



LangCon 2023



특정 도메인에 맞는 언어모델은 어떻게 만들까?

BHSN 박장원

박장원

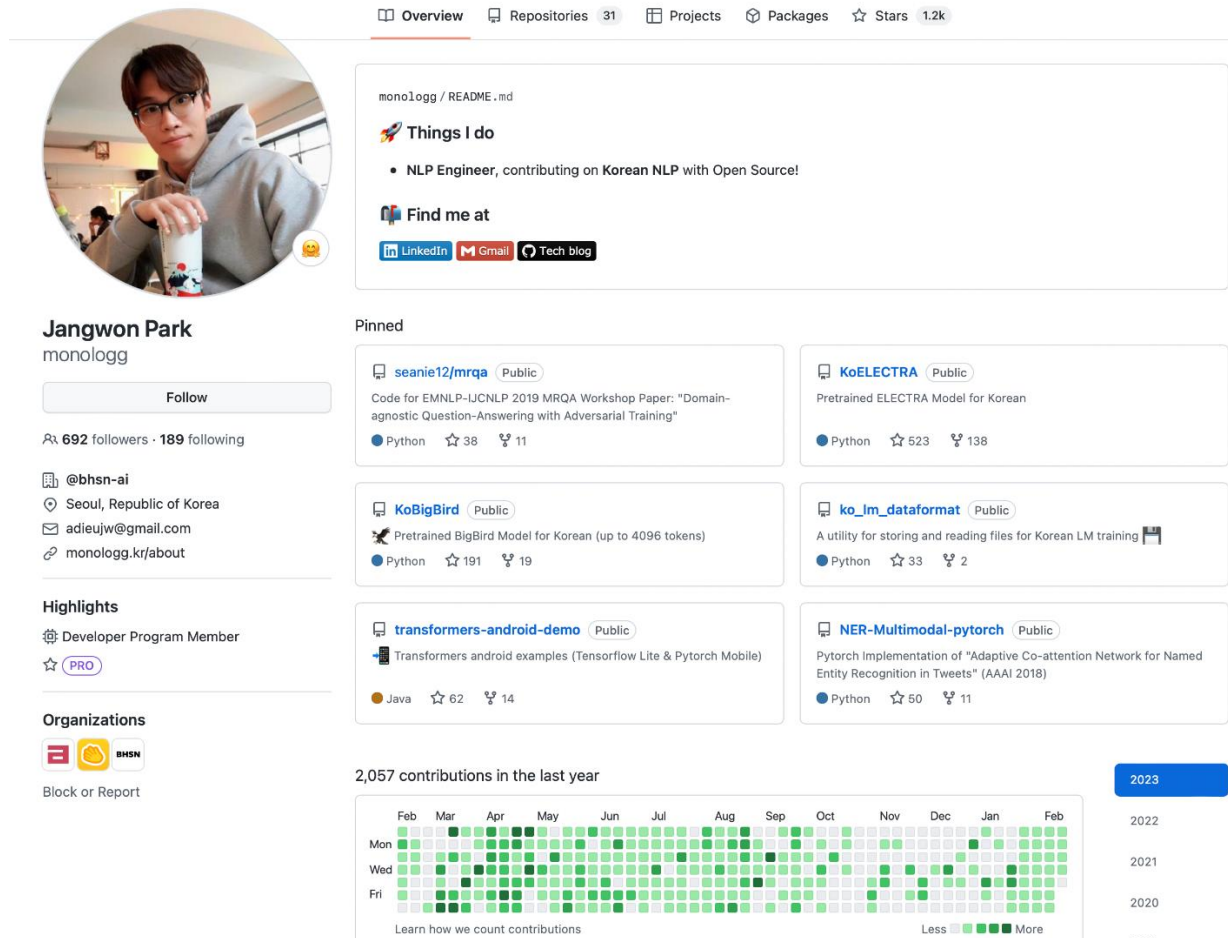
(monologg로 더 많이 아시는 거 같아요 😓)

- '한국어 언어모델'에 관심이 많습니다!

- KoELECTRA, KLUE RoBERTa 등

- **BHSN** 에서 Contract AI를 개발하고 있습니다.

- <https://www.bhsn.ai/>



GitHub profile of Jangwon Park (monologg). The profile includes a circular profile picture of a man with glasses and a grey hoodie. The header shows the name "Jangwon Park" and the username "monologg". Below the header is a "Follow" button. The profile statistics show 692 followers and 189 following. The bio lists the GitHub handle "@bhsn-ai", location "Seoul, Republic of Korea", email "adieujw@gmail.com", and website "monologg.kr/about". The "Highlights" section shows "Developer Program Member" and "PRO" status. The "Organizations" section shows the BHSN logo. The "Pinned" section lists four repositories: "seanie12/mrqa", "KoELECTRA", "KoBigBird", and "ko_lm_dataformat". The "transformers-android-demo" repository is also listed. The "Contributions" section shows a heatmap of contributions over time, with a total of 2,057 contributions in the last year. The heatmap shows contributions from February 2020 to February 2023, with a color scale from green (Less) to red (More).

Overview Repositories 31 Projects Packages Stars 1.2k

monologg / README.md

Things I do

- NLP Engineer, contributing on Korean NLP with Open Source!

