

AECBERT

협업-콘텐츠 기반의 하이브리드 추천시스템

”

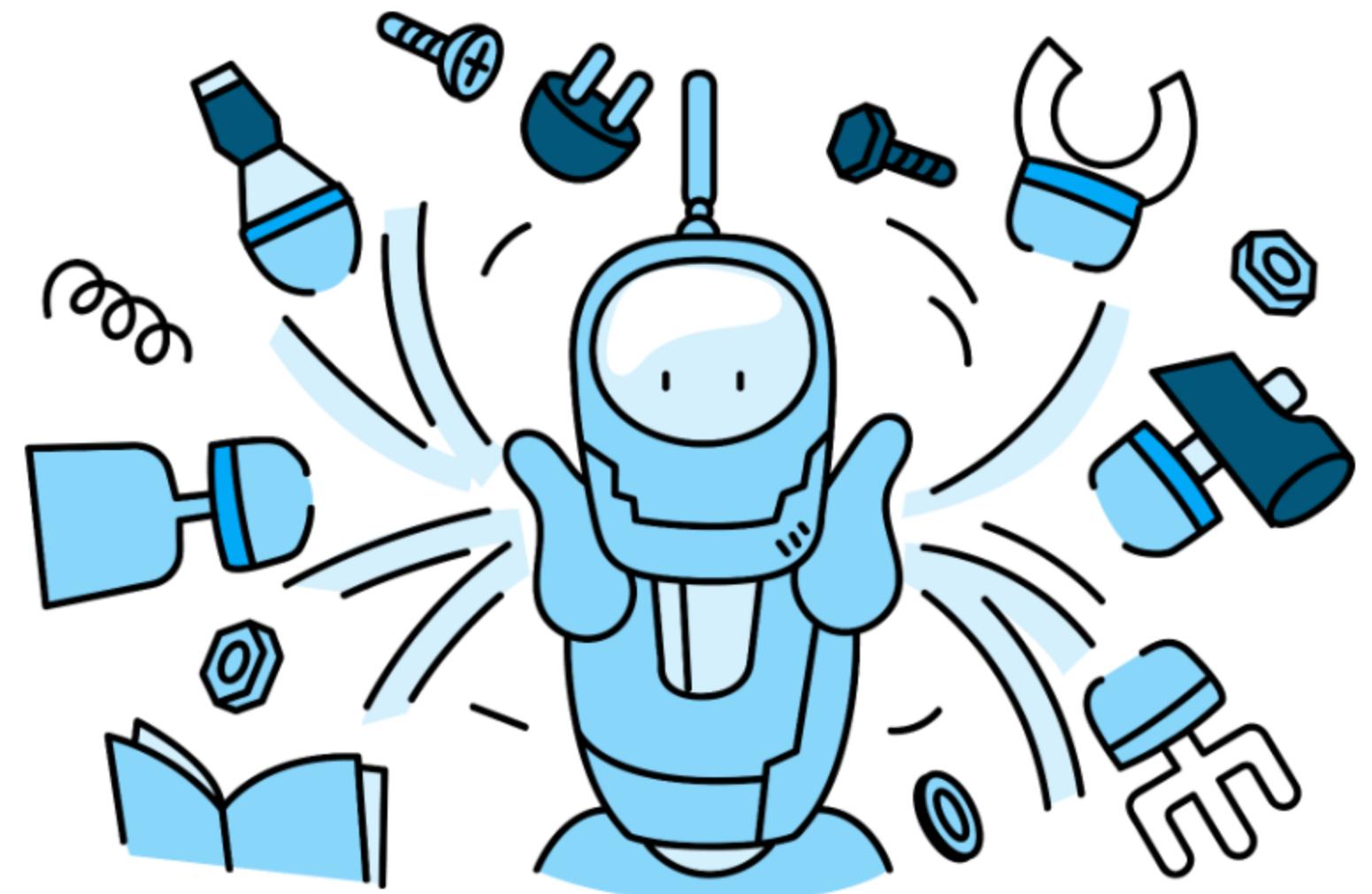
10Jobs -당신에게 드리는 10가지 직업정보

AI빅데이터융합경영학과 김서령

AI빅데이터융합경영학과 김예향

광고학전공 박지훈

행정학전공 황건하



Contents

01. 주제 선정 배경

워크넷 현황 분석

워크넷 특징 및 문제 정의

03. Modeling

데이터셋

Feature Engineering

제안 모델

Data Augmentation

02. 선행 연구

Doc2vec & BERT & AUGMENTATION
CF-DAE

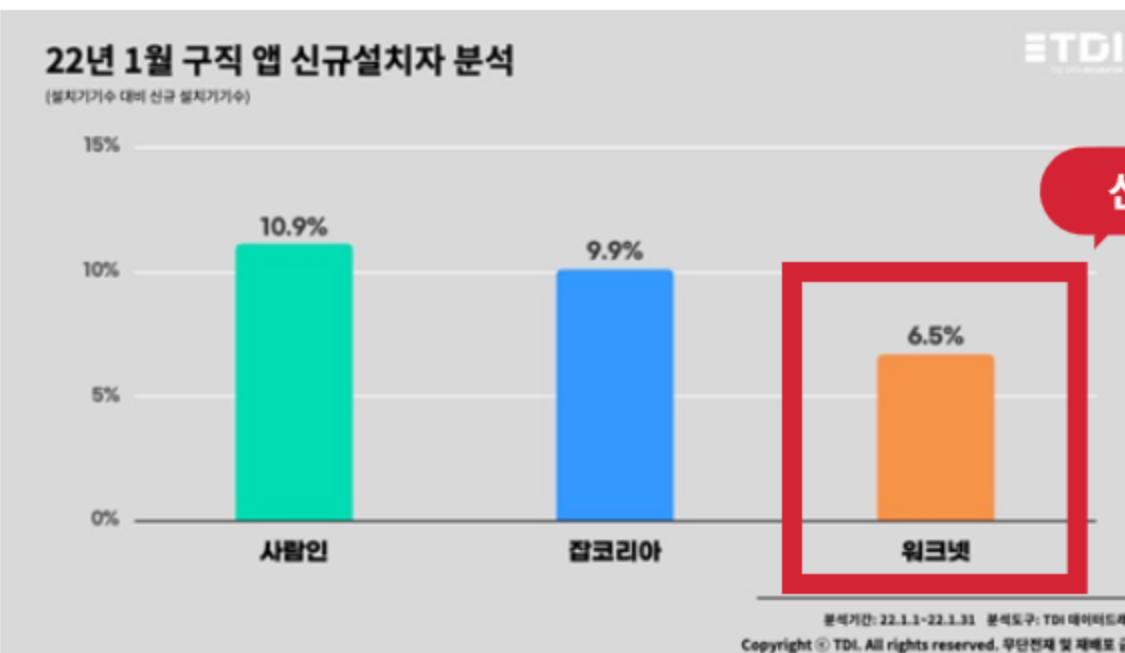
04. 실험 및 결론

Experiment

결론 및 한계

01. 워크넷 현황 분석

1. 타 구직 앱과의 사용자 수 비교

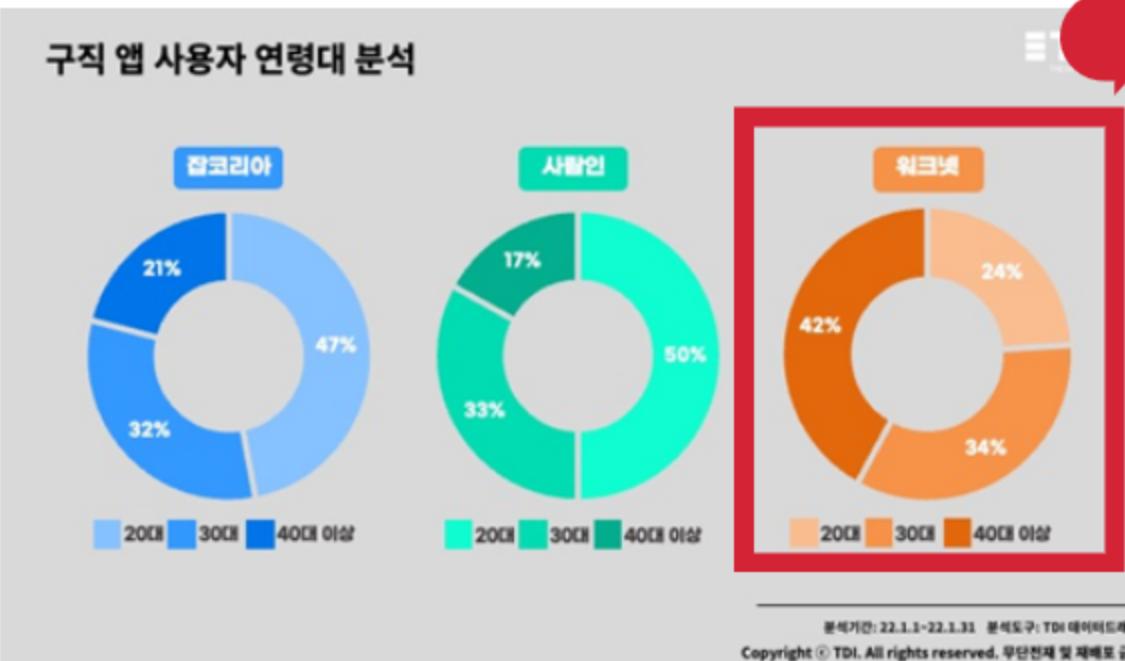


신규 설치자 수 가장 적음

3. 워크넷 채용 공고 분석

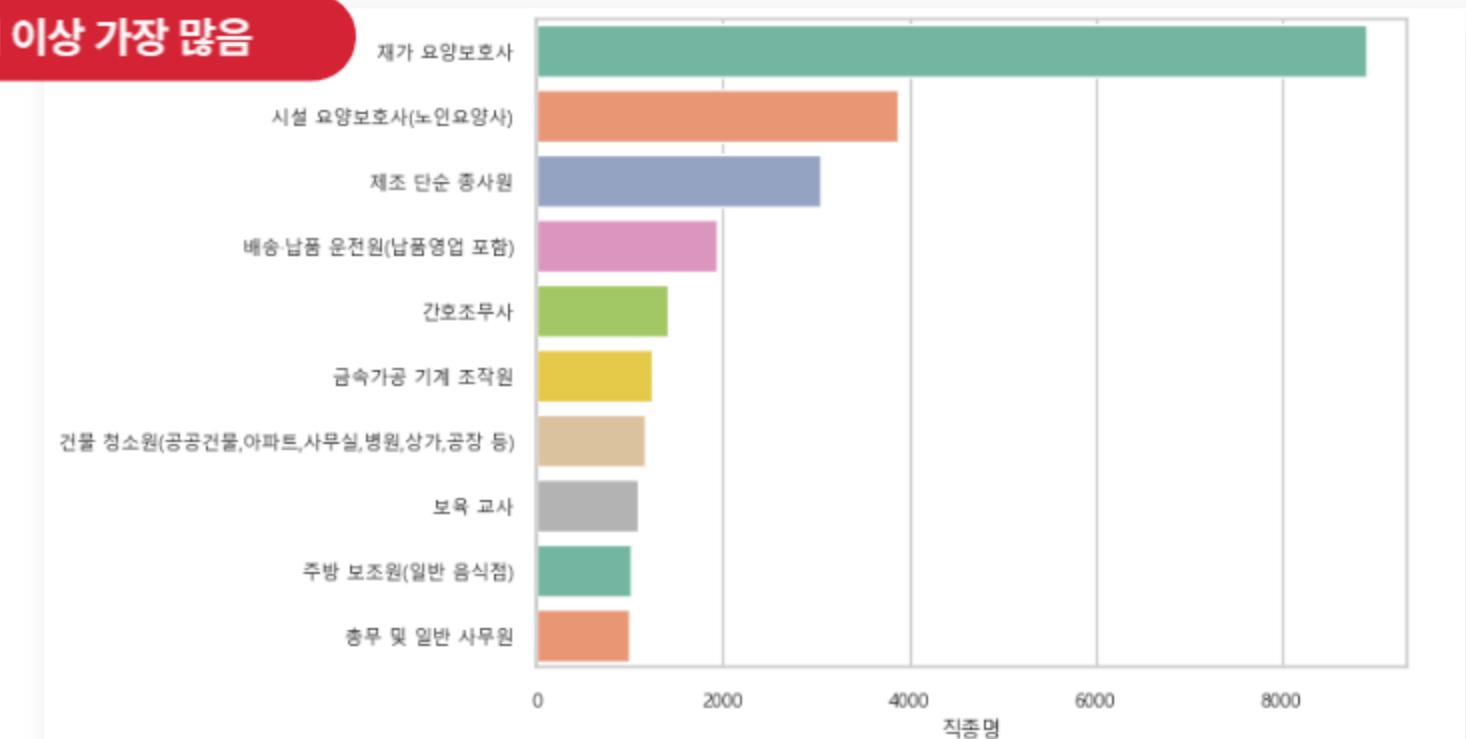


2. 타 구직 앱과의 사용자 연령대 비교



40대 이상 가장 많음

4. 워크넷 채용 직종 분석



워크넷의 사용률은 왜 저조할까?

1. 주제 선정배경

2. 선행연구

3. Modeling

4. 실험 및 결론

02. 워크넷 특징 및 문제 정의

1. 주제
선정 배경

2. 선행 연구

3. Modeling

4. 실험 및
결론

[현재 워크넷 추천 현황]

채용정보

채용정보

조회수: 682 명

오늘의 AI일자리추천

해당 후천 정보가 없습니다.

[CF]

시스템 소프트웨어 개발자
직종 희망자가 많이 본

(주)코리아티엠
[플랫폼 개발]신사옥근무/
개발자가 대우받는회사/

... 서울특별시 중구 다산로
연봉 5000만원

(주)인포이지
프로그램개발(S/W) 인력
채용공고

경기도 성남시 분당구 단...
월급 220만원

에이아이엑스 주식회사
“도약장려금” 신입 개발자
입 ...

서울특별시 금천구 가산디...
연봉 2500만원 ~ 5000...

(주)웹싱크
프로그래미 모집

워크넷 임사지원

빅데이터 추천 정보 ?

연관된 추천정보가 없습니다.

[CBF]

L 최근 업데이트 된 채용 공고는
유사한 채용 공고 추천이 이루어지지 않음

희망 직무: 개발
근무 지역: 서울

희망 직무: 개발
근무 지역: 부산

희망 직무: 의료
근무 지역: 서울

시스템 소프트웨어 개발자
직종 희망자가 많이 본

(주)코리아티엠
[플랫폼 개발]신사옥근무/
개발자가 대우받는회사/

... 서울특별시 중구 다산로
연봉 5000만원

(주)티센소프트
안드로이드, iOS App,
JAVA 시스템 개 ...

서울특별시 동대문구 탑십...
