Art Insight Lab

FNCSP 챗봇 개발 프로젝트

조영훈 김영준 이진균

목차



1. 개발 기획

- 1) 팀소개 및 업무 분담
- 2) 개발 일정
- 3) 기획방향 및 배경
- 4) 도메인 소개

2. 데이터 분석

- 1) Q&A 데이터 생성
- 2) 사용 데이터 분석

3. 챗봇 모델 구현

- 1) Fine-tuning process
- 2) Post-Generation Stage
- 3) Evaluation
- 4) Experiments

4. 결론

- 1) 애로사항
- 2) 향후 발전 방향

Part 1 개발기획



1) 팀소개 및 업무분담

조영훈

- 프로젝트 진행 총괄
- 프로젝트 일정 및 업무 분담 관리
- 룰베이스 알고리즘 개발

김영준

- 사용할 LLM 서치
- 데이터 전처리

이진균

■ 도메인 지식에 대한 리서치

공통

■ Q&A 데이터 생성

Building LLM Fine-Tuning Process

Development on Post-Generation Algorithms

Building Inference system & Demo webpage

■ 데이터증강

2) 개발 일정

| | | | | | | | | | 2023 | 3.08. | .03 ^ | - 20: | 23.0 | 9.07 | 7 | | | | | | | | |
|-------------------------|---------|--|-----|--|--|-----|--|--|------|-------|-------|-------|------|------|---|--|--|--|--|--|--|--|--|
| 프로젝트 일정 | 1주차 2주차 | | 3주차 | | | 4주차 | | | 5주차 | | | | 6주차 | | | | | | | | | | |
| FNCSP 홈페이지 내용 파악 | | | | | | | | | | | | | | | | | | | | | | | |
| 과제 분석 및 기획 | | | | | | | | | | | | | | | | | | | | | | | |
| Q&A 데이터 생성1 | | | | | | | | | | | | | | | | | | | | | | | |
| Q&A 데이터 생성2 | | | | | | | | | | | | | | | | | | | | | | | |
| LLM 파인튜닝 방법 리서치 | | | | | | | | | | | | | | | | | | | | | | | |
| Q&A 데이터를 통한 LMM 파인튜닝 | | | | | | | | | | | | | | | | | | | | | | | |
| 결과 정리 및 발표 | | | | | | | | | | | | | | | | | | | | | | | |

3) 기획 방향 및 배경

프로젝트 명

FNCSP챗봇개발프로젝트

수행 배경

FNCSP 홈페이지 내 Q&A 처리를 담당 인력이 아닌, 챗봇을 이용한 자동화의 필요성을 체감

프로젝트 목표

FNCSP Q&A 데이터를 이용한 생성형 챗봇 개발

활용 데이터

(주)벡스인텔리전스에서 제공한 FNCSP Q&A 데이터 및 생성, 증강한 데이터

4) 도메인 소개

기업의 지속성장을 위한 스마트한 솔루션 **FNCSP**





FNCSP는 기업의 지속성장을 위하여 빅데이터를 기반으로 기업의 과거부터 현재, 미래 가치를 진단하고 신용/특허와 같은 재무정보 뿐만 아니라 기업 미래 성장성을 측정할 수 있도록, 기술기반의 기업 가치인 '8대 성장지수'로 기술력을 지수화 하여 제공해주는 스마트 컨설팅 플랫폼입니다. Part 2 데이터분석



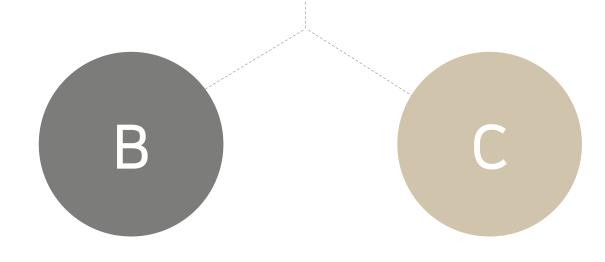


1) Q&A 데이터 생성



추가생성 데이터

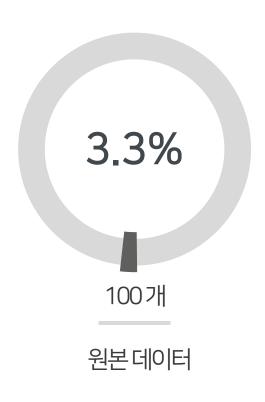
FNCSP 도메인과 관련된 적절한 Q&A 데이터 400개 생성 테스트용 질문 데이터 60개 생성

