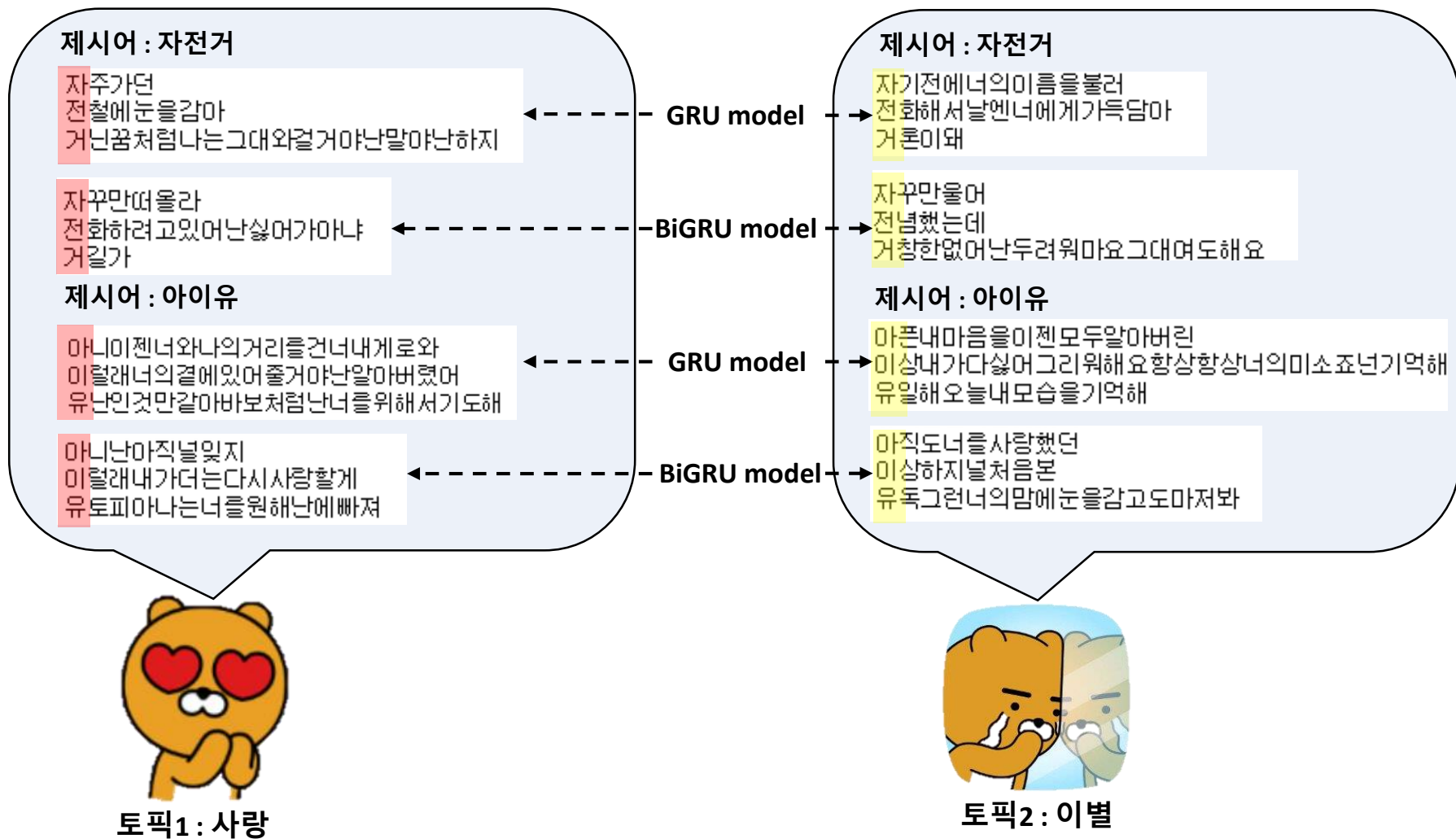
1<sup>st</sup> Sentence : 깜지 쓰느라 밤을 샀다

2<sup>nd</sup> Sentence : 시험 망했다

3<sup>rd</sup> Sentence : ...

