



# 카페업 소상공인을 위한 AI 챗봇 연구

9 조 김 세 진 , 박 선 홍 , 홍 상 욱



# CONTENTS

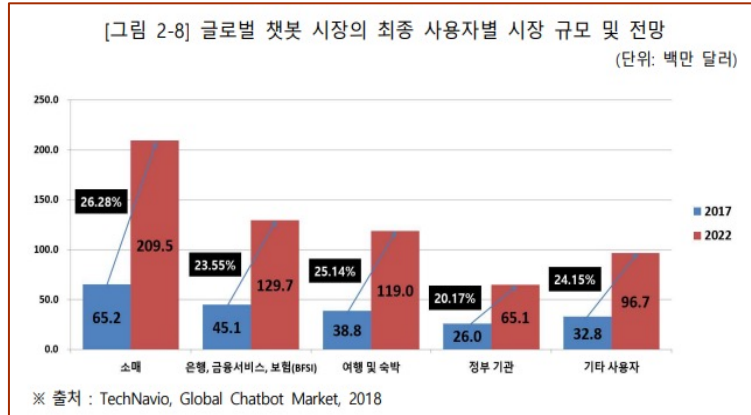
1. 연구 주제

2. 선행 연구

3. 서비스 구현

4. 제안 방법론

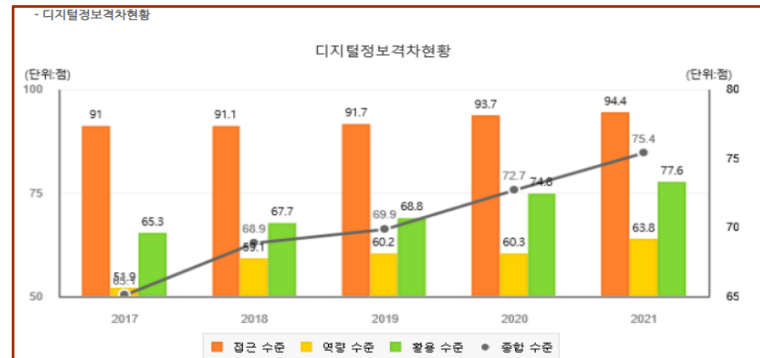
# 01. 연구 주제 - 연구 필요성



"시급 1만원에도 알바 못 구해...더 올려야 하나" 인력난 호소하는 골목상권

업력 2022.05.13. 14:17

안혜림 기자



## 1. 챗봇 활용 지속 증가

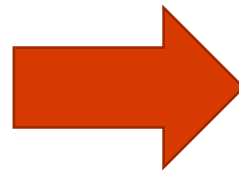
전 세계적으로 챗봇 시장은 점점 증가하는 추세이며, 특히 소매업 시장의 성장이 두드러질 것으로 전망

## 2. 인력난에 빠진 골목 상권

코로나 이후 엔데믹 전환이 되는 과정 속에서 인력부족 심화

## 3. 챗봇 서비스 진입, 디지털 장벽 존재

인력과 자원이 충분한 대기업에 비해 중소 자영업자들의 경우 데이터의 부족과 기술력 부족으로 인해 서비스 진행하지 못함



소상공인을 위한 고객 문의 챗봇 제작

# 01. 연구 주제 - 연구 가치

1

**챗봇을 이용한 신속한 고객응대 가능**

-> 혼자 일하는 소상공인들의 시간과 일손 절감가능

2

**익숙한 메신저 서비스를 이용한 서비스 제공**

-> 추가적인 앱 개발 비용 없이 저렴하게 도입가능

3

**사용자가 원하는 시간에 커뮤니케이션 가능**

-> 사용자는 자신이 편리한 시간에 질문을 보내고 답변을 받을 수 있다

## 02. 선행 연구

<p>韓國IT서비스學誌 第17卷 第3號 2018年 9月, pp.17-36</p>	<p>Journal of Information Technology Services https://doi.org/10.9716/KITS.2018.17.3.017</p>
<p>소상공인들의 고객 문의 데이터를 활용한 문의응대 챗봇의 개발 및 도입</p> <p>신민철* · 김성근* · 이 철**</p> <p>Development of Chatbot Using Q&amp;A Data of SME(Small and Medium Enterprise)</p> <p>Minchul Shin* · Sungguen Kim* · Cheul Rhee**</p>	
<p>■ Abstract ■</p> <p>In this study, we developed a chatbot (Dialogue agent) using small Q &amp; A data and evaluated its performance. The chatbot developed in this study was developed in the form of an FAQ chatbot that responds promptly to customer inquiries.</p> <p>The development of chatbot was conducted in three stages : 1. Analysis and planning, 2. Content creation, 3. API and messenger interworking. During the analysis and planning phase, we gathered and analyzed the question data of the customers and extracted the topics and details of the customers' questions. In the content creation stage, we created scenarios for each topic and sub-items, and then filled out specific answers in consultation with business owners. API and messenger interworking is KakaoTalk.</p> <p>The performance of the chatbot was measured by the quantitative indicators such as the accuracy that the chatbot grasped the inquiry of the customer and correctly answered, and then the questionnaire survey was conducted on the chatbot users. As a result of the survey, it was found that the chatbot not only provided useful information to the users but positively influenced the image of the pension.</p> <p>This study shows that it is possible to develop chatbots by using easily obtainable data and commercial API regardless of the size of business. It also implies that we have verified the validity of the development process by verifying the performance of developed chatbots as well as an explicit process of developing FAQ chatbots.</p> <p>Keyword : Chatbot, Text Mining, Artificial Intelligence, Dialog Agent</p>	
<p>Submitted : May 1, 2018      1<sup>st</sup> Revision : May 8, 2018      Accepted : June 15, 2018</p> <p>* 아주대학교 경영정보학과 석사 ** 아주대학교 경영정보학과 부교수, 교신저자</p>	

