KOREA UNIVERSITY

DSBA CS224n 2021 Study

[Lecture 15]

Integrating knowledge in Language Models

DSBA
Data Science & Business Analytics

고려대학교 산업경영공학과

Data Science & Business Analytics Lab

발표자 : 이유경

00

발표 목차

Contents

1. Introduction
2. Techniques to add knowledge to LMs
   - Add pretrained entity embeddings
   - Use an external memory
   - Modify the training data
3. Evaluating knowledge in LMs

01 Introduction

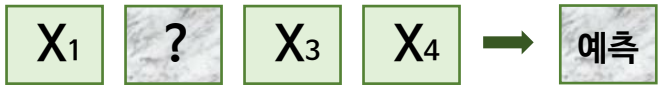
Recap : LMs

Transformer to T5 (20.5.25)
reference from XLNet

Pre-training의 대표적인 Objective

Auto Encoding

BERT는 Denoising AE라 볼 수 있음



Word sequence

$$\bar{\mathrm{x}} = [x_1, x_2, \cdots, x_T]$$

corrupted sequence

$$\hat{\mathrm{x}} = [x_1, [MASK], \cdots, x_T]$$

likelihood

$$p(\bar{\mathrm{x}}|\hat{\mathrm{x}}) \approx \prod_{t=1}^{T} p(x_t|\hat{\mathrm{x}})$$

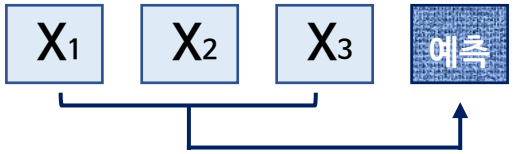
Objective function

$$Max_\theta \; logp_\theta(\bar{\mathrm{x}}|\hat{\mathrm{x}})$$

$$\approx \sum_{t=1}^{T} m_t logp_\theta(x_t|\hat{\mathrm{x}})$$

$$= \sum_{t=1}^{T} m_t log \frac{exp(H_\theta(\hat{\mathrm{x}})_t^T e(x_t))}{exp(H_\theta(\sum_{x'} exp(H_\theta(\hat{\mathrm{x}})_t^T e(x'))}$$

Auto Regressive



Word sequence

$$\mathrm{x} = [x_1, x_2, \cdots, x_T]$$

likelihood

$$p(\mathrm{x}) = \prod_{t=1}^{T} p(x_t|\mathrm{x}_{<t})$$

Objective function

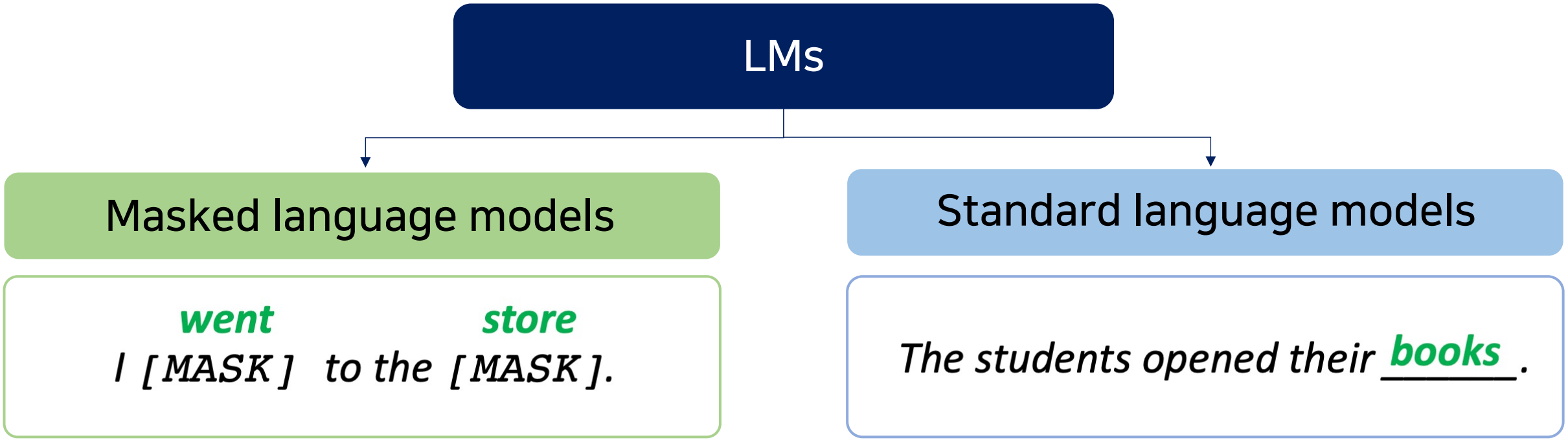
$$Max_\theta \; logp_\theta(\mathrm{x})$$

$$= \sum_{t=1}^{T} logp_\theta(x_t|\mathrm{x}_{<t})$$

$$= \sum_{t=1}^{T} log \frac{exp(h_\theta(\mathrm{x}_{1:t-1})^T e(x_t))}{exp(h_\theta(\sum_{x'} exp(h_\theta(\mathrm{x}_{1:t-1})^T e(x'))}$$

01 Introduction

Recap : LMs



01 Introduction

What does a language model know ?

