

고전소설의 현대적 재해석을 통한 연극 대본 생성 모델

자연어처리팀 | 한준호 윤세인 김나현 윤여원 권능주





목차

01

분석 배경

- 주제 선정 배경
- 데이터 수집 및 전처리

02

문체 현대화 모델링

- 이상치 탐지
- 가사체 변환
- 변환 결과 검증

03

현대어 변환 모델링

- 단어 모델링
- 현대어 변환
- CoT + NER

04

LSP 산출

- 변수 생성
- 대본화 지수(LSP) 산출
- 결과 시각화



1-1. 분석 배경

- 주제 선정 배경



주제 선정 배경

Why 고전 문학?

01 BACKGROUND

고전 문학

다양한 매체에서 끊임없이 재해석 되고 있는 **보편적 가치**를 가지고 있는 텍스트

고전 문학을 현대의 가치관에 맞게 **재창조**



주목을 이끌고

2차·3차 창작물이 만들어질 수 있는 원동력이 될 수 있음

주제 선정 배경

01 BACKGROUND

Why 생성형 AI?

생성형 AI 연구 경향을 살펴본 결과, 고전 문학은 시도해본 적 없는 **완전한 새로운 영역**



“생성형 AI와 국어국문학 연구 및 교육”은 뜨거운 감자
그러나 2023.12 기준, 고전 문학과 관련된 논의는 단 4건이며,
연구 경향 분석 과정에서 간단히 다루는 수준

Why 연극?

고전 소설의 문화 콘텐츠화

문화를 기술과 산업의 차원에서 재조명하고,
콘텐츠를 탐색하는 과정에서 전통 문화가 새롭게 주목받음.

고전 문학이 첨단 문화 콘텐츠의 산출에 기여할 핵심적인 자원으로 인식되기 시작

기존 고전 문학 문화 콘텐츠화의 한계



고전 문학 연구자의 경험·능력 면에서 콘텐츠 분야의 전문성 부족



→ 콘텐츠 관련 연구 동력을 잃어가고 있는 상황



콘텐츠의 체계와 맥락에 맞는 고전을 탐색하여 적절히 재구성하는 일에 논의가 집중 됨



→ 고전 문학이 콘텐츠의 재료이지만, 핵심 가치는 해체되어 버리는 현상 발생

기존 고전 문학 문화 콘텐츠화의 한계



고전 문학 연구자의 경험·능력 면에서 콘텐츠 분야의 전문성 부족



콘텐츠



연극(콘텐츠)의 체계와 맥락을 제대로 이해하고
그에 맞추어 고전 소설을 잘 적용한다면,
문화콘텐츠로서의 가치가 충분할 것 !

중 됨



1-2. 데이터 수집 및 전처리

- 데이터 수집 및 전처리



데이터 수집 및 전처리

01

DATA SET

고전소설 데이터셋

고전소설 오픈소스 크롤링

...	title	author	context
...	검녀(劍女)	안석경(安錫敬)	丹翁曰:“聞之湖南人曰, 蘇凝天進士, 有聲於三南, 舉以奇士目之.”
...	NULL	NULL	NULL
...
39133	화왕전(花王傳)	이이순(李頤淳)	獨不在於晚節乎(독불재어만절호):다만 끝까지 가는 절개에 있지 않은가.”

데이터 수집 및 전처리

01 DATA SET

고전소설 데이터셋

NULL 제거
중세국어 전처리
한자 제거



...	title	author	sentence	cleaned_content
...	공방전(孔方傳)	임춘	공방(孔方)의 자(字)는 관지(貫之)다.	공방의 자는 관지다.
...	공방전(孔方傳)	임춘	그는 처음 황제(黃帝) 시절에 조금 조정에 쓰였으나 워낙 성질이 굳세어 원래 세상 일...	그는 처음 황제 시절에 조금 조정에 쓰였으나 워낙 성질이 굳세어 원래 세상일에는 그..
...
10656	화왕전(花王傳)	이이순(李頤淳)	獨不在於晩節乎(독불재어만절호):다만 끝까지 가는 절개에 있지 않은가	다만 끝까지 가는 절개에 있지 않은가

최종적으로 10656행 데이터 생성

데이터 수집 및 전처리

현대소설

01

DATA SET

AI hub - 요약문 및 레포트 생성 데이터에서
문학부분만 추출

데이터 종류	원문 규모	어노테이션 규모	결과 규모		비고
			추출요약	생성요약	
뉴스기사	27,000	59,400	14,850	29,700	2~3문장 추출
			14,850		20% 추출
보도자료	20,000	44,000	11,000	22,000	2~3문장 추출
			11,000		20% 추출
역사_문화재	10,000	22,000	5,500	11,000	2~3문장 추출
			5,500		20% 추출
보고서	10,000	22,000	5,500	11,000	2~3문장 추출
			5,500		20% 추출
회의록	34,000	74,800	18,700	37,400	2~3문장 추출
			18,700		20% 추출
사설	10,000	22,000	5,500	11,000	2~3문장 추출
			5,500		20% 추출
간행물	10,000	22,000	5,500	11,000	2~3문장 추출
			5,500		20% 추출
연설문	40,000	88,000	22,000	44,000	2~3문장 추출
			22,000		20% 추출
문학	12,000	26,400	6,600	13,200	2~3문장 추출
			6,600		20% 추출
나레이션	10,000	22,742	5,640	11,371	2~3문장 추출
			5,731		20% 추출
총계	183,000	403,342	201,671	201,671	

※생성요약은 1문장 요약으로 100글자 이하로 구축

현대소설

index	title	author	context
...	맥(麥)	김남천	누가 그것을 증명할 수 있으랴! 저 회사원이 애띠고 어린 꽃 같은 색 시를 언제나 ...
...	광장 씨	이무영	"왜말 못해도 쓰일 세상이 왔으니 나도 한몫 보게 좀 해주게나." "그 야 이를 말인...
...
10799	개가(改嫁)	계용묵	개가(改嫁) 한 점의 혈육도 남기지 못한 어떤 젊은 과부가 남편의 삼년상도 치르기 ...

10800행 데이터 수집

context에 여러 문장이 포함되어 문장 별로 구분 작업



2-1. 문체 현대화 모델링

- 이상치 탐지



고전소설의 문체

주로 청자 지향형으로, 구어체적 성향을 짙게 띄고 있음



이에 따라 ① **연결 어미**로 이어지는 가사체 형식과
② **문장이 길어진다는** 특징을 가짐

ex)

유한림이 급제 후에 성혼하려고 하매, 구혼하는 규수가 많으나 좀처럼 허하지 않고
유공이 매제 두부인과 함께 성종의 모든 매파를 청하여 현철한 소저가 있는 집안을
물었으나 마땅한 상대가 없어서 좀체로 결정하지 못하였다. - 사씨남정기 中

고전소설의 문체

주로 청자 지향형으로, 구어체적 성향을 짙게 띄고 있음



이에 따라 ① **연결 어미**로 이어지는 가사체 형식과
② **문장이 길어진다는** 특징을 가짐

추가로, ③ **작자 미상 여부**에 따른 문체의 차이도 고려함

고전소설 - ① 연결 어미

[통계 자료] 연결 어미가 옛말에서 더 많이 사용된다는 근거

	현대 일상어		옛말	
	표제어 수	비율	표제어 수	비율
어미	2,820	0.583	1,360	10.61

실제 통계 자료에서 일상어, 전문어, 북한어, 방언, 옛말로 나누어 비교함
 현대 전문어는 어미로 활용되는 경우가 없으며, 북한어와 방언은 해당 사항이 없으므로
 일상어와 옛말을 비교함

고전소설 - ① 연결 어미

연결 어미의 비율

연결 어미의 비율 = (연결 어미 개수) / (형태소 개수) (문장 당)

index	cleaned_content	EC_prop	len
0	그러나 그들은 별을 ...	0.09090909090909091	18
1	거짓말 같은 일 ...	0.0	12
...
148378	새침한 얼굴이 ...	0.05970149253731343	132
148379	감격이 없고 영탄을 잊은 ...	0.12903225806451613	57