Find me at

LinkedIn Gmail Tech blog

Jangwon Park
monologg

Follow

692 followers · 189 following

@bhsn-ai
Seoul, Republic of Korea
adieujw@gmail.com
monologg.kr/about

Highlights

Developer Program Member

PRO

Organizations

BHSN

Block or Report

Pinned

seanie12/mrqa Public

Code for EMNLP-IJCNLP 2019 MRQA Workshop Paper: "Domain-agnostic Question-Answering with Adversarial Training"

Python 38 11

KoELECTRA Public

Pretrained ELECTRA Model for Korean

Python 523 138

KoBigBird Public

Pretrained BigBird Model for Korean (up to 4096 tokens)

Python 191 19

ko_lm_dataformat Public

A utility for storing and reading files for Korean LM training

Python 33 2

transformers-android-demo Public

Transformers android examples (Tensorflow Lite & Pytorch Mobile)

Java 62 14

NER-Multimodal-pytorch Public

Pytorch Implementation of "Adaptive Co-attention Network for Named Entity Recognition in Tweets" (AAAI 2018)

Python 50 11

2,057 contributions in the last year

2023

2022

2021

2020

2019

Learn how we count contributions

Less More

1. Intro



**'거대 모델을 대상으로 한 자연어처리 동향 이해' 로
하러 해요!**

후후 접수완료~ 걱정마세요~~



3 HOURS

LATER...



Matthew Carrigan
@carrigmat

...

this is [@huggingface](#), we see you out there retweeting the latest state of the art miracle of modern technology and then going home and using bert-base-uncased for the fifth year in a row

트윗 번역하기

마음에 들어요 478회 리트윗 44회 인용 4회



Matthew Carrigan
@carrigmat

...

우리는 당신이 현대 기술의 최신 기적을 리트윗하고,
집으로 돌아가서 5년 연속으로
bert-base-uncased를 사용하는
것을 보고 있습니다

트윗 번역하기

마음에 들어요 478회 리트윗 44회 인용 4회

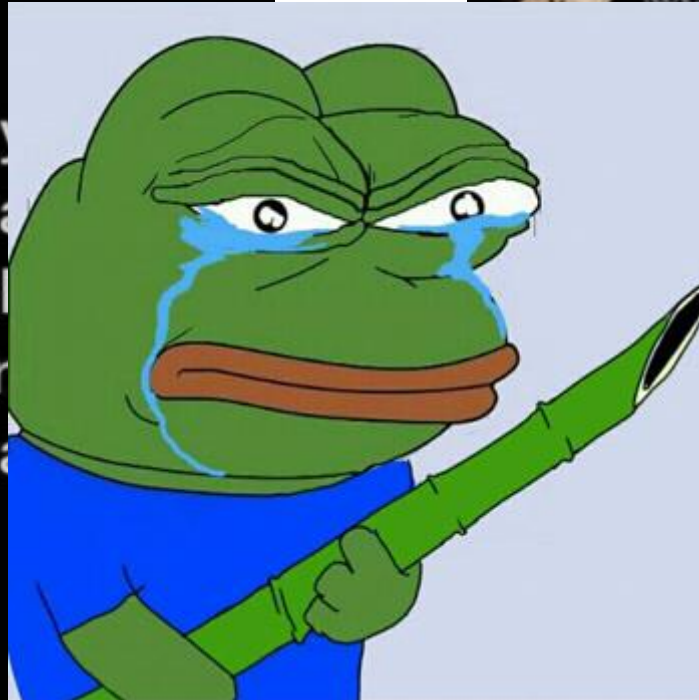


Matthew Carrigan
@carrigmat

this is [@huggingface](#), we see
there retweeting the latest sta
the art miracle of modern tec
and then going home and usin
base-uncased for the fifth ye
row

트윗 번역하기

마음에 들어요 478회 리트윗 44회 인용 4회



Matthew Carrigan
@carrigmat

당신이 현대 기술의 최신 기적
릿하고,
돌아가서 5년 연속으로
base-uncased를 사용하는
고 있습니다

마음에 들어요 478회 리트윗 44회 인용 4회

OpenAI



C.h.a.t.G.P.T.

Others



나한테는 BERT도 빅모델이야ㅠ

**99%의 현업자가
바로 적용해보고 공감할 수 있는 것을
이야기해볼까?**



**특정 도메인에서
적당한 사이즈의 모델로도
해결할 수 있는 것이 많다!**



2. 도메인 특화 언어모델

General (리뷰)

액션이랑 모든게 다 재미있는
멋난되는 영화

Positive

Domain (계약서)

제 6 조 (비밀 유지 기간)

“본 계약”은 체결일로부터 발효되며, 유효기간은 계약 체결 후 5년까지 “갑”과 “을” 쌍방은 비밀 정보 유지 의무를 부담한다. 다만 상호 서면 합의에 의하여 연장이 가능하며 계약이 중간에 종료된 경우에도 “본 계약”의 비밀 정보 유지 의무는 체결일로부터 5년간 유효하다.

제 7 조 (손해배상)

1) “갑”은 “을”의 서면 동의 없이 비밀 정보가 제3자에게 공개, 제공 또는 누설되거나 “목적사업” 외의 다른 목적이나 용도로 사용할 경우 “을”이 입은 손해를 배상해야 한다.