Predictions generally make sense, but are not all factually correct

- Why might this happen?
  - Unseen facts: some facts may not have occurred in the training corpora at all
  - Rare facts: LM hasn't seen enough examples during training to memorize the fact
  - Model sensitivity: LM may have seen the fact during training, but is sensitive to the phrasing of the prompt
    - Correctly answers "x was *made* in y" templates but not "x was *created* in y"

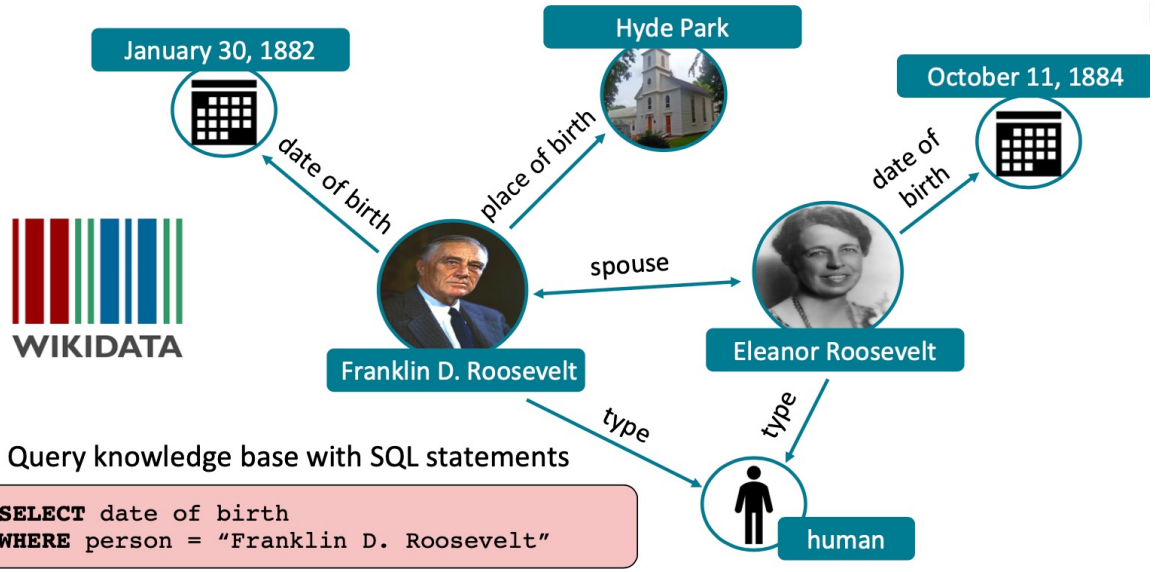
factually correct한 정보를 정확히 담기위해 LM을 다양한 방법으로 개선할 수 있음

01 Introduction

What does a language model know ?

factually correct한 정보를 정확히 담기위해 LM을 다양한 방법으로 개선할 수 있음

Querying traditional knowledge bases



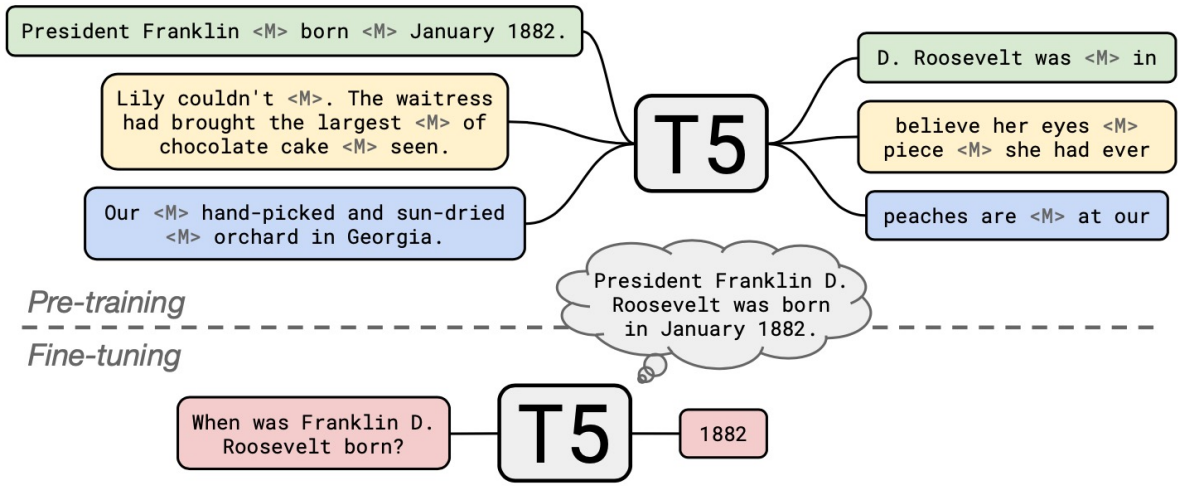
- Query knowledge base with SQL statements

```
SELECT date of birth
WHERE person = "Franklin D. Roosevelt"
```

- Manual annotation 필요함
- Complex NLP pipeline 필요함 (structured data)

Querying LMs as knowledge bases

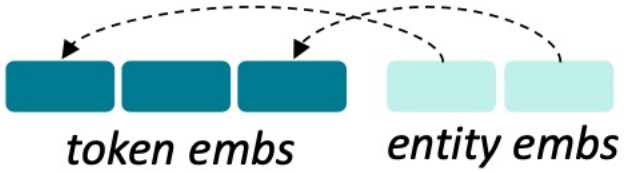
- Pretrain LM over unstructured text and then query with natural language.



