Facts as Experts에 기반한 한국어 지식 기반 질의응답

강동찬O[1], 나승훈[1], 최윤수[2], 장두성[2]

[1]전북대학교, [2] KT

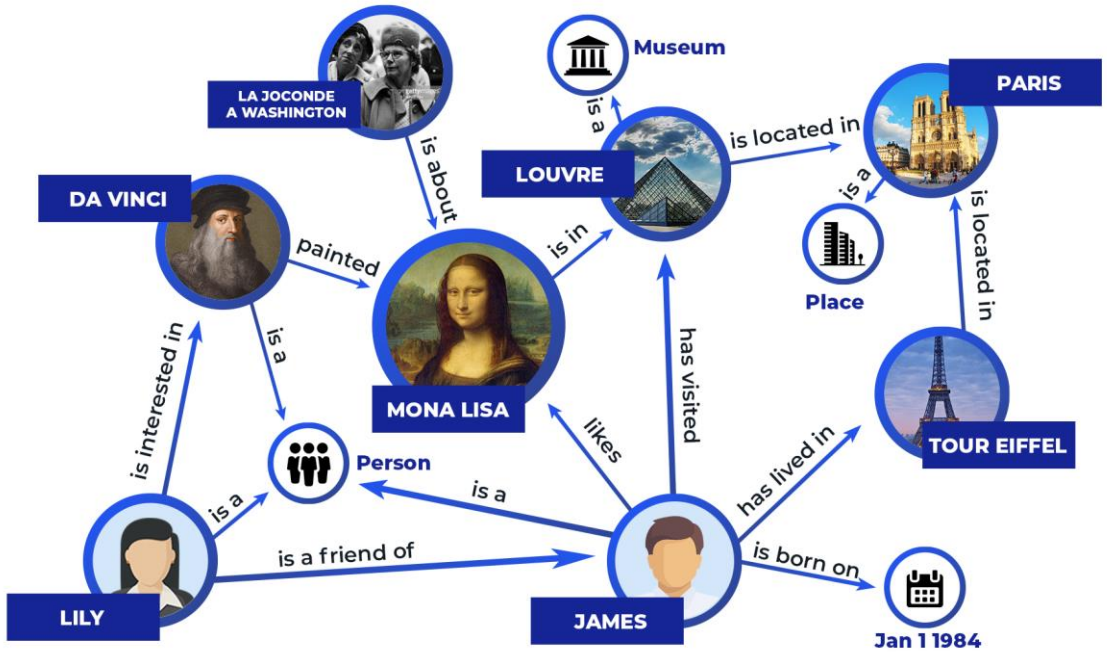
kdc1430@gmail.com, nash@jbnu.ac.kr, {yunsu.choi, dschang}@kt.com

Introduction

- Knowledge-base Question Answering (or KGQA)