Q. 본 계약의 비밀유지기간은?



Biomedical

Legal

Finance

**BioBERT, SciBERT, LegalBERT,
FinBERT, BERTweet,
PubMedBERT...**

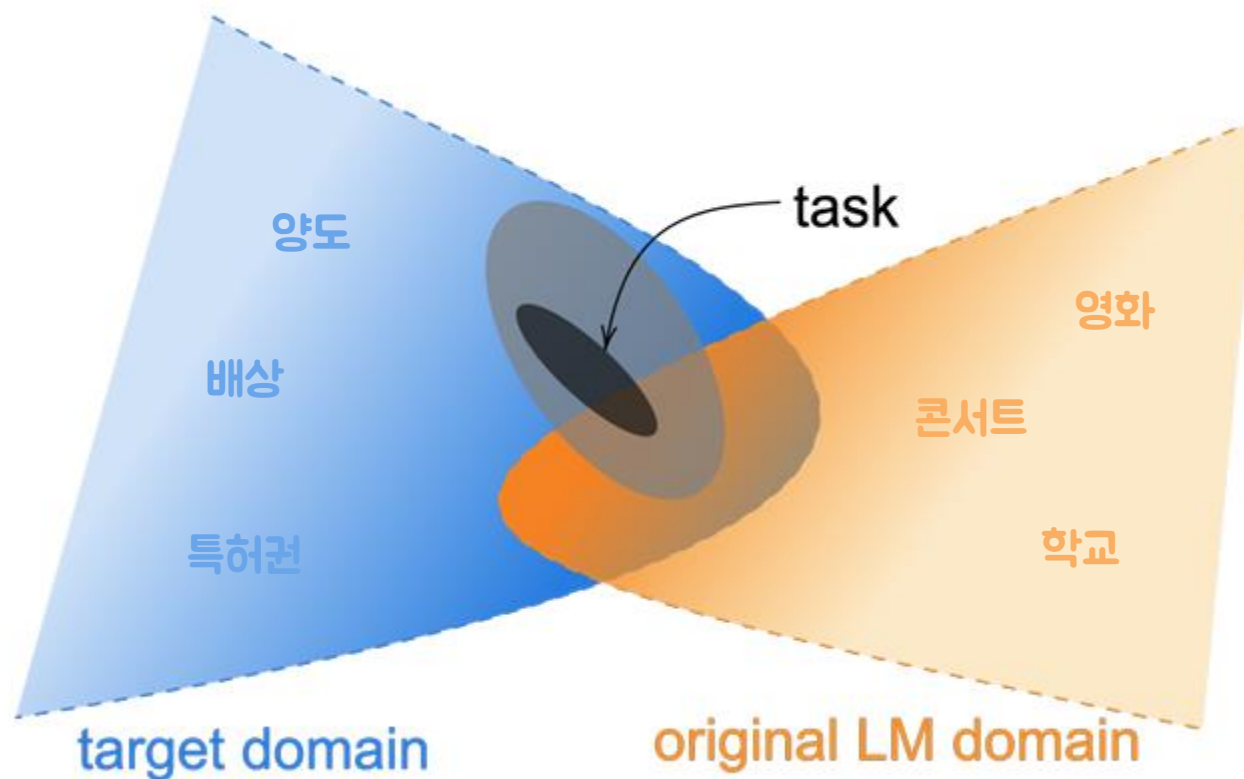
Table 28: Statistics of the pretraining corpus.

	MODU	CC-100-Kor	NAMUWIKI	NEWSCRAWL	PETITION	Total
# Sentences	167M	103M	14M	183M	5.2M	473M
# Words	1,892,814,395	1,593,887,022	265,203,602	2,716,968,038	50,631,183	6,519,504,240
size (GB)	18.27	15.46	2.52	25.87	0.53	62.65

Pretraining Corpora We gather the following five publicly available Korean corpora from diverse sources to cover a broad set of topics and many different styles. We combine these corpora to build the final pretraining corpus of size approximately 62GB. See Table 28 for overall statistics:

- **MODU** : *Modu*⁵¹ Corpus [98] is a collection of Korean corpora distributed by National Institute of Korean Languages.⁵² It includes both formal articles (news and books) and colloquial text (dialogues).
- **CC-100-Kor** : CC-100⁵³ is the large-scale multilingual web crawled corpora by using CC-Net [136]. This is used for training XLM-R [26]. We use the Korean portion from this corpora.
- **NAMUWIKI** : NAMUWIKI is a Korean web-based encyclopedia, similar to Wikipedia, but known to be less formal. Specifically, we download the dump created on March 2nd, 2020.⁵⁴
- **NEWSCRAWL** : NEWSCRAWL consists of 12,800,000 news articles published from 2011 to 2020, collected from a news aggregation platform.
- **PETITION** : Petition is a collection of public petitions posted to the Blue House asking for administrative actions on social issues. We use the articles in the Blue House National Petition⁵⁵ published from August 2017 to March 2019.⁵⁶

