한국어 BERT 모델 소개

순서

1. 공개된 한국어 BERT
2. 한국어 BERT 모델 구축 시 고려해야 하는 점
3. BERT를 넣고 돌릴 수 있는 한국어 downstream task

---

1. 공개된 한국어 BERT

- BERT를 개인이 직접 pre-training하기엔 너무 오래 걸리고 비싸다.
- 그러므로 이미 공개된 모델을 다운 받아서 우리가 풀고자 하는 task에 fine tuning하면 좋다.
- 우리가 사용할 수 있는 모델 위주로 정리해 봄. ~~(아름 헷갈림 주의)~~

BERT-base multilingual cased

(https://github.com/google-research/bert) (google)

- 104 개의 언어 처리를 위한 모델
- spec
  - Hidden size = 768, Layers = 12, Heads = 12, vocab_size = 119547
- Wordpiece tokenizer
- 이후 나온 한국어 특화 모델들에 비해 한국어 처리에서는 성능이 떨어짐.

KLUE BERT base

(https://huggingface.co/klue/bert-base)

(https://klue-benchmark.com/)

- KLUE 논문을 쓴 팀의 github에 공개된 모델
  - KLUE: 영어의 GLUE 처럼 한국어 자연어처리 모델 간 성능 비교를 위해 구축된 평가 셋
- 모두의말뭉치, CC-100, 나무위키, 뉴스 크롤링, 청와대 국민청원 크롤링 데이터로 pre-train
- morpheme-based subword tokenization
  - 형태소 단위로 초벌 토크나이징 후 BPE 알고리즘 적용
- spec
  - Hidden size = 768, Layers = 12, Heads = 12, vocab_size = 32000

Kortok

(https://github.com/kakaobrain/kortok)

(https://arxiv.org/abs/2010.02534)

- kakaobrain
- tokenization 방식에 따른 다양한 BERT 모델을 공개함.
- 한국어 위키 + 나무위키로 pre-train
- 다양한 토크나이징 방식 중에서 mecab + sentencepiece tokenization 방식의 모델이 성능이 가장 뛰어나다고 보고함.

- spec
  - Hidden size = 768, Layers = 12, Heads = 12, vocab_size = 32000, 64000

KoBERT

(https://github.com/SKTBrain/KoBERT)

- SKT
- 한국어 위키로 pre-train
- spec
  - Hidden size = 3072, Layers = 12, Heads = 12, vocab_size = 8002

KR-BERT

(https://github.com/snunlp/KR-BERT)

(https://arxiv.org/pdf/2008.03979.pdf)

- BERT 구축 시 시간, 비용 너무 많이 드니까 비교적 성능을 잘 내면서 더 작은 모델이 필요하다.
- 한국어 위키, 뉴스 기사로 pre-train
- 자체 제작 Bidirectional WordPiece tokenizaer 사용 (다른 BERT 모델들은 대부분 wordpiece 또는 sentencepiece 사용)
- 한글을 음절 단위, 자모 단위로 분해하는 tokenizing 방식 채택 → 성능 향상
- spec
  - vocab size = 16424(character), 12367(sub-character)

KcBERT

()

- 구어체에 강인하도록 네이버 뉴스 댓글 데이터로 pre-train
- spec
  - Hidden size = 768, Layers = 12, Heads = 12, vocab_size = 30000

