

자연어 처리(NLP)를 이용한 인천대 챗봇

인
채
202002961 김현주(팀
장)
202001529 유다현(팀
원)

1.

NEED

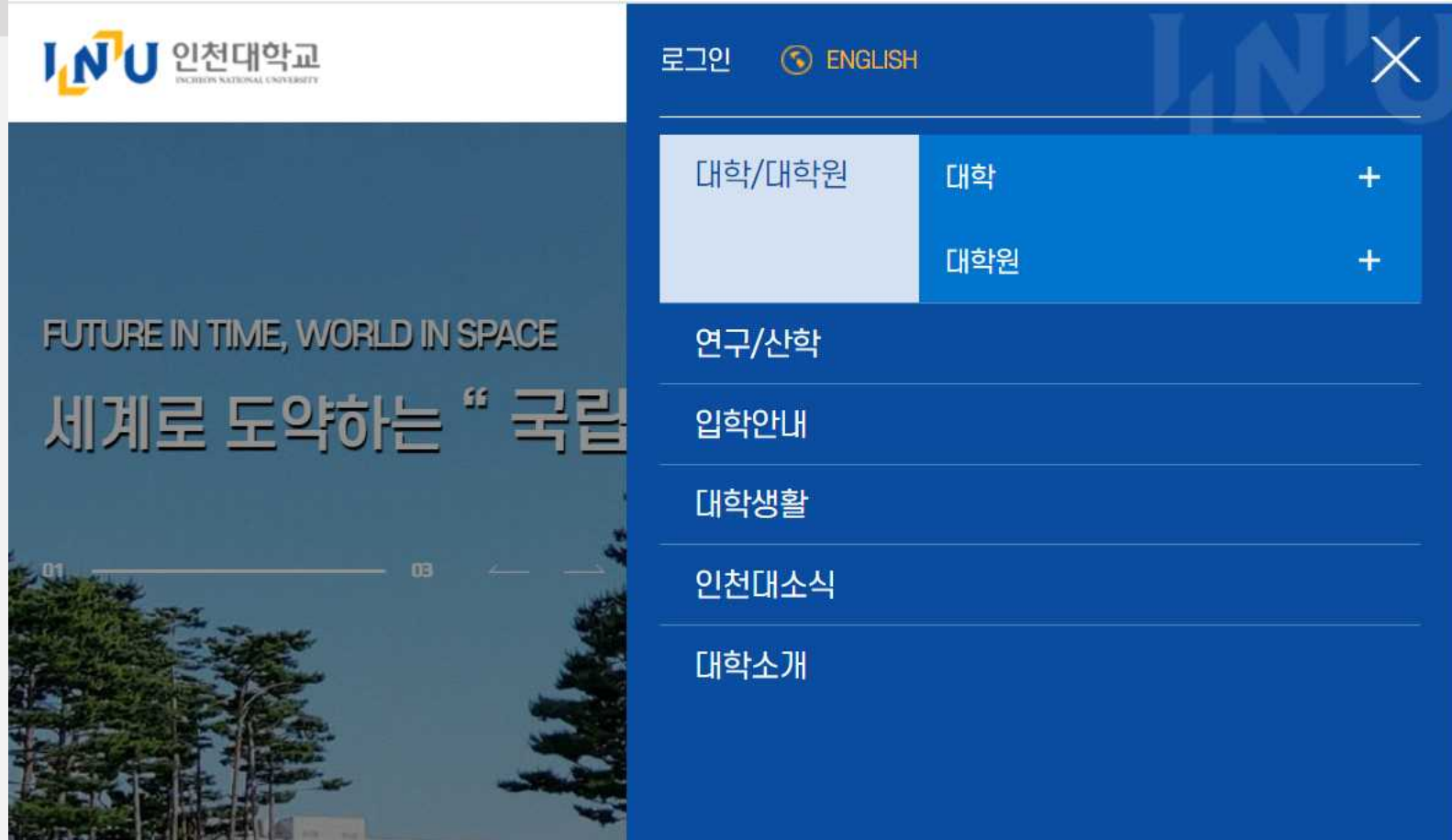


1. NEED

불편함, 요구사항

- 불필요한 웹서핑 시간 소요
- 인천대에 대한 정보를 빠르게 찾고 싶다
- 특정시간 이후에는 문의 전화가 불가능
- 전화 기피하는 사람 증가

1. NEED



2.