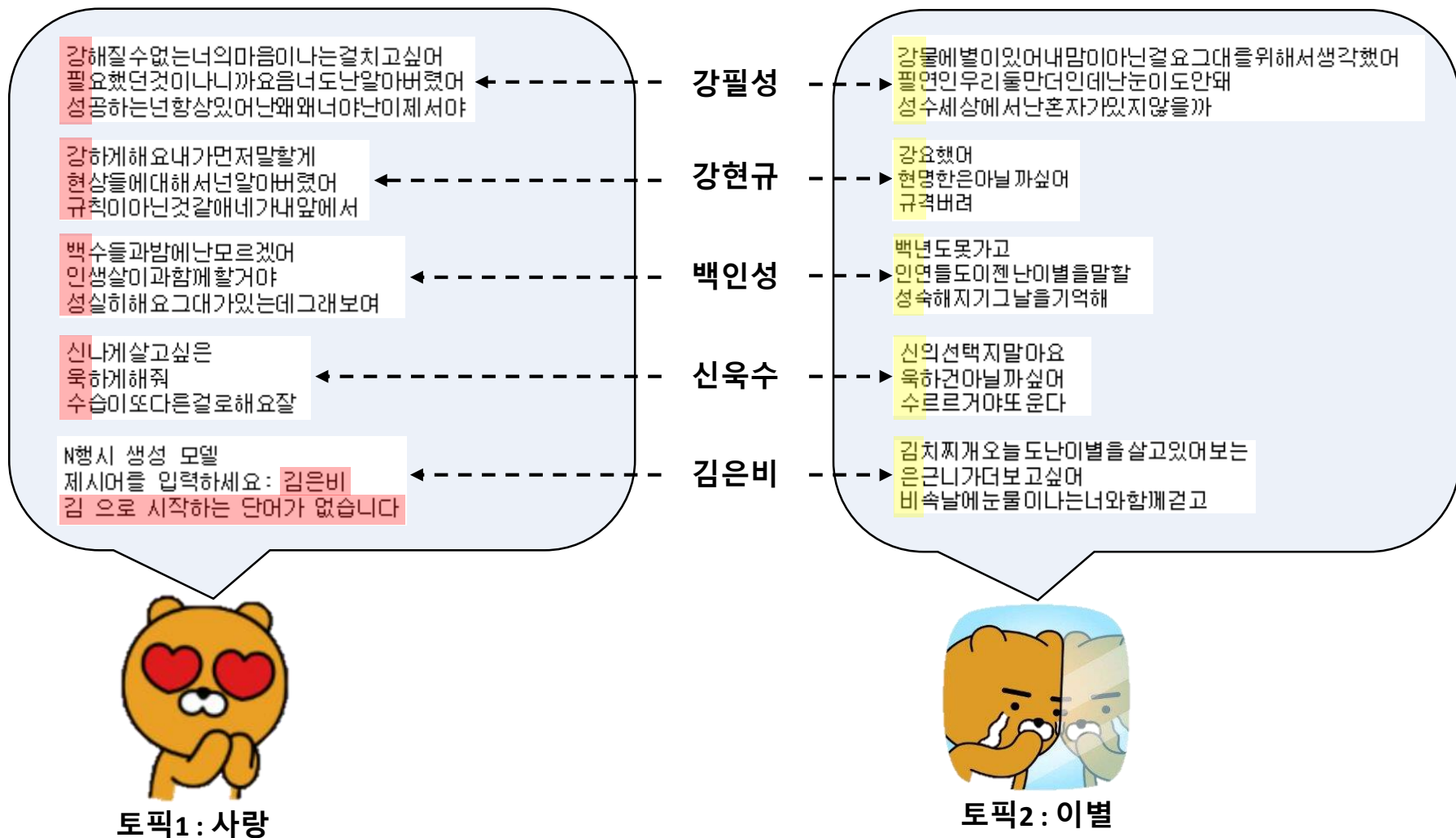
# 02 | Process

## 2.3 Result



# 02 | Process

## 2.3 Result





# 02 | Process

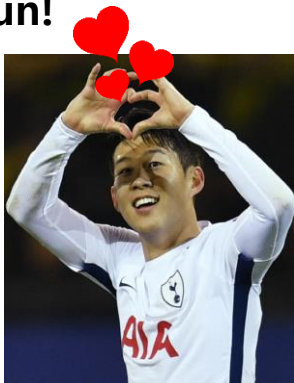
## 2.3 Result

### ➤ 두음법칙 반영한 3행시 생성

제시어 : **립스틱**  
입맞춤과밤을새너를향해  
스르르네가내게  
틱틱내가다알아버렸어

제시어 : **리니지**  
이별해서  
이렇게또생각나  
지지리간다고있는지내가싫어또생각나

### ➤ For fun!



손끝에닿을  
흥분한건나의마음이될  
민감한너의그말에대해서난너를사랑해

손가락질도좋아난좋아난좋아난싫어난싫어더많아더걸어  
흥분한내가있잖아널보며  
민감한때면됐어

손짓눈빛이되어  
흥분한말을줘  
민감한날들이 좋아난널원해또좋아지금 만나

제시어 : **정다래**  
정적이나의눈 속에비친  
다정했던날이 많이사랑해  
내게로다가와

제시어 : **고려대**  
고마워다해줄게  
여린성격이니 까요나는건 없잖아그래도 난싫어있어봐  
대단한거야난거야왜해요 그대와함께할거야



이별마저아름다워보여  
강한척했어  
인해서사람들을봐

이렇게널속이 고네게말해 줘  
강렬했던기억들은날랜잖아난줄수있어 난따라  
인자함이더돈이 좋아난널 위해서음악했어

이별해서  
강해지는걸느껴  
인해다시또생각나

## 03. Conclusion

# 03 | Conclusion

## 3.1 Limitation

- 완전하지 못한 문장 생성 결과에 대한 아쉬움

- EOS token 사용했을 때 문장 생성이 잘 되지 않음

Ex) '사랑' 으로 시작하는 문장을 생성하고 싶는데, 가사에 '사랑' 으로 끝나는 문장이 많아 바로 <EOS> 토큰 출력

- RNN 계열의 모델만을 활용

- Bert와 같은 Pre-Trained 모델 사용 시 더 완성도 있는 문장이 생성되었을 것으로 예상