한국어 형태소 분석기 바른 ai를 사용하여 연결 어미 태깅
보다 자세한 내용은 NLP 주제분석 PPT를 참고해주세요

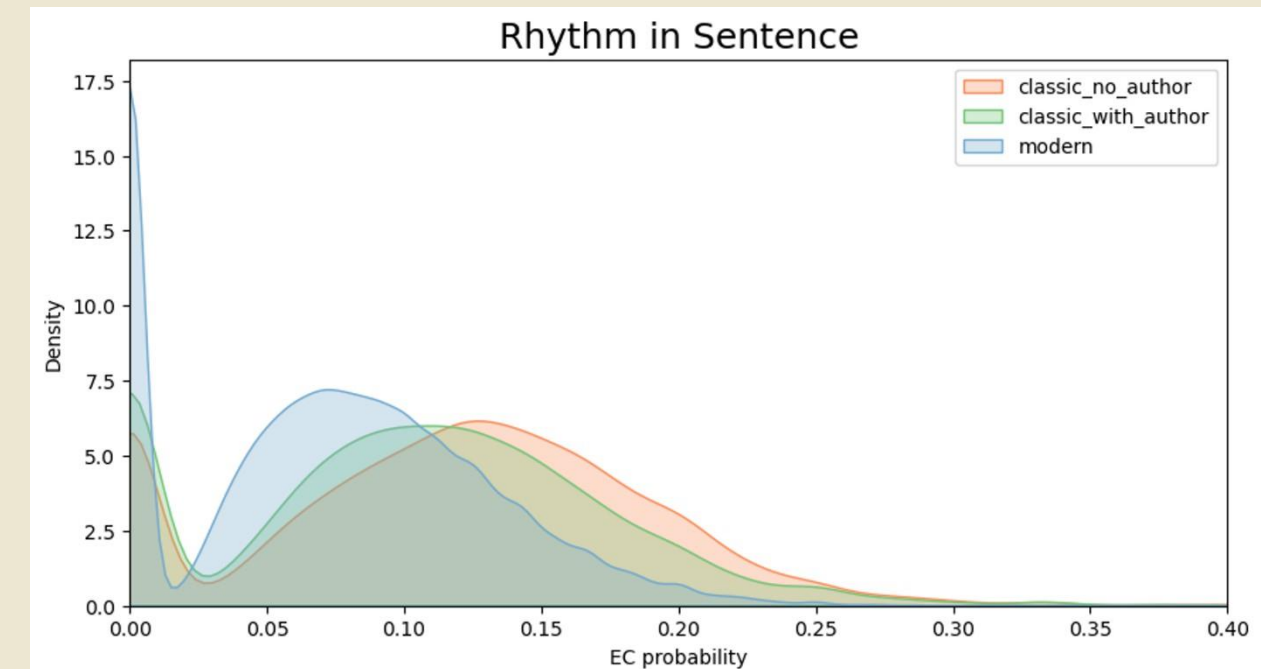
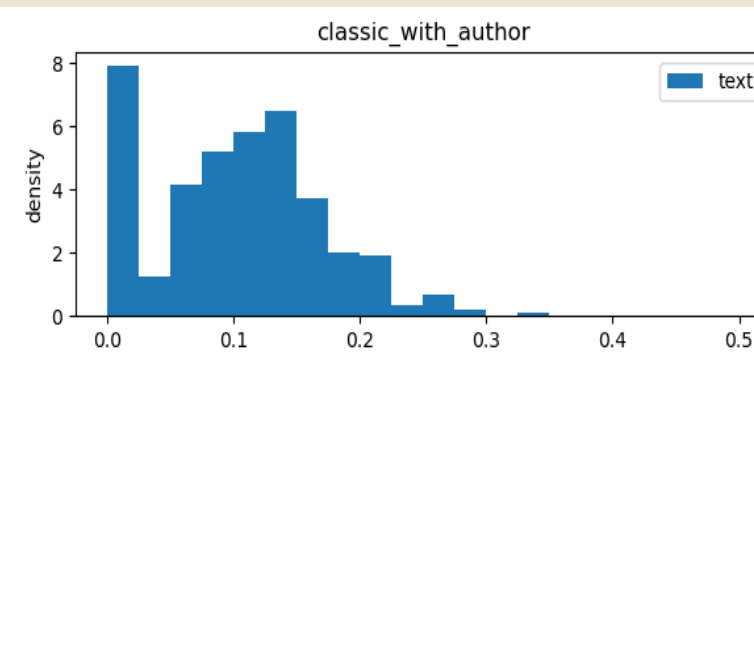
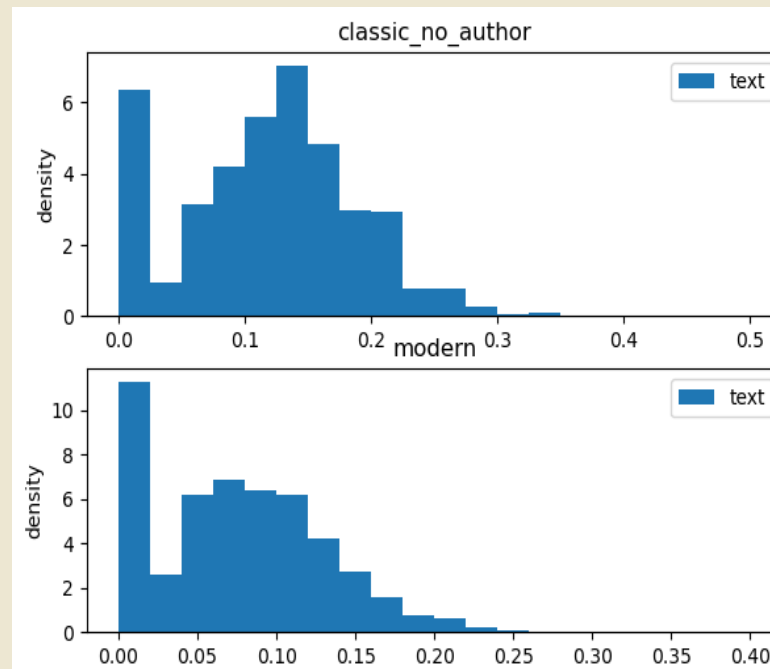


고전소설 - ② 작자 미상 여부

작자 미상의 경우 구전으로 전해져 내려 오는 경우가 많음



작자 미상인지 아닌지의 여부에 따라 **연결 어미의 비율이 달라짐**
운율감이 살아있는 텍스트가 더 많은 것이라고 판단하여 분리



이상치 탐지

02

OUTLIER
DETECTION

이상치 탐지

X
문장 길이
(len)

두 변수를 축으로
2차원 이상치 탐지 진행!

Y
연결어미 비율
(EC_prop)