월급 220만원 ~ 500만원

(주)케어비즈
(주)케어비즈 - 웹 개발자
신입직원 모집

서울특별시 성동구 아차산...
월급 250만원

시스템 소프트웨어 개발자
직종 희망자가 많이 본

(주)코리아티엠
[플랫폼 개발]신사옥근무/
개발자가 대우받는회사/

... 서울특별시 중구 다산로
연봉 5000만원 ~ 4500...

(주)티센소프트
[2023년 개발팀 공채]
JAVA,JSP 개발 ...

서울특별시 금천구 가산디...
월급 225만원 ~ 240만원

(주)무기체계연구원
SW개발자(연구직) 채용

서울특별시 동작구 등용로
연봉 4000만원 ~ 5000...

기타 보건·의료 종사원 직
종 희망자가 많이 본

롯데건설㈜
롯데건설(주) 대연3구역
주택재개발정비사업 현장

... 부산광역시 남구 천제동로...
연봉 4000만원 ~ 4500...

삼성엔지니어링(주)
삼성엔지니어링(주)평택
PJT P4초순수 보건관리 ...

경기도 평택시 청단대로
연봉 2800만원 ~ 4000...

(주)휴엔
한국巴斯프 사무보조 파견
사원 모집(업무지원팀)

전라남도 여수시 여수산단...
월급 212만원

A와 B의 근무지역이 다름에도, 비슷한 채용공고 추천
C처럼 희망 직무가 달라야 다른 채용공고를 추천

희망 직무만 고려하는 단순한 추천 알고리즘

02. 워크넷 특징 및 문제 정의

1. 주제 선정 배경
2. 선행 연구
3. Modeling
4. 실험 및 결론

사람인 구직 플랫폼

(주)플라토어학원 관심기업 등록 채용 2
플라토 어학원 영어교사 채용(초등 전담)

경력 무관(신입포함)
학력 대졸(석사) 이상
근무형태 정규직
우대사항 상세보기 ①

급여 연봉 4,000 만원 (주 40시간) 살펴보기 ①
근무일시 주 5일(월~금)
근무지역 서울 시 마포구, 마포구, 종로구, 종로구, 종로구, 경기 고양시, 고양시 덕양구, 고양시 일산동구, 고양시 일산서구, 부천시 진도 >

이어보는 AI매치 채용정보 ①
리포트미
① 채용 주니어 영어/과학 선생님 채용
경상남도 경북면 1 대졸↑ 7/4(화) [입사지원]

페스트액키즈 어학원
ITK 학습 어학원 관리 선생님 (학원강사/ 파트강사) 채용 공고

경력 무관(신입포함)
학력 대졸(석사) 이상
근무형태 정규직
우대사항 상세보기 ①

급여 단기 후 결정
직급/학력 사원
근무일시 주 5일(월~금) 09:00~18:00
근무지역 서울 서초구, 서울전체 진도 >

이어보는 AI매치 채용정보 ①
리포트미
① 채용 주니어 영어/과학 선생님 (학원강사/ 파트강사) 채용
경상남도 경북면 1 대졸↑ 7/4(화) [입사지원]

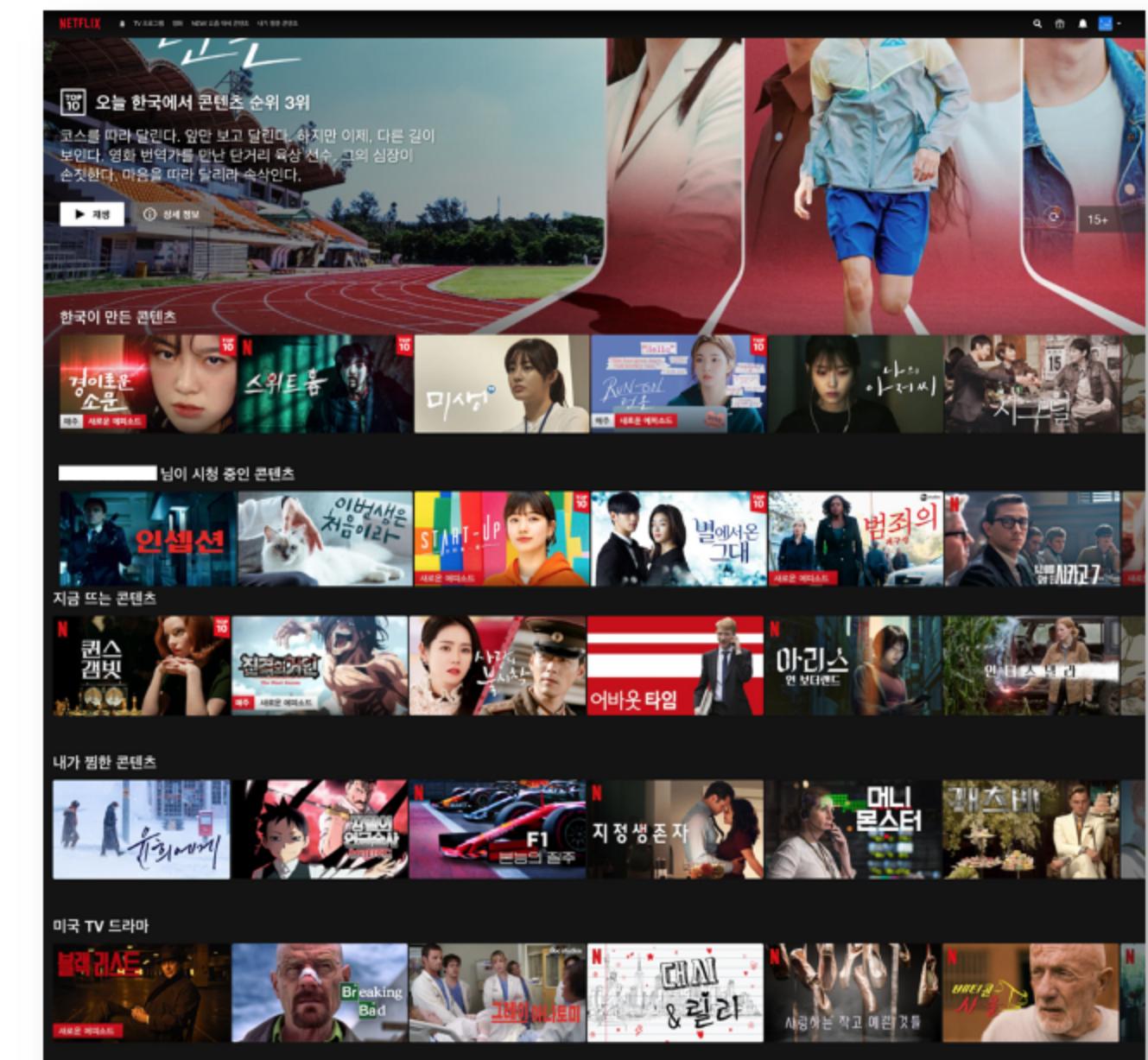
엘리트리에스비이학원(주)엘리트리
영어 유치, 초등부 담임교사 모집 (엘리트리에스비 어학원)