- 가사 외, 추가 Crawling 데이터 미 반영

- 재미와 완성도를 위해 추가 데이터 확보했으나, 미 반영

- 뉴스와 같은 텍스트 데이터 반영이 더 높은 문장의 완성도를 보여줄 것이라 예상

- 한글 문장 생성 구현과 형태소 분석기의 한계

- 형태소 분석기(w2b, ftxt)의 유사어 추출 결과가 만족스럽지 못한 경우 존재 ex)

첫 번째 시작 단어 = '아침/Noun'  
두 번째 시작 단어(word2vec) = '이쁜데/Adjective'  
두 번째 시작 단어(fasttext) = '이메일/Noun'

- 형태소 분석기 성능의 아쉬움

ex) vocab [30] 자랑할  
'할/Verb' 전형적인네가좋아지금난꿈처럼앞에그대의모습이너무눈엔

- 데이터 셋의 특성상 학습시간이 길어 다양한 hyper-parameter 튜닝 조작이 어려움

Layer (type)	Output Shape	Param #
embedding_19 (Embedding)	(None, 23, 128)	6925312
gru_19 (GRU)	(None, 23, 256)	295680
dropout_19 (Dropout)	(None, 23, 256)	0
time_distributed_19 (TimeDis)	(None, 23, 54104)	13904728

# 03 | Conclusion

## 3.2 Warm-up

- 한학기 동안 수업에서 다룬 여러가지 방법들을 실제 적용 및 학습
  - Crawling, Pos tagging, W2Vec, Fasttext, LDA, GRU, Bidirectional GRU 등 다양한 방법론 수행
  - 학습 결과가 가시적으로 도출되어 model들의 성능을 비교, 이해하기 좋았음
- N행시 문장 생성 모델은 다양한 분야에서 활용 가능
  - 가사 작사, 뉴스 생성, 이름 짓기 등등

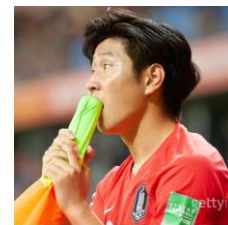
For fun!