---

2. 한국어 BERT 모델 구축 시 고려해야 하는 점

1. **영어보다 한정된 데이터의 양**
   - pre-training 시 대부분 한국어 위키, 나무위키 사용
   - 최근에는 모두의 말뭉치(https://corpus.korean.go.kr/), AIHUB를 활용할 수 있음.
2. **토크나이징**
   - 영어처럼 단순히 공백을 기준으로 단어의 경계를 나누면 성능이 별로 좋지 않다. 왜냐하면 **한국어 표기법에서는 띄어쓰기 단위가 단어의 경계가 반드시 일치하지는 않기 때문이다.** 게다가 유저들이 띄어쓰기를 잘 지키지도 않는다는 문제도 있다.
   - 한국어 같은 이단아들을 처리하기 위해 고빈도의 subword pair에서 각 subword를 하나의 토큰으로 처리하는 wordpiece, sentencepiece가 제안되었다. 그러나 이들도 간혹 조사를 잘 분리해내지 못한다는 한계가 있다.
     - 예: `철수가` 라는 조합이 코퍼스에 많고, `철수` 가 `가`와 독립적으로 출현하지 않는다면, wordpiece, sentencepiece는 `철수가`를 한 단어로 취급함.
       그러나 `철수` 는 명사, `가` 는 조사로, 한 단어로 취급해서는 안 됨.
   - 그래서 최근에는 **형태소 분석기(mecab)로 초벌 tokenizing을 한 뒤, wordpiece나 sentencepiece로 2차 tokenizing**을 하는 더 섬세한 tokenizing 방식이 제안됨. (https://arxiv.org/abs/2010.02534)

    - 이렇게 하면 형태소 분석기를 통해 조사, 어미 등의 문법형태소가 잘 분리되므로, 위에서 언급한 문제가 발생하지 않음.
    - 또한, 토큰을 자르는 단위가 더 세밀할수록 wordpiece, sentencepiece 학습 시 포함되는 vocabulary(vocab.txt)에 유의미한 형태소들을 집어넣을 수 있음.
        - 예를 들어, 조사를 잘 쪼개지 못한다면 `철수가` , `철수는` , `철수를` 등이 하나의 vocab으로 포함되어버려서, 정작 vocab에 포함되어야 할 다른 단어들이 들어갈 자리가 없어져버림.
    - mecab으로 한 뒤 왜 또 sentencepiece나 wordpiece를 하나?
        - mecab만 하면 신조어, 희귀 단어 등이 OOV(Out of vocabulary)가 됨. → OOV가 많아지면 성능 저하. 이를 예방하기 위해 OOV에 강한 sentencepiece, wordpiece를 추가로 적용. (Park et al. 2020)
- `뱃사람` 의 `ㅅ`, `춥다` 의 `ㅂ` 등과 같은 분석을 위해 자모 단위로까지 쪼개기도 함. (KR-BERT

결론: 한국어는 어절 단위(띄어쓰기 단위)로 토크나이징을 하는 것보다 형태소 분석 + wordpiece or sentencepiece를 적용하거나 자모 단위로 토크나이징을 하는 방식이 BERT의 성능을 더욱 높일 수 있음.

---

3. 한국어 downstream task

- **kornli**(https://github.com/kakaobrain/KorNLUDatasets)
    - Natual Language Inference
    - 구성: premise, hypothesis의 쌍으로 구성됨. 이들 쌍의 관계가 entailment인지, contradiction인지, neutral인지를 분류하는 태스크임.

Example	English Translation	Label
P: 저는, 그냥 알아내려고 거기 있었어요. H: 이해하려고 노력하고 있었어요.	I was just there just trying to figure it out.I was trying to understand.	Entailment
P: 저는, 그냥 알아내려고 거기 있었어요. H: 나는 처음부터 그것을 잘 이해했다.	I was just there just trying to figure it out.I understood it well from the beginning.	Contradiction
P: 저는, 그냥 알아내려고 거기 있었어요. H: 나는 돈이 어디로 갔는지 이해하려고 했어요.	I was just there just trying to figure it out.I was trying to understand where the money went.	Neutral

    - train: multinli train set + snli_1.0 train set = 942,854 line, dev: 2490 line, test: 5010 line
- **Korsts** (https://github.com/kakaobrain/KorNLUDatasets)
    - Semectic Textual similarity
    - 구성: 두 문장 쌍, label(의미적 유사도 0.0~5.0)
    - train: 5749, dev: 1500, test: 1379