### • 시도 방법

항목	내용	사용 모델
데이터 셋	펜션업체 Q&A 1066개	
데이터 분석 방법	형태소 분석	KoNLP
	토픽 분석	LSA
주제 선정	7가지 주제	
답화 작성	300개 질문 & 65개 답변 생성	MS Q&A Maker
API 서버	Flask	Flask

### • 한계점

1. 1066개의 한정된 Q&A를 기반으로 만들어 신뢰성이 떨어짐
2. 정확도는 75%로 높으나, 재이용률은 41%밖에 안됨  
→ 비슷한 답변만 나올 뿐 정확한 답변을 이끌어내지 못함



# 03. 서비스 구현 - 데이터 전처리

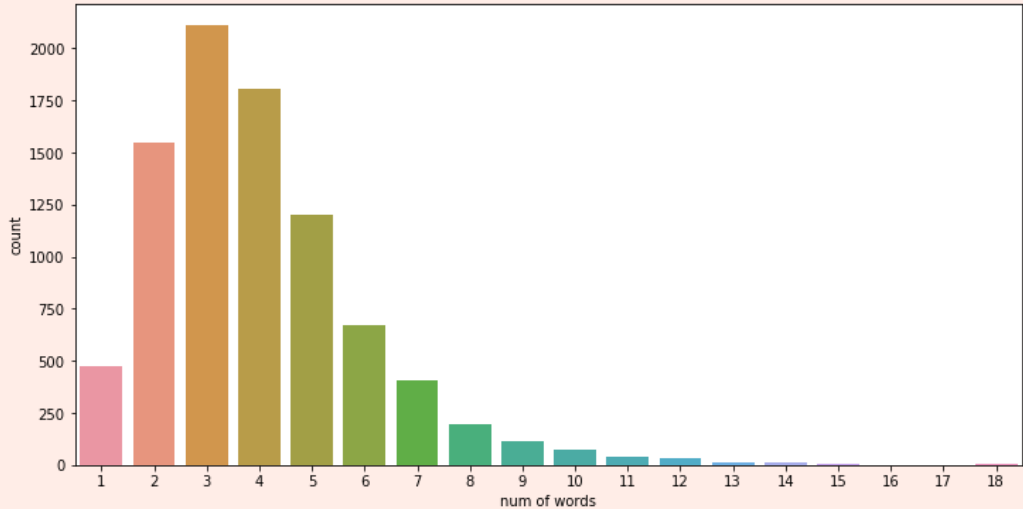
IDX	발화자	카테고리	QA 여부	...	상품명	상당번호	상당내순번
6467	C	음식점	q	...	우유식빵	617	13
6468	S	음식점	a	...	NaN	617	14

출처 : AI 허브 소상공인 고객 주문 질의 응답 텍스트 (음식점)  
 中 카페 관련 데이터  
 데이터 사이즈 : 848 row X 19 columns



SPEAKER	SENTENCE	DOMAINID	DOMAIN	개체명
고객	아이스아메리카노 하나요	F	카페	아이스 아메리카노, 하나
점원	테이크아웃 하실 건가요?	F	카페	테이크아웃