- More flexible
- But.. Hard to interpret, trust, modify

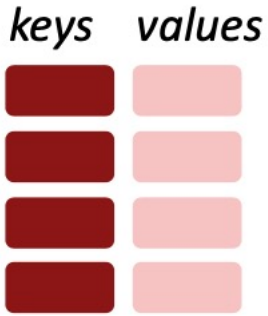
02 Techniques to add knowledge to LMs

Overview



Add pretrained entity embeddings

- ERNIE
- KnowBERT



Use an external memory

- KGLM
- kNN-LM

Modify the training data

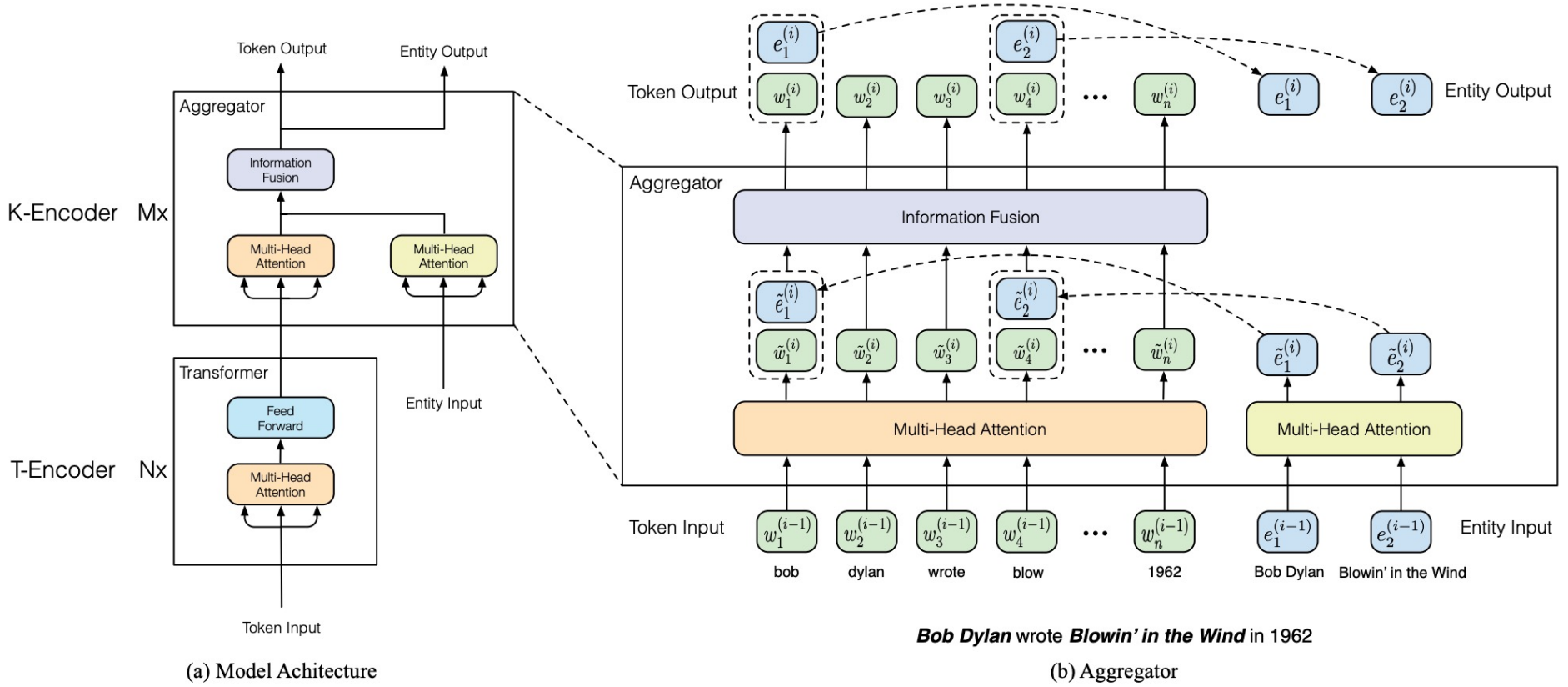
- WKLM
- ERNIE (another!), salient span masking

02 Techniques to add knowledge to LMs

Add pretrained entity embeddings

21년 8월 ERNIE 세미나 진행
[link]

- ERNIE : Enhanced Language Representation with Informative Entities
  - 참고로 ernie는 두가지 논문이 존재하므로 (칭화대, 바이두 ernie) 논문 제목을 정확히 살펴봐야함
  - 1.0, 2.0, 3.0 버전으로 지금까지 업데이트 되는 모델은 바이두 ernie이며, 뒷장에서 설명함



02 Techniques to add knowledge to LMs

Add pretrained entity embeddings

21년 8월 ERNIE 세미나 진행
[link]