Example	English Translation	Label
한 남자가 음식을 먹고 있다. 한 남자가 뭔가를 먹고 있다.	A man is eating food.A man is eating something.	4.2
한 비행기가 착륙하고 있다. 애니메이션화된 비행기 하나가 착륙하고 있다.	A plane is landing.A animated airplane is landing.	2.8
한 여성이 고기를 요리하고 있다. 한 남자가 말하고 있다.	A woman is cooking meat.A man is speaking.	0

- **NSMC**(https://github.com/e9t/nsmc)
    - naver movie review dataset
    - 영화 평점이 1(긍정), 0(부정)으로 레이블됨.
    - train: 150000, test: 50000

```
id      document        label
9976970 아 더빙.. 진짜 짜증나네요 목소리        0
3819312 흠...포스터보고 초딩영화줄....오버연기조차 가볍지 않구나        1
10265843        너무재밋었다그래서보는것을추천한다      0
```

```
9045019 교도소 이야기구먼 ..솔직히 재미는 없다..평점 조정        0
6483659 사이몬페그의 익살스런 연기가 돋보였던 영화!스파이더맨에서 늙어보이기만 했던 커스틴 던스트가 너무나도 이뻐보였다  1
```

- **PAWS-X** (https://github.com/google-research-datasets/paws/tree/master/pawsx) PAWS-X: A Cross-lingual Adversarial Dataset for Paraphrase Identification
  - 한국어를 포함하여 6개의 언어의 paraphrase identification dataset.
  - label: 두 문장이 서로 동일한 의미(1), 서로 다른 의미(0)
  - (한국어) train: 49401, dev: 2000 , test: 2000
    - 공식 repo에서 소개하는 데이터 양: 49,401 / 1965 / 1972 (데이터셋을 보면 'NS'라고 적힌 행이 몇 개 있음. NS 행 제거하면 이 숫자 나옴.)

```
id  sentence1 sentence2 label
1 1560 년 10 월 파리에서 그는 비밀리에 영국 대사 인 니콜라스 트록 모튼을 만났고 스코틀랜드를 통해 영국으로 돌아갈 여권을 요구했다. 1560 년 10 월 그는 파리
2 1975 년부터 76 년까지 NBA 시즌은 전국 농구 협회 (National Basketball Association)의 30 번째 시즌이었다.  National Basketball Associati
3 구체적인 토론, 공개 프로필 토론 및 프로젝트 토론도 있습니다.  공개 토론, 프로필 토론, 프로젝트 토론 등이 있습니다.  0
4 비교할 수있는 유량 비율을 유지할 수 있으면 결과가 높습니다.  비교 가능한 유속을 유지할 수있을 때 그 결과가 높습니다.  1
5 Akmola 지역에있는 Zerendi 지구의 좌석입니다.  Akmola 지역의 Zerendi 지구의 좌석입니다. 1
```

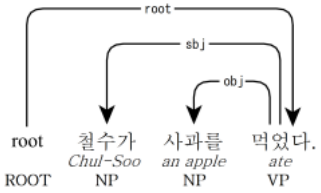
- **KorQuad(**https://korquad.github.io/)
  - 한국어 질의 응답
  - training: 60407, dev: 5774