출처 : AI 허브 한국어 데이터 中 카페 카테고리 데이터  
 데이터 사이즈 : 7858 row X 18 columns



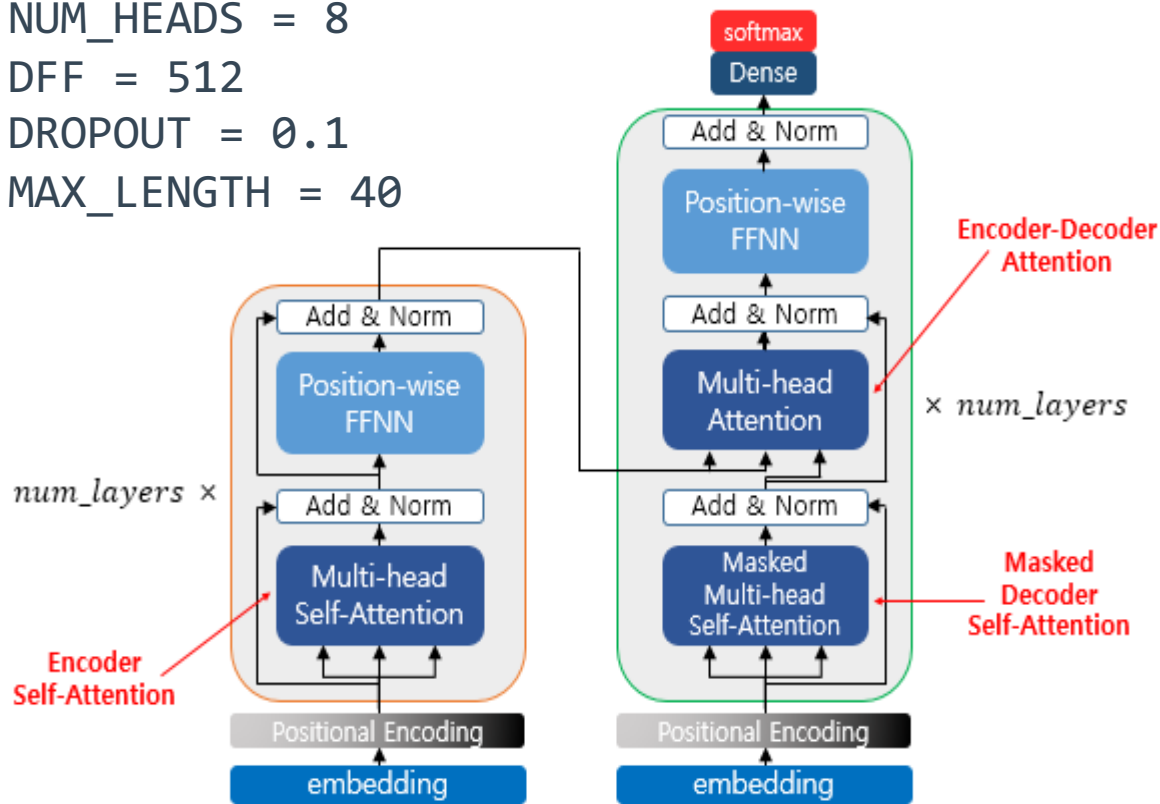
Question	Answer
아이스 아메리카노 하나요	테이크아웃 하실 건가요?
현금영수증 해드릴까요?	괜찮습니다
시럽은 얼마나 뿌려 드릴까요?	3번 뿌려주세요

데이터 사이즈 : 8706 row X 2 columns



# 03. 서비스 구현 - 모델링

NUM\_LAYERS = 2  
D\_MODEL = 256  
NUM\_HEADS = 8  
DFF = 512  
DROPOUT = 0.1  
MAX\_LENGTH = 40



**Self Attention** : 단어 간의 관계 정보를 획득

- Query (Q): 영향을 받는 단어 A를 나타내는 변수
- Key (K) : 영향을 주는 단어 B를 나타내는 변수
- Value (V) : 그 영향에 대한 가중치

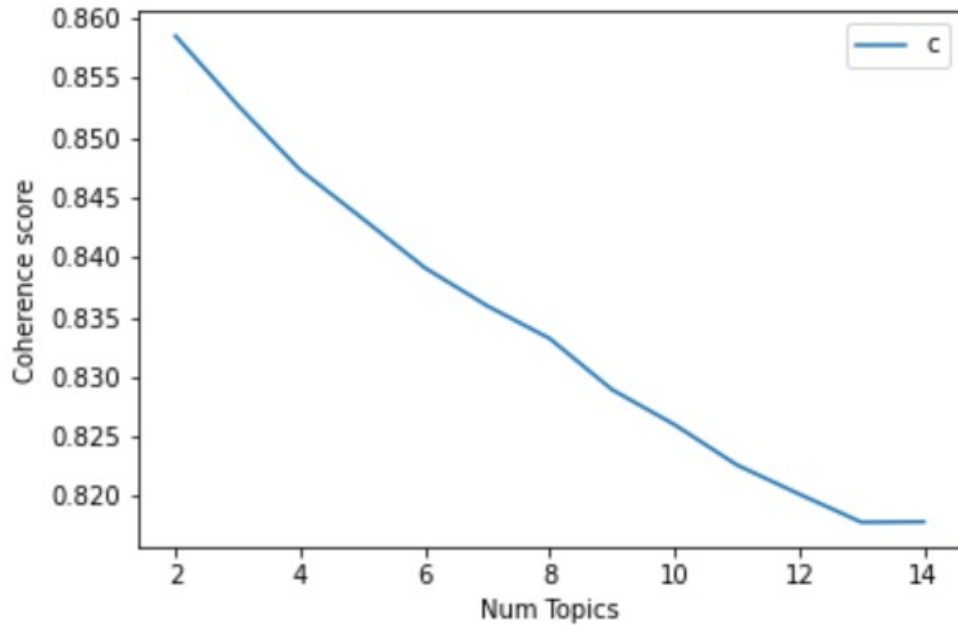
$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

**Add & Norm** : 기울기 소실 방지와 과적합 방지

- Input 값을 다시 Output 값에 더해줌으로써 기울기 소실 문제 방지
- Layer Normalization 실행



### 03. 서비스 구현 - 결과



데이터가 너무 짧아  
TOPIC 모델링 사용 불가능!

	text	Topic_Ida
7	현금영수증 해드릴까요?	1
8	저쪽에서 잠시만 대기해주세요	1
10	시럽은 얼마나 뿌려드릴까요?	1
12	오늘 이벤트 해서 커피를 주문하시면 2000원 추가하시면 브레드를 드립니다	1
13	그럼 2000원 추가해서 브레드도 같이 시킬게요	1
...	...	...
7000	우유식빵 대 얼마예요?	1
7001	우유식빵 소 얼마예요?	1
7015	아이스 홍시는 얼마인가요?	1
7055	아메리카노 뜨거운거 한잔 아이스 한잔 테이크아웃 해주세요	1
7057	아이스티도 두잔 주실래요?	1

