Project 2

뉴스 토픽 분류 AI 경진대회

Kyonggi University Applied Statistics Data Analysis Club D.N.A

1.주제

- 한국어 뉴스 헤드라인을 이용하여, 뉴스의 주제를 분류하는 알고리즘 개발

2. 배경

- 텍스트 주제를 추론하는 것은 언어 이해 시스템이 보유해야 하는 핵심 기능입니다. YNAT(주제 분류를 위한 연합뉴스 헤드라인) 데이터 세트를 활용해 주제 분류 알고리즘을 개발해 주세요.
- 국내 최초 오픈 데이터 세트인 KLUE(Korean Language Understanding Evaluation) 데이터 세트를 이용하여 다양한 언어 모델의 성능을 비교해 한국어 자연어처리 분야의 발전에 기여할 것으로 예상합니다.

3. 평가지표

- 정확도 (Accuracy)

데이터 전처리

01. Raw Data

	index	title	topic_idx
0	0	인천→핀란드 항공기 결항…휴가철 여행객 분통	4
1	1	실리콘밸리 넘어서겠다…구글 15조원 들여 美전역 거점화	4
2	2	이란 외무 긴장완화 해결책은 미국이 경제전쟁 멈추는 것	4
3	3	NYT 클린턴 측근韓기업 특수관계 조명…공과 사 맞물려종합	4
4	4	시진핑 트럼프에 중미 무역협상 조속 타결 희망	4
...	...	...	...
45649	45649	KB금융 미국 IB 스티펠과 제휴…선진국 시장 공략	1
45650	45650	1보 서울시교육청 신종코로나 확산에 개학 연기·휴업 검토	2
45651	45651	게시판 키움증권 2020 키움 영웅전 실전투자대회	1
45652	45652	답변하는 배기동 국립중앙박물관장	2
45653	45653	2020 한국인터넷기자상 시상식 내달 1일 개최…특별상 김성후	2

45654 rows × 3 columns

- index : 헤드라인 인덱스
- title : 뉴스 헤드라인
- topic_idx : 뉴스 주제 인덱스 값(label)

02. 토큰화

Konlpy의 Okt() 이용해 형태소 단위로 토큰화

```
['인천', '→', '핀란드', '항공기', '결항', '…', '휴가', '철', '여행객', '분통', '실리콘밸리', '넘어서겠다', '…', '구글', '15조원', '들여', '美', '전역', '거점', '화', '이란', '외무', '긴장', '완화', '해결', '책', '은', '미국', '이', '경제', '전쟁', '멈추는', '것', 'NYT', '클린턴', '측근', '韓', '기업', '특수', '관계', '조명', '…', '공과', '사', '맞', '물려', '종합', '시진핑', '트럼프', '에', '중미', '무역', '협상', '조속', '타결', '희망', '팔레스타인', '가자지구', '서', '16', '세', '소년', '이스라엘군', '총격', '에', '사망', '인도', '48년', '만에', '파키스탄', '공습', '…', '테러', '캠프', '폭격', '종합', '2', '보', '美', '대선', 'TV', '토론', '음담패설', '만회', '실패', '트럼프', '…', '사과', '대신', '빌', '클린턴', '공격', '해', '역효과', '푸
```

데이터 전처리

03. 불용어 제거

여러 토픽에서 겹치는 단어를 불용어로 처리 후 제거

```
['종합', '전국', '최고', '올해', '최근', '최대', '여자',
'방문', '여성', '게시판', '오후', '내일', '소식', '기자',
'목표', '이익', '특징', '전환', '이동', '판매', '내년',
'속보', '공개', '국내', '작년', '특징', '증가', '실적',
'한국', '지원', '사장', '회장', '협회', '사망', '공격',
'제재', '속보', '회의']
```