경력 무관(신입포함)
학력 대졸(석사) 이상
근무형태 정규직
우대사항 상세보기 ①

급여 단기 후 결정
직급/학력 사원
근무일시 주 5일(월~금) 09:00~18:00
근무지역 서울 서초구, 서울전체 진도 >

고도화된 알고리즘 추천과 최적화된 UX

넷플릭스 홈 화면



Mixed Hybrid + UI 최적화로 만족도 상승

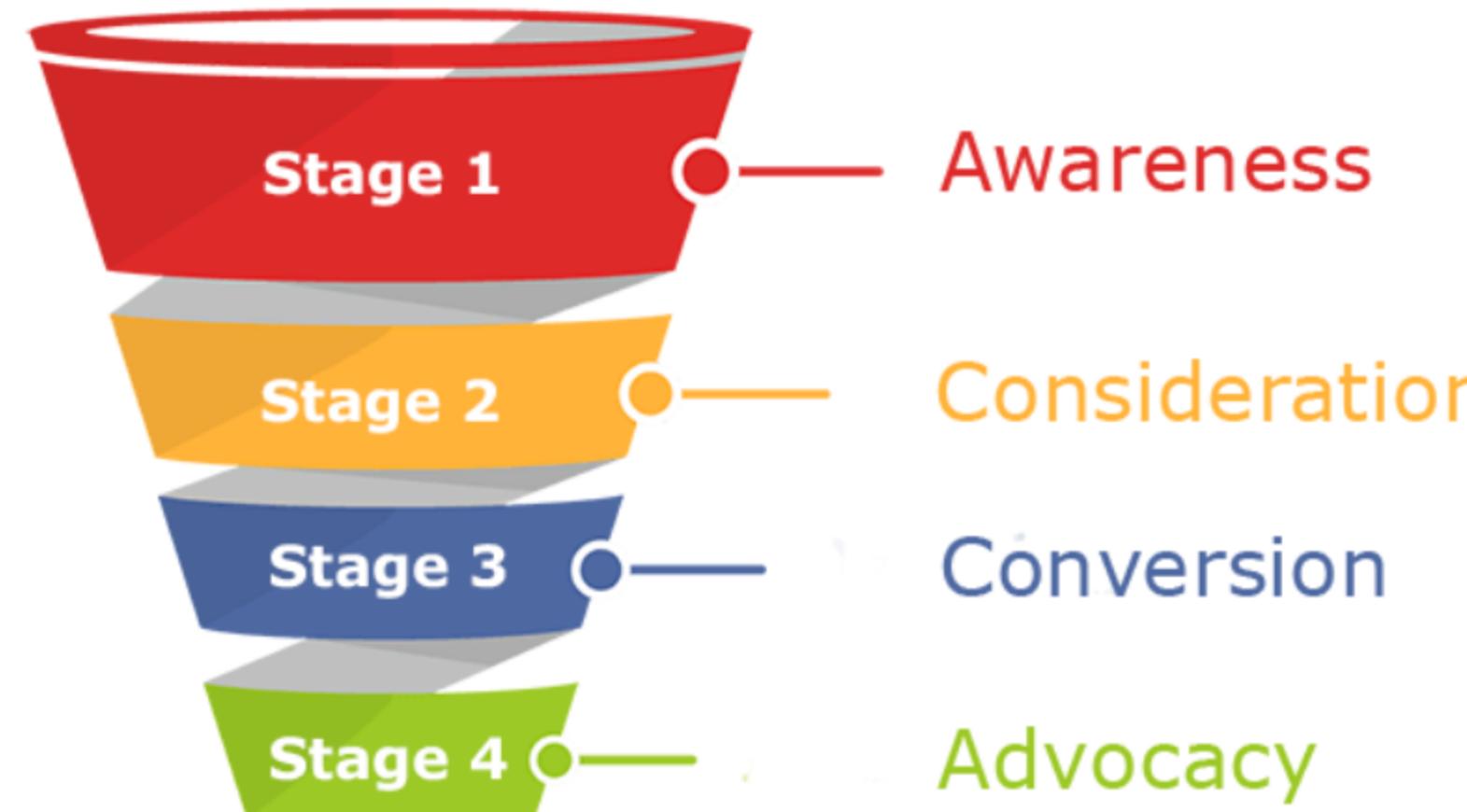


알고리즘 고도화 + Mixed hybrid로 워크넷 개선

02. 워크넷 특징 및 문제 정의

1. 주제 선정배경
2. 선행연구
3. Modeling
4. 실험 및 결론

유저의 action flow (구직 경험)



채용공고에 관심

채용공고 검색 및 조회

채용공고에 지원

워크넷은 action flow 최적화 X
유저가 선호하는 채용공고에 진입하도록
매끄러운 유도가 이루어지지 않음



유저의 action flow 최적화를 위한 추천시스템을 설계

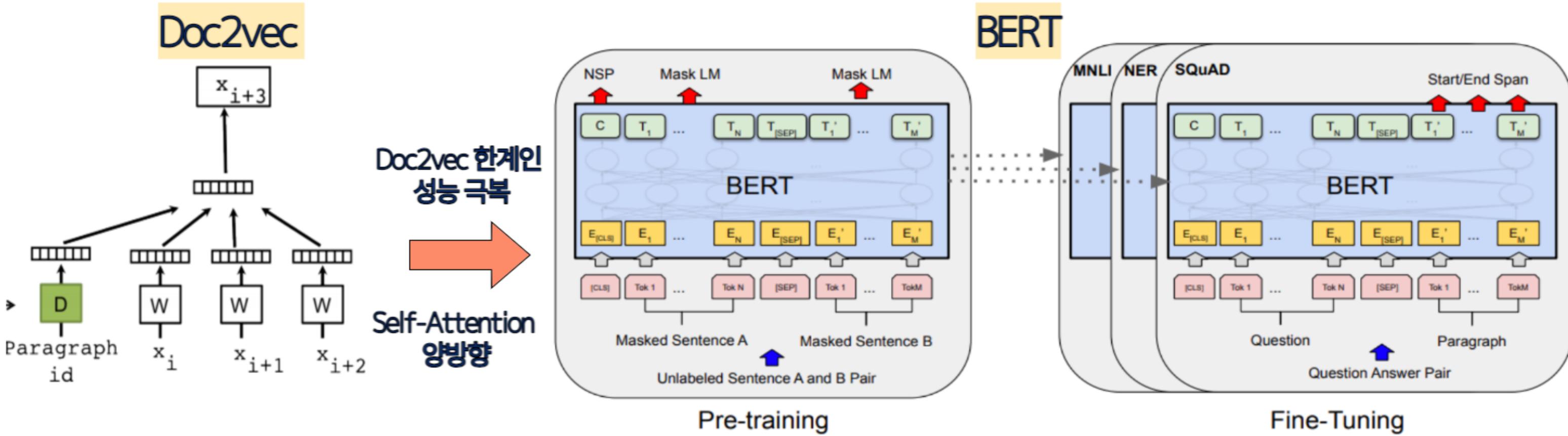
01. Doc2vec & BERT & AUGMENTATION

1. 주제
선정배경

2. 선행연구

3. Modeling

4. 실험 및
결론



BERT의 또 다른 Downstream Task로 문서를 임베딩하는 기법에 초점

Similar Embedding Replacement

모델을 통해서 학습된 Word Embedding vector를 기준으로 비슷한 정도를 판단

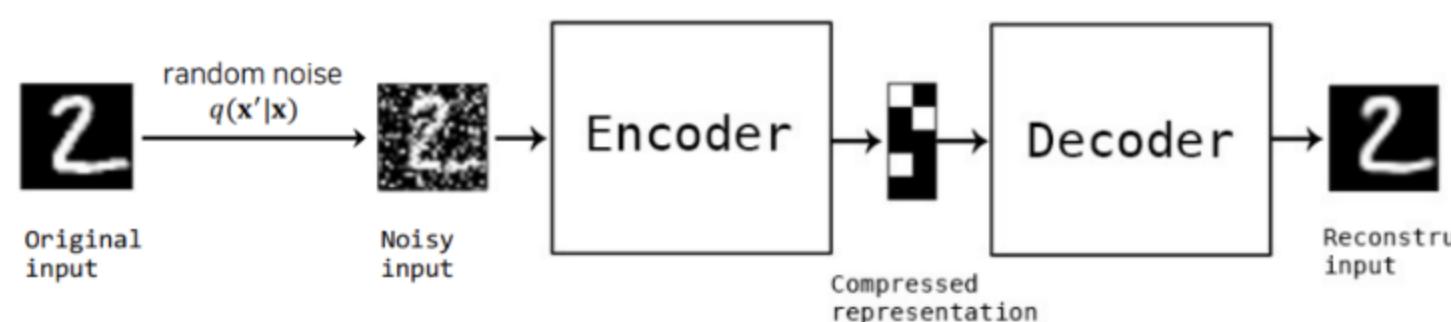
weather	ungratefulness	traffic	timewasting	talkative	swearing	stability	snobbish
rains	helped	cop	wastingmytime	Tweeters	curse	mood	smut
STORM	ungrateful	lane	colleagues	Xs	teary	sensitive	intellectual
Blizzarad	clearly	pulled	Wen	whit	qweet91	dudes	moneycars
snowed	r	speed	BruklynFinest	sheesh	swears	nigga	LoWQUI
SNOW	them	Slow	hold	TwitterJail	10	up	lifestyle
smoking	silence	showoff	sexual	services	selfishness	repetition	religious
JAYECANE	guilty	louis	box	fil	ONLY	dislike	sinners
reggie	R	rims	wonder	requests	Selfish	repeat	IAmKevinTerrell
smoking	response	sein	Preach	convos	selfish	myself	spiritual
smoke	conversation	makin	suck	TIP	stay	same	CHURCH
smokers	sending	bag	pussy	products	hit	over	FOLK