문장 길이가 짧은 경우, 연결어미의 개수가
적어도 운율감이 높다고 판단될 수
있으므로 필요한 변수

운율감을 파악하는 핵심 개념

고전소설 작가미상 데이터와 고전소설 작가O 데이터에
모두 문장길이 스케일링 후, 동일한 모델링 적용



data leakage

고전소설을 일종의 test data로 봐야 하기 때문에 문장길이의 최대/최소를 현대 소설의 것으로 사용

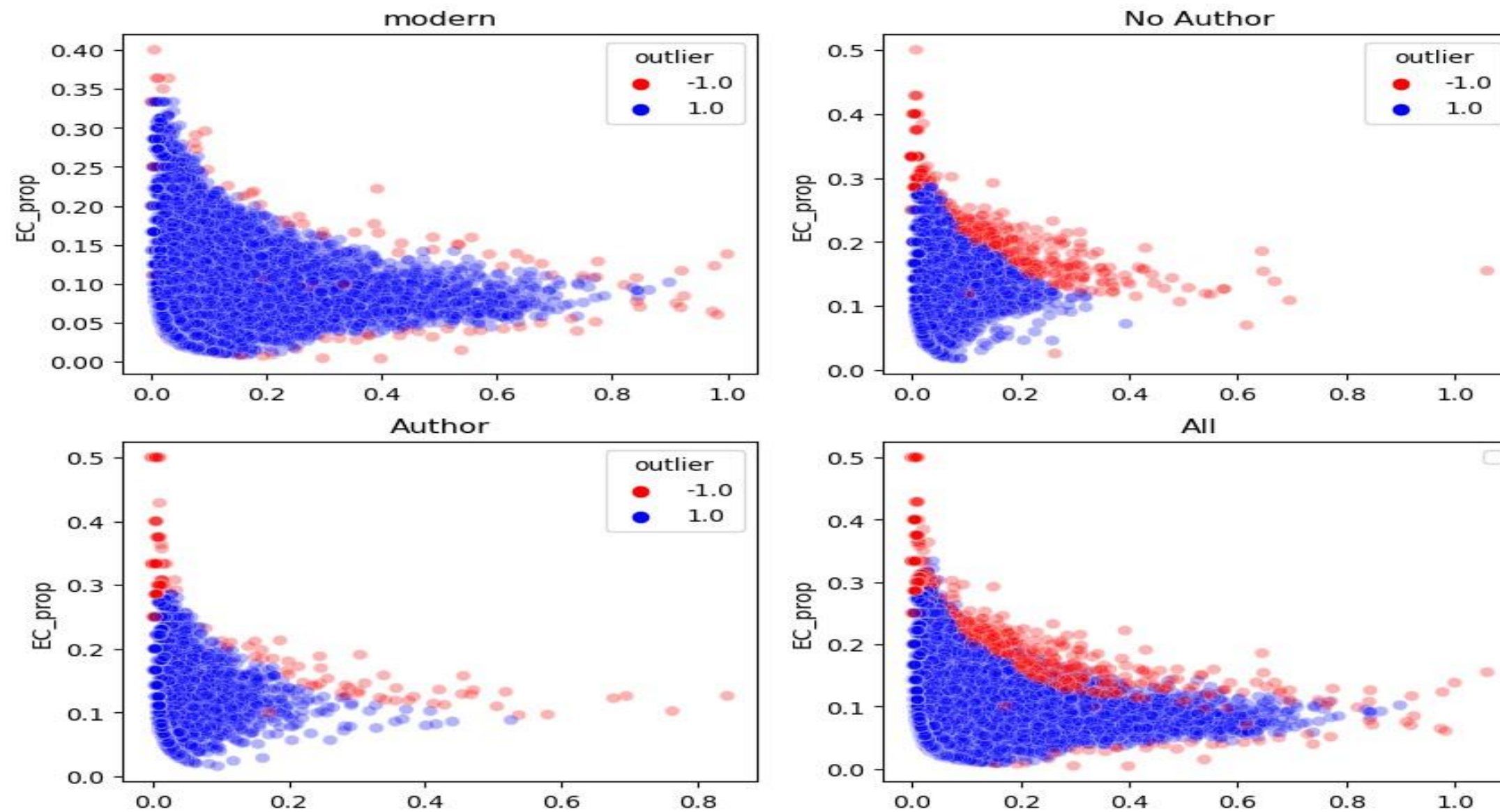
이상치 탐지

One-Class SVM

02

OUTLIER
DETECTION

Outlier Detection (0.1%)



최종적으로 현대소설과 분포가 다른

고전소설 이상치를 추출

약 600개 정도의 이상치 데이터 존재



2-2. 가사체 변환

- 가사체 변환
- 변환 결과 검증



가사체 변환

Original

이 때 천자가 하늘에서 명을 받으시니 해내가 안정치 못
하여 혹은 위왕이라 하고 혹은 국왕이라 칭하며 남서로
노략하니 일경이 진동하여 피란하는 자 무수하였는데...

Translation

???



가사체와 산문체의 특징에 대해
지도학습할 수 있는 데이터가 부재함



소량의 레이블로 추론 가능한 **Few-Shot Prompt Learning**

가사체 변환

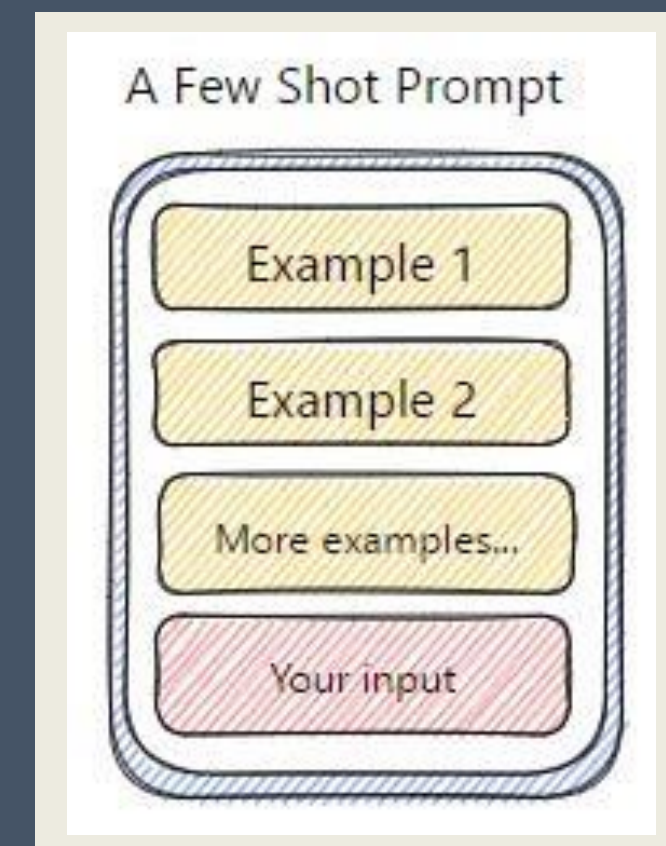
Few-Shot Prompt Learning

학습 데이터가 적은 상황에서 LLM 자체에 매개변수 변형을 하지 않고
소량의 데이터로 추론이 가능하게 만드는 기법

사전 학습 모델에게 프롬프트로 몇 가지
예시 데이터를 알려주고 해당 문제를 해결하도록 만듦



5개의 예시 제공



가사체 변환

Few-Shot Prompt

System: '너는 운율감이 느껴지는 문장인 [율문]을 운율감에 구애받지 않는 문장인 [산문]으로 바꿔주는 번역가야. 상대방이 지문을 요청하면 앞의 예시들의 형식에 맞게, 마지막 지문을 [산문]으로 바꿔주자. 모든 내용은 빠뜨리지 말고 답해주자.

모든 대답은 한국어(Korean)으로 대답해줘.

User: '[율문] 천하에 명산이 다섯이 있으니 동쪽은 동악 태산이요, 서쪽은 서악 화산이요, 남쪽은 남악 형산이요, 북쪽은 북악 향산이요, 가운데는 중악 송산이다'

Assistant: '[산문] 천하에 명산이 다섯이 있었다. 동쪽의 동악 태산, 서쪽의 서악 화산, 남쪽의 남악 형산, 북쪽의 북악 향산이 그것이다. 가운데는 중악 송산이다

...

가사체 변환

Llama-3

페이스북 팀이 개발한 대형 언어 모델(LLM)로
15T의 학습 데이터셋을 바탕으로 높은 성능을 보이는 모델

GPT-4에 준하는 성능을 보이면서
작은 크기의 오픈 소스 모델로 일반인 또한
사용 가능하다는 특징을 가지고 있음



한국어로 Fine-Tuning된
Llama-3 모델을
사용하기로 결정

가사체 변환

Original

이 때 천자가 하늘에서 명을 받으시니 해내가 안정치 못하여 혹은 위왕이라 하고 혹은 국왕이라 칭하며 남서로 노략하니 일경이 진동하여 피란하는 자 무수하였는데 장공이 그 가운데 섞이어 피난할 제 추병이 정히 위급한지라 부부 서로 해룡을 둘러업고 달아나더니 운이 다하매 부인이 울며 말하기를



Translation

[산문] 이 때 천자가 하늘에서 명을 받았다. 해내가 안정치 못하여 혹은 위왕이라 하고 혹은 국왕이라 칭하며 남서로 노략하니 일경이 진동하여 피란하는 자 무수하였다. 장공이 그 가운데 섞이어 피난할 때 추병이 정히 위급한지라 부부가 서로 해룡을 둘러업고 달아났다. 운이 다하매 부인이 울며 말하였다.

소설 지문의 경우 변동 없이 가사체로 잘 변환되는 것을
확인할 수 있었음

가사체 변환

Original

"만일 팔괘와 글을 이루고자 할진대 갑을병정은 이미
정하였으나 아랫자를 나에게 문의하시니 이는 쉬운 바라.
천지 생하시매 우리 서씨를 위하여 자회로부터 생하니, 자는
우리를 두고 이름이니 머리를 갑자라 하고 땅은 축회로부터
생하니 을축이라 하고, 만물은 인회로부터 생하니 병인이라
하여 차차 그 이류를 좇아 육갑을 지을진댄 무엇이
어렵다 하리요'



Translation

[산문] '만일 팔괘와 글을 이루고자 할진대 갑을병정은 이미
정하였으나, 아랫자를 나에게 문의하시니 이는 쉬운 바라.
천지 생하시매 우리 서씨를 위하여 자회로부터 생하니, 자는
우리를 두고 이름이니 머리를 갑자라 하고, 땅은 축회로부터
생하니 을축이라 하고, 만물은 인회로부터 생하니 병인이라
하여, 차차 그 이류를 좇아 육갑을 지을진댄 무엇이
어렵다 하리요'

그러나 인물의 대사는 제대로 바꿔주지 못했음



프롬프트에 대사 레이블을 포함하도록 개선하여 다시 시도!

가사체 변환

Few-Shot Prompt (수정)

System: '너는 운율감이 느껴지는 문장인 [율문]을 운율감에 구애받지 않는 문장인 [산문]으로 바꿔주는 번역가야. 상대방이 지문을 요청하면 앞의 예시들의 형식에 맞게, 마지막 지문을 [산문]으로 바꿔주자. 모든 내용은 빠뜨리지 말고 답해주자.

모든 대답은 한국어(Korean)으로 대답해줘.

User: [율문]"낭군의 말이 그르도다. 천하 만물이 세상에 나매 신의로써 으뜸을 삼나니, 서대주는 본래 우리로 더불어 향렬이 남과 다름이 없고 하물며 내외를 상통함도 없으되 다만 일면 교분을 생각하고 다소간 양미를 쾌히 허급하여 청하는 바를 좇았으니,"

Assistant: [산문]"낭군의 말이 틀렸습니다. 천하 만물이 세상에 태어날 때 신의를 으뜸으로 삼습니다. 서대주는 본래 우리와 향렬이 남과 다름이 없고 내외를 상통함도 없습니다. 그러나 단지 일면식의 교분을 생각해 다소 양미를 기꺼이 베풀어 청하는 바를 따랐습니다."