## 03. 서비스 구현 및 결과

```
486
487 run(['아메리카노 얼마인가요?'])
```

문제 13 출력 디버그 콘솔 텍스트 JUPYTER

2022-05-06 18:01:17.064805: I tensorflow/core/platform/cpu\_feature\_guard.cc:151] This TensorFlow binary is optimized with oneAPI Deep Neural Network Library (oneDNN) to use the following CPU instructions in performance-critical operations: AVX AVX2  
To enable them in other operations, rebuild TensorFlow with the appropriate compiler flags.  
Input: 아메리카노 얼마 인가요?  
Output: 만육천원이에요

(sesac) C:\Users\82104\Documents\2022년\_1학기\입리닝\프로젝트\_챗봇\Reference\챗봇(transformer)>

줄 487, 열 20 공백: 2 UTF-8 LF Python 3.7.0 (sesac: conda)

### 03. 서비스 구현 - 결과

```
Input: 아메리카노 얼마인가요?  
Output: 삼천원이에요  
Input: 배달 되나요?  
Output: 네 , 배달 가능합니다 .
```



아메리카노 얼마인가요?

삼천원이에요



배달 되나요?

네 , 배달 가능합니다 .



# 03. 서비스 구현 - 결과

2 > 인스턴스 > i-0a02747596c60e78f

**i-0a02747596c60e78f (우분투 서버)에 대한 인스턴스 요약 정보**  
less than a minute 전에 업데이트됨

[🔄](#) [연결](#) [인스턴스 상태 ▼](#) [작업 ▼](#)

인스턴스 ID i-0a02747596c60e78f (우분투 서버)	퍼블릭 IPv4 주소 13.209.76.246   <a href="#">개방 주소법</a>	프라이빗 IPv4 주소 10.0.0.1
IPv6 주소 -	인스턴스 상태 🟢 실행 중	퍼블릭 IPv4 DNS 13.209.76.246
호스트 이름 유형 -	프라이빗 IP DNS 이름(IPv4만 해당) ip-13-209-76-246.ap-northeast-1.compute.amazonaws.com	프라이빗 리소스 DNS 이름 응답 -
인스턴스 유형 t2.micro	탄력적 IP 주소 -	자동 할당된 IP 주소 13.209.76.246 [퍼블릭 IP]
VPC ID vpc-0cc7a411b1e19bf12	AWS Compute Optimizer 찾기 권장 사항을 위해 AWS Compute Optimizer에 옵트인합니다.   <a href="#">자세히 알아보기</a>	IAM 역할 -
서브넷 ID subnet-066c1156464c8b034	Auto Scaling 그룹 이름 -	

[세부 정보](#) | [보안](#) | [네트워킹](#) | [스토리지](#) | [상태 검사](#) | [모니터링](#) | [태그](#)

▼ 인스턴스 세부 정보 정보

플랫폼 Ubuntu(주른)	AMI ID ami-00d806d1ddaa94d5f	모니터링 비활성
플랫폼 세부 정보 Linux/UNIX	AMI 이름 ubuntu/images/hvm-ssd/ubuntu-bionic-18.04-amd64-server-20220411	종료 방지 비활성
중지 방지 비활성	시작 시간 Fri Jun 03 2022 21:35:55 GMT+0900 (한국 표준시)	AMI 위치 099720109477/ubuntu/images/hvm-

```
ubuntu@ip-13-209-76-246:~$ git clone https://github.com/keiraydev/chatbot.git
Cloning into 'chatbot'...
ubuntu@ip-13-209-76-246:~/chatbot$ python3 bot.py
Traceback (most recent call last):
  File "/home/ubuntu/chatbot/bot.py", line 7, in <module>
    from utils.Preprocess import Preprocess
  File "/home/ubuntu/chatbot/utils/Preprocess.py", line 1, in <module>
    from konlpy.tag import Komoran
ModuleNotFoundError: No module named 'konlpy'
ubuntu@ip-13-209-76-246:~/chatbot$ pip install konlpy
Defaulting to user installation because normal site-packages is not writeable
Collecting konlpy
  Downloading konlpy-0.6.0-py2.py3-none-any.whl (19.4 MB)
    19.4/19.4 MB 30.3 MB/s eta 0:00:00
Collecting numpy>=1.6
  Downloading numpy-1.22.4-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.manylinux2019_x86_64.whl (16.8 MB)
    16.8/16.8 MB 18.6 MB/s eta 0:00:00
Collecting JPype1>=0.7.0
  Downloading JPype1-1.4.0-cp310-cp310-manylinux_2_12_x86_64.manylinux2010_x86_64.manylinux2014_x86_64.whl (465.1 kB)
    465.1/465.1 KB 43.5 MB/s eta 0:00:00
Collecting lxml>=4.1.0
  Downloading lxml-4.9.0-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.manylinux2019_x86_64.whl (6.9 MB)
    6.9/6.9 MB 37.3 MB/s eta 0:00:00
Installing collected packages: numpy, lxml, JPype1, konlpy
WARNING: The scripts f2py, f2py3 and f2py3.10 are installed in '/home/ubuntu/.local/bin' which is not on PATH.
Consider adding this directory to PATH or, if you prefer to suppress this warning, use --no-warn-script-location.
Successfully installed JPype1-1.4.0 konlpy-0.6.0 lxml-4.9.0 numpy-1.22.4
ubuntu@ip-13-209-76-246:~/chatbot$ python3 bot.py
Traceback (most recent call last):
  File "/home/ubuntu/chatbot/bot.py", line 8, in <module>
    from models.intent.IntentModel import IntentModel
  File "/home/ubuntu/chatbot/models/intent/IntentModel.py", line 1, in <module>
    import tensorflow as tf
ModuleNotFoundError: No module named 'tensorflow'
ubuntu@ip-13-209-76-246:~/chatbot$ pip install tensorflow-cpu
Defaulting to user installation because normal site-packages is not writeable
Collecting tensorflow-cpu
  Downloading tensorflow_cpu-2.9.1-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.manylinux2019_x86_64.whl (207.5 MB)
    207.5/207.5 MB 1.9 MB/s eta 0:00:00
Collecting astunparse>=1.6.0
  Downloading astunparse-1.6.3-py2.py3-none-any.whl (12 kB)
Collecting protobuf<3.20,>=3.9.2
  Downloading protobuf-3.19.4-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.manylinux2019_x86_64.whl (1.1 MB)
    1.1/1.1 MB 15.3 MB/s eta 0:00:00
Collecting keras-preprocessing>=1.1.1
  Downloading Keras_Preprocessing-1.1.2-py2.py3-none-any.whl (42 kB)
    42.6/42.6 KB 8.6 MB/s eta 0:00:00
Collecting keras<2.10.0,>=2.9.0rc0
  Downloading keras-2.9.0-py2.py3-none-any.whl (1.6 MB)
    1.6/1.6 MB 21.2 MB/s eta 0:00:00
Collecting tensorboard<2.10,>=2.9
  Downloading tensorboard-2.9.0-py3-none-any.whl (5.8 MB)
    5.8/5.8 MB 43.2 MB/s eta 0:00:00
Collecting grpcio<2.0,>=1.24.3
  Downloading grpcio-1.46.3-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.manylinux2019_x86_64.whl (4.4 MB)
    4.4/4.4 MB 27.4 MB/s eta 0:00:00
Collecting termcolor>=1.1.0
  Downloading termcolor-1.1.0.tar.gz (3.9 kB)
  Preparing metadata (setup.py) ... done
```