Q: 모나리자를 그린 화가는 누구죠 ? →

A: 레오나르도 다빈치 ←

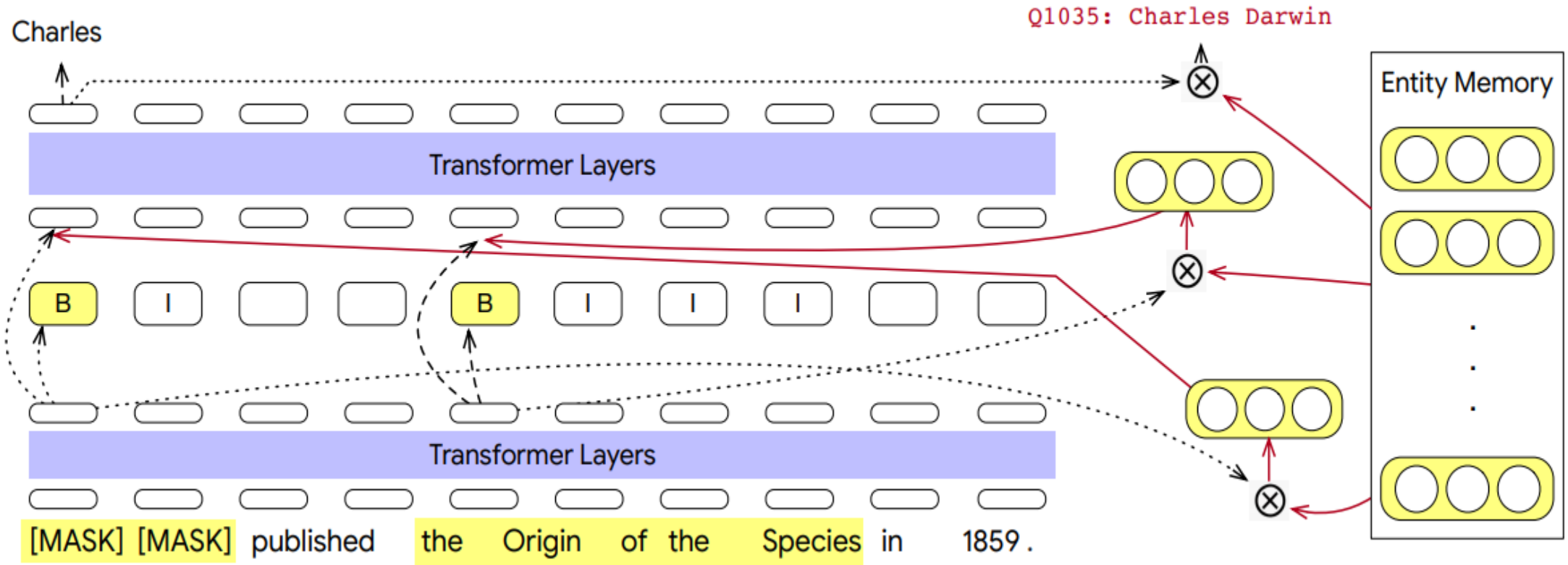


Introduction

- knowledge base vs text
  - 장점
    - 언어 독립적
    - 학습과 평가 간에 정확성, 정보의 정확성
    - 텍스트에 비해 더 가벼움
  - 단점
    - 불완전성