장점

1. 사람의 손으로 Label을 달 필요가 없음
2. 단어가 바뀐 전체 문장의 Label 자체는 비슷함

Word2vec을 통한 Embedding과
Cosine Similarity를 기준으로 KNN 적용

02. CF-DAE



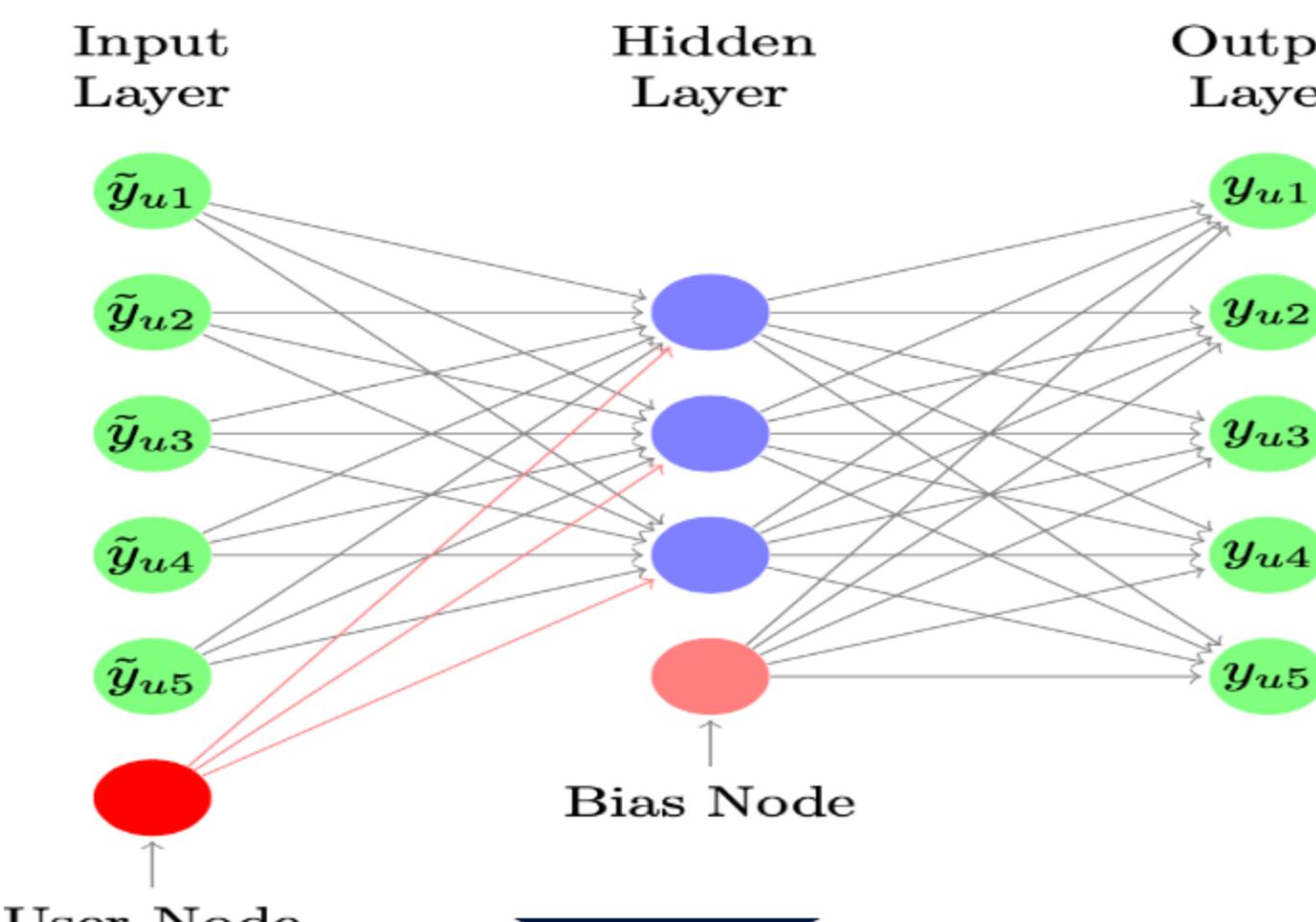
입력데이터에 임의로 노이즈한 input을 만들기위해
random noise or dropout을 추가하여 학습

1. 주제
선정배경

2. 선행연구

3. Modeling

4. 실험 및
결론



Input(Noisy Input)을 잘 복원할 수 있는 robust한 모델이
학습되어 전체적인 성능향상을 한다는 것이 DAE의 핵심

AutoRec

평점 예측 모델

그대로 input

개별유저 학습 X

CF-DAE

Top-N 추천모델

DAE로 Noisy Input

개별유저 학습 O

CDAE와 다른 모델들을 비교한 결과

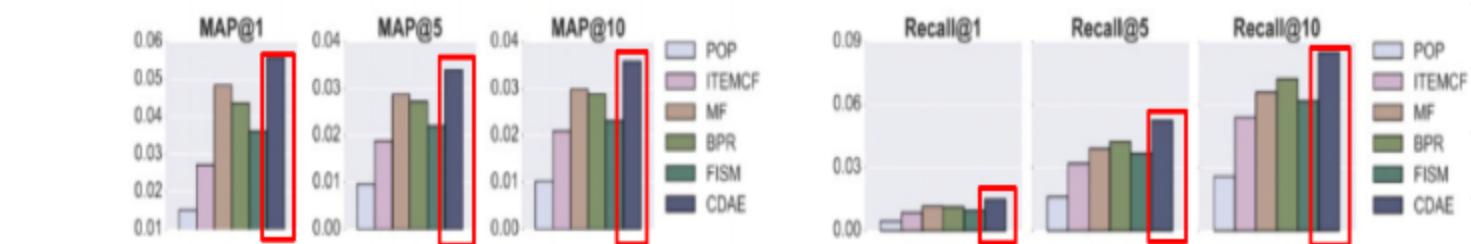


Figure 6: MAP scores with different N on the Yelp data set.

Figure 9: The Recall scores with different N on the Yelp data set.



Figure 7: MAP scores with different N on the MovieLens data set.

Figure 10: The Recall scores with different N on the MovieLens data set.

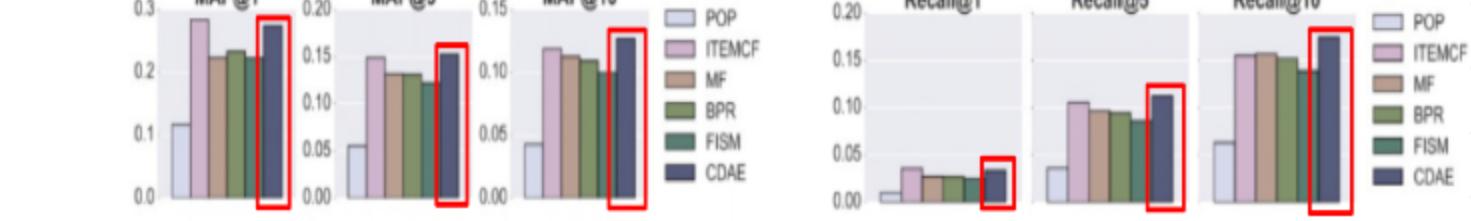


Figure 8: MAP scores with different N on the Netflix data set.

Figure 11: The Recall scores with different N on the Netflix data set.

다른 Top-N 추천모델에 비해 높은 MAP과 recall

01. 데이터셋

1. 채용공고 조회 데이터

암시적 피드백 데이터(Implicit feedback data)
조회 유무를 기반으로 구성된 이진 데이터

	0	1	2	3	4	5	6	7	8	9	...
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
1	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...
2	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	...
...
5995	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
5996	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...
5997	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
5998	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
5999	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...

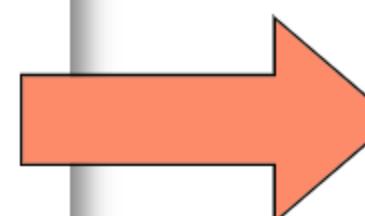
2. 채용공고 문서

워크넷 API 채용공고, 채용상세 데이터 수집

구인인증번호	관련직종	직무내용	고용 형태	모집 인원	전 과정	자격 면허	무대 조건	근무시간/현태
0 K151532306010009	중·대형 화물차 운전원(트레이일리 포함)	시내버스 경비자 두대 대형 화물차 또는 택배 운송 경비자 군 수송부 운전 ...	비정 규직	5	임 종	버스 운전 자	없음	평일 2교대 근무 모전 주 06시부터 7~8시간 근무 ...
1 K140022306010111	없음	자동차부품 고무제품 제작 사출기 조작 사상기 조작 착자기 조작 고...	정규 직	20	임 종	없음	고용 회가 제	평일 모전 8시 30분 모후 5시 30분 주 5일 근무
2 K170052306010102	종무 및 일반 사무원	전문건설 공무 보조입니다 캐드 사용 가능자 우대이며 문량산출 기성 ...	정규 직	1	임 종	없음	평일 모전 9시 00분 모후 5시 00분 주 5일 근무 평균근무시간 ...	
3 K170052306010101	기타 전기·전자 기기 설치·수리원	보일러 수리 및 설치기사	정규 직	1	임 종	없음	동전 면허증	평일 모전 8시 30분 모후 5시 30분 주 6일 근무 평균근무시간 ...
4 K150012306010195	없음	입소자 위생관리 입소자 일상생활 지원	정규 직	1	임 종	요양 보호사	없음	평일 모전 9시 00분 모후 6시 00분 주 5일 근무 평균근무시간 ...

['22년 추진실적]

구 분	추 진 실 적
취업 희망지역 20Km 이내 일자리 우선 추천(1월)	
구직자가 입사지원할 것으로 판단하는 기준점수(매칭점수) 상향(0→0.5점 5~9월)	
AI 일자리 추천 시스템에 취업 관련 데이터 학습(7월)	
AI 매칭 서비스의 알고리즘 고도화를 위한 연구용역 추진(9~12월)	
AI가 분석하는 입력변수를 기존 8종*에서 2종** 추가(11월)	
구인 직종, 구직 희망직종, 구인 근무지역, 구직 희망지역, 구인기업 제시임금, 구직자 희망임금, 구인 업종, 구직자 경력기간	
** 구직자의 입사지원 빈도, 구인기업이 입사지원을 받은 빈도	



워크넷에서 진행된 추천시스템 연구는 머신러닝 기반이며 알고리즘 고도화가 미흡한 실정이다.

