사전학습을 활용한 논문QA

팀 유나성

팀원소개

윤재웅 KAIST 자연어처리 장문이해 사전학습 아성이아빠
이현제 KAIST 자연어처리 QA Summarization Dialogue
최현진 KAIST 자연어처리 문장임베딩 멀티모달 STS

CONTENTS

01 Problem

기계독해(MRC, QA)란?
주어진 지문에서 질문에 대한 답을 찾는 태스크

논문과 같은 고도의 전문 지식을 담고 있는 텍스트를 기계 독해하는 경우,
일반 텍스트에 비해 향상된 "이해"와 "추론" 능력이 필요

조건: 과학기술분야 사전학습언어모델(KorSciBERT, KorSciElectra)를 사용한 최대성능 도출

어떻게 학습하면 논문QA의 성능을 향상시킬 수 있을까?

지문. 딥러닝 모델을 활용한 자연어 처리 분야에서 Pretext Task를 활용하여 목적 태스크(Target Task)의 성능을 향상시키는 방식은 하나의 패러다임으로 자리잡았다. 그 중 BERT 모델은 다양한 자연어 처리 분야에서 가장 성공적인 모델 중 하나이지만 문서 요약과 같은 자연어 생성 태스크에 있어서의 적용은 비교적 활발히 이루어지지 않고 있다. 이는 NLU(Natural Language Understanding) 모델이 생성 태스크에서 가지는 성능의 한계에 기인한 것으로 보인다. 그러나 사전학습된(Pre-trained) NLU모델을 활용하면 NLG(Natural Language Generation) 모델을 보다 적은 자원으로 효율적으로 만들 수 있다. 본 연구에서는 BERT와 같은 Transformer 계열의 NLU 모델을 대화 요약 태스크에 적용하는 방식을 제안하고자 한다. 구체적으로, 본 연구에서는 대화문만이 가지는 특성을 활용한 4가지 Pretext Task를 제안하며 이를 통해 생성형 대화 요약 테스크의 성능을 향상시키는 방법을 제안한다. 제안하는 방식의 검증을 위해 2가지 서로 다른 성격을 가진 데이터셋를 사용하였고 일부 실험 결과에서 ROUGE Score 기준 약 80%의 성능 향상 효과를 확인 할 수 있었다.

질문. BERT 모델은 자연어 생성 태스크에 있어서는 왜 활발히 적용되지 않고 있는가?

답. NLU모델이 생성 태스크에서 가지는 성능의 한계에 기인한 것

02 Approach

사전학습(Pre-train)은 왜 하는가?
자기지도학습 방법을 통해 레이블 데이터 없이 언어의 기본기를 학습할 수 있음
사전학습 후 레이블이 달린 데이터로 파인튜닝 시, 성능 향상 효과

논문 QA에 적합한 사전학습 방법을 고안하여,
파인튜닝(Fine-tuning)시의 성능 향상이 가능하지 않을까?

빈칸 채우기 (Masked Language Model)

BERT는 [ ] 모델이다.

1) 사전학습 2)MLM N)…

02 Approach: 사전학습

1 Sentence Order Prediction (SOP)

: 뒤바뀐 문장을 예측함으로써 지문의 문맥을 파악하는 능력을 학습
: 전체 문장의 45-50%의 순서만 변경하도록 처리

딥러닝 모델을 활용한 자연어 처리 분야에서
사전학습을 활용하여 목적 태스크의 성능을
향상시키는 방식은 하나의 패러다임으로 자리잡았다.
BERT 모델은 가장 대표적인 사전학습 모델이다.
그러나 자연어 생성 테스크에 있어서의 적용은 비교적
활발히 이루어지지 않고 있다. 이는 NLU 모델이 생성
태스크에서 가지는 성능의 한계에 기인한 것으로
보인다. 그러나 사전학습된 NLU모델을 활용하면
NLG모델을 보다 적은 자원으로 효율적으로 만들 수
있다.

순서바꾸기

딥러닝 모델을 활용한 자연어 처리 분야에서
사전학습을 활용하여 목적 태스크의 성능을
향상시키는 방식은 하나의 패러다임으로 자리잡았다.
이는 NLU 모델이 생성 태스크에서 가지는 성능의
한계에 기인한 것으로 보인다. 그러나 자연어 생성
테스크에 있어서의 적용은 비교적 활발히 이루어지지
않고 있다. BERT 모델은 가장 대표적인 사전학습
모델이다. 그러나 사전학습된 NLU모델을 활용하면
NLG모델을 보다 적은 자원으로 효율적으로 만들 수
있다.