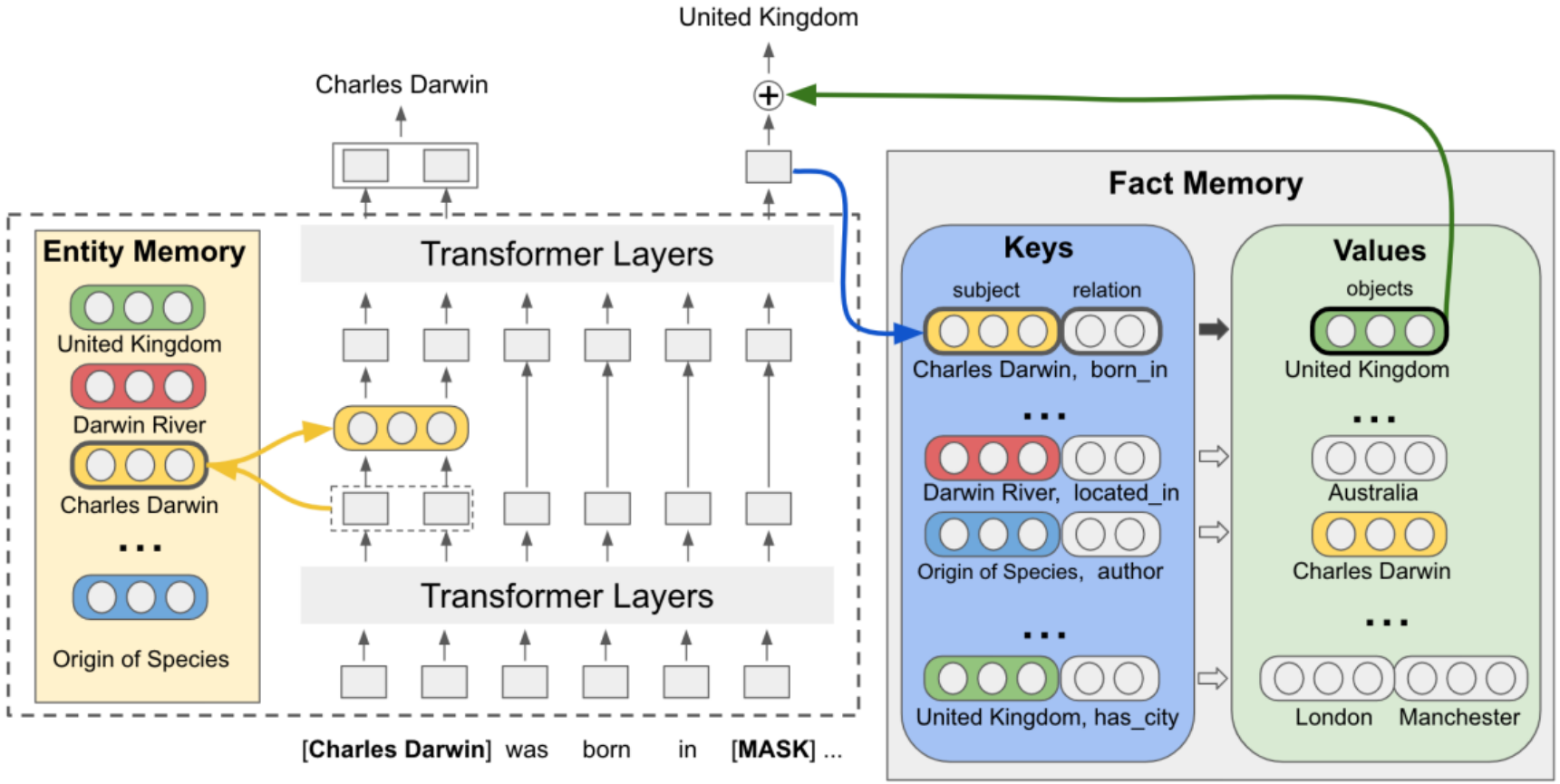
Related Works

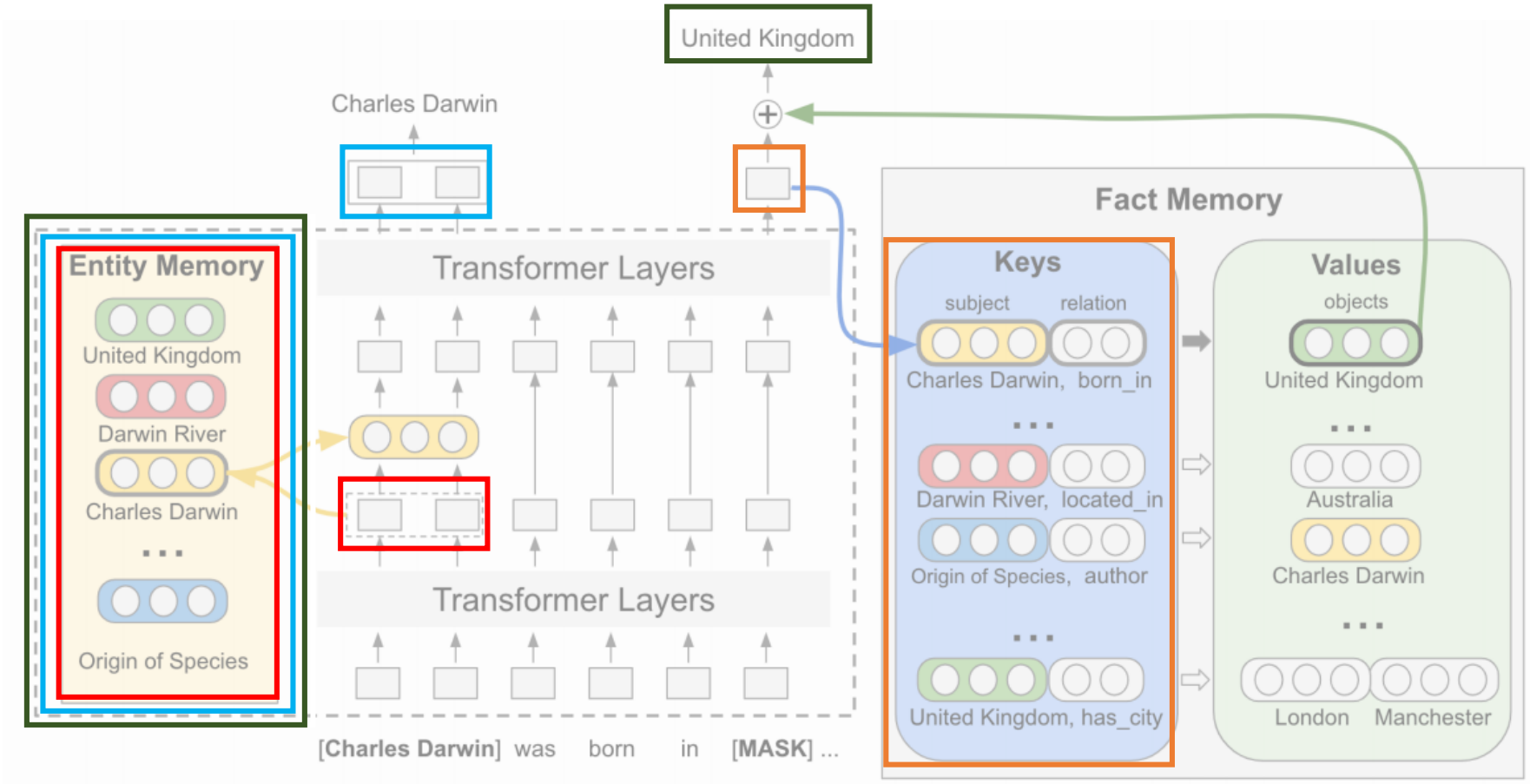
- Entities as experts: Sparse memory access with entity supervision [Fevry et al. 2020]