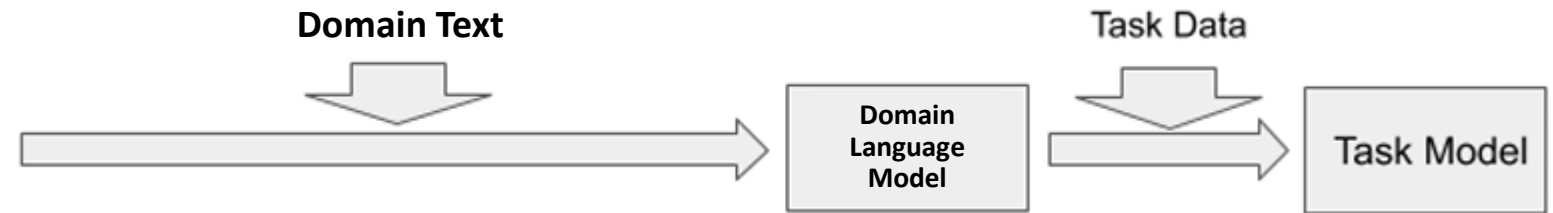
- News
- Wiki
- Book
- Web Crawl



In-Domain

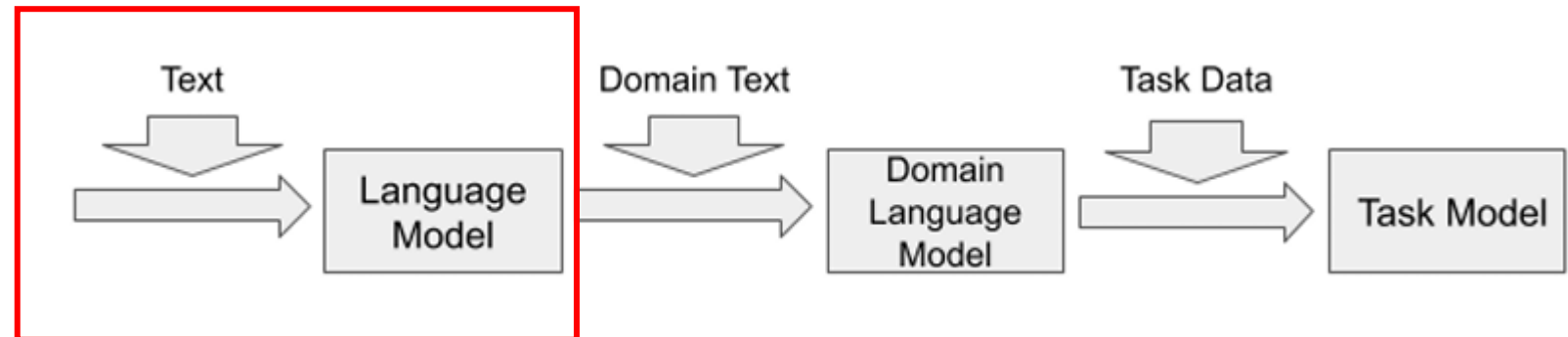
Out-Domain

Pretraining From Scratch



Continual Pretraining

<General LM (e.g. KoELECTRA)>



도메인 특화 언어모델을 만들 때
무엇을 고려해야할까?



3. Need to Consider

3.1. Data

1. 그거 진짜 Domain-Specific Corpus는 맞아?



금융 특화 언어모델을 만들고 싶어요!

1. 그거 진짜 Domain-Specific Corpus는 맞아?



금융 특화 언어모델을 만들고 싶어요!

허허 그렇군요. 말뭉치는 준비하셨나요?



1. 그거 진짜 Domain-Specific Corpus는 맞아?



금융 특화 언어모델을 만들고 싶어요!

허허 그렇군요. 말뭉치는 준비하셨나요?



올! 경제 분야 뉴스를 가져왔어요!

1. 그거 진짜 Domain-Specific Corpus는 맞아?



금융 특화 언어모델을 만들고 싶어요!

허허 그렇군요. 말뭉치는 준비하셨나요?



옹! 경제 분야 뉴스를 가져왔어요!

(띠용?!)



1. 그거 진짜 Domain-Specific Corpus는 맞아?

Table 28: Statistics of the pretraining corpus.

	MODU	CC-100-Kor	NAMUWIKI	NEWSCRAWL	PETITION	Total
# Sentences	167M	103M	14M	183M	5.2M	473M
# Words	1,892,814,395	1,593,887,022	265,203,602	2,716,968,038	50,631,183	6,519,504,240
size (GB)	18.27	15.46	2.52	25.87	0.53	62.65

Pretraining Corpora We gather the following five publicly available Korean corpora from diverse sources to cover a broad set of topics and many different styles. We combine these corpora to build the final pretraining corpus of size approximately 62GB. See Table 28 for overall statistics:

- **MODU** : *Modu*⁵¹ Corpus [98] is a collection of Korean corpora distributed by National Institute of Korean Languages.⁵² It includes both formal articles (news and books) and colloquial text (dialogues).
- **CC-100-Kor** : CC-100⁵³ is the large-scale multilingual web crawled corpora by using CC-Net [136]. This is used for training XLM-R [26]. We use the Korean portion from this corpora.
- **NAMUWIKI** : NAMUWIKI is a Korean web-based encyclopedia, similar to Wikipedia, but known to be less formal. Specifically, we download the dump created on March 2nd, 2020.⁵⁴
- **NEWSCRAWL** : NEWSCRAWL consists of 12,800,000 news articles published from 2011 to 2020, collected from a news aggregation platform.
- **PETITION** : Petition is a collection of public petitions posted to the Blue House asking for administrative actions on social issues. We use the articles in the Blue House National Petition⁵⁵ published from August 2017 to March 2019.⁵⁶

• News

• Wiki

• Book

• Web Crawl



1. 그거 진짜 Domain-Specific Corpus는 맞아?