...

가사체 변환

Original

"소생이 향자 영감의 구활지은을 입사와 소생의 수다
잔약한 명이 춘하육삭을 무고히 지낸지라. 생아자는 부모요
재생자는 대인이오니 쇼생의 부처 매양 서로 대하여 화산의
풀을 맺으며 수호의 구슬을 머금어 대인의 은혜 갚기를
원하는 바일러니, 자연 생계로 말미암아 장구지계는 없고
고식지계 뿐이라

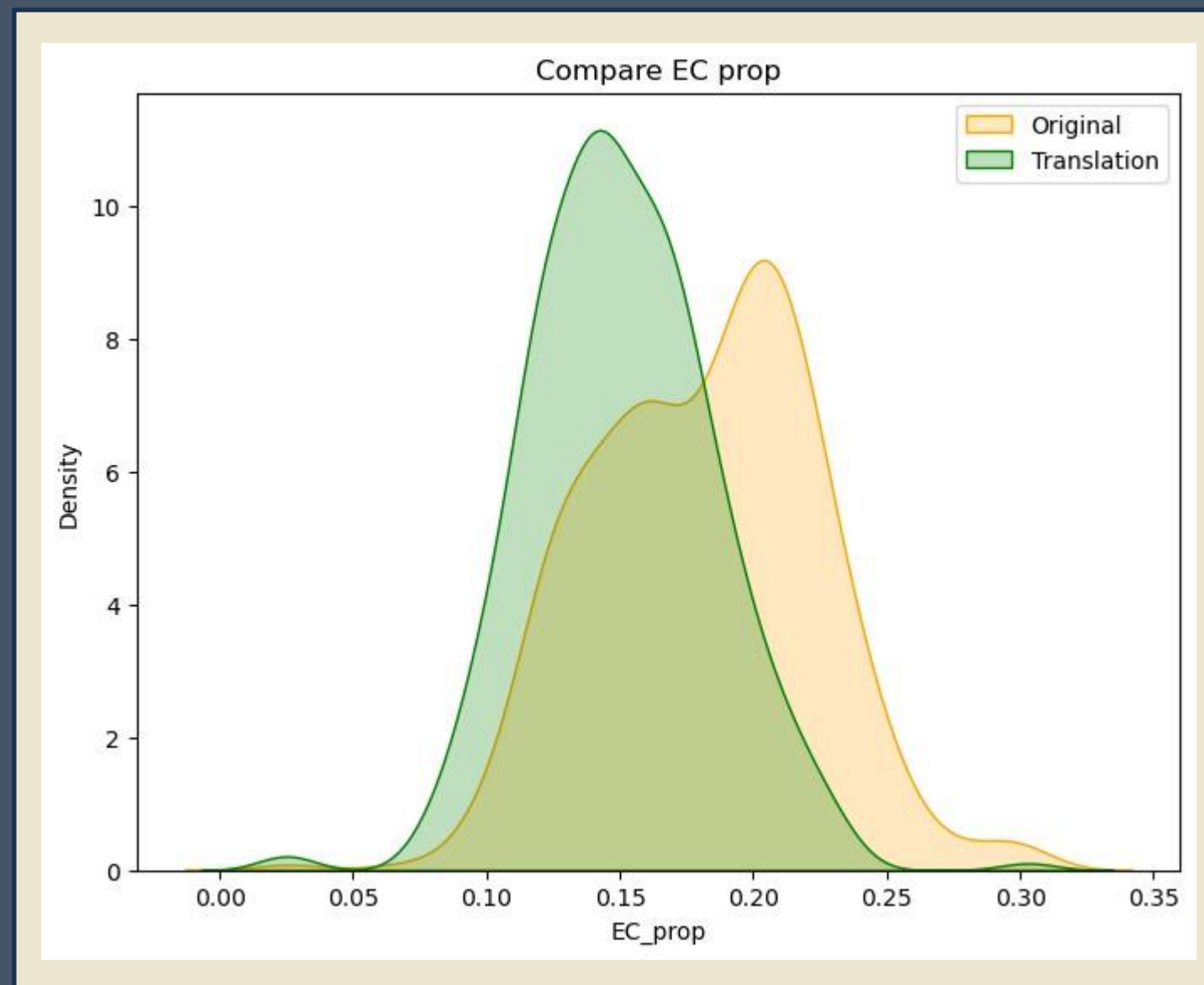


Translation

[산문] "소생이 향자 영감님의 구원한 은혜를 받은 후,
소생의 약한 몸이 봄, 여름, 가을, 겨울을 무고히 지냈습니다.
살아 있는 것은 부모요, 다시 살아나는 것은 대인이시니,
소생의 부모님과 저는 항상 서로 만나 서로를 대하여,
화산의 풀을 엮으며, 수호의 구슬을 품에 안고 대인의
은혜를 갚고자 원합니다. 그러나 저의 살림은 가난하고 몸은
연약하여, 장구한 계책은 없고, 임시변통의 방법뿐입니다.

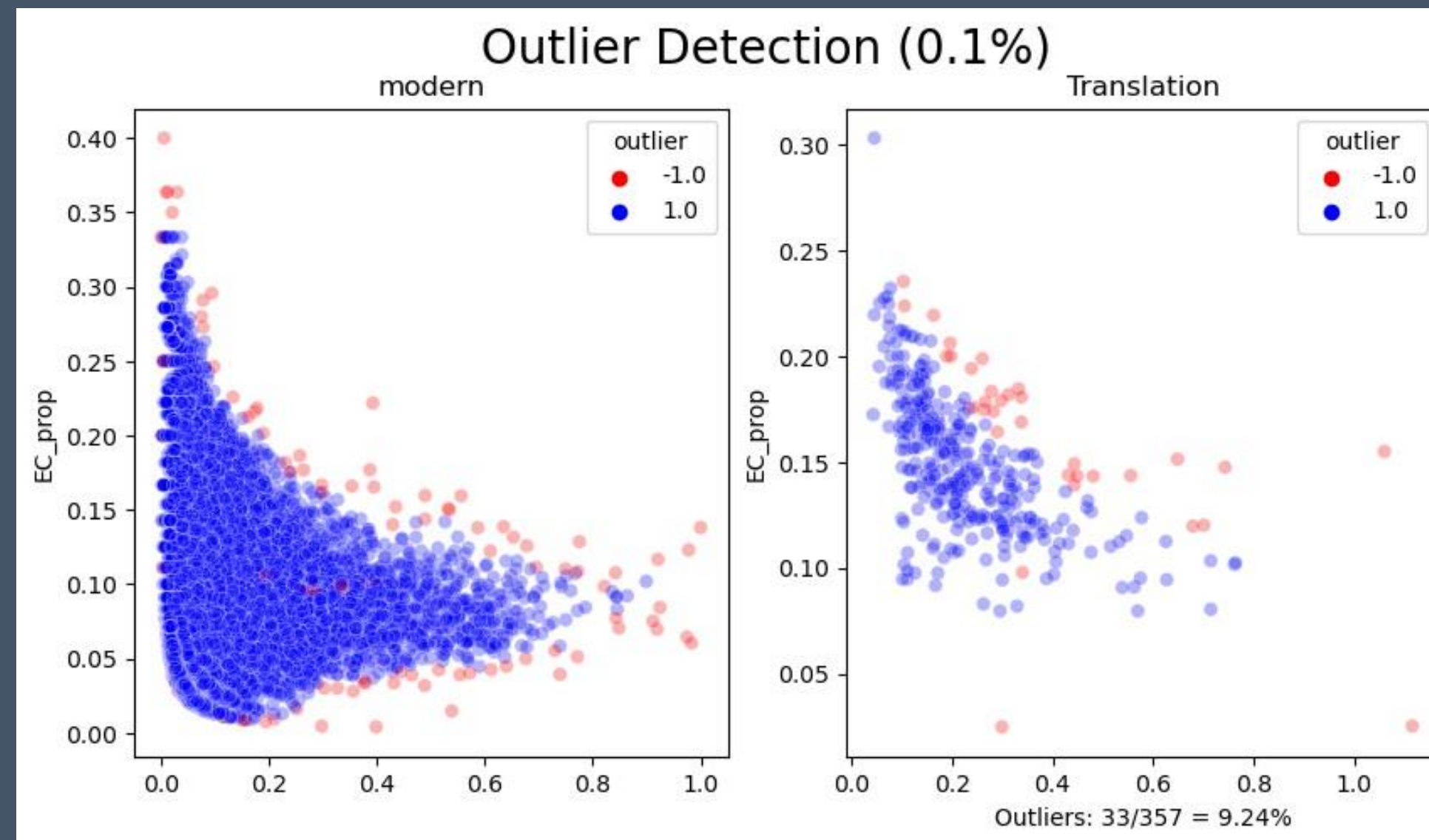
프롬프트 개선으로 가사체의 특성을 지닌 대사 또한 제대로 바꿔주는 것을 확인