- ERNIE : Enhanced Language Representation with Informative Entities
  - Contribution
    - Structured Knowledge Encoding
      - ✓ 텍스트 안에 있는 Named entity mentions과 KG의 entity align하여 정보 획득
      - ✓ 기존 모델인 TransE를 사용하여 Entity를 위한 embedding 진행
      - ✓ Entity representation를 기존 모델에 적절히 통합시킴
    - Heterogeneous Information Fusion
      - ✓ BERT Pretraining + Entity를 위한 objective 추가함
      - ✓ 추가 representation을 고려하기 위한 새로운 object를 제안하여 individual vector space를 적절히 fusion함
      - ✓ Entity representation + token representation = Knowledgeable language representation

02 Techniques to add knowledge to LMs

Add pretrained entity embeddings

21년 8월 ERNIE 세미나 진행
[link]

- ERNIE : Enhanced Language Representation with Informative Entities
  - model architecture
    - Textual Encoder (T- Encoder)

      ✓ token으로 부터 basic한 lexical, syntactic information를 추출함

    - Knowledgeable Encoder (K-Encoder)

      ✓ extra token-oriented knowledge와 textual information 결합

      ✓ Token에 대한 Heterogeneous information Representation을 생성함

      ✓ 두가지 vector space로부터 만들어지는 Representation을 united feature space로 통합함

02 Techniques to add knowledge to LMs

Add pretrained entity embeddings

21년 8월 ERNIE 세미나 진행
[link]

- ERNIE : Enhanced Language Representation with Informative Entities

Question: How do we incorporate pretrained entity embeddings from a *different embedding space*?

Answer: Learn a fusion layer to combine context and entity information.

$$\boldsymbol{h}_j = F\left(\boldsymbol{W}_t \boldsymbol{w}_j + \boldsymbol{W}_e \boldsymbol{e}_k + b\right)$$

- Text representation과 entity representation은 서로다른 embedding space로부터 생성됨
- ERNIE에서 text는 bert embedding으로, entity는 pretrained entity embedding(TransE)로 추출
- 서로 다른 embedding space를 united feature space로 결합하기 위해 추가적인 layer를 사용함

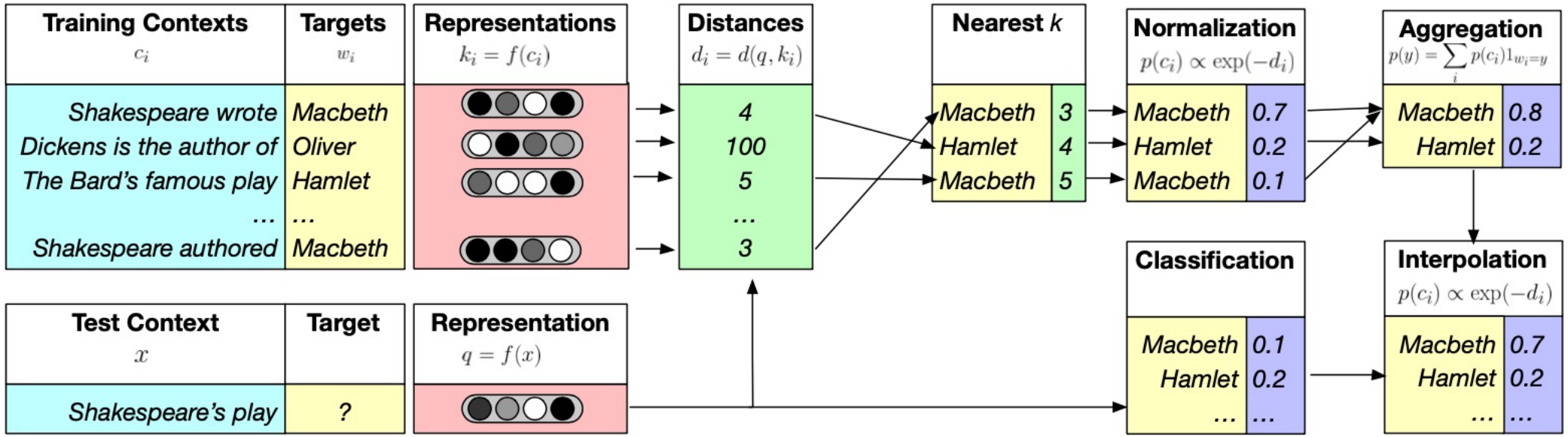
  : affine transformation – non linear transformation – new representation

02 Techniques to add knowledge to LMs

Use an external memory

- KGLM
  - 해당 논문은 좋은 방법론이나, KG의 형태에 의존적이므로 발표에서는 제외
- Nearest Neighbor Language Models (kNN-LM)
  - 의미가 유사한 text를 KNN을 통해 탐색한 후, 단어를 예측하는 방법을 통해 LM 학습
  - kNN prob.과 LM prob.을 함께 결합하여 사용 할 수 있음

$$P(y|x) = \lambda P_{kNN}(y|x) + (1-\lambda)P_{LM}(y|x)$$



02 Techniques to add knowledge to LMs

Modify the training data

- Question: Can knowledge also be incorporated implicitly through the unstructured text?
- Answer: Yes! Mask or corrupt the data to introduce additional training tasks that require factual knowledge.