AWS EWC2 서버 활용  
Git clone으로 모델 불러오기  
필수 요소 설치(서버 내)

## 03. 서비스 구현 - 결과

아메리카노

제가 할 수 있는 일이 아니에요

아메리카노 한잔 얼마

아메리카노 한잔 삼천원 입니다.

배달

이해하기 어려워요

배달 돼?

네, 배달 됩니다.

배달 가능

제가 할 수 있는 일이 아니에요

### 트랜스포머의 한계:

- 모델이 문제를 풀기에는 너무 크다!
- 충분한 크기의 데이터가 아니면 학습이 잘 안된다

## 04. 제안 방법론 – 문제 정의 및 가정

- 연구가설
  1. 닫힌 대화 챗봇 모델링을 위한 SBERT를 사용하면 성능 향상 가정
  2. 트랜스포머 파라미터 조정하면 성능 향상 가정
  3. 모델 학습 방법 변경하면 성능 향상 가정
  4. 데이터 전처리 유무에 따른 성능 차이 가정
- 평가지표  
F – Score

연구 가설	내용
1. 토픽 모델링을 위한 SBERT를 사용하면 성능 향상	(1) SBERT 사용
2. 트랜스포머 파라미터 조정하면 성능 향상 가정	(1) Head 개수 조절
3. 모델 학습 방법 변경하면 성능 향상 가정	(1) Bart Sentence Corrupt (2) BERT
4. 데이터 전처리 유무에 따른 성능 차이 가정	(1) 형태소 분리 (2) 형태소 분리 + TAG (3) 형태소 원형 복원 + 형태소 분리 + TAG

## 04. 제안 방법론 – SENTENCEBERT

- Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks (EMNLP 2019)

인용 분야	인용 기법	한계점 및 차이점
Modeling	BERT / RoBERT	semantic textual similarity(STS)와 같은 sentence-pair regression tasks에서 state-of-the-art 성능을 보임 하지만 이런 모델들은 input sentence pair가 한번에 feeding 되어야 한다는 단점이 있음
	BERT를 siamese and triplet network 형태로 바꾼 Sentence-BERT(SBERT)를 제안	이런 네트워크 구조는 문장의 의미를 sentence embedding이 효과적으로 표현할 수 있게 해주며, cosine-similarity를 통해 쉽게 유사도를 계산할 수 있게 해줌