첫번째와 마지막 문장은
순서를 변경하지 않음

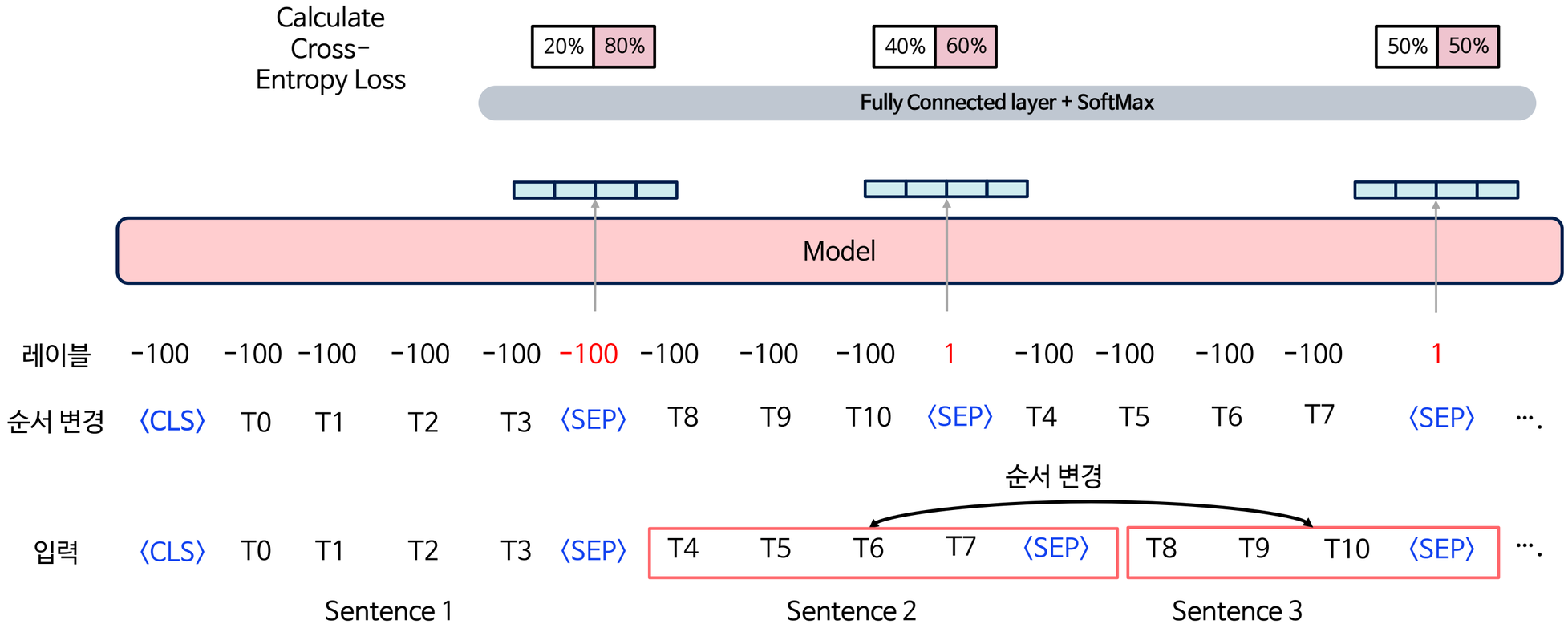
0
1
0
1
0

Label

참고 논문: Enhancing Semantic Understanding with Self-supervised Methods for Abstractive Dialogue Summarization, Interspeech 2021, 796-800

02 Approach: 사전학습

Architecture – Sentence Order Prediction



-100 : Loss 계산에서 제외, 0: 원래 순서, 1: 순서 변경됨

02 Approach: 사전학습

2 Sentence Coherence Prediction (SCP)

: 문맥에 맞지 않는 문장을 예측함으로써 지문의 맥락을 파악하는 능력을 학습
: 전체의 45-50%의 문장만이 교체되도록 처리

논문A

딥러닝 모델을 활용한 자연어 처리 분야에서 사전학습을 활용하여 목적 태스크의 성능을 향상시키는 방식은 하나의 패러다임으로 자리잡았다. BERT 모델은 가장 대표적인 사전학습 모델이다. 그러나 자연어 생성 테스크에 있어서의 적용은 비교적 활발히 이루어지지 않고 있다. 이는 NLU 모델이 생성 태스크에서 가지는 성능의 한계에 기인한 것으로 보인다. 그러나 사전학습된 NLU모델을 활용하면 NLG모델을 보다 적은 자원으로 효율적으로 만들 수 있다.