가사체 변환 검증



실제 연결어미 비율을 시각화한 결과, 원본 문장과 확연한 차이를 보임

가사체 변환 검증



현대 소설로 학습된 SVM으로 이상치 탐지 결과 약 91%가 산문체로 잘 변환됨

가사체 변환 검증

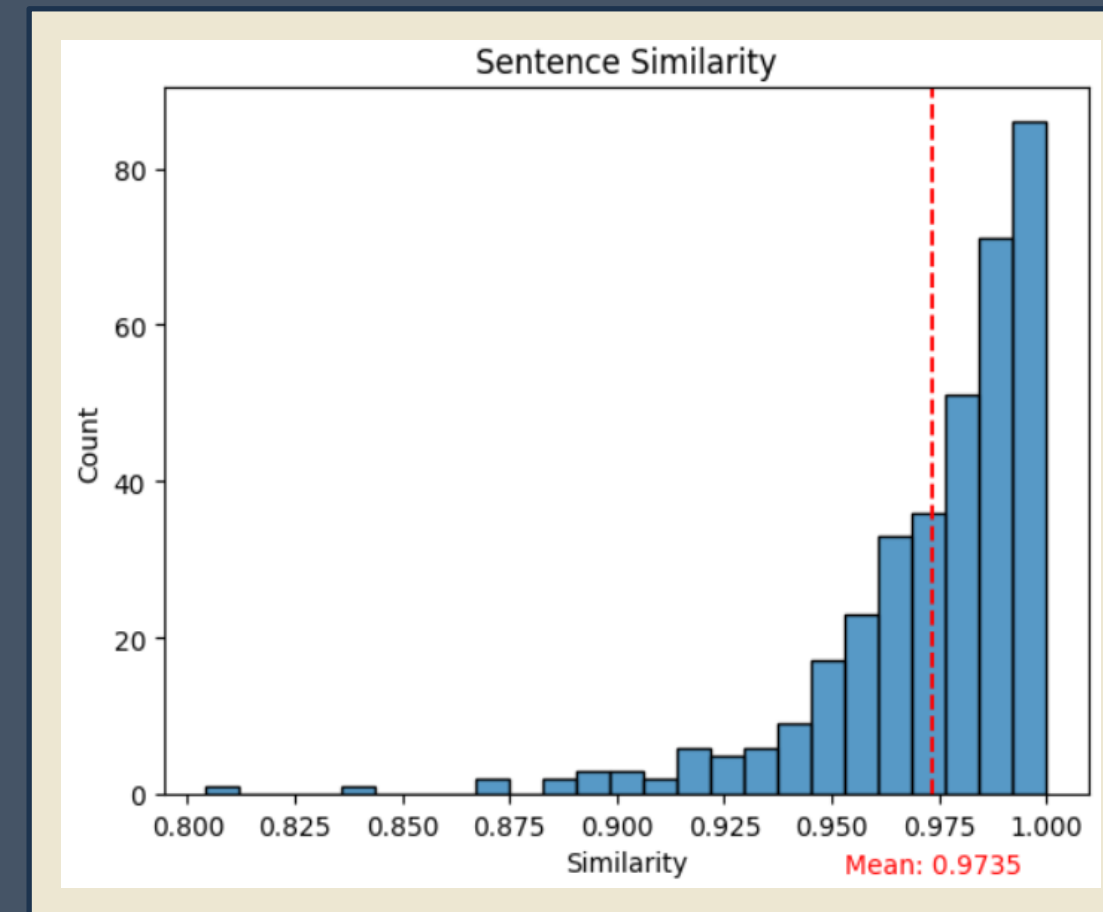
STS (Semantic Textual Similarity)

텍스트의 의미적 유사도를 측정하는 문제

문장 유사도 또한 높은 점수를 보이므로

의미 또한 보존한 채로 형식만

잘 바꿨다고 판단할 수 있음





3-1. 현대어 변환 모델링

- 단어 모델링



단어 난이도 모델링 ①

상대적 난이도

$$\text{상대적 난이도} = \frac{\text{검색량}}{\text{빈도수}}$$

검색량이 많을수록, 문서 내 빈도수가 적을수록 '어려운 어휘'로 간주함

검색량의 경우 네이버 검색광고 api,

빈도수의 경우 aihub의 [대규모 구매도서 기반 한국어 말뭉치 데이터]를 통해 데이터셋 구축

단어 난이도 모델링 ①

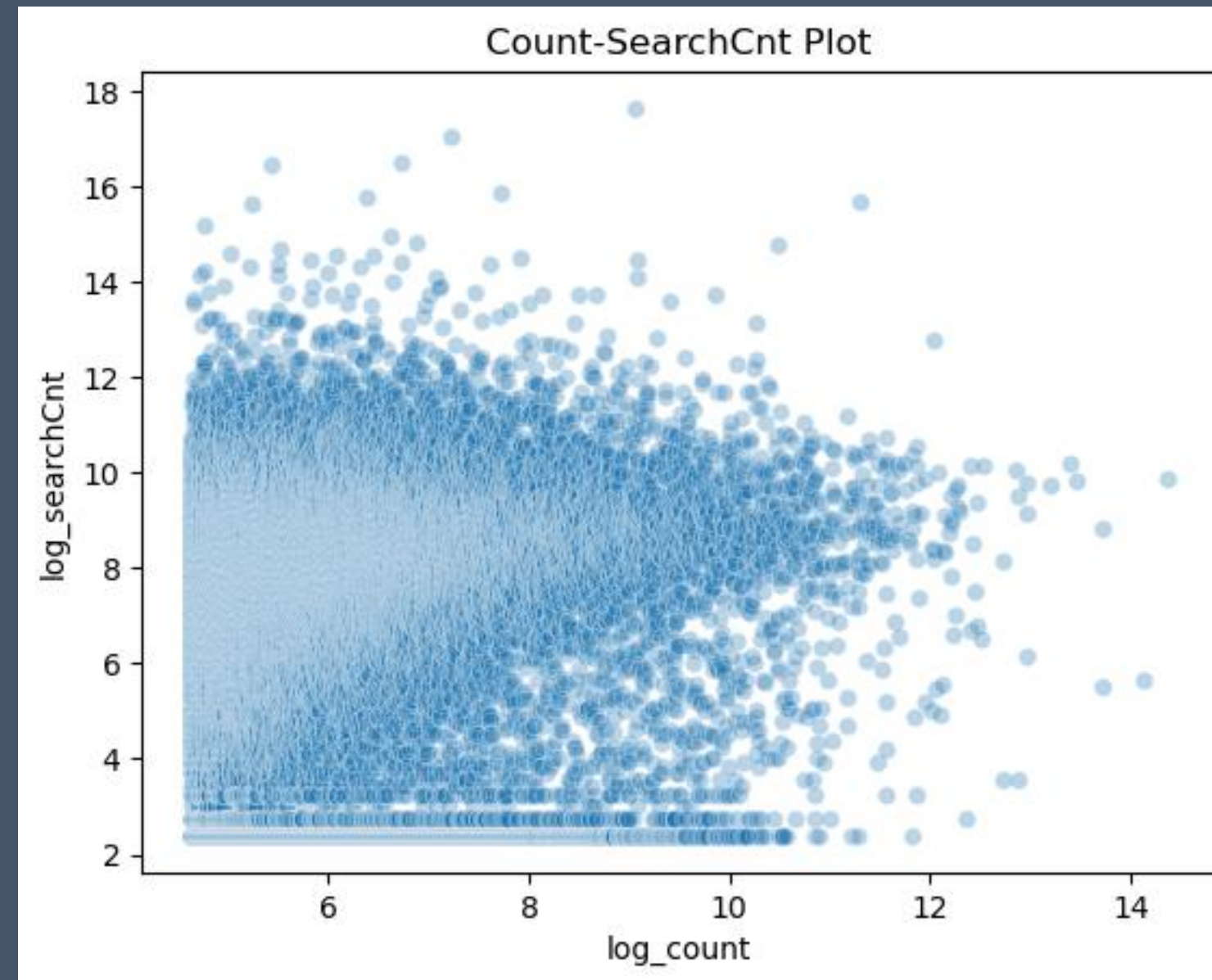
빈도수 \ 검색량	낮음	높음
	낮음	높음
낮음	Warning!	어려움
높음	쉬움	쉬움