pretraining 과정에서 자연스럽게 knowledge를 주입해보자

[의의]

direct로 knowledge를 infuse/inject하는 것 보다

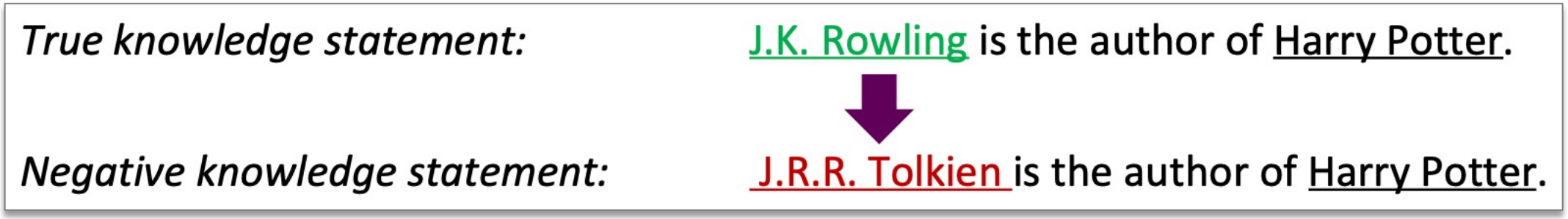
학습과정에서 자연스럽게 knowledge를 배울 수 있음

02 Techniques to add knowledge to LMs

Modify the training data

- **Pretrained Encyclopedia : Weakly Supervised Knowledge Pretrained Language Model (WKLM)**
  - 모델이 true knowledge와 false knowledge를 구분할 수 있도록 학습해보자

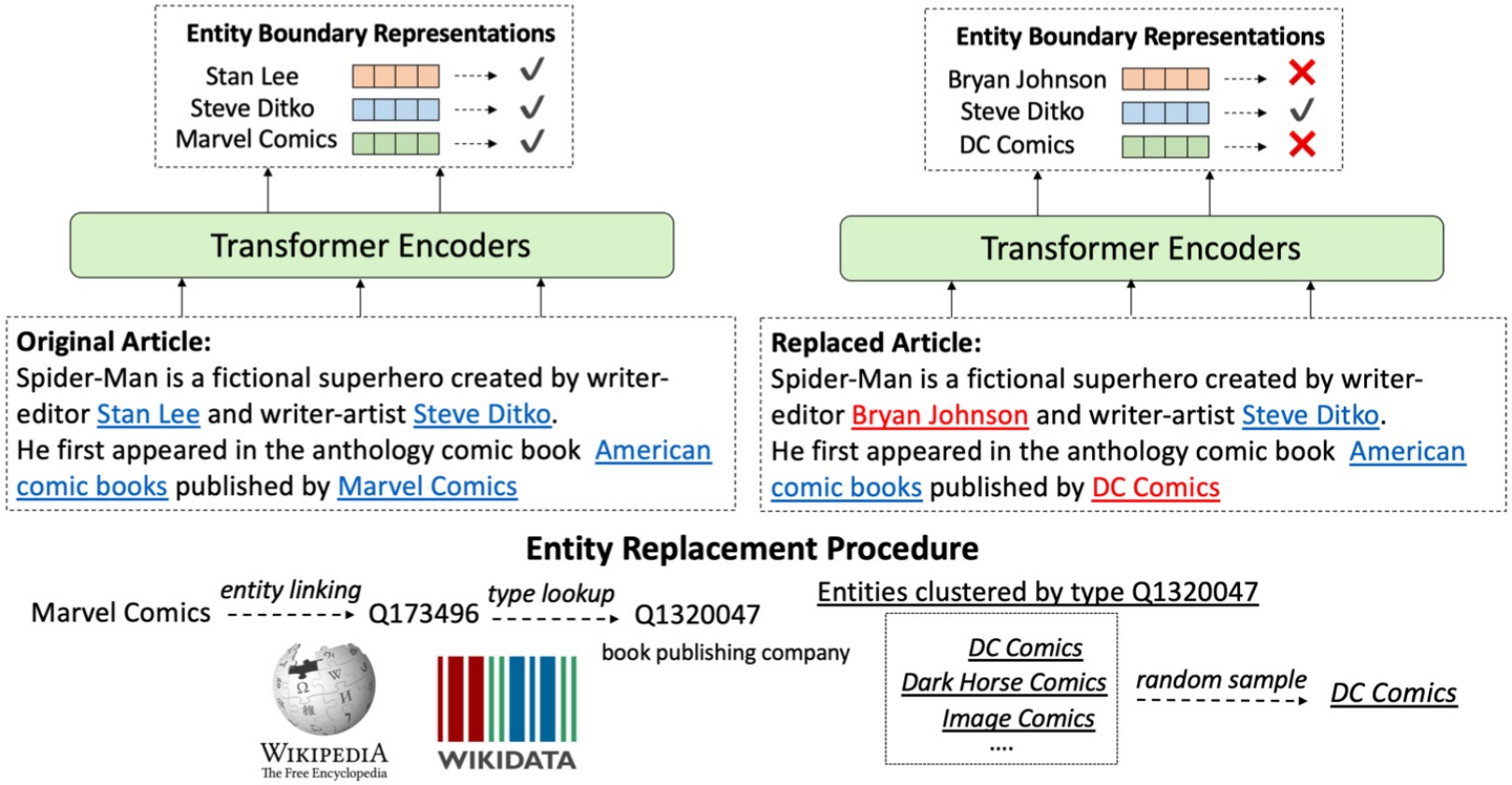
- 기존 데이터를 사용하여 false knowledge를 생성해내야함
  - 특정 entity와 동일한 type의 entity를 활용하여 기존 문장을 변경함
  - 새롭게 만들어진 문장을 negative knowledge statement라 부름