논문B

딥러닝을 이용한 암종 분류 태스크에서 어떤 부위에서 발생한 암종의 분류를 다른 부위의 같은 암종 이미지를 이용하여 빠르게 학습시키는 방법을 제안한다. BERT 모델은 가장 대표적인 사전학습 모델이다. 서로 다른 부위에서 발생한 암의 병리 이미지간 전이학습이 이루어지는지 분석하여 그 가능성을 검증하였다. 먼저 폐암 병리이미지를 정상, 샘암종, 편평세포암종 3가지로 분류하는 모델을 학습시키고, 해당 모델을 기반으로 대장암 병리 이미지를 정상, 샘암종 2가지로 분류하는 모델을 만들었다.

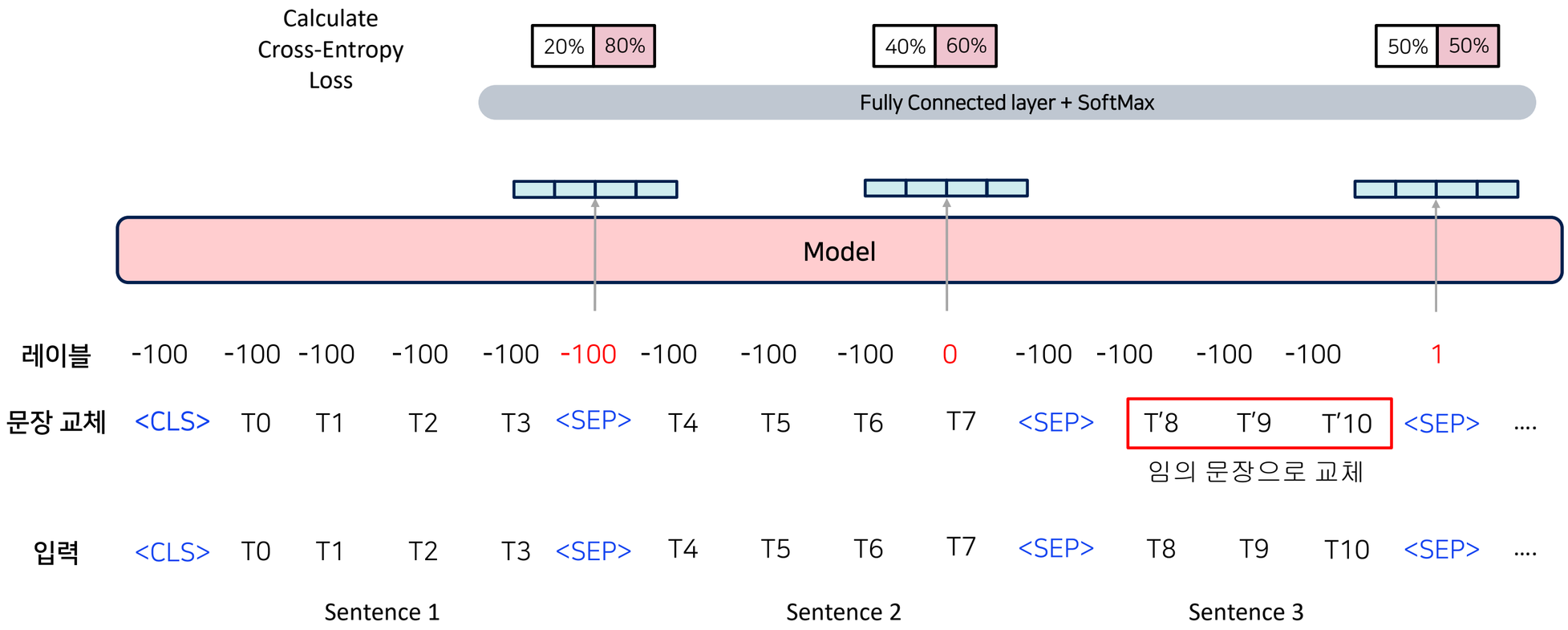
첫번째와 마지막 문장은 원래 문장 유지

0
1
0
0
0
Label

참고 논문: Enhancing Semantic Understanding with Self-supervised Methods for Abstractive Dialogue Summarization, Interspeech 2021, 796-800