Related Works

- Facts as experts: Adaptable and interpretable neural memory over symbolic knowl-edge [Verga et al. 2020]



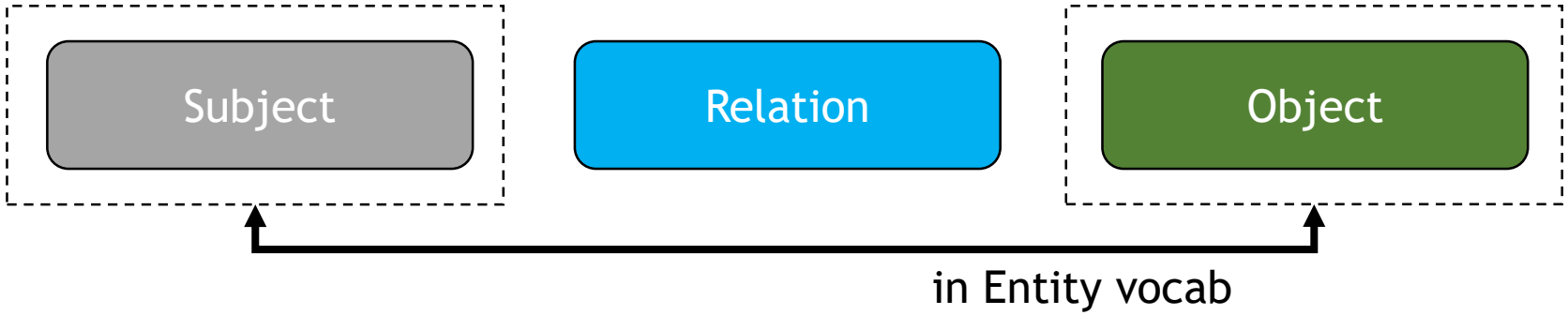


$$loss_{pretrain} = loss_{ent} + loss_{ctx} + loss_{fact} + loss_{ans}$$
$$loss_{finetune} = loss_{fact} + loss_{ans}$$

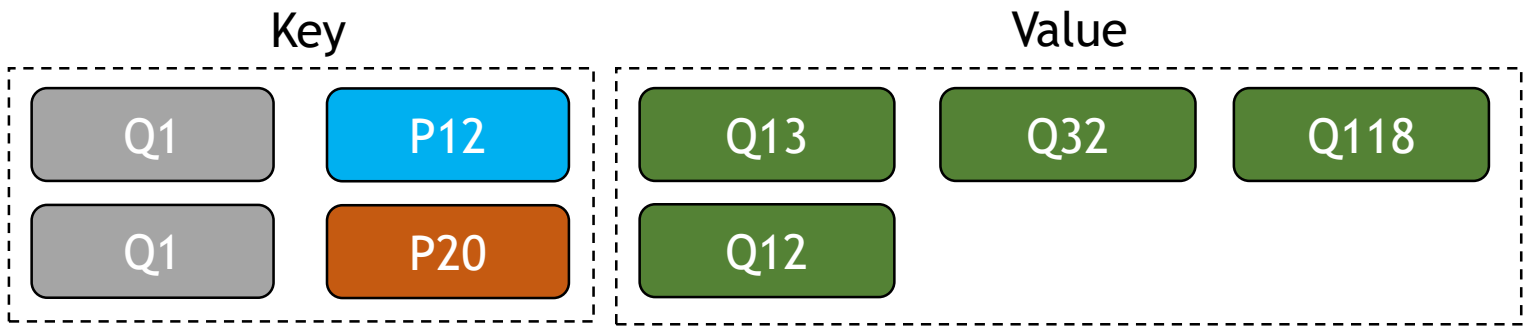
Experiment

- **Data**
  - **Pretraining Corpus**
    - 2020-05-01 위키피디아 기사 덤프
  - **Knowledge base**
    - 2020-10 위키데이터 덤프
  - **Entity vocab**
    - 한국어 위키피디아 기사 텍스트에서 한번이라도 링킹 된 약 40만 개의 엔티티 어휘를 모두 사용하였음.

Knowledge Base



- 트리플은 전체 트리플 중 Subject 와 Object 모두 Entity vocab 안에 있는 경우만 취함
- 관계 어휘의 경우 이렇게 수집된 트리플에 나타나는 모든 관계를 바탕으로 생성하였다.
- 이렇게 수집 된 트리플을 바탕으로 (개체1, 관계), (개체2 리스트(최대 길이 32)) 형태의 지식 베이스를 만들었다.