표 1. KorQuAD 질문 유형

유형 1. 구문 변형 (56.4%)
Q: 김영하 소설가가 제 1 회 문학동네 작가상을 수상한 작품으로, 96 년 발표된 장편소설은 무엇인가?
1995 년 단편 <거울에 대한 명상>을 계간 《리뷰》에 발표하며 작품활동을 시작하였고 이듬해 96 년 장편《나는 나를 파괴할 권리가 있다》로 제 1 회 문학동네 작가상을 수상하였다.
유형 2. 어휘 변형 – 유의어 활용 (13.6%)
Q: 증강현실은 서덜랜드가 무엇을 발전시킨 것을 시작으로 연구가 **시작**되었는가?
이반 서덜랜드가 see-through HMD를 발전시킨 것을 **시초**로 하여 연구되기 시작한 증강현실은 …
유형 3. 어휘 변형 – 일반상식활용 (3.9%)
Q: **해외**에서 활동하는 Kayip, Superdrive와 함께 결성한 프로젝트 그룹의 이름은?
영국에서 활동하고 있는 Kayip, **베를린**에서 활동하고 있는 Superdrive 와 함께 프로젝트 그룹 '모텟'을 결성 …
유형 4. 여러 문장의 근거 종합적 활용 (19.6%)
Q: 클레멘스가 명예의 전당에 입성하지 못한 이유는?
… 이 투수들이 **클레멘스를 제외하고 모두 명예의 전당에 올랐기 때문**이다. 클레멘스만이 … 받았다. **그는** 경기력 향상 **약물 사용에 연루되어 있기 때문에** 입성 여부가 불확실하다.
유형 5. 논리적 추론 요구 (3.6%)
Q. 대한민국 제17대 대통령 선거 당시 후보로 등록했으나 예비경선의 경선 후보로 뽑히지 못한 사람 중 법무부와 관련 있는 사람은?
정동영 전 열린우리당 의장, …, **천정배** 전 **법무부 장관**, … 등이 후보로 등록하였고… 예비경선으로 정동영, 손학규, 이해찬, 한명숙, 유시민 후보가 경선 후보로 결정되었다.
유형 6. 기타 출제 오류 (2.9%)
Q. 티베트 고원에서 발원하는 강은?
… 강들이 티베트 고원에서 발원하는데, 창장, 황허, 인더스, 사틀루즈, 창포 (…), 메콩, 이라와디, 살윈 강 등이 포함된다.

kornli, korsts, nsmc, paws-x, korquad: kortok github에 공개된 코드에 BERT 모델을 넣어서 구동 가능

- **KLUE**(https://arxiv.org/pdf/2105.09680.pdf)
  - github에 공개된 코드에 BERT 모델을 넣어서 구동 가능
  - Dependency parsing(DP)
    - 단어 간 의존 관계 찾기 (head, dependents)



    - 화살표의 방향: head → dependent
    - '먹었다'가 '사과를' 및 '철수가'의 head. '사과를', '철수가'는 '먹었다'의 dependents
    - heaed의 인덱스를 찾고, relation class를 예측하는 것.
  - RE (Relation Extraction)
    - 텍스트 내의 entity 간의 의미적 관계, subject entity와 object entity 간의 관계

```
{
        "guid": "klue-re-v1_dev_00005",
        "sentence": "왕손의 나이 6세에 왕손사부를 임명하는 것이 관례였지만 영조는 1757년에 두 왕손을 가르칠 왕손교부를 초빙, 임명하였다.",
        "subject_entity": {
            "word": "영조",
            "start_idx": 31,
            "end_idx": 32,
            "type": "PER"
        },
        "object_entity": {
            "word": "1757년",
            "start_idx": 35,
            "end_idx": 39,
            "type": "DAT"
        },
        "label": "no_relation",
        "source": "wikipedia"
    },
```

  - yant(younhap news agency dataset for topic classification
    - 뉴스 기사 제목을 통해 토픽을 예측함

      `"5억원 무이자 융자는 되고 7천만원 이사비는 안된다"` → `경제`

  - STS (Semantic Textual Similarity)
    - korsts와 목적 동일
  - NLI (Natural Language Inference)
    - kornli와 목적 동일
  - WoS (Wizard of Seoul)
    - Multi domain dialogue state tracking
    - 멀티턴 대화에서의 Dialogue states 예측

Table 22: An example of dialogue state tracking in our WoS. Note that all dialogue states are cumulative in the actual dataset and that we only track states in the user turns.