N 뉴스 | 연예 | 스포츠 | 날씨 | 프리미엄

언론사별 | 정치 | **경제** | 사회 | 생활/문화 | IT/과학 | 세계 | 랭킹 | 신문보기 | 오피니언 | TV | 팩트체크

02.16(목)

경제
금융
증권
산업/재계
중기/벤처
부동산
글로벌 경제
생활경제

① 헤드라인 뉴스

13 "폭탄 고지서" 공공요금 속도조절 · 전기·가스로 추가인상 늦춘다지만 >

 가스·전기료 인상 방침 고수한 정부...중산층 지원 없지만 尹 "속도조...
정부가 서민부담 최소화를 목표로 폭과 속도를 조절하되 가스·전기요금 정상화는 변
함 없이 추진하겠다는 의지를 재차 분명히 했다. 취약계층에 대한 지원을 보 ...
뉴스1

'폭탄 고지서'에 공공요금 속도조절...공기업 손실 더 커진다 중앙일보 | 💬 10+

'폭탄 고지서'에 속도조절 "촉선 후 더 큰 폭탄" 우려도 JTBC | 💬 30+

정치 + **경제** + 사회 + 생활/문화 + IT/과학 + 세계

1. 그거 진짜 Domain-Specific Corpus는 맞아?

(PT = RoBERTa)

	PT	News	Reviews	BioMed	CS
PT	100.0	54.1	34.5	27.3	19.2
News	54.1	100.0	40.0	24.9	17.3
Reviews	34.5	40.0	100.0	18.3	12.7
BioMed	27.3	24.9	18.3	100.0	21.4
CS	19.2	17.3	12.7	21.4	100.0

Figure 2: Vocabulary overlap (%) between domains. PT denotes a sample from sources similar to RoBERTa's pretraining corpus. Vocabularies for each domain are created by considering the top 10K most frequent words (excluding stopwords) in documents sampled from each domain.

Vocabulary Overlap

- 총 4개의 도메인
 - News, Reviews, BioMed, Computer Science
- News, Reviews -> 유사

1. 그거 진짜 Domain-Specific Corpus는 맞아?

Dom.	Task	RoBA.	DAPT	\neg DAPT
BM	CHEMPROT	81.9 _{1.0}	84.2 _{0.2}	79.4 _{1.3}
	†RCT	87.2 _{0.1}	87.6 _{0.1}	86.9 _{0.1}
CS	ACL-ARC	63.0 _{5.8}	75.4 _{2.5}	66.4 _{4.1}
	SCIERC	77.3 _{1.9}	80.8 _{1.5}	79.2 _{0.9}
NEWS	HYP.	86.6 _{0.9}	88.2 _{5.9}	76.4 _{4.9}
	†AGNEWS	93.9 _{0.2}	93.9 _{0.2}	93.5 _{0.2}
REV.	†HELPFUL.	65.1 _{3.4}	66.5 _{1.4}	65.1 _{2.8}
	†IMDB	95.0 _{0.2}	95.4 _{0.2}	94.1 _{0.4}

점수차가 크지 않음! 🙄🙄

Table 3: Comparison of RoBERTa (RoBA.) and DAPT to adaptation to an *irrelevant* domain (\neg DAPT). Reported results are test macro- F_1 , except for CHEMPROT and RCT, for which we report micro- F_1 , following [Beltagy et al. \(2019\)](#). We report averages across five random seeds, with standard deviations as subscripts. † indicates high-resource settings. Best task performance is boldfaced. See §3.3 for our choice of irrelevant domains.

2. Corpus는 어느 정도의 양이 적당한가?

일단 많으면 많을수록
무조건 좋습니다!



2. Corpus는 어느 정도의 양이 적당한가?

Corpus의 최소 요구량은 정답이 없습니다.
다만 최근(2022.12)에 이런 논문은 있었어요.



2. Corpus는 어느 정도의 양이 적당한가?



Dataset	BERT	PubMedBERT	4GB	8GB	12GB
NCBI-disease	84.3	87.8	87.7	87.9	88.0
HoC	79.0	82.3	81.1	82.5	81.4
PubMedQA	54.4	55.8	54.9	53.4	55.2

Table 2: Performance comparison of pre-trained language models. The models are evaluated on the tasks using the same fine-tuning process. All of our experimental models are pre-trained for 67K steps.

3. 데이터의 총 사이즈를 직접 계산해보세요

농민신문

‘임차 농지’ 공익직불금 사각지대 여전

입력 2023.02.15 오전 5:02 | 기사원문

하지혜 기자

2 4

PICK 1

계약서 없는 농가 상당수...농식품부 예외 사항 제시
농업경영체 등록정보엔 등재 못해 사실상 ‘신청 불가’



이미지투데이

“밭린 땅에서 농사짓고 세금까지 낸 세월이 열만대, 임대차계약서 없이는 직불금을 못 준다니 포기했죠, 뭐.”