04. 품사 태깅

불용어 제거한 후 각 형태소에 대해 품사 부착

```
('사과', 'Noun'), ('대신', 'Noun'), ('빌', 'Verb'), ('클린
턴', 'Noun'), ('공격', 'Noun'), ('해', 'Verb'), ('역효과',
'Noun'), ('푸틴', 'Noun'), ('한반도', 'Noun'), ('상황', 'No
un'), ('진전', 'Noun'), ('위', 'Noun'), ('한', 'Josa'),
('방안', 'Noun'), ('김정은', 'Noun'), ('위원장', 'Noun'),
('과', 'Josa'), ('논의', 'Noun'), ('특검', 'Noun'), ('면죄
부', 'Noun'), ('받은', 'Verb'), ('트럼프', 'Noun'), ('스캔
들', 'Noun'), ('보도', 'Noun'), ('언론', 'Noun'), ('맹공',
'Noun'), ('…', 'Punctuation'), ('국민', 'Noun'), ('의', 'J
osa'), ('적', 'Noun'), ('日', 'Foreign'), ('오키나와', 'Nou
n'), ('서', 'Josa'), ('열린', 'Verb'), ('강제', 'Noun'),
```

→ 같은 형태의 단어라도 의미가 다를 수 있어 의미 구분 가능

05. TF-IDF 벡터화

상대적으로 빈도수가 낮은 단어의 가중치를 올려 중요한 단어 파악

EDA

IT과학

경제

사회

EDA

생활문화

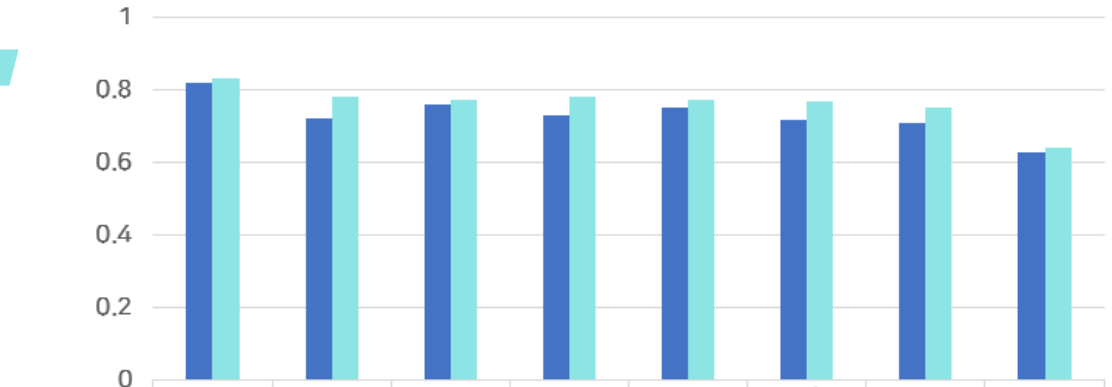
세계

스포츠

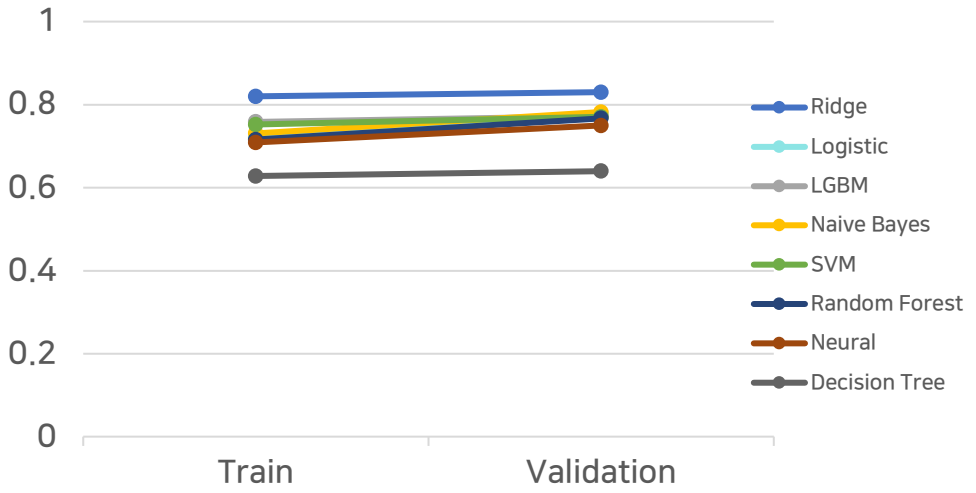
정치

전처리 비교

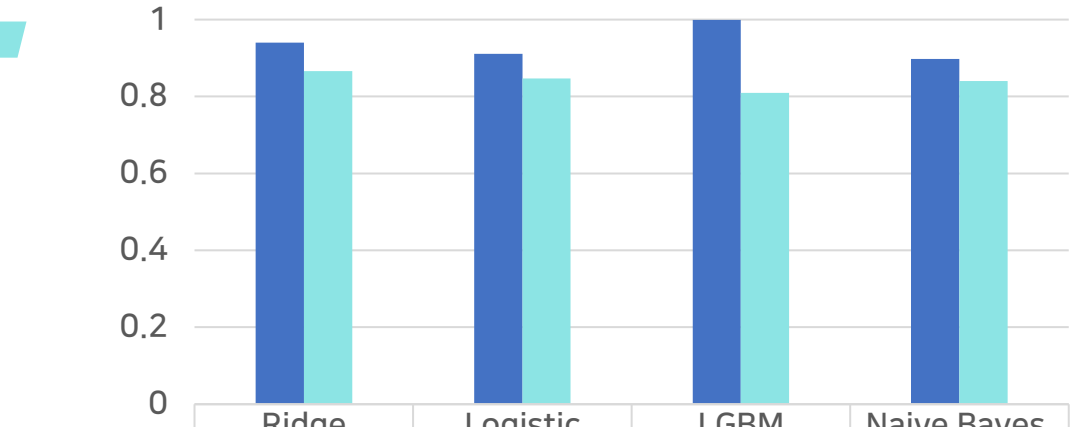
명사 Only



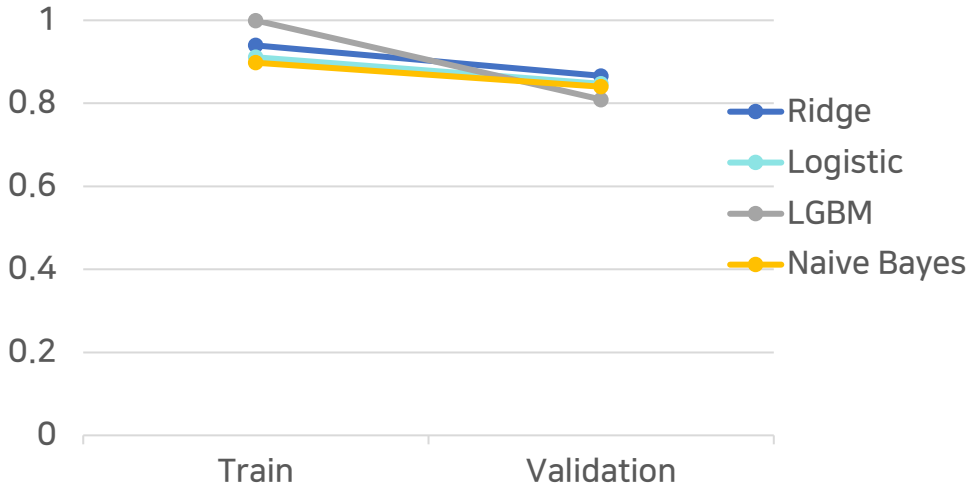
	Ridge	Logistic	LGBM	Naive Bayes	SVM	Random Forest	Neural	Decision Tree
Train	0.82	0.72	0.7585	0.731	0.752	0.716	0.709	0.628
Validation	0.8301	0.7788	0.772	0.782	0.7708	0.767	0.75	0.64