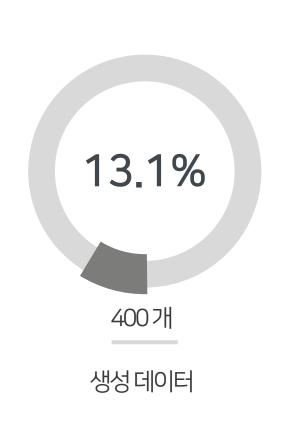


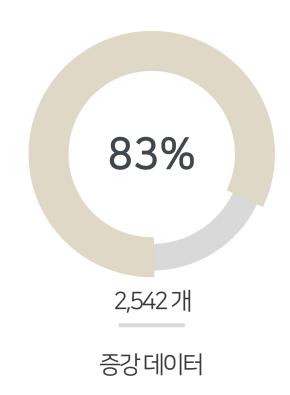
증강데이터

Chat GPT를 이용 Q&A 데이터 500개를 구어체, 문어 체 등 다양한 어조로 변경 총합 3,042개 데이터로 증강

Part 2 2)사용데이터분석







Part 3 모델구현



Load LLM (Korea Univ. KULLM)



Quantization (4Bit, double mode)



LoRA configuration

○ KULLM (구름): Korea University Large Language Model



KULLM(구름)은 고려대학교 NLP & AI 연구실과 HIAI 연구소가 개발한 한국어 Large Language Model (LLM) 입니다. 구름 프로젝트는 한국어 모델 뿐만 아니라, 데이터 셋까지 전면 공개하여 한국어 LLM 생태계에 기여하고자 하였습니다.

Quantize 🤗 Transformers models

AutoGPTQ Integration

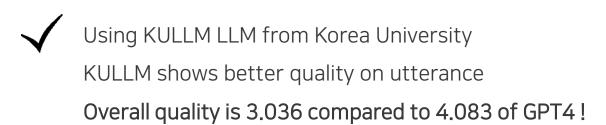
Transformers has integrated optimum API to perform GPTQ quantization on language models. your model in 8, 4, 3 or even 2 bits without a big drop of performance and faster inference speed! hardwares.

PEFT: Parameter-Efficient Fine-Tuning of Billion-Scale Models on Low-Resource Hardware

Load LLM (Korea Univ. KULLM) Quantization (4Bit, double mode)

LoRA configuration

01 Load LLM with 4Bit Quantization



| Туре | Base- model | Model | 이해 가 능성 (0 - 1) | 자연스러 움 (1 - 3) | 맥락 유 지 (1 - 3) | 흥미롭 기 (1 - 3) | 지시어 사용 (0- 1) | 전반적인 품질 (1- 5) |
|--------|-----------------------|------------------|-----------------------|-------------------|----------------------|---------------------|---------------------|----------------------|
| Closed | GPT3.5- turbo | GPT-3.5 | 0.980 | 2.806 | 2.849 | 2.056 | 0.917 | 3.905 |
| Closed | GPT-4 | GPT-4 | 0.984 | 2.897 | 2.944 | 2.143 | 0.968 | 4.083 |
| Open | Polyglot- ko-12.8b | KoAlpaca v1.1 | 0.651 | 1.909 | 1.901 | 1.583 | 0.385 | 2.575 |
| Open | LLaMA-7b | koVicuna | 0.460 | 1.583 | 1.726 | 1.528 | 0.409 | 2.440 |
| Open | Polyglot- ko-12.8b | KULLM v2 | 0.742 | 2.083 | 2.107 | 1.794 | 0.548 | 3.036 |

Load LLM (Korea Univ. KULLM) Quantization (4Bit, double mode)

LoRA configuration

01 Load LLM with 4Bit Quantization

4Bit Quantize allow for us to save & train LLM more efficiently Leveraging double quantization for better performance

This allow for LLM to select another dtype in computation!