전북의 시골엔 농가 김모씨는 올해도 1322㎡(400평) 규모의 임차 농지에 대해 기본형 공익직불금을 신청하지 못했다. 직불금 신청에 필요한 임대차계약서가 없어서다. 그는 오래전 농지 소유주가 갑자기 사망한 후 땅을 물려받은 자녀들과 연락이 닿지 않아 임대차계약서를 쓰지 못했다. 그러다 2021년부터 임대차계약서를 제출해야 기본형 공익직불금을 신청할 수 있게 되면서 임차 농지에 대한 직불금을 아예 포기했다.

김씨는 “주위에 중증이나 소유주가 여러명인 땅은 임대차계약서를 쓰지 못한 농가가 적지 않다”며 “예전엔 임차 농지의 경작사실확인서를 신청 서류로 내면 직불금을 받을 수 있었는데, 이젠 똑같이 농사를 지어도 직불금을 받지 못하니 씁쓸하다”고 토로했다.

2023년 기본형 공익직불금 신청이 시작된 가운데 올해부터는 공익직불제의 사각지대가 해소될 것이란 기대감이 높다. 지난해 법 개정으로 올해부터 ‘2017~2019년 중 직불금을 받은 이력이 없는 농지’도 기본형 공익직불금을 신청할 수 있게 됐기 때문이다. 그러나 전체 농가의 절반에 달하는 임차농가들 사이에선 아직도 제도에 구멍이 있다는 의견이 제기된다.

문제는 임대차계약서다. 농지 소유자의 사망이나 행방불명, 해외 거주 등으로 소유권자가 불확실한 땅은 현실적으로 임대차계약을 맺기 어렵다. 소유관계가 복잡한 중증 땅이나 공동 소유 농지도 마찬가지다.

이런 문제가 제기되면서 농림축산식품부는 2021년부터 기본형 공익직불사업 시행지침에 나뉘는 해결책을 제시해왔다. 소유권자가 불확실한 농지는 경작자의 재산세납부서나 재산세납부자와의 계약서 등을 제출하면 임대차계약서를 갈음할 수 있게 해준 것이다. 중증 소유 농지도 실경 작자에게 땅을 임대한다는 내용이 담긴 중증 회의록을 임대차계약서 대신 제출할 수 있도록 했다.

이 지침대로라면 김씨 같은 임차농들도 직불금을 신청할 수 있지만 실제로 그렇지 않다. 사실상 임대차계약서가 없는 임차 농지는 농업경영체 등록정보에 등재할 수 없기 때문이다. 기본형 공익직불금 지급 대상은 ‘농업경영체 육성 및 지원에 관한 법률’에 따라 농업경영체 등록정보에 등록된 농지로 정해져 있다. 농업경영체 등록정보에 임차 농지를 등재하려면 임대차 현황 내용이 담긴 농지대장을 제출해야 한다. 그런데 이 농지대장에 임차 농지를 신고하려면 결국 임대차 계약서가 필요하다.

이에 대해 농업경영체 등록을 담당하는 국립농산물품질관리원은 임대차계약서를 통해 ‘농지법’상 적법한 임차 농지만 농업경영체 등록정보에 등재하는 건 당연하다는 입장이다.

현행 ‘농지법’은 징집·질병 등 불가피한 상황을 제외한 사적인 임대차를 원칙적으로 금지한다. 1996년 1월1일 이후 취득한 농지는 한국농어촌공사를 통해 임대차계약을 맺어야 한다.

농관원 관계자는 “지난해 농식품부 국정감사에서조차 불법 임차 농지의 직불금 수급문제가 도마 위에 올랐다”면서 “불법 임대차와 직불금 부정수급을 막기 위해선 임대차계약서를 바탕으로 적법성을 따져야 한다”고 말했다.

현장에선 합법적인 임대차계약을 맺고 싶어도 현실적으로 힘들다는 목소리가 나온다.

한 면사무소의 직불금 담당자는 “농지은행을 통해 임대차계약을 맺는 데 번거로움을 느끼거나 8년 이상 자경 때 양도소득세를 감면받는 조항 때문에 임대차계약서를 써주지 않는 농지 소유자들이 여전히 있다”며 “불가피하게 불법 임차농으로 몰린 농민들은 억울할 수밖에 없다”고 지적했다.

• UTF-8 기준

• 한글, 한자: 3 Byte

• 영어, 숫자, 공백: 1 Byte

• 좌측의 뉴스는

4293Byte = 0.004 MB

• 1GB를 모으려면 256000개의 뉴스가 필요함

• 그러나 모든 뉴스가 예시만큼 길지 않습니다ㅠ

2. Corpus는 어느 정도의 양이 적당한가?

**Domain-Specific Corpus를 모을 것이라면
꼭 미리 계산하고 **각을 재보세요!****

(생각보다 많이들 안 해보십니다ㅠ)



3.2. Tokenizer



데이터를 도저히 많이 못 모으겠네요 ㅠ.ㅠ

그러면 Tokenizer라도 고쳐볼까요?



1. 일단 UNK이 발생하는 것 먼저 보자!