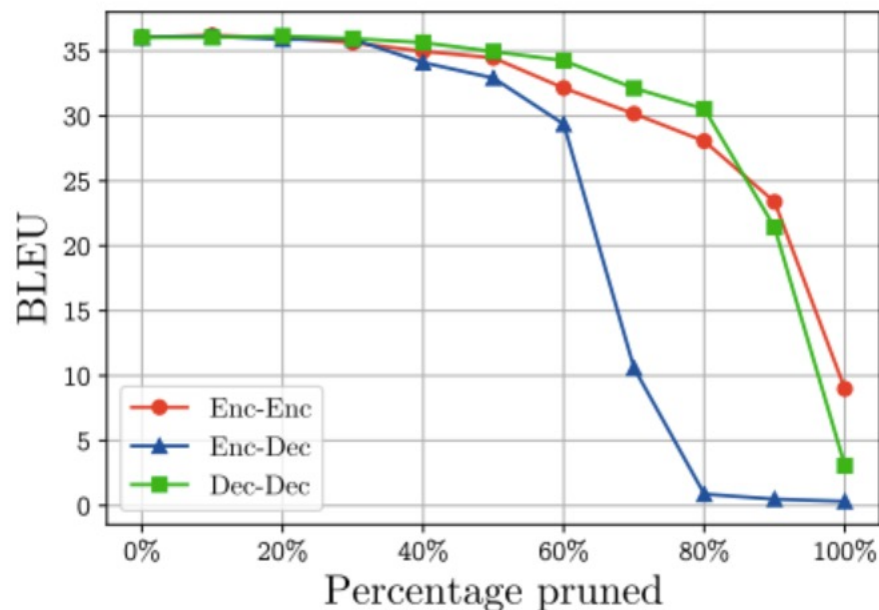


# 04. 제안 방법론 – Parameter 조정

Are Sixteen Heads Really Better than One?

Head 개수 조정 : Transforemr의 Head 개수에 따라 성능에 차이가 있을 수 있다.

논문에 따르면 Head 개수를 제거한 Layer들에서 성능의 유의미한 하락은 없지만 혹시 적은 데이터에서는 어떻게 될까?



Layer	Enc-Enc	Enc-Dec	Dec-Dec
1	<u>-1.31</u>	<u>0.24</u>	-0.03
2	-0.16	0.06	0.12
3	0.12	0.05	0.18
4	-0.15	-0.24	0.17
5	0.02	<u>-1.55</u>	-0.04
6	<u>-0.36</u>	<u>-13.56</u>	0.24

Table 2: Best delta BLEU by layer when only one head is kept in the WMT model. Underlined numbers indicate that the change is statistically significant with  $p < 0.01$ .

Layer	Layer
1	-0.01%
2	0.10%
3	-0.14%
4	-0.53%
5	-0.29%
6	-0.52%
7	0.05%
8	-0.72%
9	-0.96%
10	0.07%
11	-0.19%
12	-0.12%

Table 3: Best delta accuracy by layer when only one head is kept in the BERT model. None of these results are statistically significant with  $p < 0.01$ .

<각 레이어에서 head를 하나만 남겨둔 채 추론했을 때 베스트 결과를 기재함>

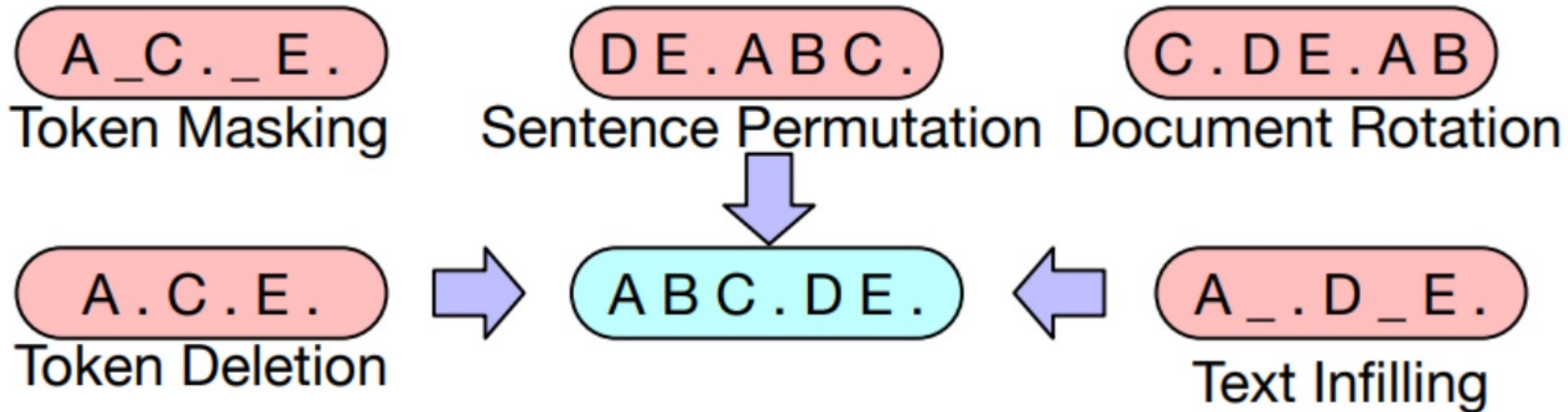
Head \ Layer	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	0.03	0.07	0.05	-0.06	0.03	<u>-0.53</u>	0.09	<u>-0.33</u>	0.06	0.03	0.11	0.04	0.01	-0.04	0.04	0.00
2	0.01	0.04	0.10	<u>0.20</u>	0.06	0.03	0.00	0.09	0.10	0.04	<u>0.15</u>	0.03	0.05	0.04	0.14	0.04
3	0.05	-0.01	0.08	0.09	0.11	0.02	0.03	0.03	-0.00	0.13	0.09	0.09	-0.11	<u>0.24</u>	0.07	-0.04
4	-0.02	0.03	0.13	0.06	-0.05	0.13	0.14	0.05	0.02	0.14	0.05	0.06	0.03	-0.06	-0.10	-0.06
5	<u>-0.31</u>	-0.11	-0.04	0.12	0.10	0.02	0.09	0.08	0.04	<u>0.21</u>	-0.02	0.02	-0.03	-0.04	0.07	-0.02
6	0.06	0.07	<u>-0.31</u>	0.15	-0.19	0.15	0.11	0.05	0.01	-0.08	0.06	0.01	0.01	0.02	0.07	0.05

<기계번역 BLUE 스코어, baseline = 36.05 / p<0.01로 유의한 경우 밑줄 표시>

## 04. 제안 방법론 - 학습 방법

Bart - Sentence Corrupt

학습 방법 : 모델로 하여금 학습을 어렵게 만들면 학습이 잘 될 것이다.