APPROA

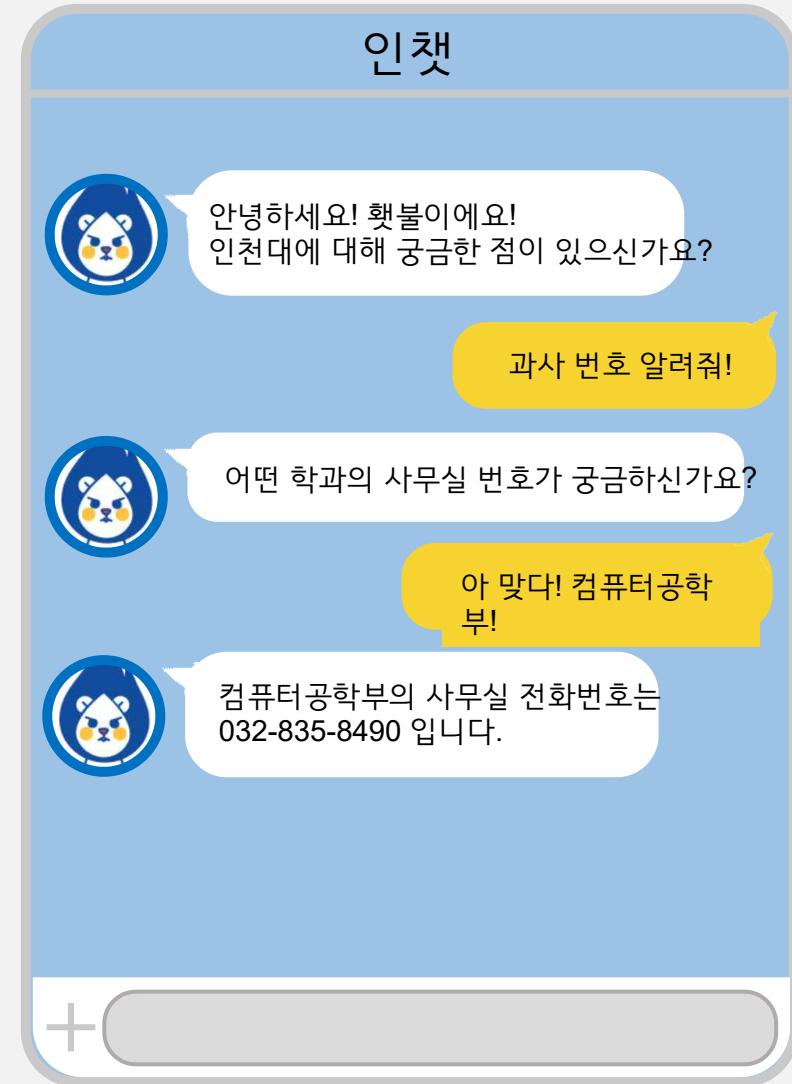
CH

...

2. APPROACH

불편함 해소 방법

- 딥러닝 기반의 챗봇 어플
- 질문에 답변만 주는 챗봇이 아닌
문서를 검색해 기계를 통해 답변을 주는 챗봇



2. APPROACH

기계독해 (MRC, Machine Reading Comprehension)

- 기계가 사람처럼 문서를 읽고 이해할 수 있게 하는 기술
- 질문과 질문의 정답이 포함된 문서에서 정답의 위치를 파악
- 지문(Passage)-질문(question)-답변(Answer) 데이터셋 구축 필요
- 데이터가 적으면 성능이 좋지 않다

2.