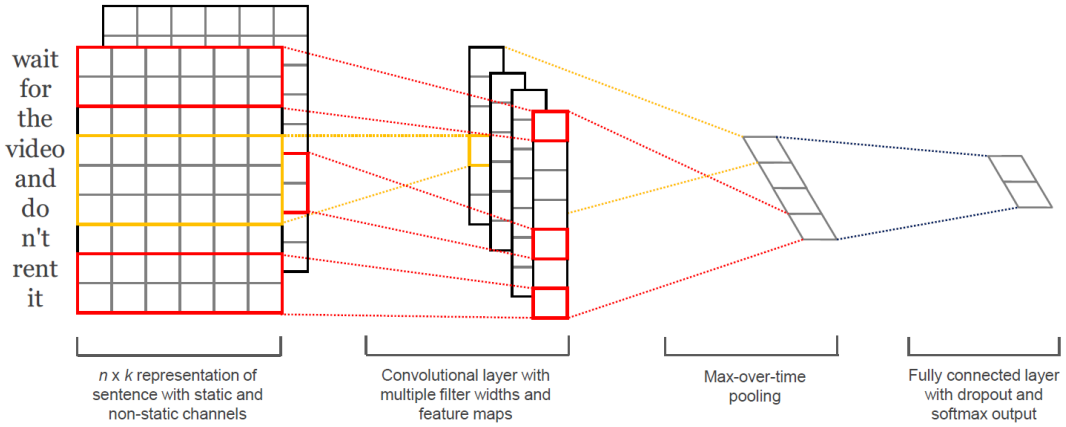
모든 품사

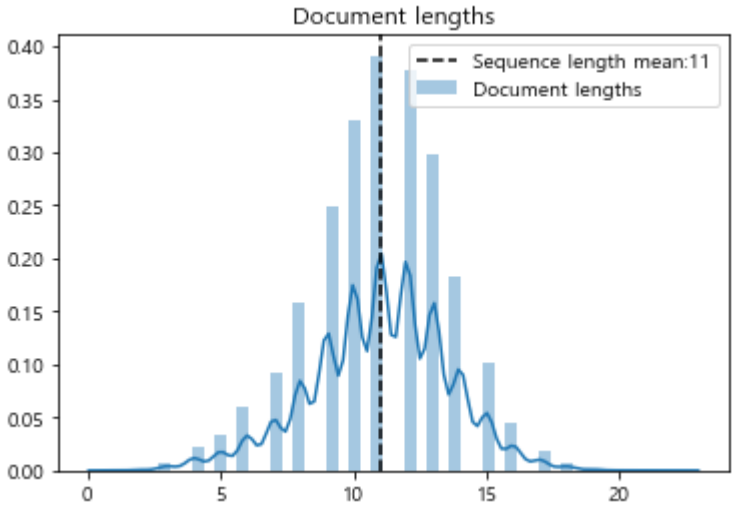


	Ridge	Logistic	LGBM	Naive Bayes
Train	0.9396	0.91099	0.99925	0.89787
Validation	0.86597	0.8469	0.80902	0.83989