Utterances (English Translations)	Dialogue States
User: 안녕하세요. (Hello.)	-
Sys: 네. 안녕하세요. 무엇을 도와드릴까요? (Hello. How can I help?) **User**: **서울 중앙**에 위치한 **호텔**을 찾고 있습니다. 외국인 친구도 함께 갈 예정이라서 원활하게 **인터넷을 사용할 수 있는 곳**이었으면 좋겠어요. (I'm looking for a **hotel** at the city **center**. I'm going with a foreign friend, so easy access to the **Internet should be available**.)	**Hotel-area**: center **Hotel-type**: hotel **Hotel-internet**: yes
Sys: 네 확인해보겠습니다. 혹시 추가로 필요하신 사항이 있으실까요? (Sure, let me check. Do you need anything else?) **User**: 음... 예약 인원은 총 **8명**이고요. 아, **가격대는 크게 상관 없습니다**. (Hmm.. I want to reserve for **8 people**. Ah, the **price range doesn't matter**.)	**Hotel-area**: center **Hotel-type**: hotel **Hotel-internet**: yes **Hotel-book people**: 8 **Hotel-price range**: dontcare
Sys: 네, 확인 감사합니다. 숙박을 원하시는 요일과 기간 같이 확인 부탁드립니다. (Great, thanks for confirming. Please let us know when and how long you want to stay.) **User**: 아, 중요한 걸 깜빡했네요. **일요일**에 **2일**간 예약하고 싶습니다. (Right, I forgot an important thing. I would like to book for **two days** from **Sunday**.)	**Hotel-area**: center **Hotel-type**: hotel **Hotel-internet**: yes **Hotel-book people**: 8 **Hotel-price range**: dontcare **Hotel-book day**: Sunday **Hotel-book stay**: 2

- NER(Named Entity Recognition
  - 개체명 인식

```
## klue-ner-v1_dev_00000-wikitree <경찰:OG>은 또 성매매 알선 자금을 관리한 <박:PS>씨의 딸(<32:QT>)과 성매매 여성 <김:PS>모(<33:QT>)씨 등 <16
경  B-OG
찰  I-OG
은  O
  O
또  O
  O
성  O
매  O
매  O
  O
알  O
선  O
  O
자  O
금  O
을  O
  O
관  O
리  O
한  O
  O
박  B-PS
씨  O
의  O
  O
딸  O
(  O
3 B-QT
2 I-QT
)  O
과  O
  O
성  O
매  O
매  O
  O
여  O
성  O
  O
김  B-PS
모  O
(  O
3 B-QT
3 I-QT
)  O
씨  O
  O
등  O
  O
1 B-QT
6 I-QT
```

```
명  I-QT
을  O
  O
같  O
은  O
  O
혐  O
의  O
로  O
  O
불  O
구  O
속  O
  O
입  O
건  O
했  O
다  O
. O
```

- Machine Reading Comprehension (MRC)
  - 질의응답


```
"title": "BMW 코리아, 창립 25주년 기념 ‘BMW 코리아 25주년 에디션’ 한정 출시",
            "paragraphs": [
                {
                    "context": "BMW 코리아(대표 한상윤)는 창립 25주년을 기념하는 ‘BMW 코리아 25주년 에디션’을 한정 출시한다고 밝혔다. 이번 BMW 코
                    "qas": [
                        {
                            "question": "말라카이트에서 나온 색깔을 사용한 에디션은?",
                            "answers": [
                                {
                                    "text": "뉴 740Li 25주년 에디션",
                                    "answer_start": 666
                                },
                                {
                                    "text": "뉴 740Li 25주년",
                                    "answer_start": 666
                                }
                            ],
                            "question_type": 2,
                            "is_impossible": false,
                            "guid": "klue-mrc-v1_dev_01891"
                        }
                    ]
                }
            ],
            "news_category": "자동차",
            "source": "acrofan"
        },
```