APPROACH

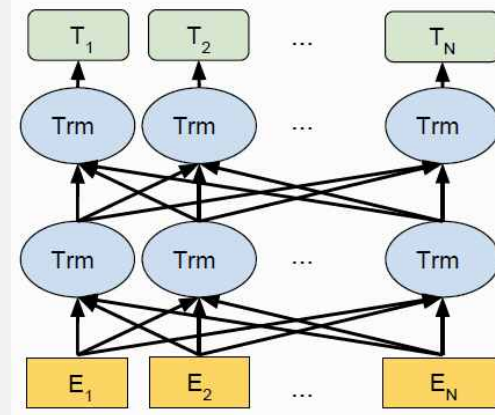
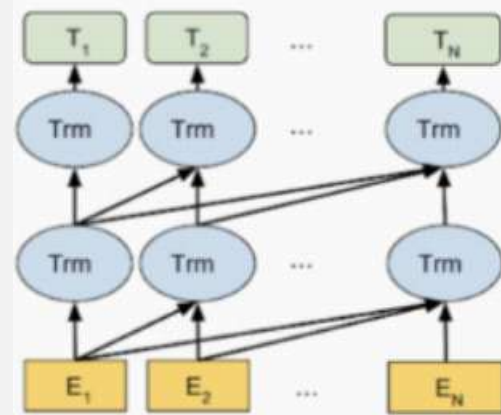
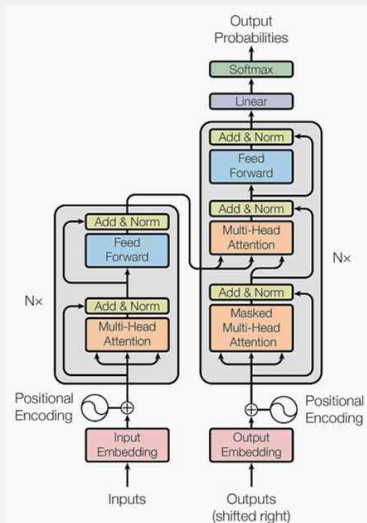
언어 처리 모델

Transformer

GPT2

BERT

MODEL



INPUT

Question

Word

Question, Paragraph

OUTPUT

Answer

Paragraph

Answer

2.

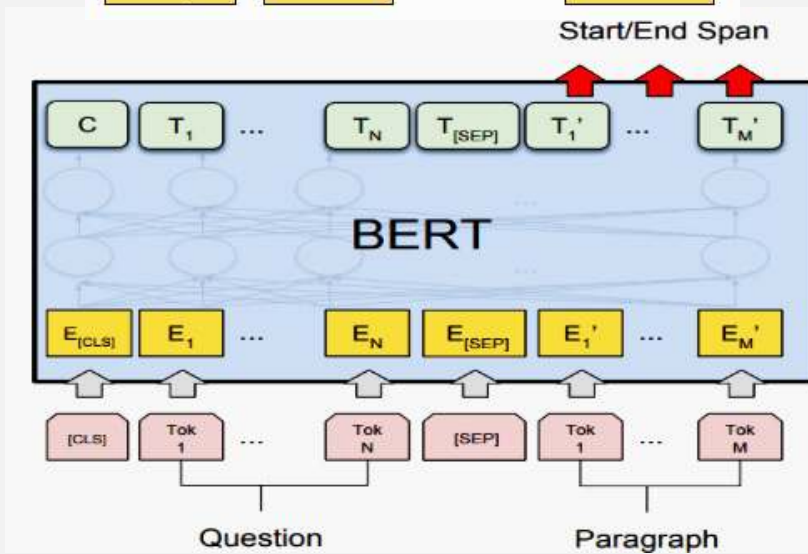
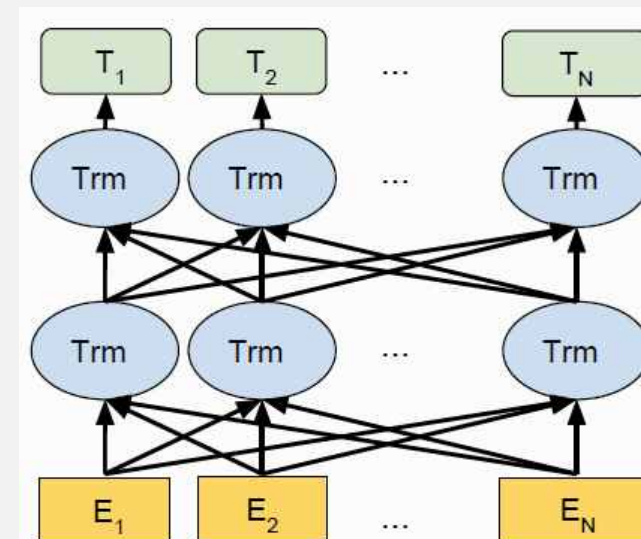
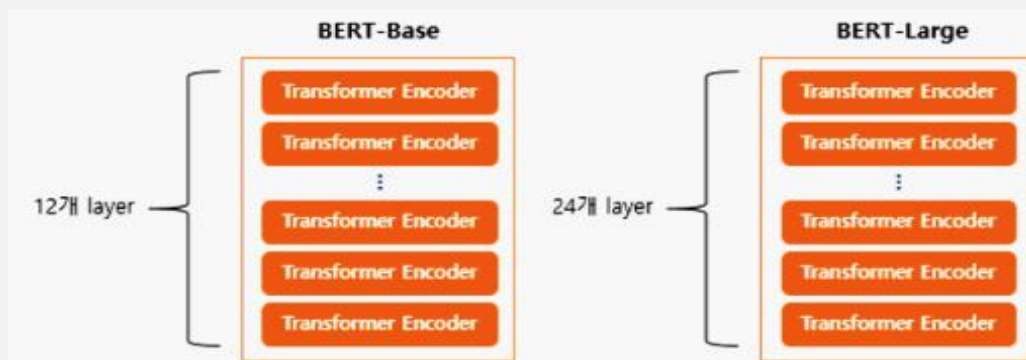
언어 처리 모델