01 Savemode

02 Computation mode

4Bit (normal float)

16Bit (brain float)



This technique let model be trained with

more precision in computational mode

Load LLM (Korea Univ. KULLM)



Quantization (4Bit, double mode)

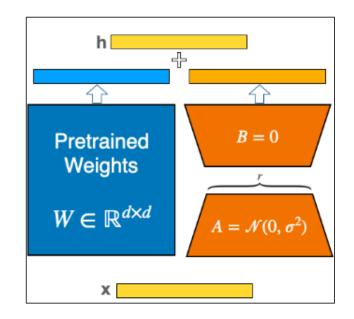


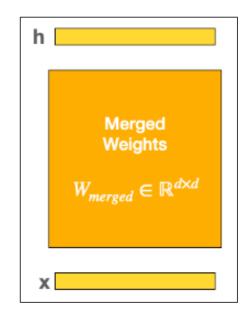
LoRA configuration

02 Apply LoRA Configuration



We use LoRA fine-tuning method out of many PEFT methods
LoRA uses decomposed matrix for fine-tuning on specific layers
We can adjust the rank(dimension) of matrix,
This let us control the computational cost!





2) Post-Generation Stage

Training



Inference



Post-Generation

But, How to deal with Hallucination?



Most of models which have decoding layers cause hallucination
This is fundamental problem on this kind of models

Even GPT also has hallucination problem



한국의 동화 '콩쥐 팥쥐' 전에 대해서 알려주세요

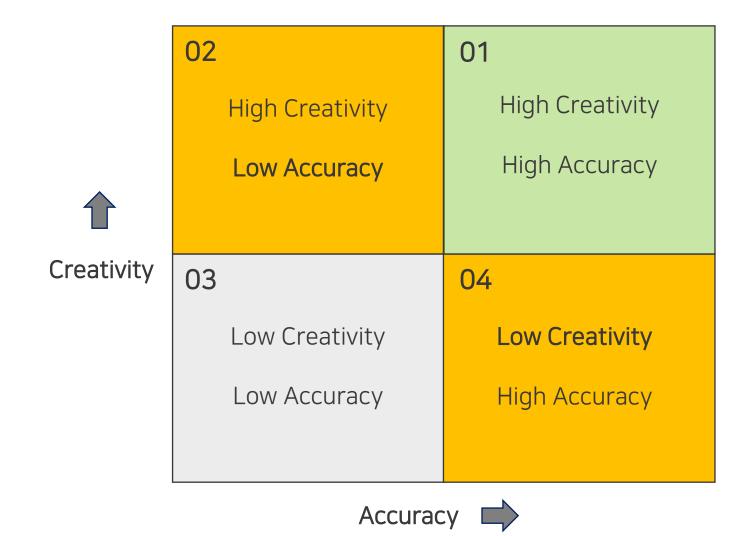


'콩쥐 팥쥐'는 한국의 전래동화 중 하나로, 두 마리의 쥐가 서로 협력하여 어려 △ ▽ 움을 극복하는 이야기입니다.