02 Techniques to add knowledge to LMs

Modify the training data

- Pretrained Encyclopedia : Weakly Supervised Knowledge Pretrained Language Model (WKLM)
  - 모델이 true knowledge와 false knowledge를 구분할 수 있도록 학습해보자



02 Techniques to add knowledge to LMs

Modify the training data

- **Pretrained Encyclopedia : Weakly Supervised Knowledge Pretrained Language Model (WKLM)**
  - 모델이 true knowledge와 false knowledge를 구분할 수 있도록 학습해보자
  - Entity replacement loss

$$\mathcal{L}_{entRep} = \mathbb{I}_{e \in \mathcal{E}^+} \log P(e \mid C) + (1 - \mathbb{I}_{e \in \mathcal{E}^+}) \log(1 - P(e \mid C))$$

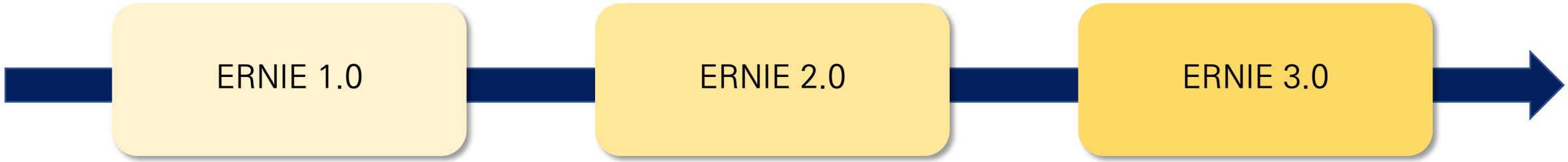
  - Total loss

$$\mathcal{L}_{WKLM} = \mathcal{L}_{MLM} + \mathcal{L}_{entRep}$$

02 Techniques to add knowledge to LMs

Modify the training data

- **ERNIE : Enhanced Representation through Knowledge Integration**
  - 바이두에서 발표한 ERNIE
  - 최근 3.0을 발표하며 지속적으로 발전하는 중



- Knowledge masking strategy로 정보 주입
- 중국어에 대해서 실험

- 여러가지 TASK를 학습하는 Continual Multitask training framework 제안
- 영어 모델도 발표함

- Universal Representation
- Task specific Representation
- Long sequence text 다룸

02 Techniques to add knowledge to LMs

Modify the training data

- **ERNIE : Enhanced Representation through Knowledge Integration**
  - 바이두에서 발표한 ERNIE
  - 최근 3.0을 발표하며 지속적으로 발전하는 중

- **두 가지 ERNIE의 차이점**
  - 칭화대 ERNIE

    : entity 정보를 주입하기위해 pretrained entity embedding을 사용함
  - 바이두 ERNIE

    : entity 정보를 주입하기 위해 별도의 entity embedding을 사용하지 않고 masking 방법을 통해 information 주입
  - 두가지 다 사용해본 경험상 바이두 ERNIE가 보다 적용 관점에서 유연함

02 Techniques to add knowledge to LMs

Modify the training data

- **ERNIE : Enhanced Representation through Knowledge Integration**
  - 바이두에서 발표한 ERNIE
  - 최근 3.0을 발표하며 지속적으로 발전하는 중
  - Knowledge masking strategy로 사전학습 수행함
  - 크게 basic level, entity level, phrase level masking이 존재함
    - Basic level Masking : 15%의 랜덤 확률로 token을 마스킹함
    - Entity level Masking : named entity에 속하는 token들을 한번에 마스킹함
    - Phrase level Masking : 여러개의 단어를 묶은 phrase들을 한번에 마스킹함
  - 사전 학습 과정에서 highlevel knowledge를 미리 학습하여 기존 LM보다 수준 높은 representation 생성함

Sentence	Harry	Potter	is	a	series	of	fantasy	novels	written	by	British	author	J.	K.	Rowling
Basic-level Masking	[mask]	Potter	is	a	series	[mask]	fantasy	novels	[mask]	by	British	author	J.	[mask]	Rowling
Entity-level Masking	Harry	Potter	is	a	series	[mask]	fantasy	novels	[mask]	by	British	author	[mask]	[mask]	[mask]
Phrase-level Masking	Harry	Potter	is	[mask]	[mask]	[mask]	fantasy	novels	[mask]	by	British	author	[mask]	[mask]	[mask]