BERT

(Bi-directional Encoder Representations from

Transformers)

- 구글의 오픈 소스
- 텍스트를 양방향으로 확인하여 자연어 처리하는 모델
- 인간보다 더 높은 정확도를 보임
- Transformer의 encoder를 여러 개 쌓아 올린 구조
- 다음 문장 예측과 마스킹 방법으로 학습



2.

APPROACH

언어 처리 모델

한국어 언어처리모델 KorBERT

- ETRI에서 개발한 한국어의 특성을 반영해 개발한 BERT 언어모델
- 구글보다 많은 한국어 데이터를 기반으로 개발
- 구글의 BERT보다 평균 4.5% 성능 우수
- 한국어에 맞게 어절 단위가 아닌 형태소 단위로 wordpiece

Tokenizing

	(Google) Word Piece 기반 한국어 언어 모델	(ETRI) Word Piece 기 반 한국어 언어 모델	(ETRI) 형태소 기반 한 국어 언어 모델
EM	80.82 %	80.70 %	86.40 %
F1	90.68 %	91.91 %	94.18 %

2. APPROACH

CdQA 파이프라인 (Closed-domain Question on Answering)

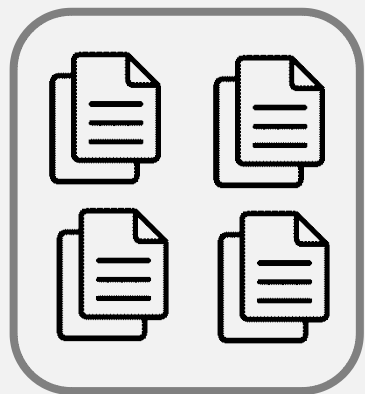
- 도메인 특정 질의응답
- 질문과 문서의 유사도를 기준으로 문서를 선정하여 질문에 대한 답을 추론
- INPUT과 OUTPUT이 “문장과 문장” 형식으로 챗봇에 활용
- 시스템 상에 답변에 대한 정확도 포함
- <https://github.com/cdqa-suite/cdQA>

2.

CdQA 파이프라인

컴퓨터 공학부 졸업 학점 알려줘!