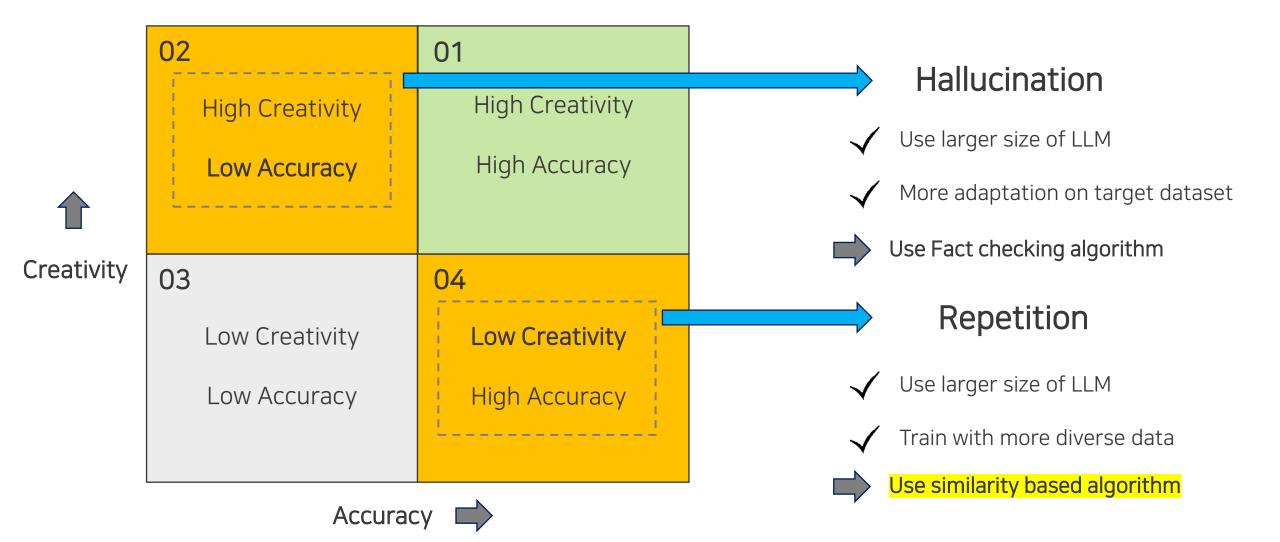
이 이야기는 매우 가난한 삶을 살던 콩쥐와 팥쥐가 각각 콩과 팥을 모아 아껴서 겨울을 버티며 살아가던 중, 갑자기 온갖 재난이 찾아오면서 그들의 고통이시작됩니다. 그러나 콩쥐와 팥쥐는 서로 협력하여 어려움을 극복하고, 결국 이웃들과 함께 풍성한 생활을 할 수 있게 됩니다.

이 이야기는 작은 것이라도 노력과 협력을 통해 큰 것을 이룰 수 있다는 가치를 강조하며, 서로의 도움을 받아 어려움을 극복하는 중요성을 보여줍니다. 또한, 이 이야기는 한국의 전통 문화와 가치관을 담고 있어, 한국어 교육과 한국 문화 이해에 중요한 역할을 하고 있습니다.