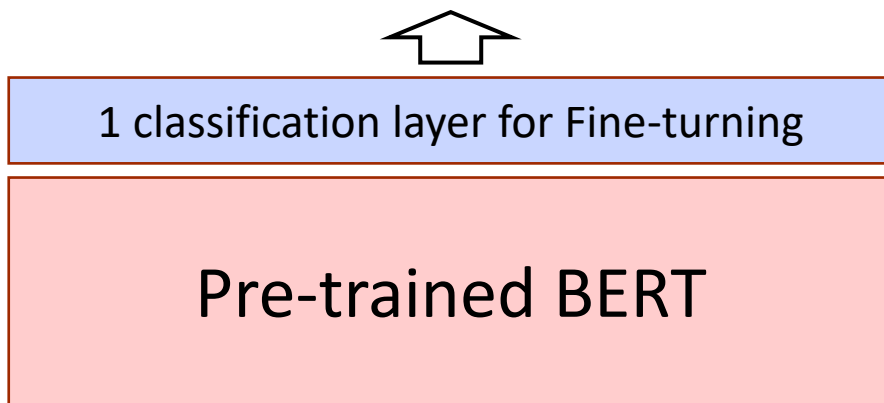
- Token Masking : 랜덤한 토큰이 샘플링 되어 [MASK] 토큰으로 치환 -> [MASK]토큰을 맞춰야 함
- Sentence Permutation : 마침표를 기준으로 문장으로 분류 후 섞임 -> 원래의 순서를 맞춰야 함
- Document Rotation : 랜덤한 토큰이 문서의 시작점이 됨 -> 원래 문서의 시작점을 찾아야 함
- Text Infilling : Span Length를 추출한 길이 만큼의 토큰을 하나의 [MASK]로 토큰화-> 토큰의 개수를 맞춰야 함
- Token Deletion : 랜덤한 토큰을 제거 -> Input의 없어진 위치와 토큰을 맞춰야 함

## 04. 제안 방법론 – BERT

학습 방법: Bi-Directional(양방향)로 학습하면 성능이 향상될 것이다. -> BERT

- BERT는 bi-directional Transformer로 이루어진 언어모델
- 잘 만들어진 BERT 언어모델 위에 1개의 classification layer만 부착하여 다양한 NLP task를 수행
- 영어권에서 11개의 NLP task에 대해 state-of-the-art (SOTA) 달성

"Jimi" Hendrix was an American rock guitarist, singer, and songwriter.



....  
James Marshall "Jimi" Hendrix  
was an American rock guitarist,  
singer, and songwriter.  
....

Who is Jimi Hendrix?

SQuAD v1.1 dataset leaderboard

Rank	Model	EM	F1
	Human Performance Stanford University (Rajpurkar et al. '16)	82.304	91.221
1 Oct 05, 2018	BERT (ensemble) Google AI Language <a href="https://arxiv.org/abs/1810.04805">https://arxiv.org/abs/1810.04805</a>	87.433	93.160
2 Oct 05, 2018	BERT (single model) Google AI Language <a href="https://arxiv.org/abs/1810.04805">https://arxiv.org/abs/1810.04805</a>	85.083	91.835

# 04. 제안 방법론 – Data Preprocessing

데이터 전처리: 데이터 전처리 방식이 학습에 영향을 미칠 것이다.

- BERT 성능에 영향을 미치는 요인

- Corpus 사이즈
- Corpus 도메인
- Corpus tokenizing (어절, BPE, 형태소)
- Vocab 사이즈

전처리	예제
어절 + BPE (22만 vocab)	멕시코 ##부터 브라질 ##까지 그리고 카리브 해 섬 지역에서 발견된다 ##. 8 ##종으로 이루어져 있다.
어절 + BPE (4만 vocab)	
형태소 분리	멕시코 부터 브라질 까지 그리고 카리브 해 섬 지역에서 발견된다. 8 종 으로 이루어져 있다 .
형태소 분리 + 형태소 tag	멕시코/NNP 부터/JX 브라질/NNP 까지/JX 그리고/MAJ 카리브/NNP 해/NNG 섬/NNG 지역/NNG 에서/JKB 발견/NNG 된/XSV 다/EF ./SF 8/SN 종/NNG 으로/JKB 이루어져/VV 있/VX 다/EF ./SF
형태소 분리 + 원형 복원	멕시코 부터 브라질 까지 그리고 카리브 해 섬 지역에서 발견된다 . 8 종 으로 이루어져 이 있다 .
형태소 분리 + 동사 원형 복원	멕시코 부터 브라질 까지 그리고 카리브 해 섬 지역에서 발견된다 . 8 종 으로 이루어지 다 있다 .