02 Approach: 사전학습

Architecture – 2 Sentence Coherence Prediction



-100 : Loss 계산에서 제외, 0: 원본 문장, 1: 교체된 문장

02 Approach: 사전학습

3 Keyword Prediction

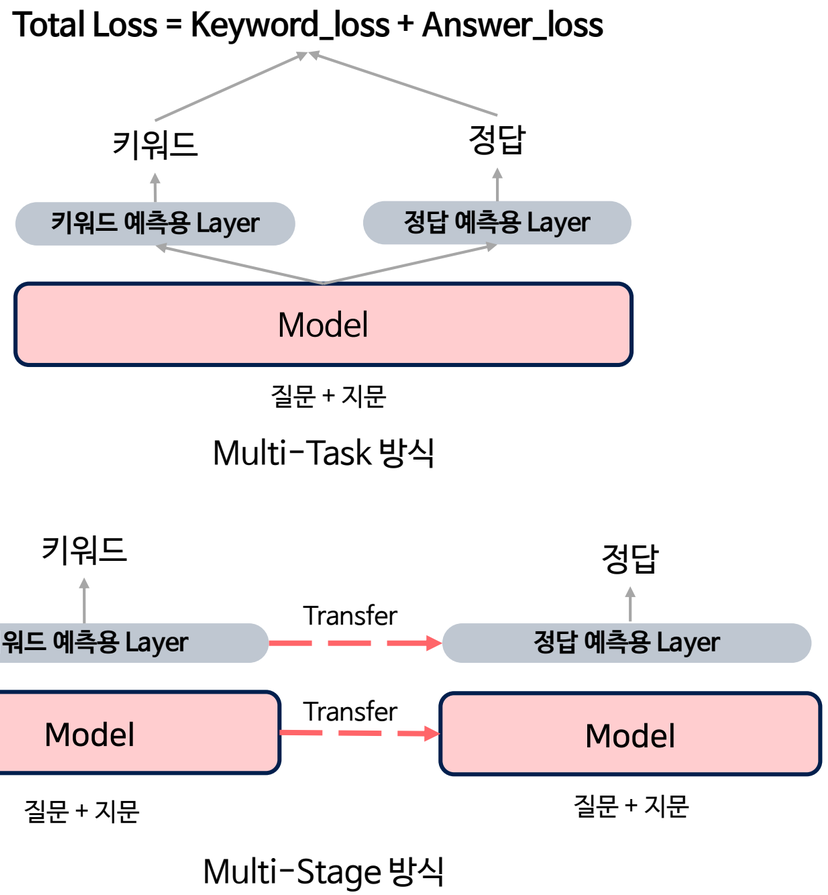
: 국내논문QA 데이터셋에 주어진 Keyword를 예측하도록 하여 지문 이해도를 높임

지문. 딥러닝 모델을 활용한 자연어 처리 분야에서 사전학습을 활용하여 목적 태스크의 성능을 향상시키는 방식은 하나의 패러다임으로 자리잡았다. BERT 모델은 가장 대표적인 사전학습 모델이다. 그러나 자연어 생성 태스크에 있어서의 적용은 비교적 활발히 이루어지지 않고 있다. 이는 NLU 모델이 생성 태스크에서 가지는 성능의 한계에 기인한 것으로 보인다. 그러나 사전학습된 NLU모델을 활용하면 NLG모델을 보다 적은 자원으로 효율적으로 만들 수 있다.

질문. BERT 모델은 자연어 생성 태스크에 있어서는 왜 활발히 적용되지 않고 있는가?