CNN

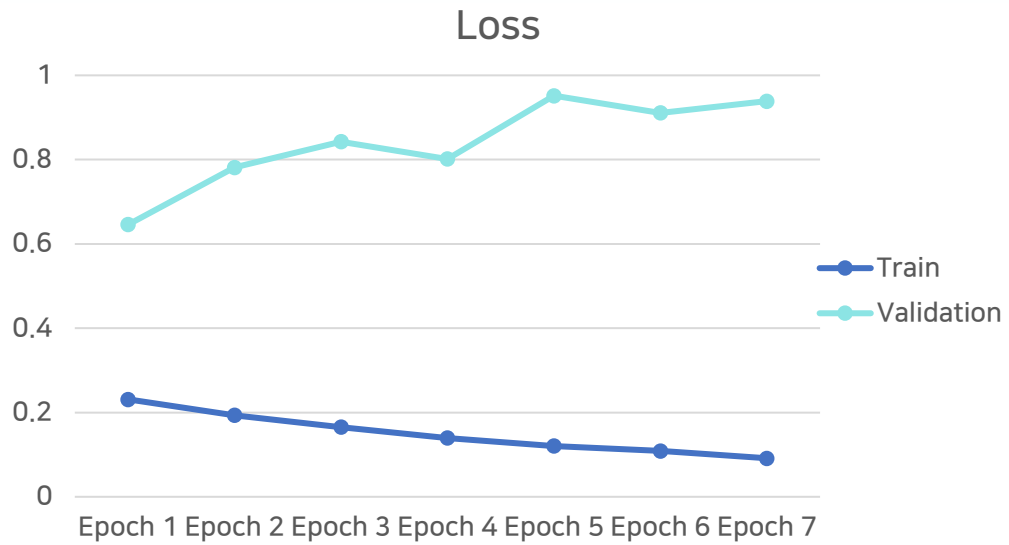


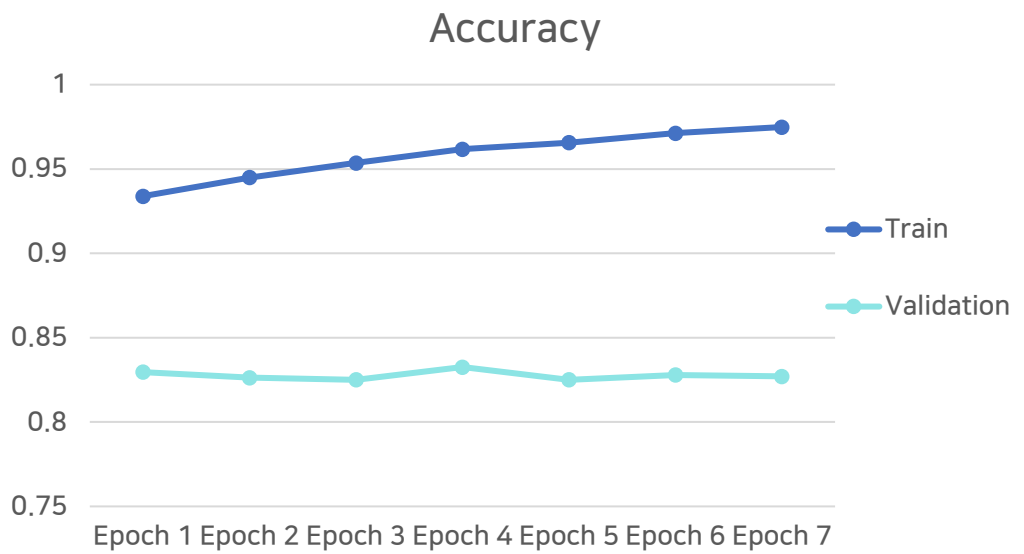


< 뉴스 헤드라인 별 문장의 길이 >

```
[[0. 0. 0. ... 0. 1. 0.]
 [0. 1. 0. ... 0. 0. 0.]
 [0. 0. 1. ... 0. 0. 0.]
 ...
 [0. 0. 0. ... 1. 0. 0.]
 [0. 0. 0. ... 0. 0. 1.]
 [1. 0. 0. ... 0. 0. 0.]]
(41088, 7)
```