즉, 충분한 연구가치를 가진다

1. 주제
선정배경

2. 선행연구

3. Modeling

4. 실험 및
결론

02. Feature Engineering

■ 불필요한 Feature 삭제 및 전처리

구인인증번호	관련직종	직무내용	고용 형태	모집 인원	전 공	자격 면허	우대 조건	근무시간/형태	연금4대보험	장애인 편의시설	회사명	임금 형태	최소임금 액	근무 지역	근무 형태	최소 학력	경력	고용형태코드	2단계 직종코드
0 K151532306010009	증·대형 화물차 운전원(트레일러 포함)	시내버스 경력자 우대 대형 화물차 또는 택배 운송 경력자 군 수송부 운전 ...	비정규직	5	없음	버스 운전자	없음	평일 2교대 근무 오전 06시부터 7~8시간 근무 ...	국민연금 고용보험 산재보험 의료보험	0	(주)미래교통	월급	3000000	경기 구리시	주6일 근무	학력 무관	관계 없음	20 079	
1 K140022306010111	없음	자동차부품 고무제품 제작 사출기 조작 사상기 조작 착자기 조작 고...	정규직	20	없음	없음	고용 허가 제	평일 오전 8시 30분 오후 5시 30분 주 5일 근무	국민연금 고용보험 산재보험 의료보험	0	에스티엠주식회사	시급	9620	경북 군위군	주5일 근무	학력 무관	관계 없음	10 112	
2 K170052306010102	총무 및 일반 사무원	전문건설 공무 보조 입니다 캐드 사용 가능자 우대이며 물량산출 기성 ...	정규직	1	없음	없음	없음	평일 오전 9시 00분 오후 5시 00분 주 5일 근무 평균근무시간 ...	국민연금 고용보험 산재보험 의료보험	0	(주)구상건설	월급	2000000	대전 유성구	주5일 근무	고졸	관계 없음	11 018	
3 K170052306010101	기타 전기·전자 기기 설치·수리원	보일러 수리 및 설치기사	정규직	1	없음	없음	운전면허증	평일 오전 8시 30분 오후 5시 30분 주 6일 근무 평균근무시간 ...	국민연금 고용보험 산재보험 의료보험	0	경동나비엔대전 유성중앙대리점	월급	2200000	대전 유성구	주6일 근무	학력 무관	신입	10 091	
4 K150012306010195	없음	입소자 위생관리 입소자 일상생활 지원	정규직	1	없음	요양 보호사	없음	평일 오전 9시 00분 오후 6시 00분 주 5일 근무 평균근무시간 ...	국민연금 고용보험 산재보험 의료보험	0	평북영락양로원	월급	2010580	인천 연수구	주5일 근무	학력 무관	관계 없음	10 038	

■ Feature 삭제 기준:

1. 각 데이터의 고유한 값(Url, 상세주소)

2. 모든 데이터에 공통적으로 출현하여 분석에 의미가 없는 feature들 삭제

ex) 사업자등록번호, 최대학력, 정보제공처, 워크넷 url, 워크넷 모바일 url, 근무지 우편주소, 근무지 도로명주소, 근무지 기본주소, 근무지 상세주소, 최종수정일, 우대조건

■ 전처리 수행

1. 결측치 제거

2. 특수문자 제거

3. Feature 통일 및 분리 등

1. 주제 선정 배경

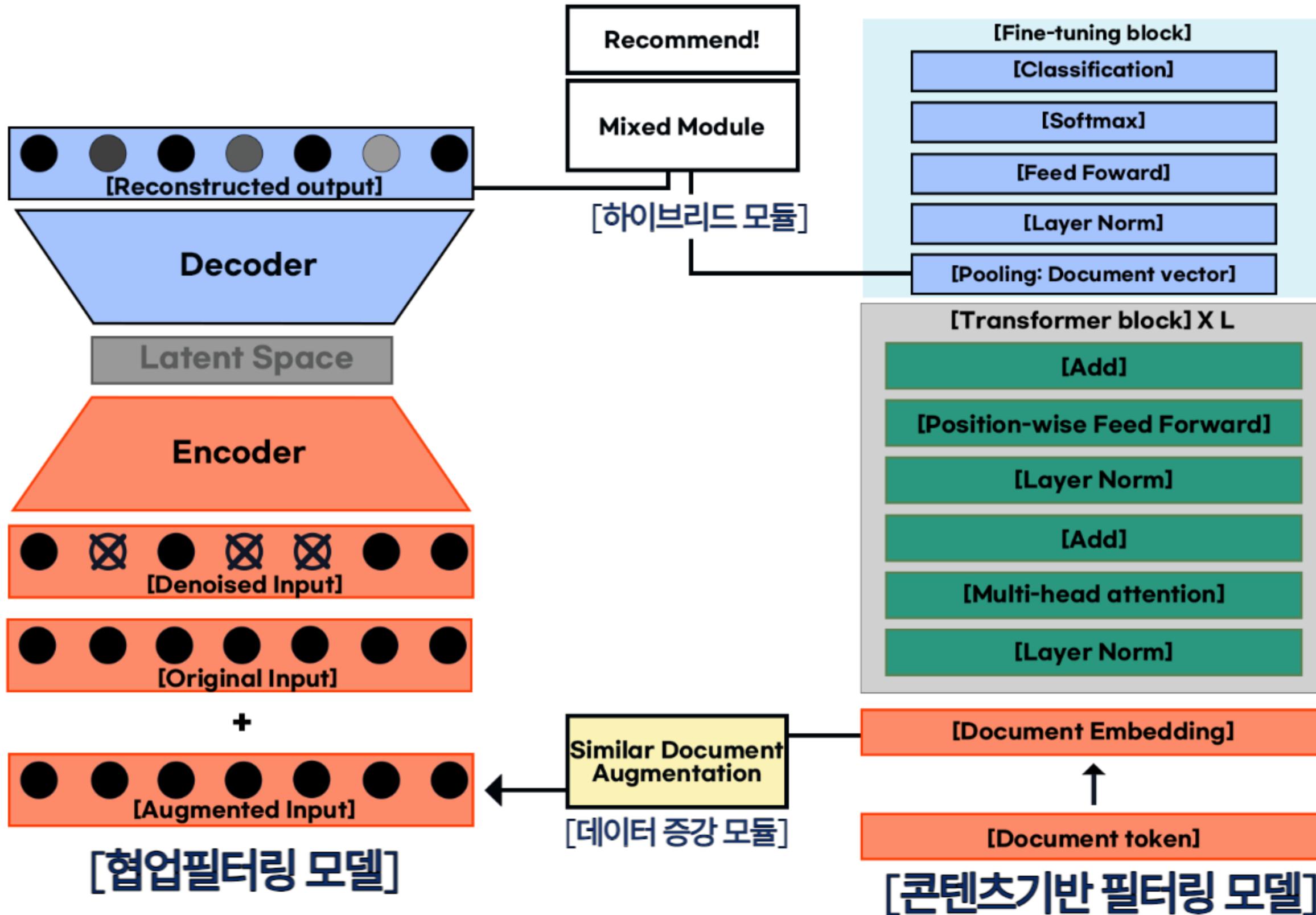
2. 선행 연구

3. Modeling

4. 실험 및 결론

03. 제안 모델

AECBERT : Architecture



03. 제안 모델

Phase 1 : Content based filtering

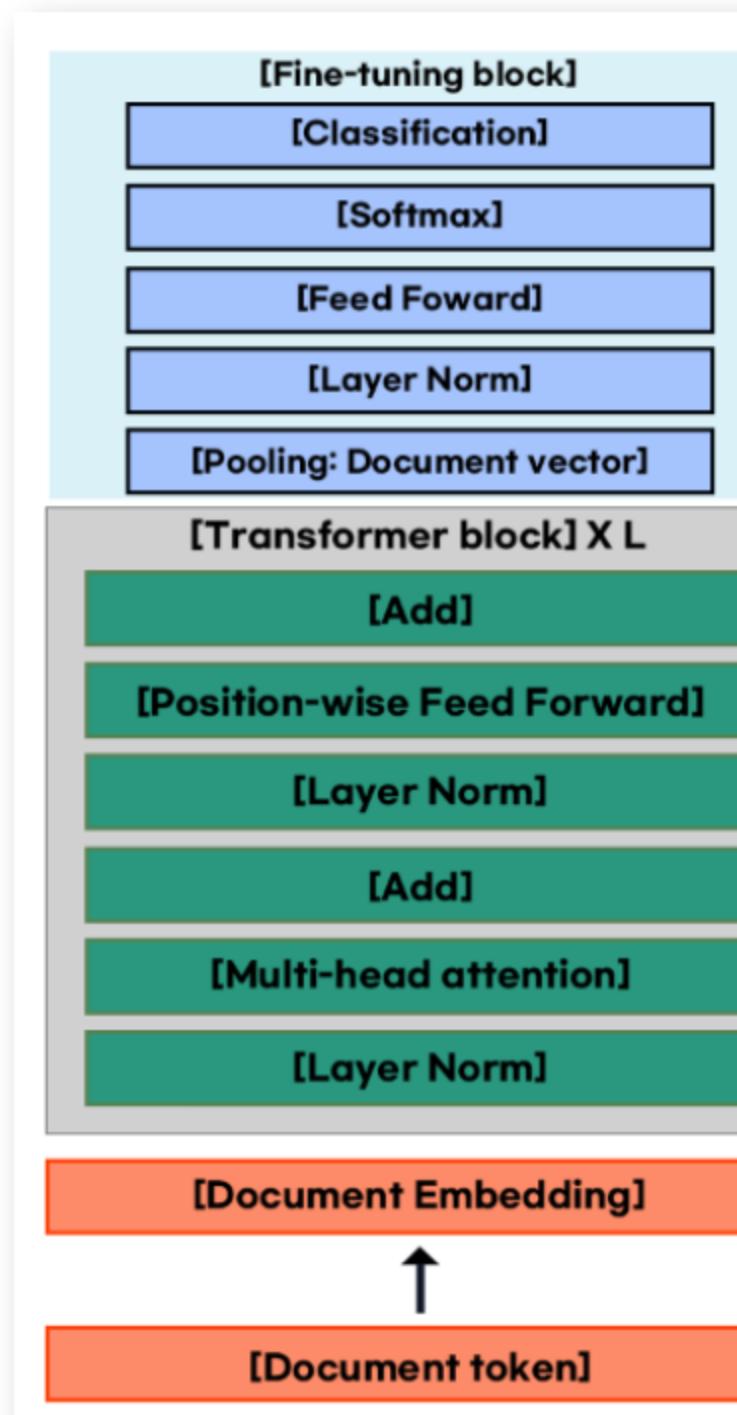
Pretraining : Masked Language Model

1. 주제
선정배경

2. 선행연구

3. Modeling

4. 실험 및
결론



■ Document Representation: 문서에 구성된 각 토큰 임베딩의 평균값

■ No Positional Embedding:

- 기존 BERT 모델에서는 단어 간의 순서 정보를 포착하기 위한 Positional Embedding을 사용하였다.
하지만 채용 공고 데이터셋은 단어 간의 종속성이 아닌 feature 간의 종속성을 나타내기 때문에
각 토큰 간의 순서 정보를 담지 못한다. 때문에 Positional Embedding을 제거한다.
Ex. 요양보호사 + 월급 200만원

Input & Output

Trainset : {item id : [T1, T2, T3 ... Tn]}

input : [Document Token]

output : Hidden state: (item, sequence length, embeded size)