Figure 2: Different masking level of a sentence

02 Techniques to add knowledge to LMs

Modify the training data

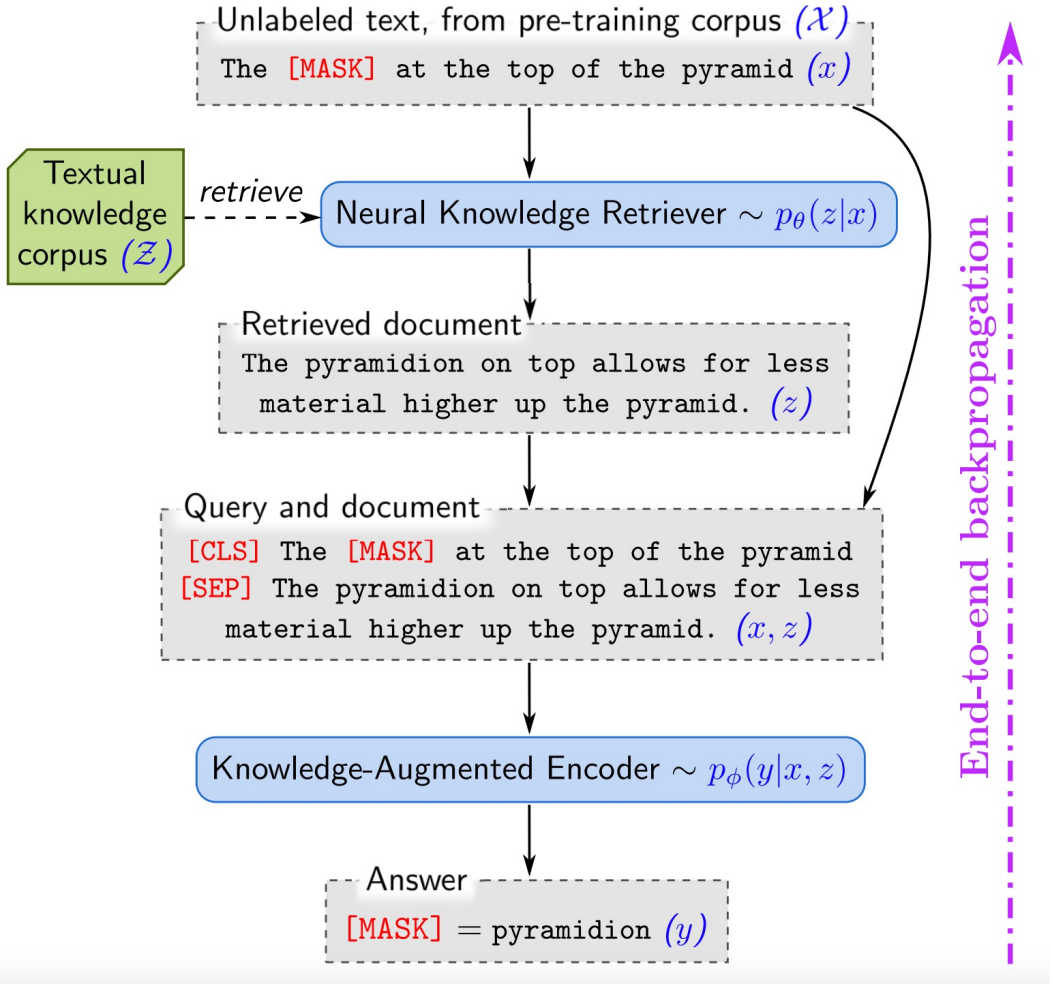
- **Background**
  - Bert 이후 masking strategy를 변경한 다양한 모델이 제안됨
  - 그 중에서도 span 단위로 masking을 수행하는 모델 등장

- **Salient span masking**
  - Span masking은 대부분 비슷한 방법론을 취하고 있음 (SpanBERT, MASS, BART, Pegasus 등)
  - Named entity를 사용하여 salient spans 생성 (such as “United Kingdom” or “July 1969”)
  - 간단하지만 entity 정보를 확실히 학습할 수 있는 좋은 방법론
  - QA task에서 두각을 드러냄

02 Techniques to add knowledge to LMs

Modify the training data

- REALM : Retrieval-Augmented Language Model Pre-Training.



- TL;DR
  - Retriever도 학습하면 QA 성능이 매우 높아짐
  - Retriever 와 reader를 한번에 학습(Joint training)할 수 있음
  - Retriever를 Pretraining에서 수행하는 모델 제안

- Main Contribution
  - Retriever와 Reader를 한번에 학습하는 E2E 모델
  - Query을 넣어(input), 답(output)을 찾는 과정을 두 단계로 분리
    - Neural Knowledge Retriever
      - Query -> Query 의 답이 될만한 document를 찾음
    - Knowledge-Augmented Encoder
      - Retrieved Document -> Answer
  - Pretraining과 Finetuning(ODQA)을 모두 진행함

02 Techniques to add knowledge to LMs

Modify the training data

- REALM : Retrieval-Augmented Language Model Pre-Training.

1. **Pretrained LM의 능력과 한계**
   - PLM은 Pretrain 단계에서 이미 large corpora로 학습되므로 대량의 정보를 포함하고 있음
   - 대부분의 PLM은 Cloze task로 학습을 진행하기 때문에 Mask를 예측하는 과정에서 언어를 이해할 뿐만 아니라 정보를 습득함
   - 하지만, PLM이 정보를 저장하는 방식은 **"implicitly"** 함
     - Network에 어떤 knowledge가 학습되어 있는지 알 수 없음
     - 더 많은 knowledge를 학습하기 위해서는 model size를 증가 시켜야 하며, 계산 비용이 상당함

2. **Explicit하게 Knowledge를 학습 및 저장하는 모델 필요**
   - Textual knowledge retriever를 통해 기존 PLM을 보다 해석 가능하고 explicit하게 knowledge를 학습하는 모델로 개선
   - 즉, Retriever 과정이 pretraining에 포함되어있는 형태임
   - 문장 -> Retriever -> 정답을 찾아낼 수 있는 새로운 모델 구조를 제안함

02 Techniques to add knowledge to LMs

Modify the training data

- REALM : Retrieval-Augmented Language Model Pre-Training.