< 원-핫 인코딩 >

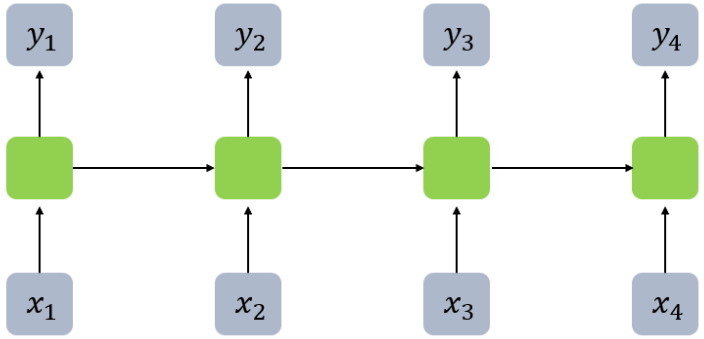




양방향 LSTM

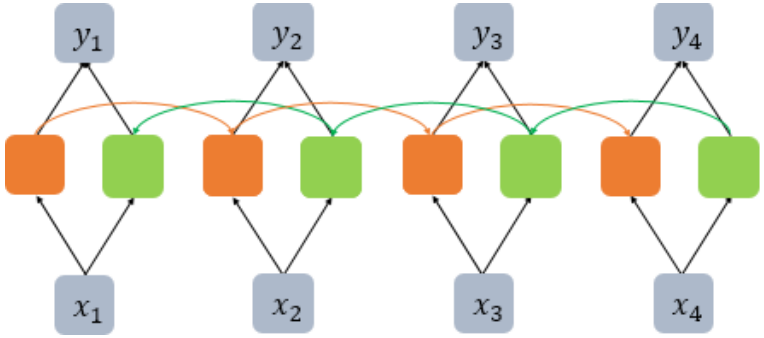
단방향 LSTM

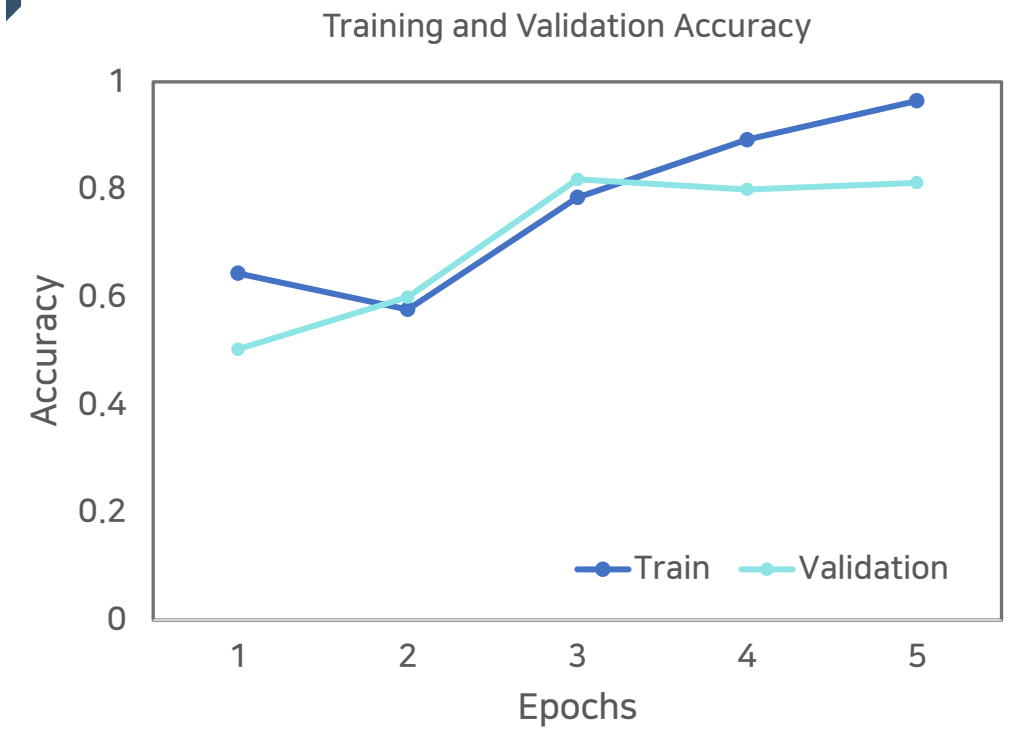
- 단어 순서가 가지는 문맥 정보가 한 방향으로만 학습.
- 뒤에 오는 단어의 영향을 받게 되는 경우 학습 불가능.



양방향 LSTM

- 양방향 순서 모두 학습 가능.
- 양방향으로 학습 후 두 결과를 합침.





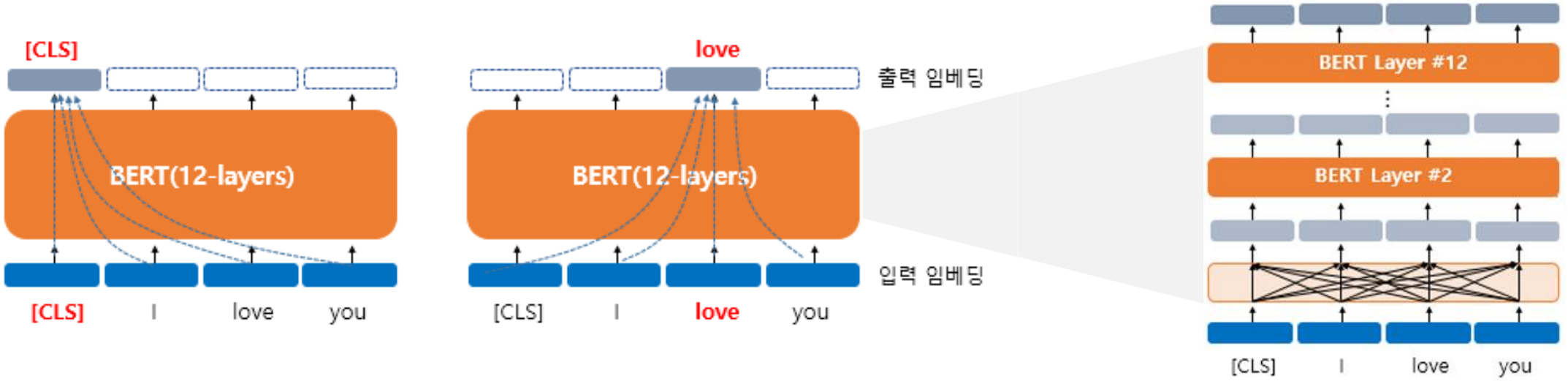
Test Accuracy : 0.7071193866

Bert

BERT(Bidirectional Encoder Representations from Transformers)

: 2018년, 구글이 개발한 사전 훈련된 모델

→ 위키피디아(약 25억 개) & BooksCorpus(약 8억 개)