그러나, 앞에서 정의한 상대적 난이도는 빈도 수가 낮지만 검색량이 적은 데이터는 반영하지 못함



2차원 클러스터링을 통해 군집화

단어 난이도 모델링 ①



그러나 2차원 시각화 결과, 거리 기반 클러스터링은 할 수 없었음



단어 난이도 모델링 ②

단어 모델링 변수

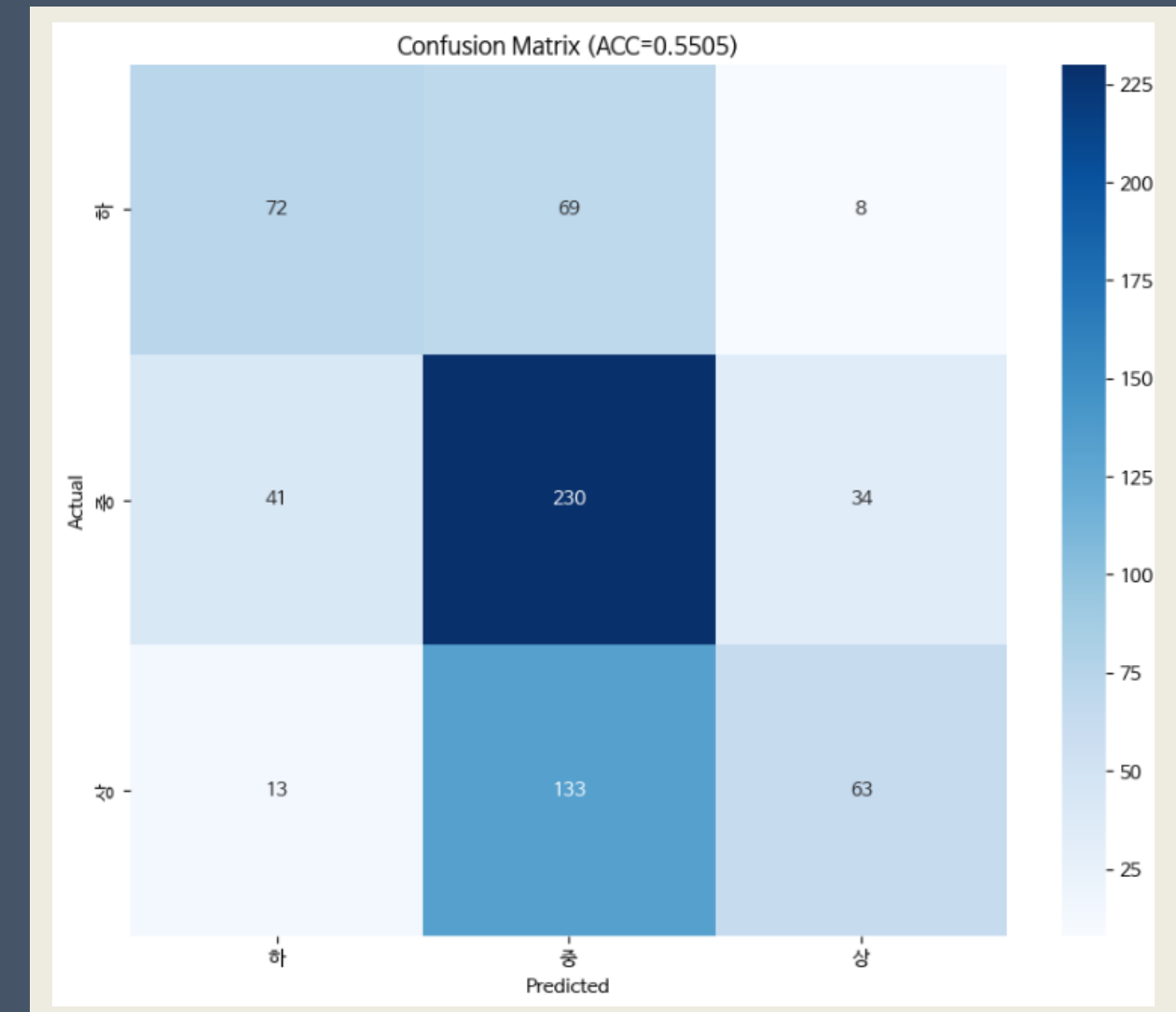
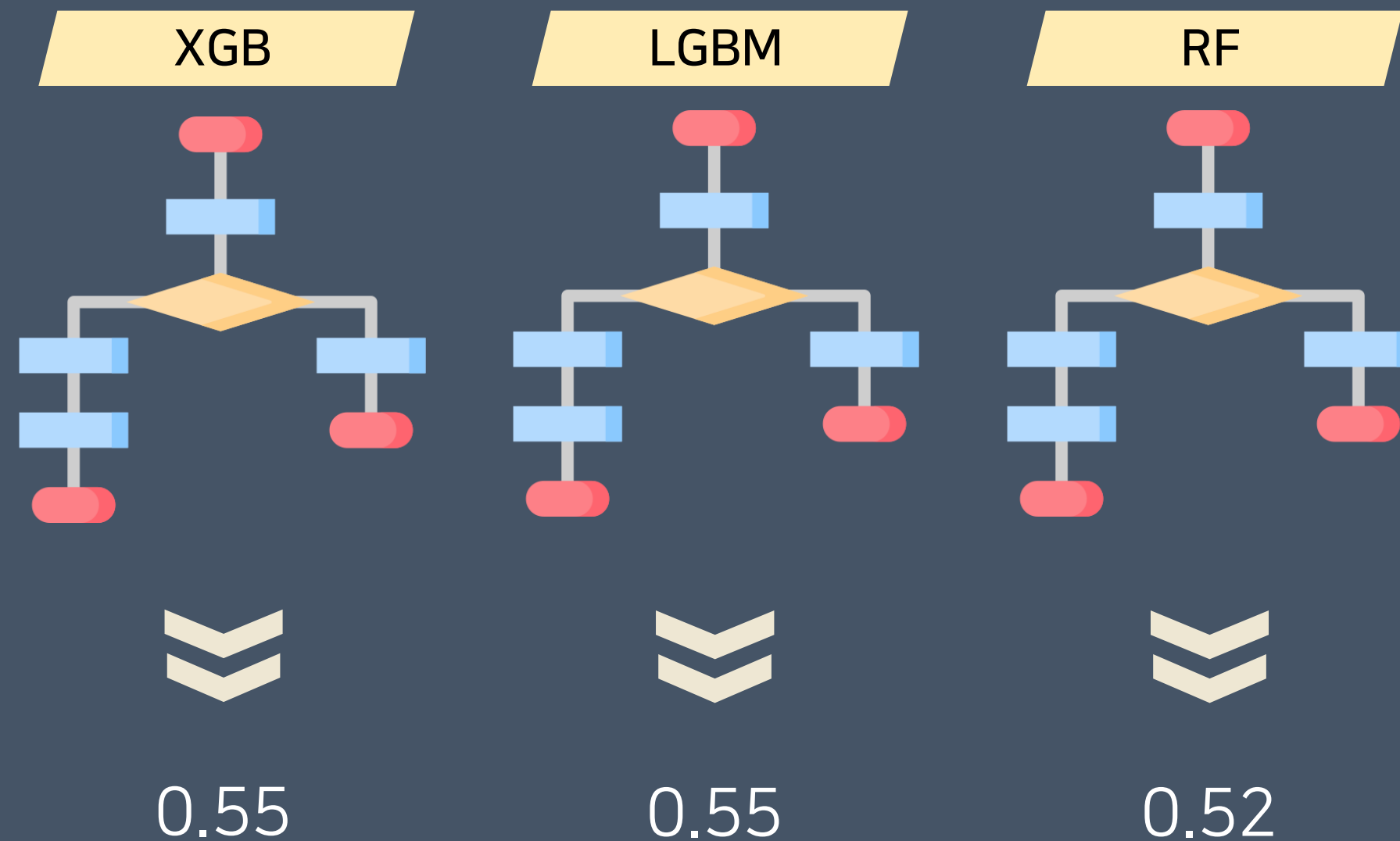
빈도 수
임베딩

단어 모델

XGBoost, LightGBM
Neural Network
Tree-NN Ensemble

각 단어의 빈도 수와 임베딩을 추가하여 모델링을 진행함

단어 난이도 모델링 ②



트리 모델의 결과 0.55에 가까운 낮은 정확도를 보임



3-2. 현대어 변환

- 현대어 변환
- CoT + NER



현대어 변환

현대어(easy sentence)로 변환하는 작업

GPT를 활용하여 어려운 단어를 인식

title	cleaned_content	complex_words_and_annot
숙영낭자전	... 죄가 없는 숙영낭자 는 안색이 조금도 변하지 않고 구김이 없는 목소리를 말하였다.	* 어려운 단어: 숙영낭자 , 구김 숙영낭자: 어린 아들 구김: 표정의 변화나 감정을 나타내는 표시
유충렬전 (劉忠烈傳)	하였거늘 천자 상소 를 보시고 대노하여 조정 에 내리어 보라 하신다. 이때 정한담 최일귀, 강희주의 상소를 보고 대분하여 즉시 궐내 에 들어가 여쭙오되,	* 어려운 단어: 상소 , 조정 , 정한담 상소: 상대방의 심정이나 의견을 묻거나 알아보는 것 조정: 상대방의 의견을 따라 행동하거나 수용함 정한담: 상대방의 의견이나 말을 따라서 행동하는 것