Masked Language Modeling

지정한 mask 비율(p)에 따라 마스킹을 진행

모든 input값을 맞히는 것이 아닌 mask된 input만 맞히는 학습을 한다

Input: $[v_1, v_2, v_3, v_4, v_5]$ $\xrightarrow{\text{randomly mask}}$ $[v_1, \text{[mask]}_1, v_3, \text{[mask]}_2, v_5]$

Labels: $\text{[mask]}_1 = v_2, \text{[mask]}_2 = v_4$

03. 제안 모델

Phase 1 : Content based filtering

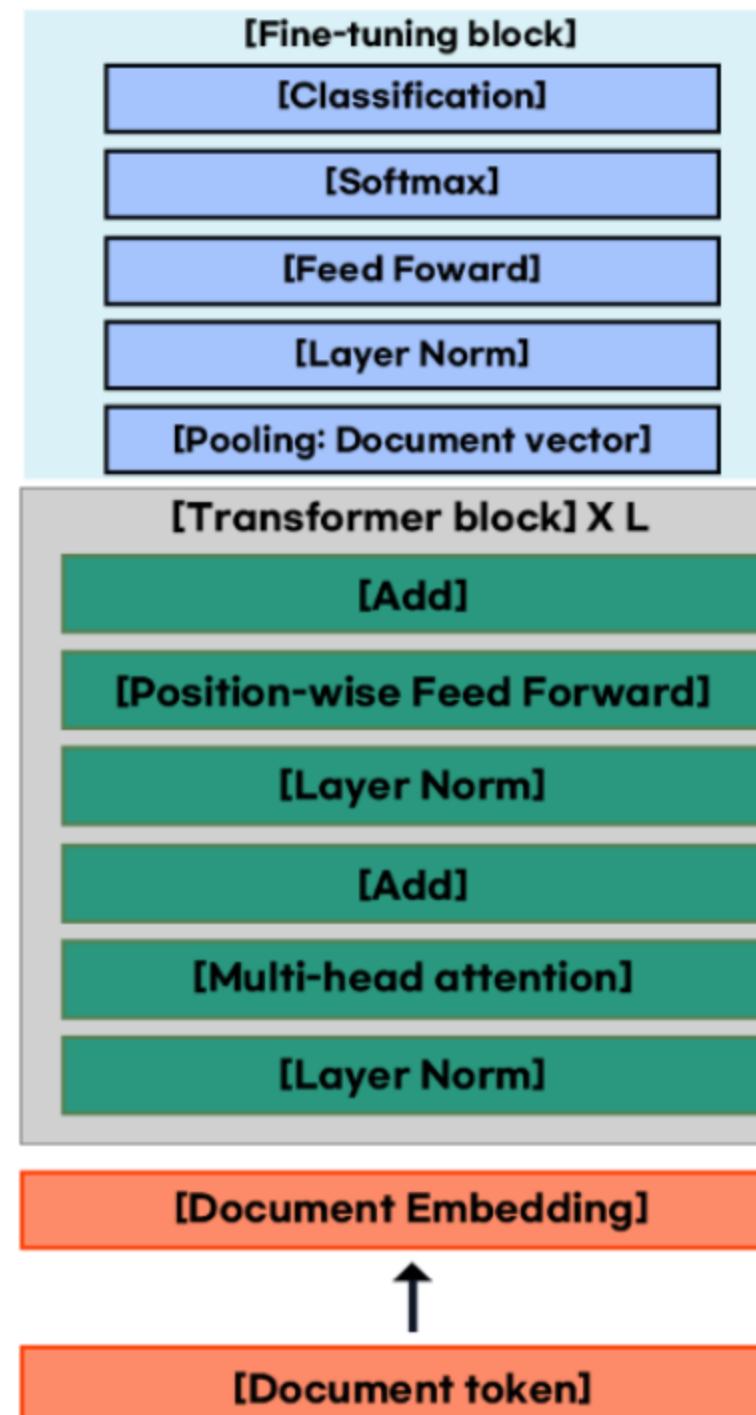
Fine-tuning : Document classification

1. 주제
선정배경

2. 선행연구

3. Modeling

4. 실험 및
결론



■ Document Representation vector를 얻는 방법
Document를 구성하는 Token들을 집계하여 문서를 표현

[Doc1]에 대한 임베딩 추출

Doc1	0.25	0.53	-0.17	-0.47
------	------	------	-------	-------

생산	0.15	-0.24	0.44	0.73
서울	-0.52	0.66	0.27	-0.89
경력	0.79	0.56	0.09	-0.5
정규직	0.32	0.61	-0.91	-0.22

H1	0.42	0.58	0.672	0.52
H2	0.51	-0.40	0.19	0.69
H3	0.25	0.76	-0.9	-0.7
H4	-0.25	0.23	-0.14	-2.1

DownSteam Task에 사용한
Hidden state별 성능

Layer	Test error rates(%)
Layer-0	11.07
Layer-1	9.81
Layer-2	9.29
Layer-3	8.66
Layer-4	7.83
Layer-5	6.83
Layer-6	6.83
Layer-7	6.41
Layer-8	6.04
Layer-9	5.70
Layer-10	5.46
Layer-11	5.42
First 4 Layers + concat	8.69
First 4 Layers + mean	9.09
First 4 Layers + max	8.76
Last 4 Layers + concat	5.43
Last 4 Layers + mean	5.44
Last 4 Layers + max	5.42
All 12 Layers + concat	5.44

가장 마지막 4개의 hidden state를 이용하는 것이
단어 간의 contextualization이 잘 되어있기
때문에 fine-tuning 시에도 성능이 좋게 나왔다

03. 제안 모델

Phase 1 : Content based filtering

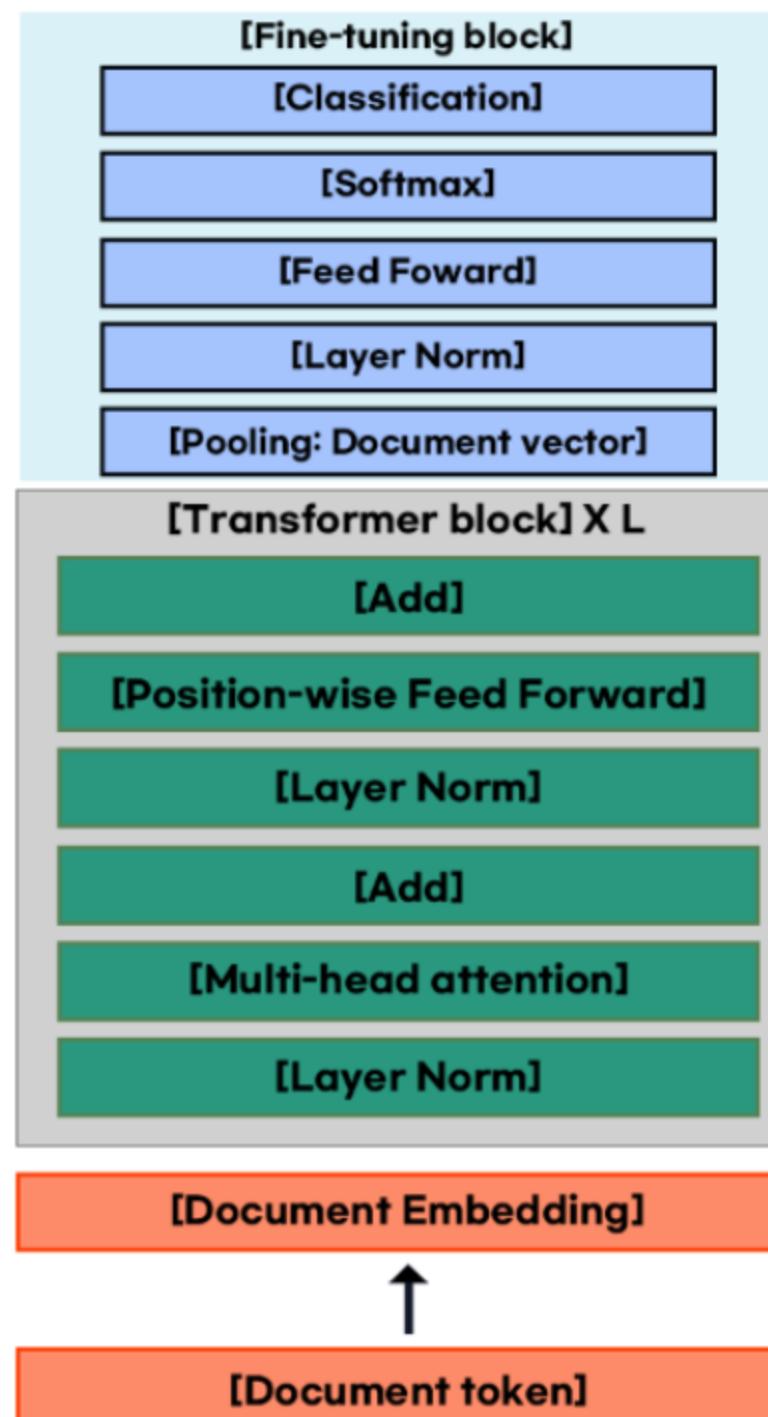
Fine-tuning : Document classification

1. 주제
선정배경

2. 선행연구

3. Modeling

4. 실험 및
결론



■ pretraining에서 얻은 Document embedding vector에 대한 학습 진행
학습 목표: 같은 분야의 채용공고가 비슷한 공간에 임베딩 되도록 한다
Pretraining에서는 각 Word간의 관계를 학습한 벡터들의 합으로 문서를 표현할 수 있었다

■ Fine Tuning Process

여기까지의 단계로 충분하게 채용공고간의 유사도를 구할 수 있지만
임베딩을 정교화하는 Fine-Tuning 학습을 추가로 제안한다
Input: 각 토큰의 벡터가 Pooling된 Document Representation
Output: 채용공고의 직종 분류 확률
Label: 직종코드 (112개 Class)
ex) 어업종사자, 낙농사육, 건설·채굴