➡ 모든 단어들을 참고해 문맥을 반영한 출력 임베딩을 얻음

Bert

서브워드 토크나이저(WordPiece) 사용

: 단어보다 더 작은 단위로 쪼개어 입력에 사용

```
['[CLS]인천→핀란드 항공기 결항…휴가철 여행객 분통[SEP]',
 '[CLS]실리콘밸리 넘어서겠다…구글 15조원 들여 美전역 거점화[SEP]',
 '[CLS]이란 외무 긴장완화 해결책은 미국이 경제전쟁 멈추는 것[SEP]',
 '[CLS]NYT 클린턴 측근韓기업 특수관계 조명…공과 사 맞물려종합[SEP]',
 '[CLS]시진핑 트럼프에 중미 무역협상 조속 타결 희망[SEP]']
```

- [CLS] : 문장이 시작됨 의미
- [SEP] : 문장과 문장 사이라는 것 의미
- ## : 단어의 중간부터 등장하는 서브워드라는 것을 알려주기 위함

※ BERT는 한국어에 특화되어 있지 않아 단어 집합에 존재하는 단어가 없음

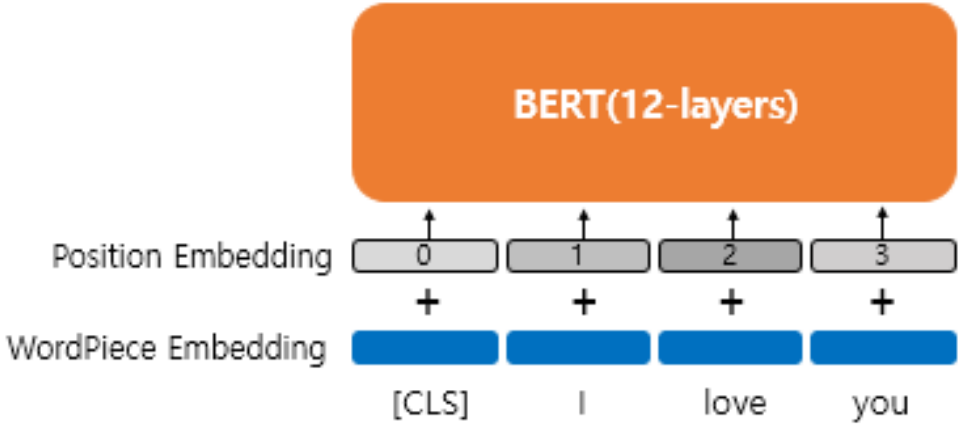
→ 한국어는 KoBERT

```
['[CLS]', '인', '##천', '##→', '##핀', '##란드', '항', '##공', '##기', '결', '##항', '[UNK]', '휴', '##가', '##철', '여', '##행', '##객', '분', '##통', '[SEP]']
['[CLS]', '실', '##리', '##콘', '##밸', '##리', '넘', '##어', '##서', '##겠', '##다', '[UNK]', '구', '##글', '15', '##조', '##원', '들', '##여', '美', '전', '##역', '거', '##점', '##화', '[SEP]']
['[CLS]', '이란', '외', '##무', '긴', '##장', '##완', '##화', '해', '##결', '##책', '##은', '미국', '##이', '경', '##제', '##전', '##쟁', '멈', '##추', '##는', '것', '[SEP]']
['[CLS]', 'NY', '##T', '클', '##린', '##턴', '측', '##근', '韓', '기', '##업', '특', '##수', '##관', '##계', '조', '##명', '[UNK]', '공', '##과', '사', '맞', '##물', '##려', '##종', '##합', '[SEP]']
['[CLS]', '시', '##진', '##핑', '트', '##럼', '##프', '##에', '중', '##미', '무', '##역', '##협', '##상', '조', '##속', '타', '##결', '희', '##망', '[SEP]']
['[CLS]', '팔', '##레스', '##타', '##인', '가', '##자', '##지', '##구', '##서', '16', '##세', '소', '##년', '이', '##스', '##라', '##엘', '##군', '총', '##격', '##에', '사', '##망', '[SEP]']
```

포지션 임베딩(Position Embedding)

: 포지셔널 인코딩을 사용해 단어의 위치 정보를 표현

(위치에 따라 다른 값을 가지는 행렬을 만들어 단어 벡터들과 더하는 방법)