- BERT 학습

- 코퍼스: 6.3GB, 약 6.7억 어절
- 배치 사이즈: 128
- 학습 steps:  $\approx 400,000$

- Vocab size:  $\approx 30,000$

- KorQuAD로 평가

모델	전처리	Tokenizing	Exact matching	F1
BERT <sub>base</sub>	X	Word piece	67.23	86.39
BERT <sub>morp</sub>	형태소 분리	형태소 단위	84.46	92.13
BERT <sub>morp+wp</sub>	형태소 분리	Word piece	<b>85.94</b>	93.35
BERT <sub>morp+tag</sub>	형태소 분리 + Tag	형태소 단위	81.81	91.86
BERT <sub>morp_r+tag+wp</sub>	형태소 원형복원 + 형태소 분리 + Tag	Word piece	83.62	<b>93.60</b>

## 04. 실험 및 결과 - 시험 시나리오 결과

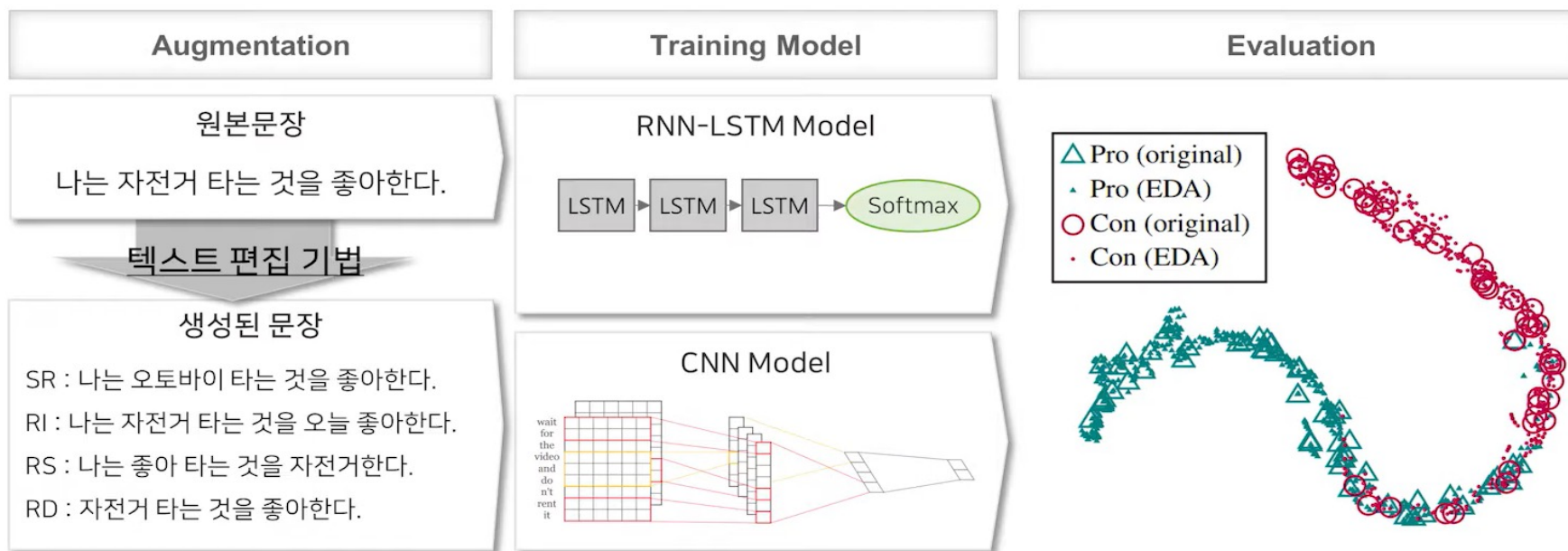
분야	제안 기법	실험 시나리오
Preprocessing	형태소 처리	(1) 사용 X (2) 형태소 분리 (3) 형태소 분리 + TAG (4) <b>형태소 원형 복원 + 형태소 분리 + TAG</b>
Modeling	파라미터 조절	(1) <b>파라미터 조절</b> (2) 사용하지 않음
Modeling	SBERT	(1) BERT (2) RoBERT (3) <b>SBERT</b> (4) Bart

## 04. 실험 및 결과 - 개선 및 제언 방향

### [ TEXT AUGMENTATION ]

1. Synonym Replacement(SR) : 특정 단어를 유의어로 교체
2. Random Insertion(RI) : 임의의 단어를 삽입
3. Random Swap(RS) : 문장 내 임의의 두 단어 위치를 변경
4. Random Deletion(RD) : 임의의 단어를 삭제

- 
1. 데이터가 적을 때, 전체 데이터를 사용한 것과 비슷한 정확도를 가지며,
  2. 데이터가 많을 때에도 미미하게 성능 향상



# 참고자료

1. <https://arxiv.org/pdf/1810.04805.pdf>  
J Devline etc. (2018). pre-training of deep bidirectional transformers for language understanding
2. <https://arxiv.org/pdf/1905.10650.pdf>  
Paul Michel etc. (2019). Are Sixteen Heads Really Better than One?
3. <https://arxiv.org/pdf/1908.10084.pdf>  
Nils Reimers and Iryna Gurevych (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks
4. <https://arxiv.org/pdf/1706.03762.pdf>  
Ashish Vaswani etc. (2017). Attention is all you need



감사합니다!