Knowledge Base

	2020-10 Wikidata
Triples	2,221,332
KB key length	1,676,541
Entity vocab size	407,977
Relation vocab size	912

Pretraining data

Paragraph: **매트릭스**는 **1999년** 개봉한 **미국**의 **SF 액션 영화**이다.

Example1: **[MASK]**는 **1999년** 개봉한 **미국**의 **SF 액션 영화**이다.
Example2: **매트릭스**는 **[MASK]** 개봉한 **미국**의 **SF 액션 영화**이다.
Example3: **매트릭스**는 **1999년** 개봉한 **[MASK]**의 **SF 액션 영화**이다.
Example4: **매트릭스**는 **1999년** 개봉한 **미국**의 **[MASK]**이다.

- N개의 멘션이 있는 128 토큰 길이의 단락에서는 N개의 사전학습 예제가 나오는 방식
- 여기서 각 다른 색의 멘션은 **컨텍스트 멘션, 정답 멘션** 이라고 하자
- 만약 KB 안에 <**subject**, relation, **object**>의 트리플이 있다면 1개 이상 있다면 검색 Key의 정답은 <**subject**, relation>, ... 이다. (distant supervision setting)
- 그렇지 않은 경우 검색 key의 정답은 <NULL, NULL> 이다.

Pretraining data

#pretrain paragraph	899,564
#pretrain examples	4,780,377

- 정확히 토큰 경계에 맞는 예제만을 사용하고, 연산 부담을 줄이고 고정메모리를 사용하기 위해 최대 맨션 개수(=10)를 설정 했다.

Model architecture

Initial transformer layers	4
Final transformer layers	8
Entity vocab size	407,977
Entity embedding size	176
Relation vocab size	912
Relation embedding size	96
Parameters	180M

Finetuning data

Question: 매트릭스는 언제 개봉했어 ? [MASK]

Finetuning data

	Train	Dev	Test
Neural KBQA	3749	242	428

KB-answerable	Train	Dev	Test
2020-10 Wikidata	3588	234	415
2019-04 Wikidata	3626	239	419

- 뉴로 심볼릭 관계 모델을 이용한 지식베이스 질의 응답 [이영훈 et al '20] 에서 사용한 Neural KBQA 데이터 셋
- 지식베이스의 경우 같은 논문에서 사용했던 지식베이스(**2019-04 Wikidata**) 또한 파인튜닝에 사용하여 실험을 진행

Finetuning data

	2020-10 Wikidata	2019-04 Wikidata
Triples	2,221,332	1,764,200
KB keys	1,676,540	1,337,510

KB-answerable	Train	Dev	Test
2020-10 Wikidata	3588	234	415
2019-04 Wikidata	**3626**	**239**	**419**

- 예전의 지식 베이스는 최근의 지식 베이스에 비해 80% 트리플 만을 가지고 있음에도 불구하고, KT 질의 응답 데이터 셋이 생성되었을 당시와 더 가까워 상대적으로 과거의 트리플을 많이 가지고 있음.

Finetuning Result

모델	Full	Answerable
Neural KBQA(+TransE)[이영훈 et al]	**83.17**	83.17
FAE + 2020-10-KB	81.30	83.85
FAE + 2019-04-KB	82.00	**84.57**

Finetuning Result

$$loss = loss_{fact} + loss_{ans}$$ (1)원논문

$$loss = \boldsymbol{\alpha} * loss_{fact} + (1 - \boldsymbol{\alpha}) * loss_{ans}$$ (2)제안

$\boldsymbol{\alpha}$: null probabilty

모델	Full	Answerable
FAE + 2020-10-KB + (1)	79.67	82.16
FAE + 2020-10-KB + (2)	**81.30**	**83.85**
FAE + 2019-04-KB + (1)	81.50	83.21
FAE + 2019-04-KB + (2)	**82.00**	**84.57**

Conclusion & Future work

- 본 연구에서는 FAE모델을 한국어 도메인에서 사전학습 및 파인튜닝하여 성능을 도출하였다. 또한 최근과 이전의 지식베이스를 이용하는 설정에서, 파인튜닝 시에 사용하는 손실함수를 보완하여 성능 개선 효과를 확인하고 이전연구와 비교하였다.
- 후속 연구에서는 추가적인 데이터와 함께 FAE의 효율적인 사전학습과 멀티홉 질의응답 데이터셋에서의 적용을 연구하고자 한다.