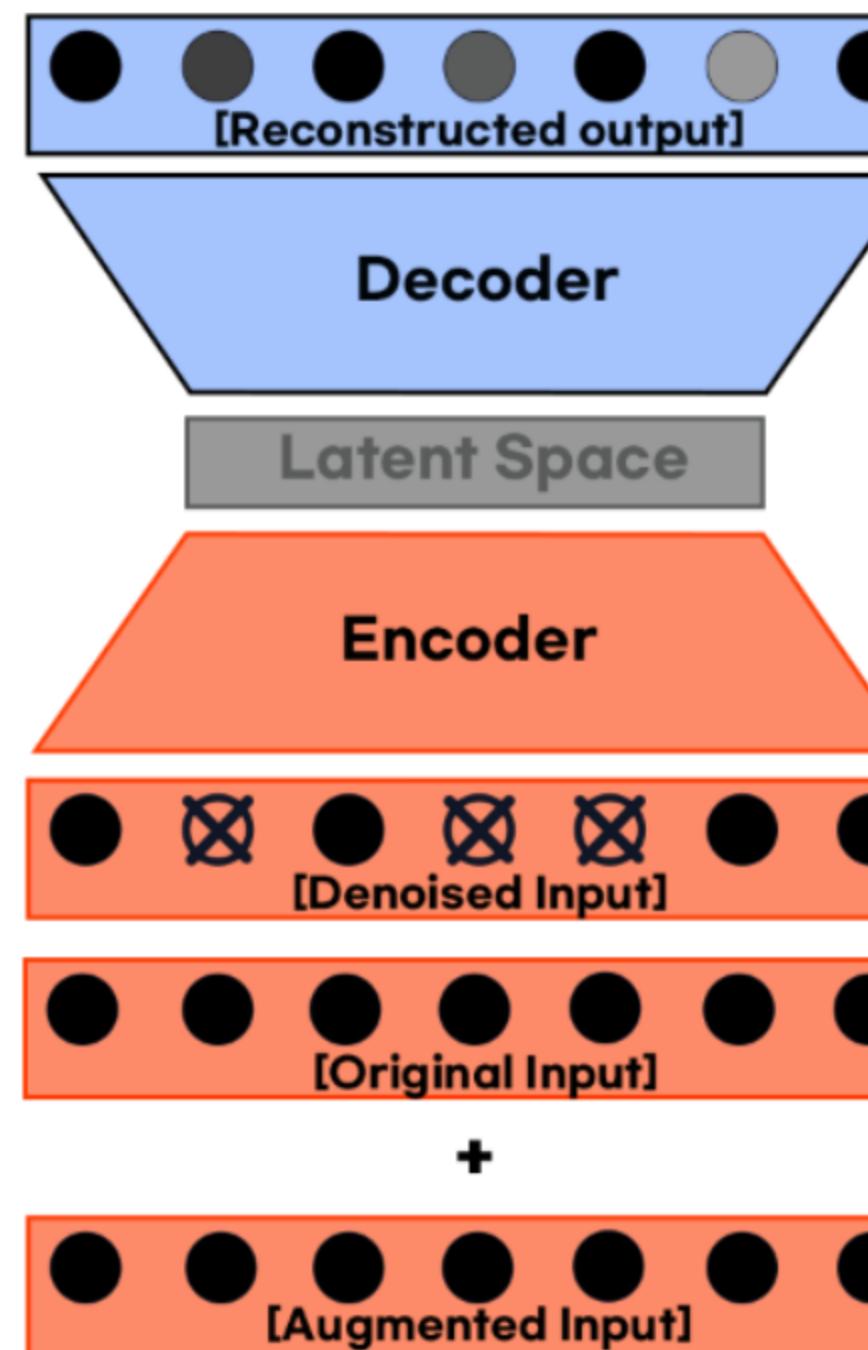
03. 제안 모델

1. 주제
선정배경

2. 선행연구

3. Modeling

4. 실험 및
결론



[Main Idea]

증강된 유저 데이터를 이용하여 Reconstruct하되 일반화 가능성을 높이자

$y_u \in R^N$: Ground Truth

$\tilde{y}_u \in R^N$: Noised vector

$\tilde{y}_u + V_u$: Input

$z_u = h(W^T \tilde{y}_u + V_u + b)$: Latent Space

$\hat{y} = f(W'^T z_u + b')$: Output

04. Data Augmentation

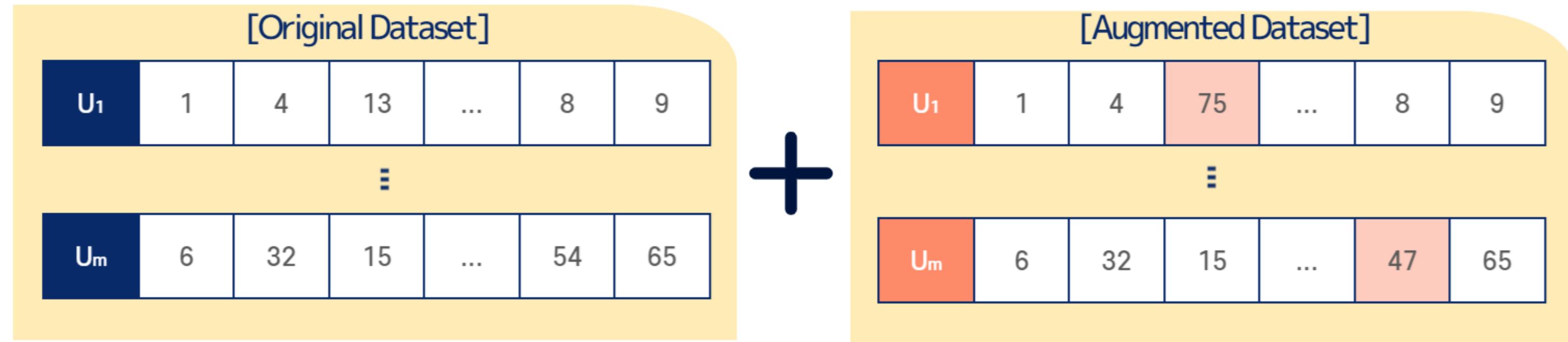
SDA(Similar Document Augmentation)

1. 주제
선정배경

2. 선행연구

3. Modeling

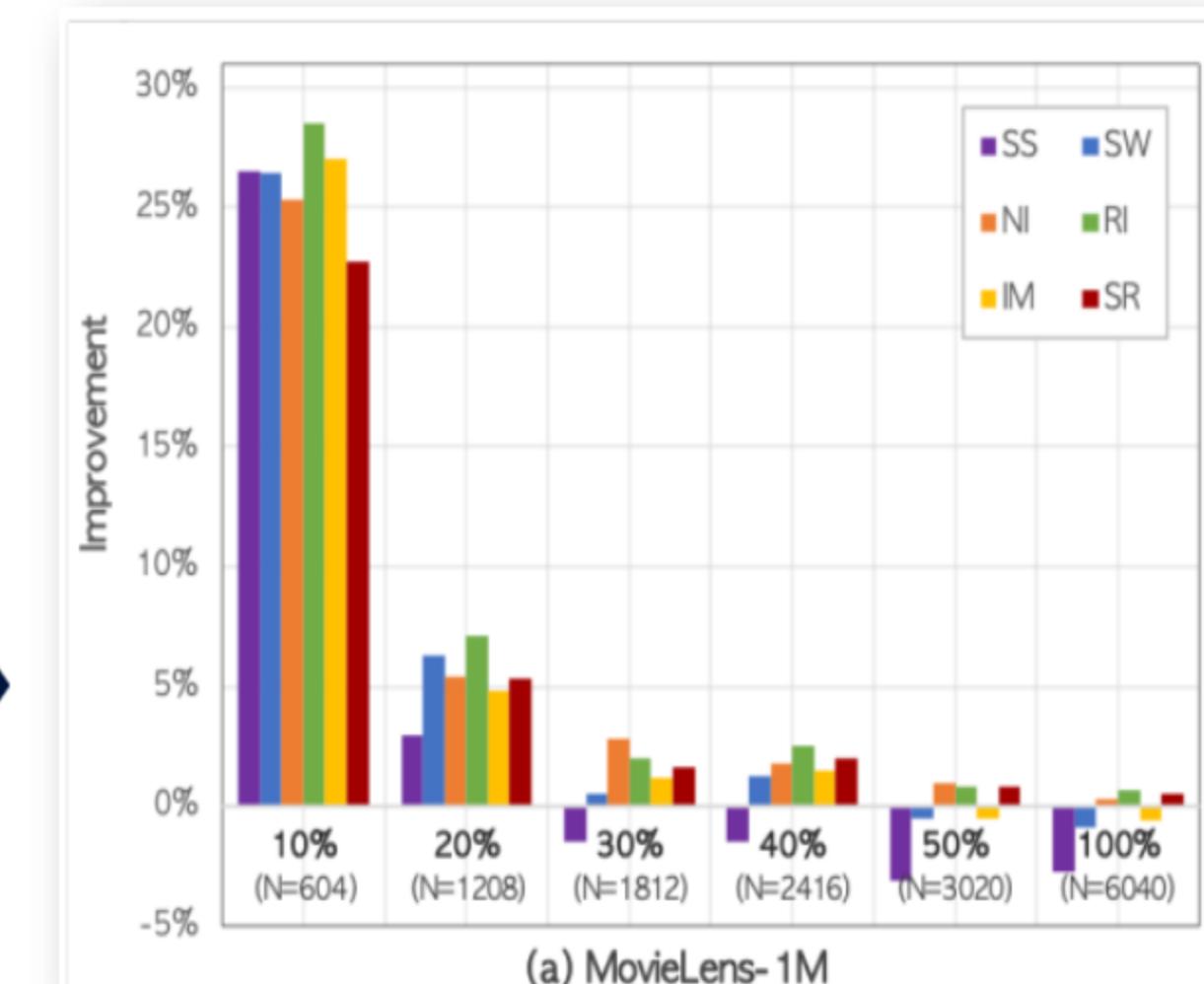
4. 실험 및
결론



고려해야 할 파라미터: N_aug

N_aug는 데이터 증강 규모를 의미하며
한 유저의 피드백 데이터를 몇 배수로
늘릴 것인지에 대한 파라미터이다

- 추천시스템과 자연어의 성질이 비슷하여
유사한 아이템으로 교체해도 문제가 없다고 판단
- 순차적 추천시스템 데이터 연구 결과에 의하면
데이터 수집이 제한된 상황에서
Data Augmentation을 하게 되면 성능 Good



01. Experiment

Phase 1 : Content based filtering

[DoBERT]

1. Document를 구성하는 토큰들에 대해 MLM 학습
2. 각 Word Embedding vector를 pooling -> Document Embedding
3. 채용공고가 속한 직무 분류를 예측하는 학습으로 Fine-Tuning

[Loss Function] : Cross Entropy Loss

*pretrain과 fine-tuning

$$L(\hat{y}, y) = - \sum_k^K y^{(k)} \log \hat{y}^{(k)}$$

[HyperParameter]

Layer_num: 12

Max_length: 256

*평균 시퀀스 길이 : 208

Embedding_dim: 768

Num_head: 12

*BERT의 preset을 따름

[Baseline] : Doc2Vec

[Metric] : Coverage, Hit Ratio@k

$$\text{Coverage}_{item} = \frac{n}{N} * 100$$

$$HR = \frac{U_{hit}}{U_{all}}$$

[Long tail theory]