1. 사용자 질문에 대한 TF/IDF 벡터 계산 (TF/IDF : 유사도 계산의 위한 가중치)



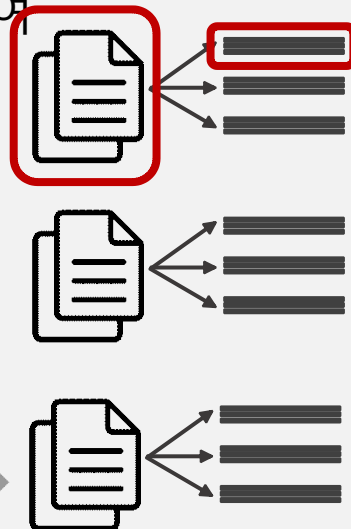
문서들의 집합

Retriever

검색 단계

2. 질문과 문서간의 유사도를 계산

3. 유사도 점수 부여



4. 고득점 문서 및 문장 선택

Reader

독해 단계
(KorBERT 모델)



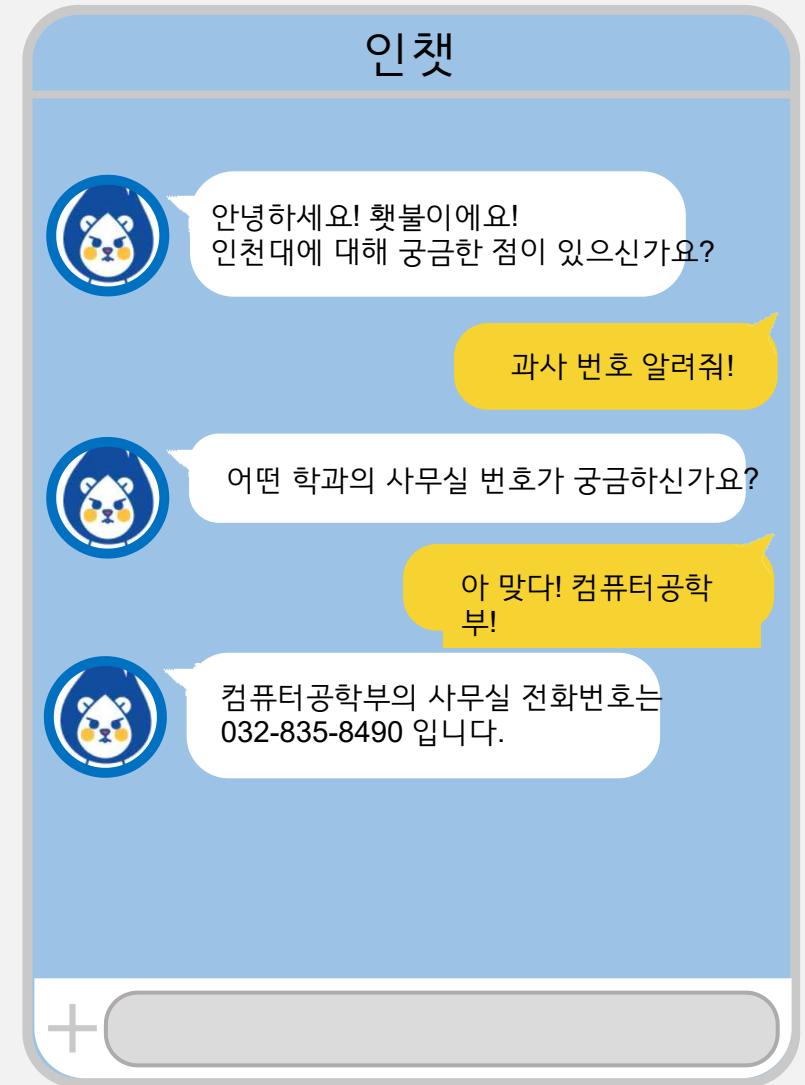
컴퓨터 공학부의 졸업학점은 총 140학점이며,
30~55학점의 교양 학점과
72학점 이상의 전공 학점이 필요합니다.

2. APPROACH

추가 기능

기억 기능 -> slot filling

- 모든 슬롯이 다 채워질 때까지 질문
- 보통 시나리오 기반의 챗봇에서 사용



2.