현대어 변환

현대어(easy sentence)로 변환하는 작업

추출한 어려운 단어를 다시 GPT에 제공해
Few-Shot Prompt Engineering 진행

title	cleaned_content	complex_words_and_annot	easy_sentence
숙영낭자전	... 죄가 없는 숙영낭자 는 안색이 조금도 변하지 않고 구김이 없는 목소리를 말하였다.	* 어려운 단어: 숙영낭자 , 구김 숙영낭자: 어린 아들 구김: 표정의 변화나 감정을 나타내는 표시	... 죄가 없는 아들 은 얼굴색이 변하지 않 고 감정을 나타내는 표정이 없는 목소리로 말했다.
유충렬전 (劉忠烈傳)	하였거늘 천자 상소 를 보시고 대노하여 조정 에 내리어 보라 하신다. 이때 정한담 최일귀, 강희주의 상소를 보고 대분하여 즉시 궐내 에 들어가 여쭙오되,	* 어려운 단어: 상소 , 조정 , 정한담 상소: 상대방의 심정이나 의견을 묻거나 알아보는 것 조정: 상대방의 의견을 따라 행동하거나 수용함 정한담: 상대방의 의견이나 말을 따라서 행동하는 것	하였는데 상대방의 의견을 물어보고 받아들여라 하셨다. 이때 최일 귀와 강희주가 상대방의 의견을 듣고 즉시 궁중에 들어가 물어보았다.

GPT 활용의 문제점

- 👉 한자가 너무 많을 경우, 환각(Hallucination) 발생
- 👉 인명과 지명은 고유명사이므로 유지해야 하지만, 어려운 단어로 판단
- 👉 Few-shot Learning만으로는 GPT에게 고전을 바꾸기 위한 충분한 정보를 제공하지 못함



깊은 추론이 가능하도록 CoT & NER 진행

CoT

CoT (Chain-of-Thought)

추론 과정을 상세히 제시하여 LLM의 다중 추론 능력 향상

Standard prompting gives examples of the correct answer

Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ❌

CoT prompting pairs each example with an associated rationale

Chain-of-Thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✅

Solution + Chain of Thought

LLM learns (via prompting) to generate a rationale with its answer

CoT

NER (Named Entity Recognition)

텍스트에서 고유한 의미의 개체(entity)를 인식(recognition)하는 것으로
사람, 장소, 기관, 날짜 등의 명명된 개체를 텍스트로 식별하는 작업



어려운 단어로 태깅되었던 **인명, 지명**들을 **NER**을 통해서
어려운 단어 리스트에서 제거!

CoT



CoT 프롬프트를 통해 **인명, 지명들을 판단하고 바꾸지 않도록** 유도



각설, 차시 등 **소설의 장면 전환 지표를 바꾸지 않도록** 유도

CoT

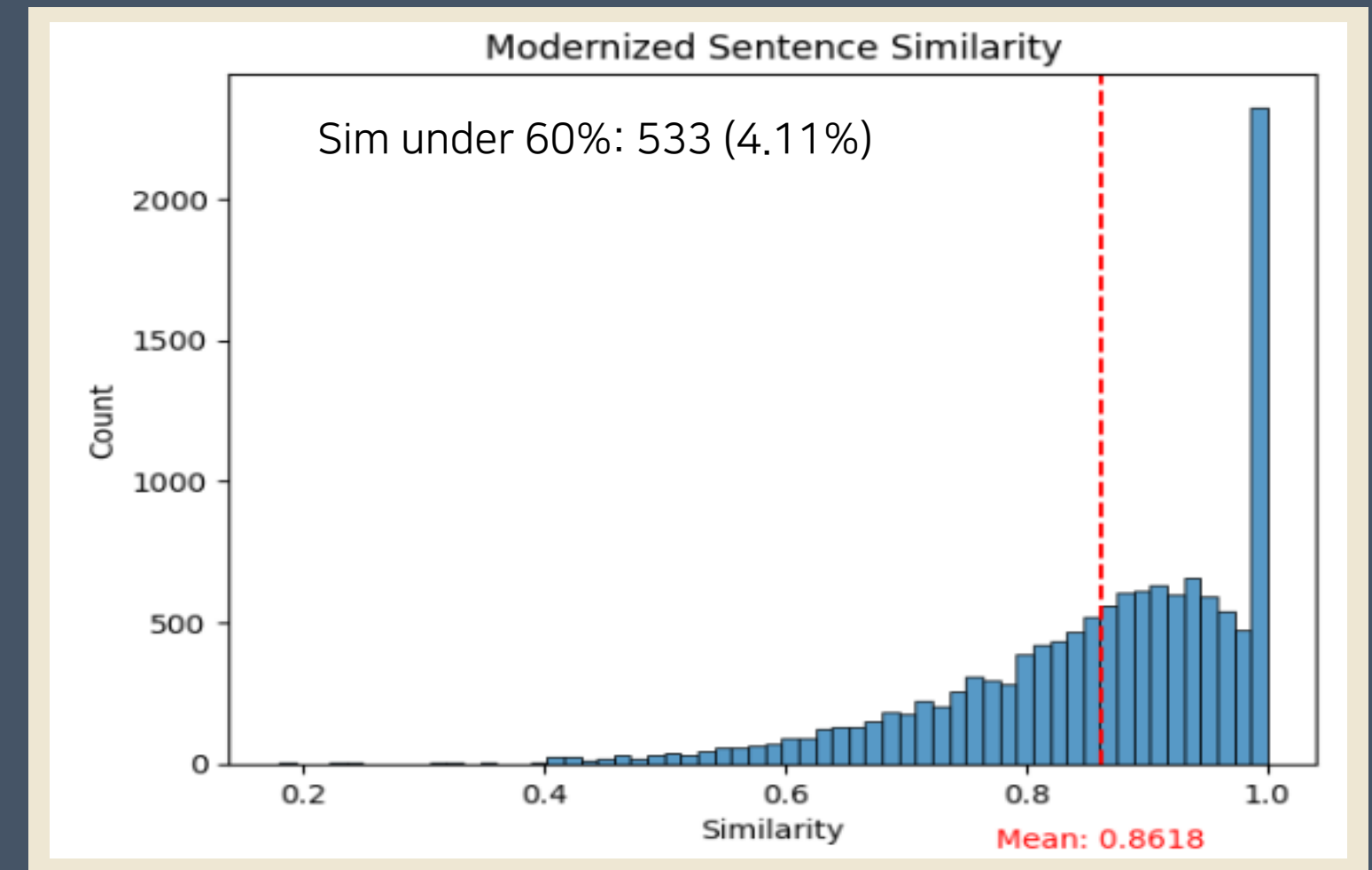
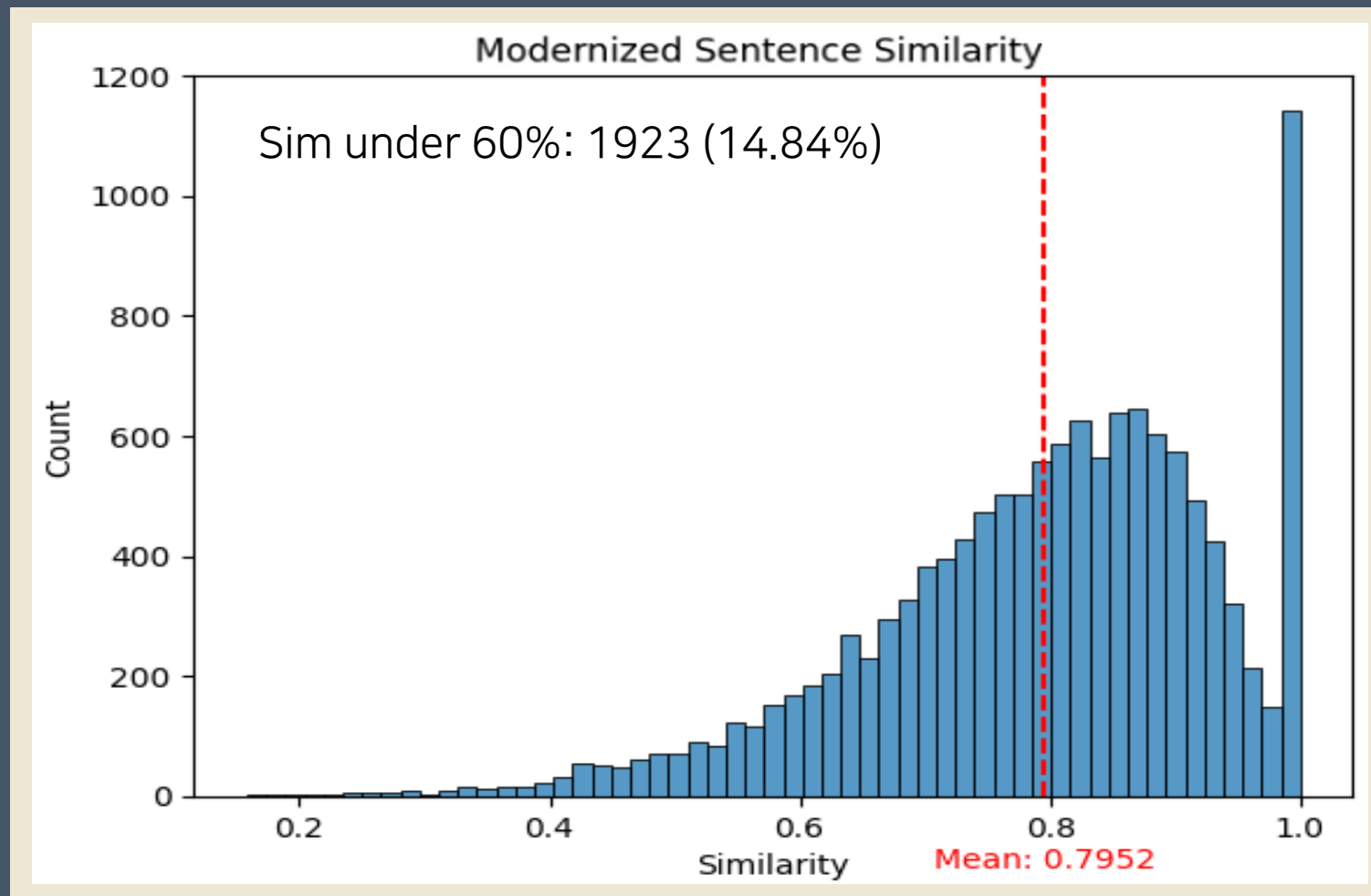
CoT with NER 현대어 변환 결과