2) Post-Generation Stage



2) Post-Generation Stage



2) Post-Generation Stage

Character Based

Valid character

Valid URL

Word Based

Stopwords

Number of URLs

Pattern Based

Numbering

Sentence Completion

Indicator of last sentence



Levenshtein Distance



Too Simple But, Very Fast

Sentence1



Deletion

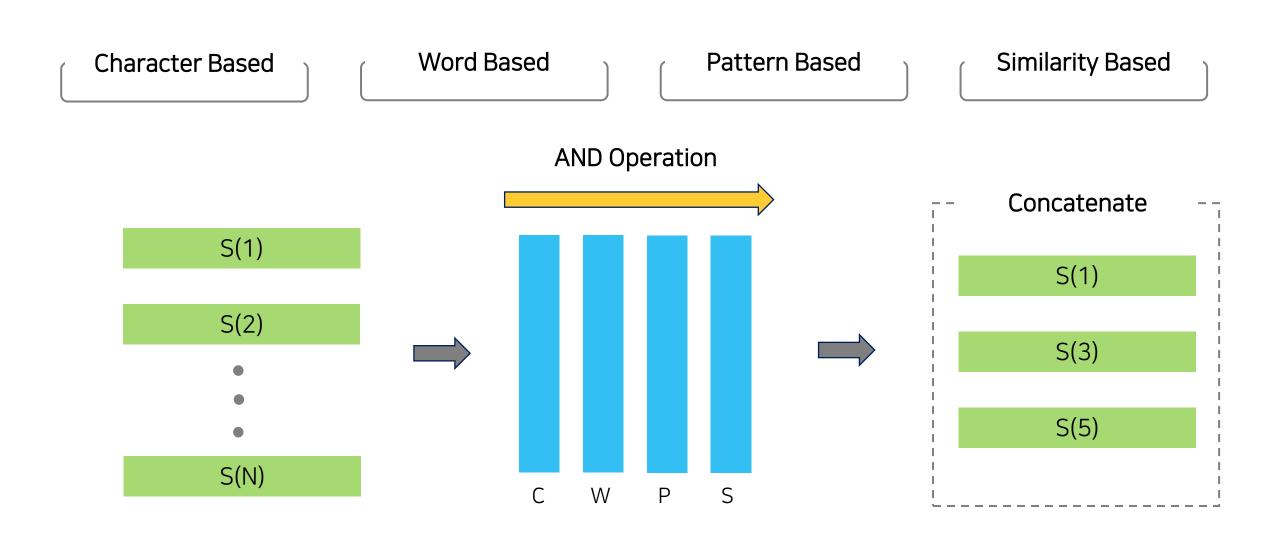
Insertion

Modification



Sentence2

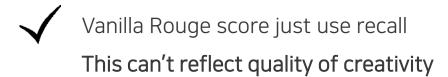
2) Post-Generation Stage



3) Evaluation

02 01 **High Creativity High Creativity** High Accuracy Low Accuracy Creativity 03 04 Low Creativity **Low Creativity** Low Accuracy High Accuracy Accuracy

Rouge-N





Average Rouge-N F1-Score

$$\sum_{n=1}^{3} F1Rouge(response, reference, n)$$

This can penalizes unnecessary generated words

3) Evaluation

Average Rouge-N F1-Score

 $\sum_{n=1}^{3} F1Rouge(response, reference, n)$



Before Post-Generation Stage

Score: 0.59966





After Post-Generation Stage

Score: 0.68416



23.41% Improved!

Score: 0.28186

질문: 검색 안 되는 기업 등록 요청

답변: 해당 부분은 고객 센터로 문의 바랍니다.

문제가 해결이 안되신 경우 FNCSP고객센터로 문의 남겨주시면 빠르게 응대 도와드리겠습니다.

문의 남겨주시면 신속하게 처리 도와드리겠습니다.

문의 페이지로 이동

문의하기

(https://fncsp.org/SU/Community/Post/Detail/2025957)

문의하기

(https://fncsp.org/SU/Community/Post/Detail



Post-Generation Stage



Score: 0.49022

해당 부분은 고객 센터로 문의 바랍니다.

문제가 해결이 안되신 경우 FNCSP고객센터로 문의 남겨주시면 빠르게 응대 도와드리겠습니다.

Answer

해당 부분은 고객 센터로 문의 바랍니다.

4) Experiment 1 – Level of Adaptation



Prompt



아이디와 비밀번호가 정확한데 로그인이 안돼요.

Trainable 20%

Trainable 40%

Trainable 60%

Trainable 80%

FNCSP는 기업의 재무정보를 분석하여 기업의 가치를 평가하는 인공지능 기반의 기업가치 평가 서비스입니다. 현재까지 약 1,000여개의 기업이 서비스를 이용하고 있습니다. 현재까지의 이용실적으로 볼 때, 서비스의 유용성은 매우 높다고 할 수 있습니다.

https://fncsf.com/sc/

FNCSP 기본기능은 [홈페이지 좌측 상 단 메뉴바 호출 기능, 홈페이지 하단 좌 측 중간 영역에 기업명 입력 기능, 홈페 이지 하단 최종 점수 표출 기능]입니다. 각 기능에 대한 설명을 원하시면 고객센 터로 연락주시면 도와드리겠습니다.

1. 홈페이지 호출 기능
(https://fncsp.org/SU/SSign1020.
do)

FNCSP는 기업회원과 일반회원으로 나 뉩니다. 일반회원으로 가입하려면 기업 회원으로 가입하거나, 둘 중 하나를 선 택해서 진행하시면 됩니다.