정답. NLU 모델이 생성 태스크에서 가지는 성능의 한계

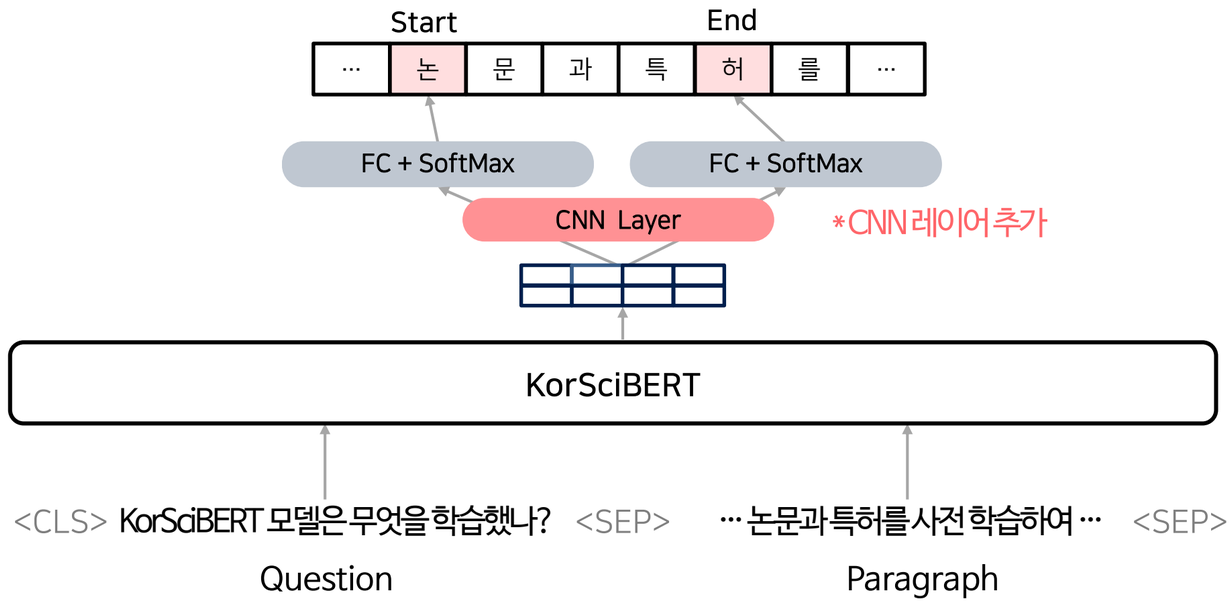
키워드. 생성 태스크



02 Approach: 아키텍쳐 개선

4 CNN 추가

KorSciBERT의 아웃풋에 CNN레이어를 추가하여 예측 정밀도를 높이고자 함



참고자료

<Sentence Embedding 성능 비교> *논문 발췌

Model	Spearman (Pearson)
Not fine-tuned	
BERT [CLS]-token embedding	6.43 (1.70)
BERT Avg. pooled token embedding	47.29 (47.91)
ALBERT [CLS]-token embedding	0.86 (4.57)
ALBERT Avg. pooled token embedding	47.84 (46.57)
Fine-tuned on STSb	
BERT [CLS]-token embedding	12.96 (7.49)
BERT Avg. pooled token embedding	55.76 (54.90)
SBERT	84.66 (84.86)
CNN-SBERT	85.72 (86.15)
ALBERT [CLS]-token embedding	37.98 (27.89)
ALBERT Avg. pooled token embedding	61.06 (60.41)
SALBERT	74.33 (75.26)
CNN-SALBERT	**82.30 (83.08)**

* Evaluation of BERT and ALBERT Sentence Embedding Performance on Downstream NLP Tasks, International Conference on Pattern Recognition (ICPR 2020), January 2021

03 Experiments

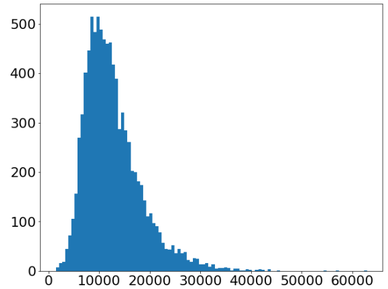
1 Data

국내논문 QA 데이터셋

기계가 과학기술 문헌을 읽고 이해하는 능력을 평가하기 위한
질의응답 데이터셋

크기. 824,337개
구성. 10만 개 추출 (Train 8만개, Validation 1만개, Test 1만개)
- 잘못 레이블된 데이터 제거
- 정답과 무관한 특수문자 기호 제거

특징.
평균 논문 길이: 12,700자
평균 질문 길이: 76자
평균 정답 길이: 30자



논문 길이 분포도

활용.
정답이 포함된 문단 (약 1000자≒KorSciBERT의 최대 인풋 길이)을
추출하여 학습에 활용
→ 학습 속도 90% 이상 향상
평가 시에는 전문 입력 (Sliding Window 방식)

사전학습.
추출한 데이터를 바탕으로 사전학습용 데이터 생성하여 사용
(Appendix 표1)

03 Experiments

2 Setup

NVIDIA V100 x4
Pytorch, HuggingFace Library
CentOS 7.9

Stage	LR (AdamW)	Batch	Epoch	Training Duration	Test Duration
Pre-training	5.00E-05	64	3	1.5 Hours	0.5 Hours
QA Fine-tuning	4.00E-05	64	5	5 hours	**3 Hours**

03 Experiments

3 Evaluation Metric

평가 지표 : EM, F1 사용
· Exact Match (EM) : 정답 텍스트의 어절과 예측 텍스트 어절 간의 단순 비교
(정답 1, 오답 0 으로 계산)

· F1 Score: 정답 텍스트와 예측 텍스트 어절 간의 정밀도(P) 와 재현율(R)을 구해서 F1 점수 계산

03 Experiments

4 Results

표1. 사전학습방법별 성능 평가

Method	Test		성능 차이 (평균)
	F1	EM	
Baseline (No Pre-training)	83.32	45.60	0.00
Sentence Order Prediction (SOP)	85.90	49.59	+3.285
Sentence Coherence Prediction (SCP)	85.94	49.52	+3.270
Keyword Prediction (KP)	85.98	49.42	+3.240