인기 아이템만 추천한다면 성능은 당연히 올라갈 것이지만 추천의 Coverage는 적어질 것이다
아이템 속성 기반의 접근으로 Coverage를 증가시킬 수 있다

01. Experiment

Phase 2: Collaborative filtering

[Denoising AutoEncoder]

1. DoBERT로 학습된 아이템(문서) 임베딩을 통한 데이터 증강
2. Input에 대하여 Random하게 noise를 추가
3. 손상된 데이터를 Reconstruct하는 학습을 진행

[Loss Function] : Cross Entropy Loss

*pretrain과 fine-tuning

$$L(\hat{y}, y) = - \sum_k^K y^{(k)} \log \hat{y}^{(k)}$$

$$\arg \min_{W, W', V, b, b'} \frac{1}{U} \sum_{u=1}^U \mathbb{E}_{p(\tilde{y}_u | y_u)} [\ell(\tilde{y}_u, \hat{y}_u)] + \mathcal{R}(W, W', V, b, b')$$

[Metric] : Recall@k, MAP@k

$$Recall = \frac{TP}{TP + FN}$$

@k의 의미 : k개의 아이템을 추천했을 때

[Control Overfitting]

Data Augmentation은 기존 유저의 원본 데이터를 기준으로

여러 시퀀스를 만들기 때문에 특정 유저의 패턴을 외워버리는 과적합을 유발할 수 있다

때문에 Dropout과 Noise를 부여하는 하이퍼 파라미터의 적절한 처리가 필요하다

*기존 연구에서는 데이터 증강 시 dropout 비율을 50%로 증가시킴

[Baseline] : Vanilla Autoencoder, Variational Autoencoder

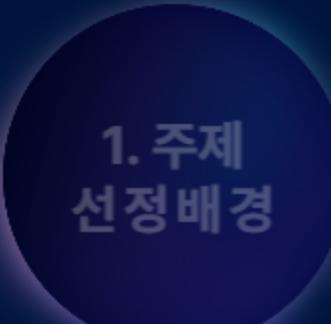
1. 주제
선정배경

2. 선행연구

3. Modeling

4. 실험 및
결론

02. 결론 및 한계



1. 주제
선정배경



2. 선행연구



3. Modeling



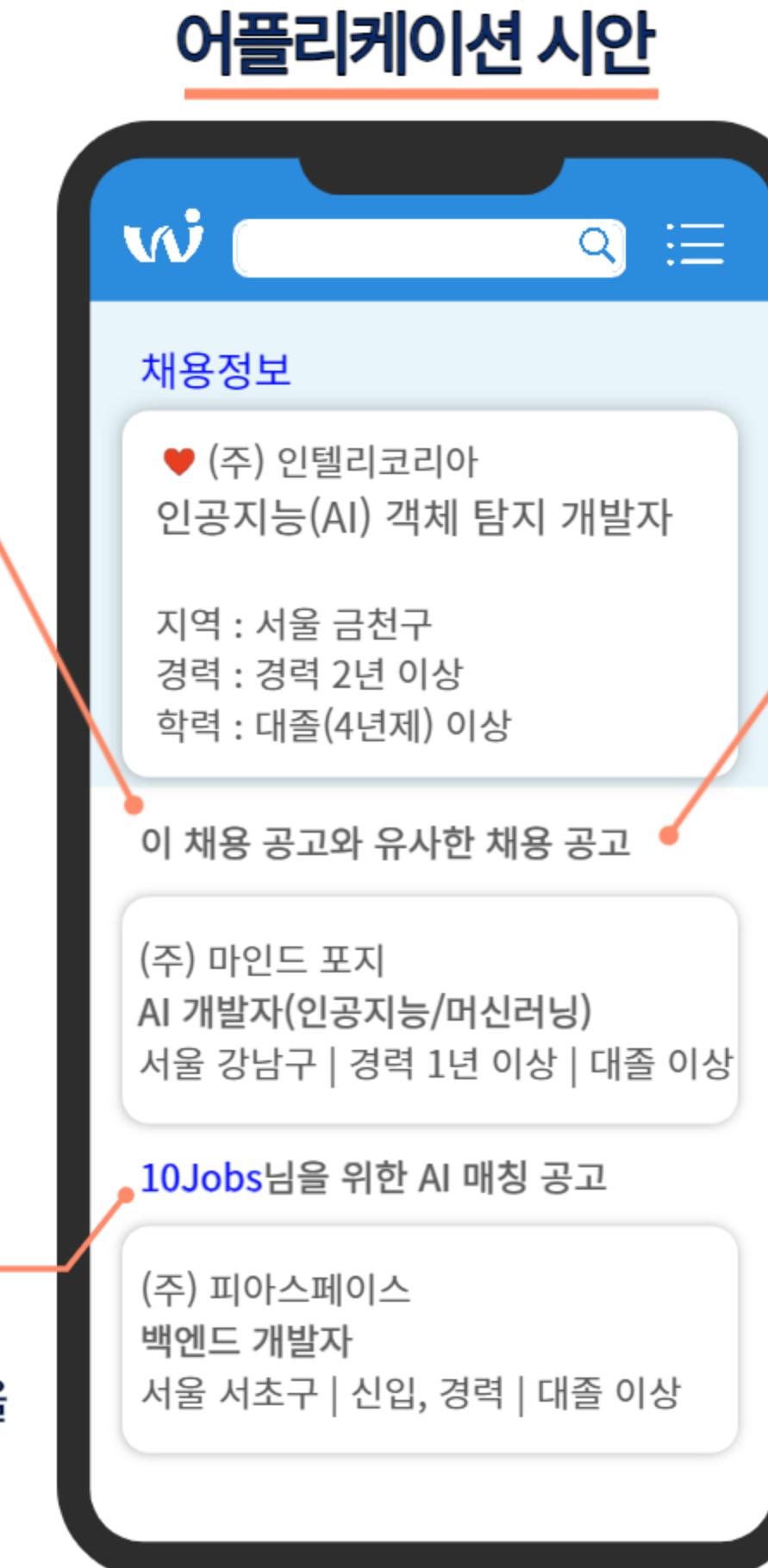
4. 실험 및
결론

Content Based Filtering **DoBERT**

Word2vec 수준의 임베딩이 아닌
고성능의 임베딩 기법을 사용함으로
유사한 아이템 추천에 좋은 퍼포먼스를 보임

Collaborative Filtering **Denoising AutoEncoder**

Input 데이터에 적절하게 노이즈를
추가함으로 증강된 데이터의 Overfitting을
적극적으로 제어 할 수 있음



Similar Document Augmentation

CBF의 임베딩 정보를 활용하여
유저 로그 데이터에 대해
효율적인 증강이 가능함

Mixed Hybridization

웹 공간에 두 모델의 추천을 동시 제공하며
모델이 가지는 치명적 약점을 보완할 수 있음
(CF-유저 데이터 부족, CBF-개인화 부족 등)

한계

1. 주제
선정배경

2. 선행연구

3. Modeling

4. 실험 및
결론

1. BERT의 사전학습 모델에 필요한 대량의 Corpus가 없음 [환경적인 요인]

-> 현재 컴퓨팅 환경에서는 불가능함

2. BERT 속도를 더 올릴 수 있었지만 그러지 못함 [속도 개선]

-> 학습속도가 느리다는 단점이 있었는데, 가속화 방법이 필요함

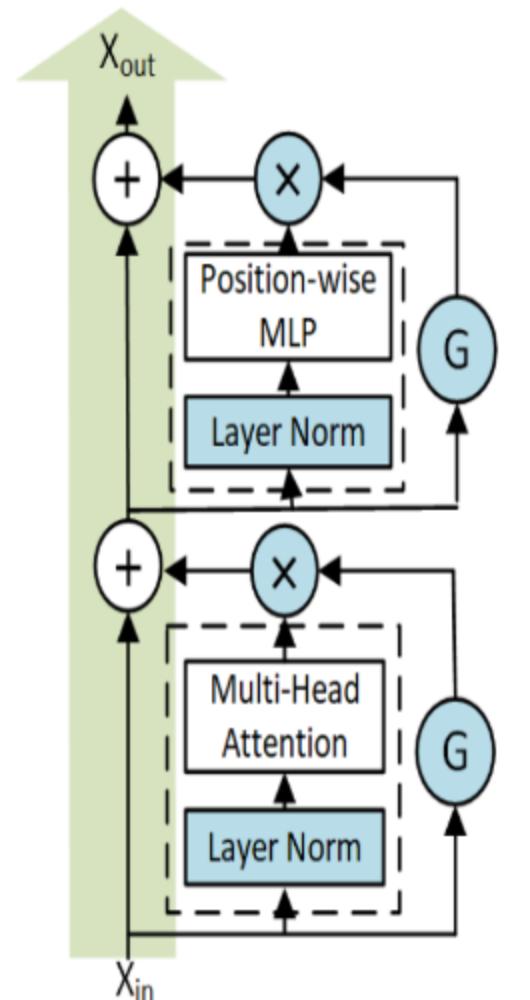
-> **Progressive Layer Dropping: Switchable Transformer**

Accelerating training of transformer-based language models with progressive layer dropping

3. Sparsity Data [제어 한계]

-> 7만개의 채용 공고는 Sparsity Data로 학습의 비효율성을 야기함

-> 추후 Negative Sampling, Feature Selection이 필요함



Reference

1. 주제
선정배경

2. 선행연구

3. Modeling

4. 실험 및
결론

- [1] Zhang, Minjia, and Yuxiong He. "Accelerating training of transformer-based language models with progressive layer dropping." *Advances in Neural Information Processing Systems* 33 (2020): 14011-14023.
- [2] Song, Joo-yeong, and Bongwon Suh. "Data Augmentation Strategies for Improving Sequential Recommender Systems." *arXiv preprint arXiv:2203.14037* (2022).
- [3] Wang, William Yang, and Dyi Yang. "That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using# petpeeve tweets." *Proceedings of the 2015 conference on empirical methods in natural language processing*. 2015.
- [4] Le, Quoc, and Tomas Mikolov. "Distributed representations of sentences and documents." *International conference on machine learning*. PMLR, 2014.
- [5] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).
- [6] Wu, Yao, et al. "Collaborative denoising auto-encoders for top-n recommender systems." *Proceedings of the ninth ACM international conference on web search and data mining*. 2016.
- [7] Sun, Chi, et al. "How to fine-tune bert for text classification?." *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20*. International Publishing, 2019.
- [8] Adhikari, Ashutosh, et al. "Docbert: Bert for document classification." *arXiv preprint arXiv:1904.08398* (2019).
- [9] Çano, Erion, and Maurizio Morisio. "Hybrid recommender systems: A systematic literature review." *Intelligent Data Analysis* 21.6 (2017): 1487-1524.