손 흥 민



손가락질도 좋아난 좋아난 좋아난 싫어난 싫어 더 많아 더 걸어  
홍분한 내가 있잖아 날 보며  
민감한 때면 됐어

이 강 인



이렇게 널 속이 고네 게 말해 줘  
강렬했던 기억 들은 날 랜 찰아 난 줄 수 있어 난 따라  
인자함이 더 돈이 좋아 난 널 위해서 음악했어

- 프로젝트의 의의는....

# Appendix

# Appendix

## Topic Modeling Result

Topic	0	1	2	3	4	5	6	7	8	9
Naming	사랑	사랑	사랑	-	사랑	놀이	성경	-	사랑	성경
Word1	오늘	날아가	축하	살다	꿈결	도대체	합니다	있어요	따뜻했던	주님
Word2	우리	나비	받아줘	몰라서	좋아해요	감아도	없어요	크리스마스	내리고	라라라
Word3	좋아	만나게	주길	빠졌어	캄캄한	솔로	주님	줘요	날까	하나님
Word4	마음	좋아서	생일	견네	빠져들어	여기저기	잠들지	떠나간	찬란했던	예수
Word5	지금	알겠어	꾸고	않겠지만	데려가줘	함께해	인도	부르고	까요	찬양

Topic	10	11	12	13	14	15	16	17	18	19
Naming	이별	사랑	이별	이별	-	-	이별	이별	사랑	이별
Word1	슬프게	흔들	기억	예뻐서	라라라	입니다	주의	조금	당신	있나요
Word2	기다리지	떠올려	다시	어떡해야	빛속	보고싶다	기다릴게	혹시	기특	괜찮아요
Word3	기다릴게	세계	우리	잊지마	리다	목금	와줘	누가	멋진	아프지
Word4	행복해야	비추는	시간	묻는다	갈을까	주르륵	않니	안아줘	우릴	있어서
Word5	울면	봄바람	눈물	꽃길	불타는	필요없어	그만하자	싶다	벚꽃	숨결

# Appendix

## Topic Modeling Result

Topic	20	21	22	23	24	25	26	27	28	29
Naming	이별	-	-	이별	-	-	이별	-	-	놀이
Word1	행복하길	머물	걷는다	여인	연애	연애	리라	걸어요	나나	하겠어
Word2	숨쉬는	사나이	소망	그리워서	그만해	없었어	흘날리는	쏟아져	내버려	축제
Word3	그리네	있구나	고운	감싸	미안해요	없는걸	파는	잡아요	닭아	행복은
Word4	합시다	가줘	헤어	헤어져	놀자	없단	했어요	두근거리	잊어야	부족해
Word5	않을게요	있었나	나올	나타나	춤추는	민을	울었어	아름다워	고파	거부

Topic	30	31	32	33	34	35	36	37	38	39
Naming	-	이별	사랑	사랑	-	-	-	-	-	-
Word1	너희	인걸	예뻐	루루	회색	완벽해	좋아요	없어요	돼요	봅시다
Word2	소원	떠나가지	해볼까	자유롭게	멈춘	아이스크림	있을까요	퍼고	자유까지	데리러
Word3	할래	지워도	미치겠어	속삭여줘	잠든	되어줄게	소나기	뛰는	없네요	사랑
Word4	가득	애원	두근거리	와줘요	불어오는	사라지는	평화	놀라운	믿어요	추고
Word5	미쳐	있어주세	안아줄래	초콜릿	왔다	돌아오면	부시	불러도	날아올라	아픈데