Bert

어텐션 마스크(Attention Mask)

※ 어텐션 : 디코더(결과를 출력)에서 출력단어를 나타낼 때 매 시점마다 전체 입력 문장을 참고하는 연산

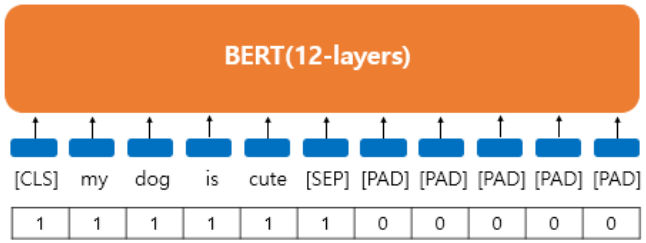
불필요하게 패딩 토큰에 대해 어텐션을 하지 않도록 실제 단어와 패딩 토큰을 구분하는 입력

```
array([   101,   9904, 100929,  22695,  12030,   8843,  13764,  12508,
        17196,  12424,  10250,  24982,   9448,  10954,   9638,  12605,
        17342,  96720,  17360,   9761,  45465,  10530,   9405,  89292,
          102,      0,      0,      0,      0,      0,      0,      0])
```

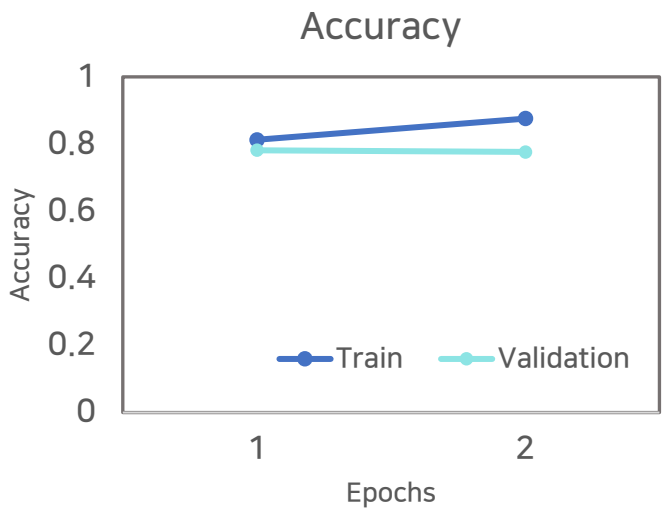
< Padding 예시 >

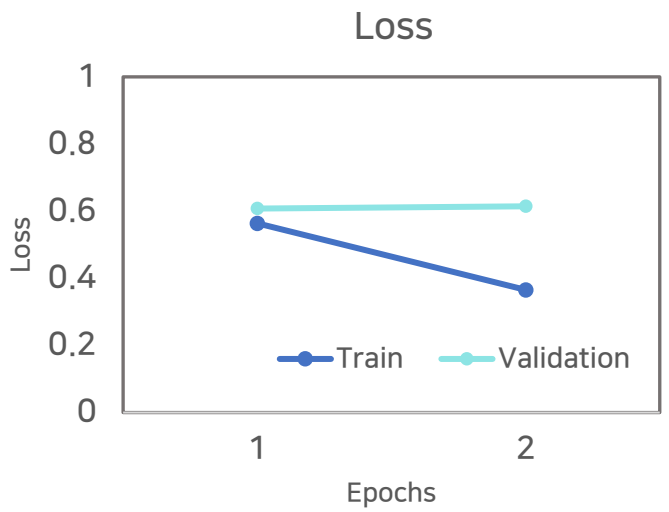
```
[1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0,
 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0,
 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0,
 1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0]
```

< 어텐션 마스크 예시 >



< BERT의 최종 입력 >





Test Data

Public 점수 : 0.8313253012

Private 점수 : 0.7939115199

최종 결론

Ridge	양방향 LSTM	Bert
Test Data	Test Data	Test Data
Public 점수 : 0.8256297919	Public 점수 : 0.7071193866	Public 점수 : 0.8313253012
Private 점수 : 0.8024529128	Private 점수 : 0.6798072711	Private 점수 : 0.7939115199

최종 점수 : PUBLIC : 0.83132, PRIVATE : 0.80245

Thank you

감사합니다!

Kyonggi University Applied Statistics Data Analysis Club **D.N.A**