title	cleaned_content	complex_words_and_annot	NE_tagged
금방울전 (金鈴傳)	당부하는지라. 이 때 도적이 오다가 해룡을 보고 죽이려 하다가 그 중에 장삼이란 도적이 말리며,	* 어려운 단어: 당부, 해룡, 장삼 당부: 맡은 일, 책임 해룡: 사람 이름 장삼: 사람 이름	[{'캐릭터': '해룡'}, {'캐릭터': '장삼'}]
NE_tagged_set	complex_words_converted	easy_sentence	similarity
['장삼', '해룡']	* 어려운 단어: 당부 당부: 맡은 일, 책임	맡은 일을 하고 있는 중이었어. 이때 도적이 오다가 해룡을 보고 죽이려 하다가 그 중에 장삼이란 도적이 말리며,	0.94516

↓
어려운 단어(complex_words_and_annot)에서
인명·지명(NE_tagged) 제거

↓
cleaned_content와 easy_sentence 간 유사도

CoT



고전 소설 문장과 현대어 번역 문장 간 유사도 평균이
0.79에서 0.86으로 상승하여 의미를 잘 보존하였다는 것을 확인함



4-1. LSP 산출

- 변수 생성



LSP 개요

잠재 대본 가능성(LSP, Latent Script Possibility)

NLP팀이 새롭게 고안해 낸 지표!

대본으로 바꿀 만한 장면을 추출하는 것이 최종 목표



한 소설을 여러 개의 장면으로 분할



장면으로 변환하기에 적합한지 판단

LSP 개요



장면 후보

주제 함축, 갈등 상황, 소설의 구성 단계 등



소설 지문 전체를 확보하지 못하여 데이터 자체의 한계 발생

[발단-전개-위기-절정-결말]의 형식은 기각, 절정이 될 만한 장면만 확인

변수 생성

Y (Score)

LSP

Latent Script Possibility
잠재 대본 가능성

X (Variables)

Interaction

In-degree, Out-degree

Diversity

Character Variation

Diversity

Sentiment Variation

Proportion

Scene Proportion

소설에서 대본화 가능한 장면을 추출하기 위해 **Interaction, Diversity, Proportion**을 고려함

각 변수 생성에 대한 자세한 내용은 NLP 주제분석 PPT를 참고해주세요



4-2. 대본화 지수 산출

- 대본화 지수(LSP) 산출
- 결과 시각화



대본화 지수(LSP)

생성된 X변수를 바탕으로, 각 변수마다의 함수를 생성하여 **최종 목적함수**를 제작

$$S^* = \operatorname{argmax}_{S' \subset S} Q(S')$$

$$Q(S') = INT(S') + CV(S') + SV(S') + PROP(S')$$

INTERACTION
IN / OUT-
DEGREE

CHARCTER
VARIATION

SENTIMENT
VARIATION

PROPORTION

대본화 지수(LSP)

$$Q(S') = INT(S') + CV(S') + SV(S') + PROP(S')$$

모든 Term에 대한 구현 후, 최종 목적함수 계산!

Ex) 열녀춘향수절가의 LSP SCORE TOP1

```
"열녀춘향수절가(烈女春香守節歌)": [  
  {  
    "Rank": "Top-1",  
    "Scene": "Scene #18",  
    "LSP_score": 3.4583,  
    "IntScore": 0.8333,  
    "CvScore": 1.0,  
    "SvScore": 0.8333,  
    "PropScore": 0.7917,  
    "sentence": "죽창문을 열어보니 아름다운 달빛이 방  
안으로 들어왔지만, 어린 아이가 홀로 앉아  
달을 보며 묻는 말이 있었어 ...(중략)"
```

임진록에서는

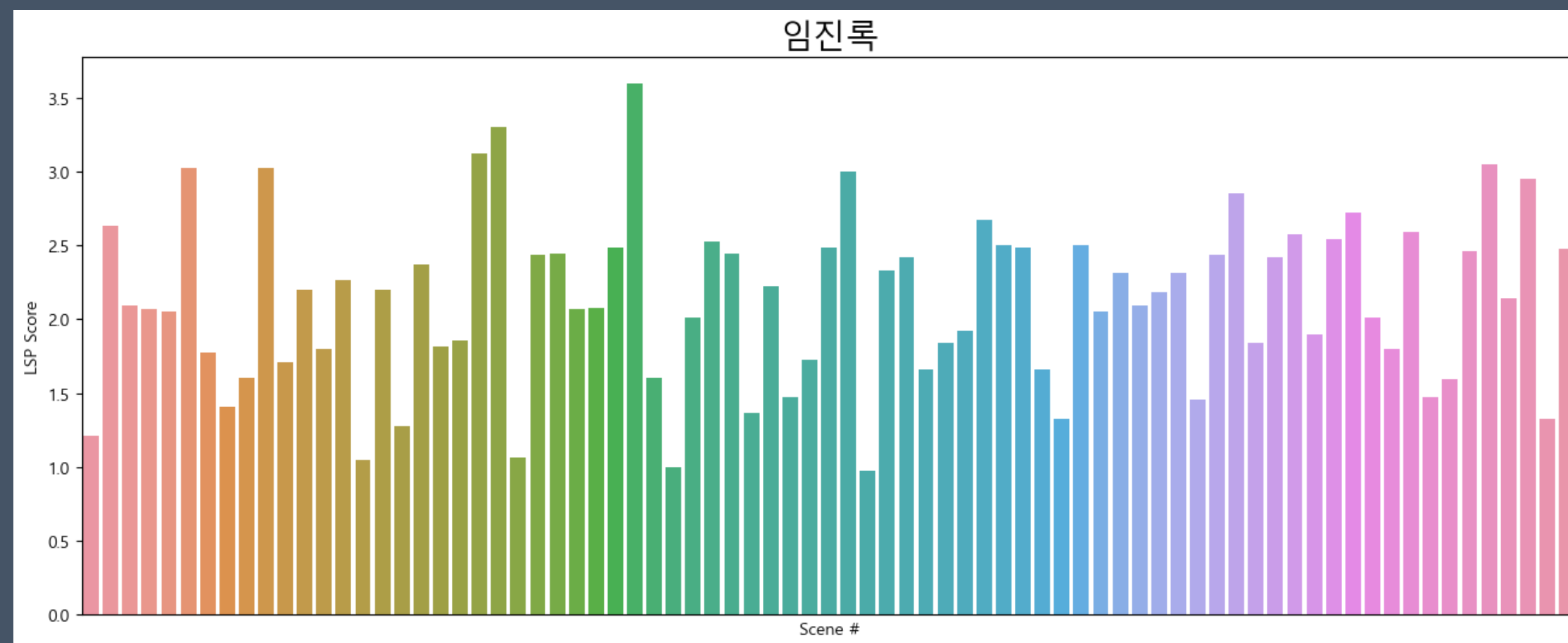
[전쟁 도중 장면], [이순신 장군의 죽음 장면]

열녀춘향수절가에서는

[춘향이 옥에 갇혀 이몽룡을 그리워하는 장면] 등

우리의 직관과 부합하는 하이라이트 장면이 뽑힘

결과 시각화



소설 [임진록]을 예시로 보면, 시각화 예시 각 장면마다 다양한 LSP Score 값을 지님



LSP가 **장면의 대본화 특징을 잘 반영한다**고 판단할 수 있음