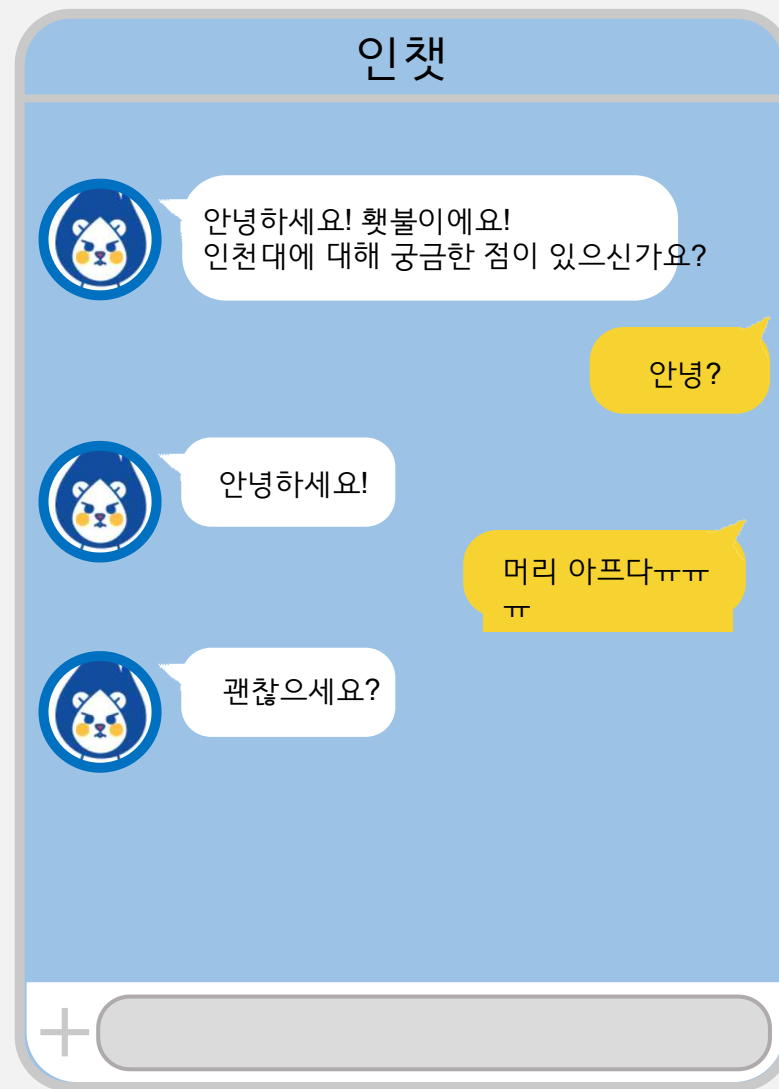
APPROACH

추가 기능

일반적인 질의응답 -> ETRI의 WIKIQA

API
유사도가 1 미만일 경우 WIKIQA API 이용

- 위키백과 문서에서 정답과 신뢰도 및 검색 단락을 추론하여 제공하는 API
- Open domain의 잡담 기능 추가



2. APPROACH

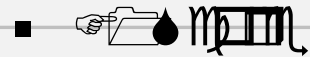
기계독해 성능평가


- EM(Exact Match)

- 예측값과 정답이 완전히 똑같은 경우에만 1점, 하나라도 다른 경우 0점

$$\frac{\text{부여맞은 sample 수}}{\text{전체 sample 수}}$$

-



-  예측값과 실제 정답 간의 정밀도와 재현율의 조화평균

- 0점과 1점 사이의 부분점수를 받을 수 있음

2.

CdQA 파이프라인에 서 제공하는 Evaluating model


Evaluating models

In order to evaluate models on your custom dataset you will need to annotate it. The annotation process can be done in 3 steps:

1. Convert your pandas DataFrame into a json file with SQuAD format:

```
from cdqa.utils.converters import df2squad

json_data = df2squad(df=df, squad_version='v1.1', output_dir='.', filename='dataset-name')
```




2. Use an annotator to add ground truth question-answer pairs:

Please refer to our [cdQA-annotator](#) , a web-based annotator for closed-domain question answering datasets with SQuAD format.

3. Evaluate the pipeline object:

```
from cdqa.utils.evaluation import evaluate_pipeline


evaluate_pipeline(cdqa_pipeline, 'path-to-annotated-dataset.json')
```



4. Evaluate the reader:

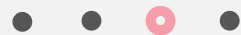
```
from cdqa.utils.evaluation import evaluate_reader

evaluate_reader(cdqa_pipeline, 'path-to-annotated-dataset.json')
```




3.