- OOV (Out Of Vocabulary), UNK (Unknown Token)
- e.g. 나는 오늘 **똥교**에 갔다
-> ['나', '##는', '오늘', '**[UNK]**', '갔', '##다']

1. 일단 UNK이 발생하는 것 먼저 보자!

```
class WordpieceTokenizer(object):  
    """Runs WordPiece tokenization."""
```

```
def __init__(self, vocab, unk_token, max_input_chars_per_word=100):  
    self.vocab = vocab  
    self.unk_token = unk_token  
    self.max_input_chars_per_word = max_input_chars_per_word
```

```
def tokenize(self, text):  
    """
```

Tokenizes a piece of text into its word pieces. This uses a greedy longest-match-first algorithm to perform tokenization using the given vocabulary.

For example, `input = "unaffable"` will return as output `["un", "##aff", "##able"]`.

Args:

text: A single token or whitespace separated tokens. This should have already been passed through `*BasicTokenizer*`.

Returns:

A list of wordpiece tokens.

```
    """
```

1. 일단 UNK이 발생하는 것 먼저 보자!

```
output_tokens = []
for token in whitespace_tokenize(text):
    chars = list(token)
    if len(chars) > self.max_input_chars_per_word:
        output_tokens.append(self.unk_token)
        continue

    is_bad = False
    start = 0
    sub_tokens = []
    while start < len(chars):
        end = len(chars)
        cur_substr = None
        while start < end:
            substr = "".join(chars[start:end])
            if start > 0:
                substr = "##" + substr
            if substr in self.vocab:
                cur_substr = substr
                break
            end -= 1
        if cur_substr is None:
            is_bad = True
            break
        sub_tokens.append(cur_substr)
        start = end

    if is_bad:
        output_tokens.append(self.unk_token)
    else:
        output_tokens.extend(sub_tokens)
return output_tokens
```

최소 단위인 Character에서
일치하는 게 없으면
전부 [UNK] 처리가 됨ㄷㄷ



2. Tokenizer를 만드는 3가지 방법

1. KoELECTRA Vocab을 그대로 쓴다
2. Vocab를 아예 새로 만든다 (새로운 도메인의 Corpus를 이용하여)
3. 기존 KoELECTRA Vocab에 새로운 단어를 추가

2. Tokenizer를 만드는 3가지 방법

Continual Pretraining이 가능!

1. KoELECTRA Vocab을 그대로 쓴다
2. Vocab를 아예 새로 만든다 (새로운 도메인의 Corpus를 이용하여)
3. 기존 KoELECTRA Vocab에 새로운 단어를 추가

2. Tokenizer를 만드는 3가지 방법

1. KoELECTRA Vocab을 그대로 쓴다

Pretraining From Scratch가 바람직함

2. Vocab를 아예 새로 만든다 (새로운 도메인의 Corpus를 이용하여)

3. 기존 KoELECTRA Vocab에 새로운 단어를 추가

2. Tokenizer를 만드는 3가지 방법

1. KoELECTRA Vocab을 그대로 쓴다
2. Vocab를 아예 새로 만든다 (새로운 도메인의 Corpus를 이용하여)
3. 기존 KoELECTRA Vocab에 새로운 단어를 추가

Vocabulary Expansion

-> 기존 35,000개인 KoELECTRA Vocab에 단어를 더 추가

그러면 Vocab Size는
어디까지 키우는 게 제일 좋을까요?



3. Vocabulary Expansion

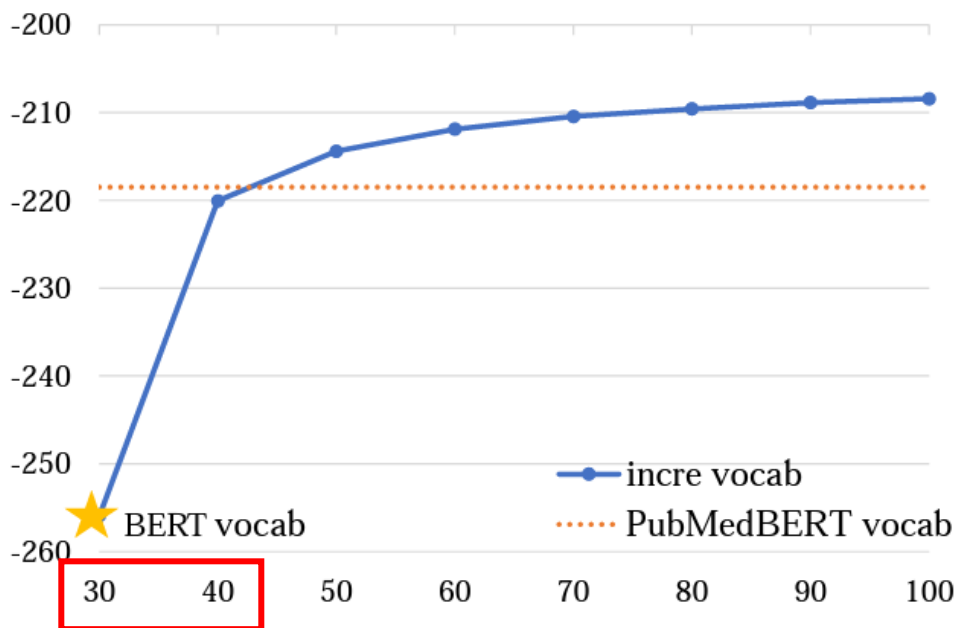


Figure 4: The $P(D)$ of different vocab sizes under biomedical domain. We use the BERT's vocabulary as the 30k vocabulary without vocabulary expanding. The PubMedBERT vocabulary is also 30k.