단, 기업회원으로 가입하더라도 일반회 원으로 가입한 것으로 간주됩니다. 단, 기업회원으로 가입 후 탈퇴하거나 다른 사람에게 양도하면 일반회원으로 전환되지 않습니다. FNCSP는 회원의 비즈니스 환경 정보를 종합적으로 분석하여 기업의 체질 개선 및 성장 지원을 목표로 서비스를 제공합니다.

해당 기업이 제공하는 제품 및 서비스가 산업 분류 코드와 어떤 관련이 있는지 종합적으로 분석합니다.

앞으로 서비스를 통해 어떤 서비스를 제 공할지 최종적으로 완성된 서비스를 토 대로 서비스를 기획합니다.

해당 서비스는 기업 회원만 이용할 수 있으며, 자세한 내용은 홈페이지를 통해 확인하실 수 있습니다.

4) Experiment 2 – LLM Comparison





FNCSP의 핵심기능을 알려줘.

01 Polyglot

02 KoAlpaca

03 KULLM

FNCSP의 핵심 기능은
[데이터 몰, 컨설팅 보고서,
고객 센터]로 구성되어
있습니다.
데이터 몰에서는 기업 정보와 관 련된 데이터를 다운로드 하고 분석할 수 있습니다. FNCSP의 핵심 기능은 회원님들의 사업 기회 포착과 공 유를 가능하게 해줍니다. 자세한 내용은 홈페이지에서 확인하시기 바랍니다. FNCSP 홈페이지는 개인과 기업이 공존할 수 있는 산업구조혁신을 위한 기업데이터 분석 전문 스마트 컨설팅 플랫폼 입니다.

4) Experiment 3 – Inference Performance

| / | Prompt | 아이디와 비밀번호가 정확한데 로그인이 안돼요 |
|----------|--------|--------------------------|
| V | Prompt | 아이니와 미월면오가 정확인데 도그인이 안되 |

| Trials | 01 A100 | 02 V100 |
|--------|----------|----------|
| | | |
| Trial1 | 13.843 초 | 15.003 초 |
| Trial2 | 13.797 초 | 15.043 초 |
| Trial3 | 13.894 초 | 15.166 초 |
| Trial4 | 13.89 초 | 15.012 초 |
| Trial5 | 13.736 초 | 15.103 초 |



There is just about 8% performance gain but, cost is twice as higher…?

Part 4 결론



4) 애로사항

01

제공되지 않는 기능에 대한 질문의 경우 답변이 너무 단조로워, 문맥이 풍부하게 만들어지지 않는 문제가 있었다.

02

특정한 도메인의 Q&A데이터 같은 경우 질문과 답이 정해져 있는 때문에 정확성과 문장의 유창함을 모두 갖춘 생성형 모델을 만드는 것이 어려웠다.

03

직접 데이터를 보고 수정해야 할 부분들이 많아 어려움을 많이 겪었다. 전처리 단계에서는 질문에 대한 답변을 정확하게 하기 위한 특수기호 제거, 행간 여백 제거, 특정 단어 리스트화, 링크 묶기 등의 작업들이 있었고 후처 리 단계에서는 할루시네이션을 방지하기 위한 하이퍼 파라미터 튜닝, 알고리즘 개발 작업들이 있었다.

4) 향후 발전 방향

01

학습 데이터에 좀 더 일상적인 Q&A 데이터나 경영, 경제 관련 데이터를 추가한다면 제대로 답변하지 못하는 질문에도 일반적인 답변을 도출해낼 수 있을 것으로 예상된다.

02

문서임베딩 모델을 같이 활용함으로써 데이터와 모델의 부적합을 상쇄할 수 있도록 아키텍처를 고도화 하면 더 좋을 것으로 예상된다.

03

하이퍼 파라미터 튜닝 및 데이터 퀄리티 개선을 통해 Accuracy와 Creativity를 모두 잡을 수 있는 생성형 모델을 개발한다면 GPT와 유사한 수준의 모델을 개발할 수 있을 것으로 예상된다.

챗봇 시연

Thank You