표2. QA 난이도별 성능 평가

난이도	Method	F1	EM	성능 차이 (평균)
하	Baseline	87.03	52.66	
	Ours	88.61	54.93	+1.925
중	Baseline	85.57	46.24	
	Ours	87.74	50.54	+3.235
상	Baseline	76.99	37.37	
	Ours	81.16	42.57	+4.685

*Ours: SCP Best Model

하: 키워드와 응답이 한 문장 내에 존재
상: 키워드와 응답이 다른 문장에 존재

- Baseline은 논문QA 데이터셋으로 파인튜닝만 진행 후 평가
- 제안한 3가지 사전학습 방법 모두 성능 개선 효과가 있었음
- SOP와 SCP는 Unsupervised Method임에도 불구하고 KP 이상의 성능 향상
- 난이도가 높은 문제일수록 제안한 사전학습 방법에서의 효과가 뛰어남

03 Experiments

4 Results Detail

표3. 구현기법별 성능비교

Method	Test		성능 차이 (평균)
	F1	EM	
Baseline (No Pre-training)	83.32	45.60	0.00
SOP	85.90	49.59	+3.29
SOP + CNN	85.53	48.20	+2.41
SOP + Augment	83.98	47.48	+1.27
SCP	84.65	47.65	+1.69
SCP + CNN	84.77	48.18	+2.02
SCP + CNN + Augment	85.94	49.52	+3.27
KP (Multi Stage)	84.92	47.84	+1.92
KP (Multi Task)	84.27	47.36	+1.36
KP (Multi Stage) + CNN	85.98	49.42	+3.24

CNN추가의 효과

- SCP와 KP에서 성능 향상

Data Augmentation

- SOP와 SCP에서 학습 데이터를 5배로 증가시켜 학습 진행
- SCP에서 성능 향상

KP의 경우 Multi Stage 학습에서 더 큰 성능 향상

- Multi Stage: Key Prediction 훈련 후, 논문QA 데이터셋 훈련
- Multi Task: Key Prediction 훈련과 논문QA 훈련을 동시에 진행

04 Conclusion

1 3가지 사전학습 방식 모두 논문QA 성능 향상에 도움

SOP, SCP, KP 모두 성능 향상 효과가 있었음
별도의 레이블링이 필요 없는 자기지도학습으로 논문과 같은 전문지식 QA Task 성능 향상

2 난이도가 높은 문제일수록 효과가 극명

제시된 데이터셋에서 높은 난이도의 문제일수록 추가 사전학습의 효과가 크게 나타남을 확인
일반 텍스트보다 논문과 같은 전문 지식 QA에서의 효과성을 입증