BENEFIT



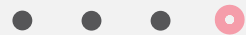
3. BENEFIT

이익

-  시간 운영
- 시간과 장소에 구애받지 않는다
- 불필요한 웹서핑 시간 감소
- 단순 UI를 이용하여 누구나 이용 가능, 편리

4.

COMPETITI ON



4.

COMPETITION

인챗



인혁이

기계독해 방식, 기억방식

시나리오 방식

안드로이드 어플

웹

(입)재학생들이 정말 필요한 정보

외국인 (신)입학생 정보

[c2] cupmo, 2023-10-12

기계독해:

AI 애플리케이션이 스스로 문제를 분석하고 질문에; 최적화된 답안을 찾아내는 기술

[c1] cupmo, 2023-10-12

시나리오 방식이란?

고객이 주어진 상황에 맞는 질문을하면 그에 대한 답변을 한다.

한정적임

질문의 형식마저 정해져있어야지 그에 대한 답을 한다.

인챗

c11



안녕하세요! 햇불이에요!
인천대에 대해 궁금한 점이 있으신가요?

과사 번호 알려줘!



어떤 학과의 사무실 번호가 궁금하신가요?

아 맞다! 컴퓨터공학부!



컴퓨터공학부의 사무실 전화번호는
032-835-8490 입니다.

인혁이

컴퓨터공학부 졸업요건 알려줘

오전 11:54



인혁이

인천대학교는 10개의 단과대학에서 외국인 학생들을 모집합니다

- ◆ 신입생은 생명공학부를 제외한 모든 학과에서 모집
- ◆ 편입생은 생명공학부, 한국화전공, 서양화전공, 체육학부, 운동건강학부를 제외한 모든 학과에서 모집
- ◆ 소비자학과 신입학, 2학년 편입은 글로벌정경대학 소비자학과, 3학년 편입은 자연과학대학 소비자아동학과로 신청해야 함

잘가

오전 11:56



인혁이

제가 잘 모르는 내용이에요 😊
다시한번 질문해 줄래요?

🗣️ 다른 언어로 변경해 대화 할 수도 있어요!

If you want to change the language, click the button in the top left.

[c11] cupmo, 2023-10-15

문장으로 질문은 가능하지만 그 문장이 챗봇에서 예시를 준 문장이어야만 답변가능하고 즉흥적인 문장은 답변하지 못합니당

4.

c10

COMPETITION

인챗



기계독해 방식, 기억방식

안드로이드 어플

(입)재학생들이 정말 필요한 정보

INU챗봇



시나리오 방식

카카오톡 플러스 친구

학식, 학교의 날씨 등 소량의 정보

c9

[c10] cupmo, 2023-10-12

기계독해:

AI 얹고리즘이 스스로 문제를 분석하고 질문에; 최적화된 답안을 찾아내는 기술

[c9] cupmo, 2023-10-12

시나리오 방식이란?

고객이 주어진 상황에 맞는 질문을하면 그에 대한 답변을 한다.

한정적임

질문의 형식마저 정해져있어야지 그에 대한 답을 한다.

인챗

c11



안녕하세요! 햇불이에요!
인천대에 대해 궁금한 점이 있으신가요?

과사 번호 알려줘!



어떤 학과의 사무실 번호가 궁금하신가요?

아 맞다! 컴퓨터공학부!



컴퓨터공학부의 사무실 전화번호는
032-835-8490 입니다.

INU챗봇

학식 알려줘

안녕

오전 11:22



INU봇

이해하기 어려워요

오전 11:22

2023년 10월 15일 일요일 >

과사번호알려줘

오후 12:49



INU봇

총 학생회 SNS 링크입니다.
제휴업체를 확인하세요

INSTAGRAM

FACEBOOK

플러스친구

오후 12:49

안녕

오후 12:50



INU봇

무엇을 원하시는지 잘 모르겠어요

오후 12:50

컴퓨터공학부 전화번호 알려줘

오후 12:50

[c11] cupmo, 2023-10-15

문장으로 질문은 가능하지만 그 문장이 챗봇에서 예시를 준 문장이어야만 답변가능하고 즉흥적인 문장은 답변하지 못합니당

발표 들어주셔서 감사합니다