Given a domain-specific corpus D , the occurrence probability of corpus D is formulated as:

$$P(D) = \sum_{\mathbf{x}}^{|D|} \log(P(\mathbf{x})), \quad (3)$$

where \mathbf{x} represents tokenized sentence in corpus D .

- Occurrence Probability라는 심플한 방법을 사용
 - 각 token의 출현빈도의 곱 -> 로그 합
- 30K -> 40K 로 갈 때 가장 효과적이고, 그 이후부터는 점차 줄어듦
 - 최적의 사이즈는 본인의 기준에 맞게 판단하면 됨!

Vocab Expansion만 하고
Continual Pretraining을 하지 않아도
성능이 오른 연구 결과가 있다.



3. Vocabulary Expansion

BioMed	CS	News	Reviews
[inc, ub, ated] → incubated	[The, orem] → Theorem	[t, uesday] → tuesday	[it, 's] → it's
[trans, fect] → transfect	[L, em, ma] → Lemma	[ob, ama] → obama	[that, 's] → that's
[ph, osph, ory] → phosphory	[vert, ices] → vertices	[re, uters] → reuters	[sh, oes] → shoes
[mi, R] → miR	[E, q] → Eq	[iph, one] → iphone	[doesn, 't] → doesn't
[st, aining] → staining	[cl, ust, ering] → clustering	[ny, se] → nyse	[didn, 't] → didn't
[ap, opt, osis] → apoptosis	[H, ence] → Hence	[get, ty] → getty	[can, 't] → can't
[G, FP] → GFP	[Seg, mentation] → Segmentation	[inst, agram] → instagram	[I, 've] → I've
[pl, asm] → plasm	[class, ifier] → classifier	[bre, xit] → brexit	[b, ought] → bought
[ass, ays] → assays	[Ga, ussian] → Gaussian	[nas, daq] → nasdaq	[you, 'll] → you'll
[ph, osph, ory, lation] → phosphorylation	[p, olyn] → polyn	[ce, o] → ceo	[kind, le] → kindle

Table 4: Samples of token sequences with large JSD between base and domain corpora sequence distributions; all of these sequences were added during AT to the Roberta-Base tokenizer.

‘incubated’ 란 단어가 추가되면

-> ‘inc’, ‘ub’, ‘ated’ 의 embedding의 평균으로 초기화!

3. Vocabulary Expansion

AT = Adaptive Tokenization

Domain	Task	RoBERTa	DAPT	TAPT	DAPT + TAPT	AT (Mean)	AT (Proj)		State-of-the-art (in 2020)
BioMed*	ChemProt	81.9 _{1.0}	<u>84.2_{0.2}</u>	82.6 _{0.4}	84.4_{0.4}	83.6 _{0.4}	83.1 _{0.3}		84.6
	RCT	87.2 _{0.1}	<u>87.6_{0.1}</u>	87.7 _{0.1}	87.8_{0.1}	87.5 _{0.4}	<u>87.6_{0.3}</u>		92.9
CS*	ACL-ARC	63.0 _{5.8}	<u>75.4_{2.5}</u>	67.4 _{1.8}	75.6_{3.8}	70.1 _{2.0}	68.9 _{1.6}		71.0
	SciERC	77.3 _{1.9}	<u>80.8_{1.5}</u>	79.3 _{1.5}	81.3 _{1.8}	<u>81.4_{0.4}</u>	81.2 _{1.2}		81.8
News	HyperPartisan	86.6 _{0.9}	88.2 _{5.9}	90.4 _{5.2}	90.0 _{6.6}	<u>93.1_{4.2}</u>	91.6 _{5.5}		94.8
Reviews	IMDB	95.0 _{0.2}	95.4 _{0.1}	95.5 _{0.1}	95.6_{0.1}	95.4 _{0.1}	<u>95.5_{0.1}</u>		96.2

Table 2: Results of different adaptive pretraining methods compared to the baseline RoBERTa. AT with mean subword and projective initializations are denoted as AT (Mean) and AT (Proj) respectively. Stddevs are from 5 seeds. Results for DAPT, TAPT, DAPT+TAPT, and state-of-the-arts are quoted from [Gururangan et al. \(2020\)](#). The highest non-state-of-the-art result is bolded, since the state-of-the-art functions as a performance ceiling, leveraging both domain-specific pretraining and an adapted tokenizer. The best of the three approaches which utilize only source and domain domain data before fine-tuning (i.e., DAPT and AT) is underlined. *Due to restrictions on accessible papers in S2ORC, The BioMed and CS pretraining corpora used were respectively 33% and 74% smaller than the versions in [Gururangan et al. \(2020\)](#). Note that state-of-the-art numbers are current at the time of [Gururangan et al. \(2020\)](#), and are from the following works: ChemProt: S2ORC-BERT [Lo et al. \(2020\)](#), RCT: Sequential Sentence Classification [Cohan et al. \(2019\)](#), ACL-ARC: SciBert [Beltagy et al. \(2019\)](#), SciERC: S2ORC-BERT [Lo et al. \(2020\)](#), HyperPartisan: Longformer [Beltagy et al. \(2020\)](#), IMDB: XLNet Large [Yang et al. \(2019\)](#).



<https://bhsn.ai>



감사합니다

LangCon 2023