3 사전학습 방법에서 디테일한 설정의 차이가 결과를 좌우

몇%의 문장을 섞는가
첫 번째와 마지막 문장은 반드시 Fix

05

Future Work

1 다른 사전학습 모델에서의 활용

KorSciBERT 모델 외의 다른 모델들(KorSciELECTRA, BERT, RoBERTa 등)에서도 동일한 효과가 나타나는가

2 다른 Task, Dataset에서의 활용

제시한 사전학습 방법은 General한 텍스트에도 활용이 가능한 방법
논문QA가 아닌 다른 태스크나 데이터셋에서도 동일한 효과가 나타나는가

3 논문 작성

해당 연구 결과를 논문으로 작성하여 자연어처리 학회에 제출

06 Demo

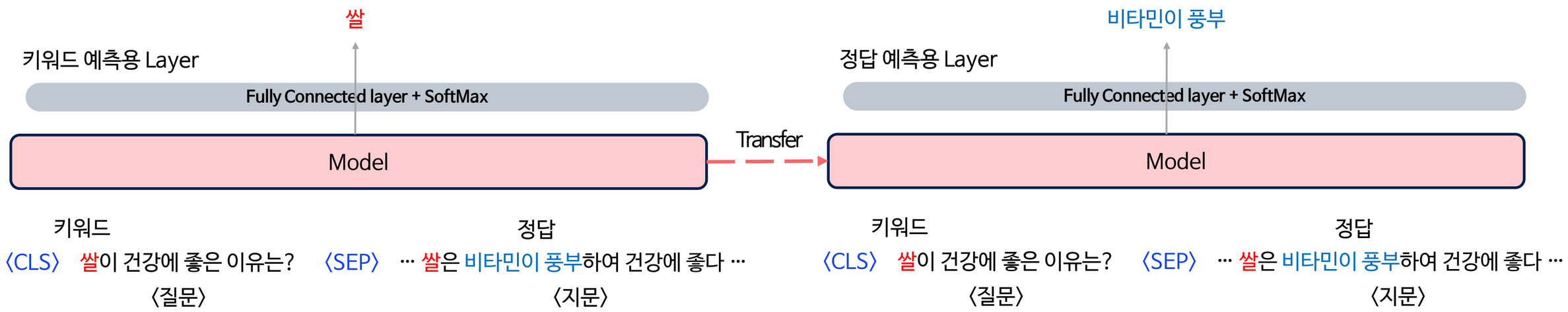
데모 시연

THANK YOU

Q & A

Appendix

Architecture – 3 Keyword Prediction (Multi-Stage)

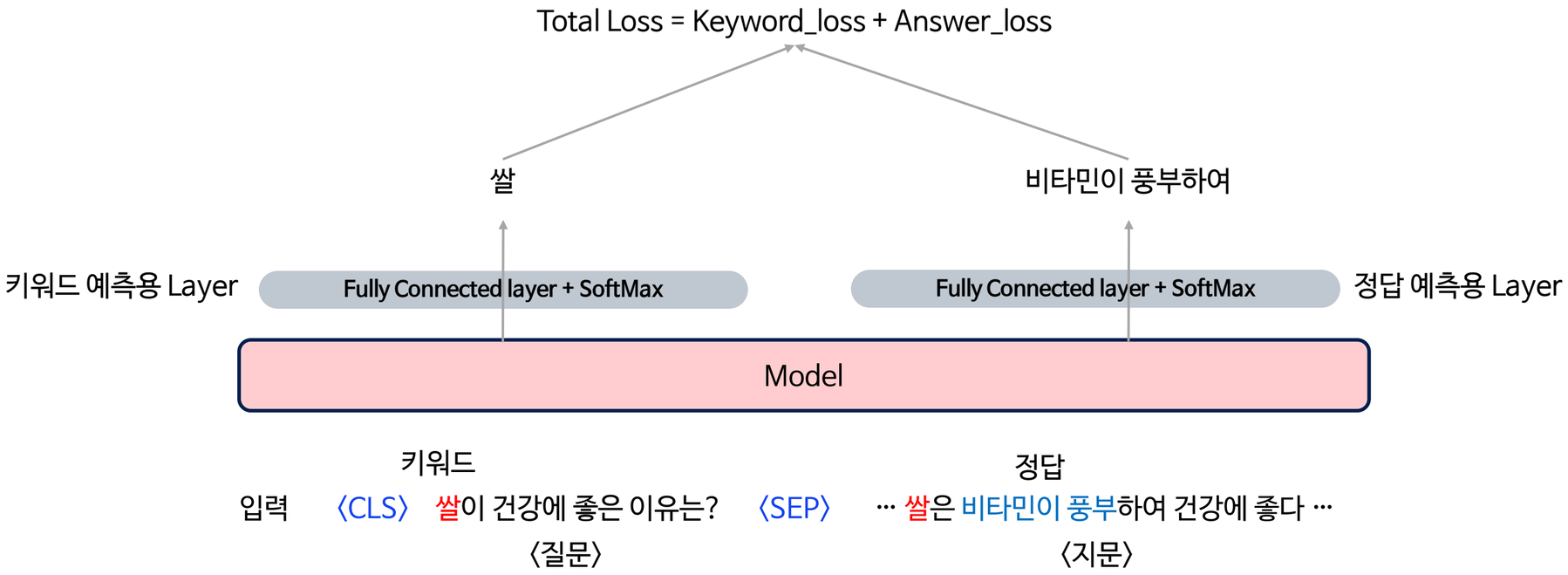


Stage 1

Stage2

Appendix

Architecture – 3 Keyword Prediction (Multi-Task)

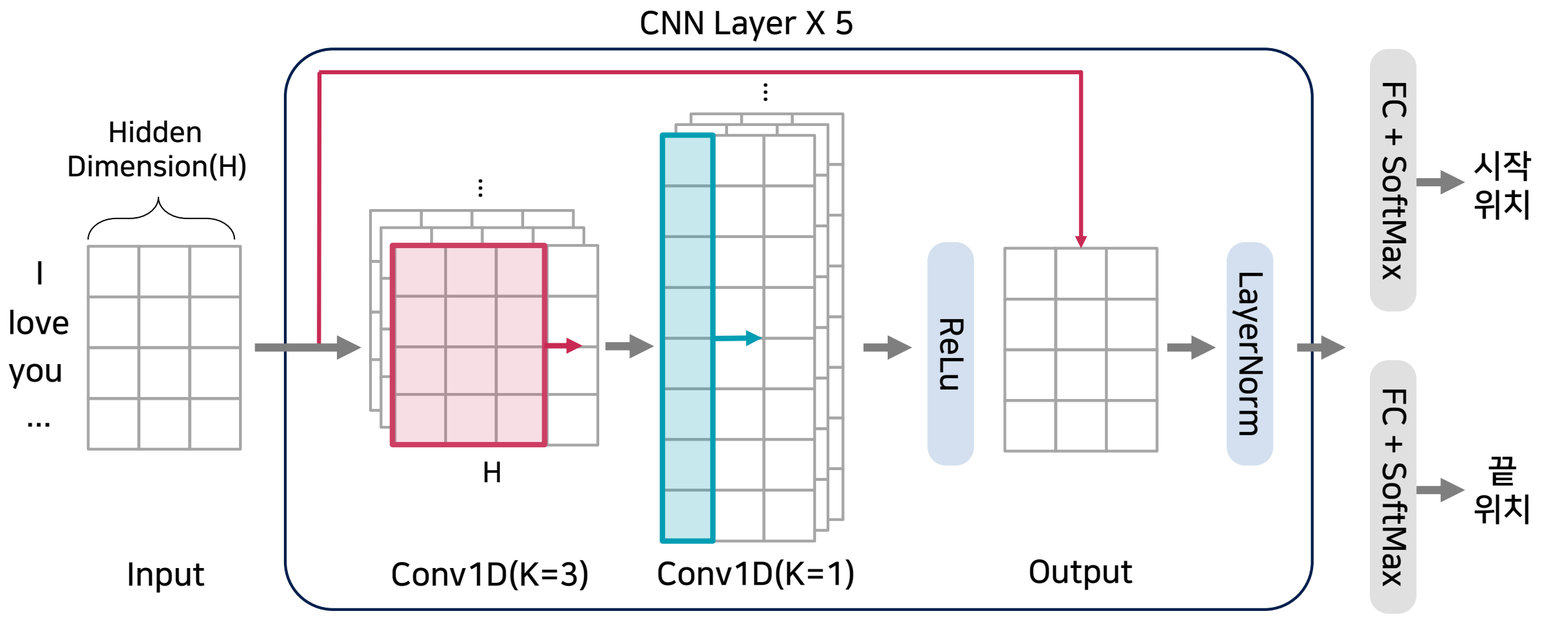


Appendix

Architecture -4 CNN Detail

그림1. CNN 아키텍쳐 디테일

Input Shape을 유지하면서 근접 벡터간 연관 정보를 학습하도록 설계
CNN Layer는 5개층, 1D Convolution 과 ReLu, LayerNorm, Residual Connection 이 적용됨



Appendix
Dataset Detail

표1. 사전학습 용 생성 데이터 건수

Method	Stage	Train	Validation	Test
지도 학습	QA Fine-tuning	78859	9822	9168
	Keyword-prediction	78859	9822	-
자기지도 학습	Sentence Order prediction	38870	4889	
	Sentence Order prediction 5x	194350	4889	-
	Sentence Coherence prediction	38870	4889	-
	Sentence Coherence prediction 5x	194350	4889	-

Appendix
Results Detail

표2. 학습 방법별 결과 비교

Good

Method	Dev		Test		Dev		Test		Avg.
	F1	EM	F1	EM	F1	EM	F1	EM	
Baseline (No Pre-training)	83.73	44.42	83.32	45.60	0.00	0.00	0.00	0.00	0.00
SOP	84.36	44.83	85.90	49.59	0.63	0.41	2.59	3.98	1.90
SOP + Augment	84.16	44.83	83.98	47.48	0.44	0.41	0.67	1.88	0.85
SOP + CNN	84.53	44.31	85.53	48.20	0.80	-0.11	2.21	2.60	1.38
SCP	84.30	45.09	84.65	47.65	0.57	0.67	1.33	2.05	1.16
SCP + Augment + CNN	84.71	45.51	85.94	49.52	0.98	1.09	2.63	3.92	**2.15**
SCP + CNN	84.67	44.90	84.77	48.18	0.95	0.48	1.46	2.57	1.36
KP (Multi Stage)	84.19	44.35	84.92	47.84	0.47	-0.07	1.61	2.24	1.06
KP (Multi-Task)	84.04	45.17	84.27	47.36	0.31	0.75	0.95	1.76	0.94
KP (Multi-Stage) + CNN	84.74	45.04	85.98	49.42	1.01	0.62	2.67	3.82	2.03