- Previous methods incorporated knowledge explicitly through pretrained embeddings and/or an external memory.
- Question: Can knowledge also be incorporated implicitly through the unstructured text?
- Answer: Yes! Mask or corrupt the data to introduce additional training tasks that require factual knowledge.

주장은.. 하기 나름 ..

02 Techniques to add knowledge to LMs

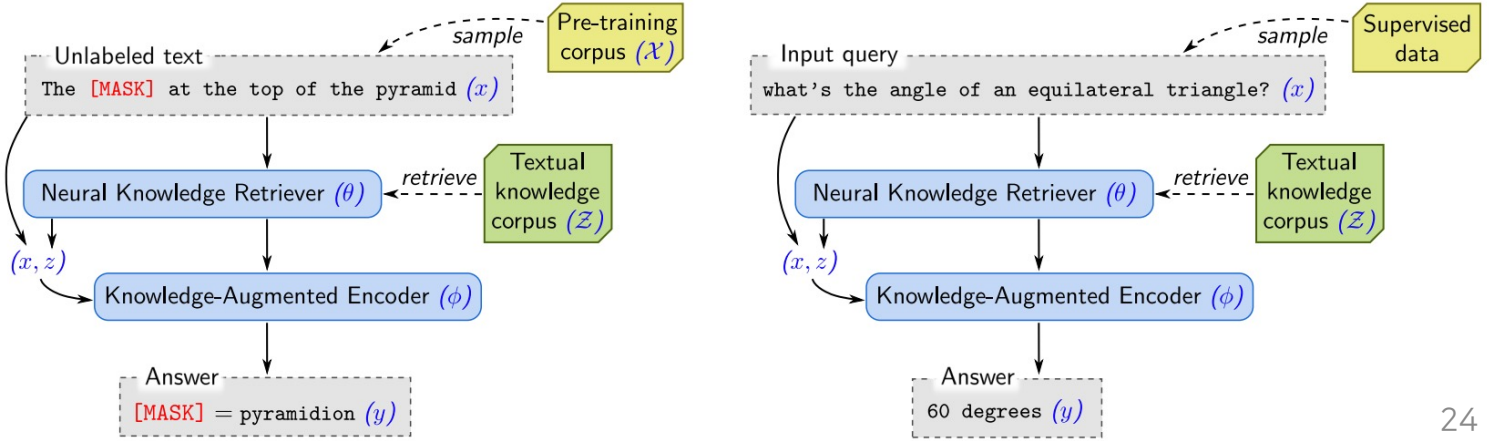
Modify the training data

- **REALM : Retrieval-Augmented Language Model Pre-Training.**
- **Main Idea**
  - Original QA : Query(x)를 넣어 Answer(y)를 찾겠어
  - REAML
    - Step 1 : Query(x)를 넣어 Retrieved document(z)를 찾고
    - Step 2 : Query(x)와 Retrieved document(z)를 넣어 Answer(y)를 찾겠어
- **Model Architecture**
  - Neural knowledge retriever (Step 1)
  - knowledge-augmented encoder (Step 2)
- **Training Process**
  - Unsupervised (Pretraining)
  - Supervised training (Finetuning)
    - QA Task

$$p(y|x) = \sum_{z \in \mathcal{Z}} p(y|z,x)\, p(z|x).$$

Step 2 Step 1

All knowledge corpus



03 Evaluating knowledge in LMs

LAMA

- LAnguage Model Analysis (LAMA) Probe
  - 동일한 setting에서 학습하여 어떤 LM이 가장 많은 정보를 포함하는지 비교함 (RE와 QA에 한정)
  - 하나의 benchmark로서 factual and commonsense knowledge를 probe함
  - Unsupervised BERT가 factual knowledge를 학습한다고 주장함

- Generate cloze statements from KG triples and question-answer pairs
- Compare LMs to supervised relation extraction (RE) and question answering systems
- **Goal:** evaluate knowledge present in existing pretrained LMs (this means they may have different pretraining corpora!)

Mean precision at one (P@1)

BERT struggles on N-to-M relations

Corpus	DrQA	RE baseline	fairseq-fconv	Transformer-XL	ELMo	ELMo (5.5B)	BERT-base	BERT-large
Google-RE	-	7.6	2.6	1.6	2.0	3.0	9.8	**10.5**
T-REx	-	**33.8**	8.9	18.3	4.7	7.1	31.1	32.2
ConceptNet	-	-	3.6	5.7	6.1	6.2	15.6	**19.2**
SQuAD	**37.5**	-	3.6	3.9	1.6	4.3	14.1	17.4

LMs are NOT finetuned!

49

03 Evaluating knowledge in LMs

LAMA

- LAMA-UHN
  - LAMA에서 relational knowledge없이 답변 가능한 예제 모두 제거하고 새로운 데이터 생성
  - LAMA-UHN에는 entity가 있어야 답변 가능한 데이터만 남음
  - 이 경우 LAMA대비 성능이 저하됨
  - BERT가 entity name의 surface form에 지나치게 의존하는것을 보임

프랑스 사람,
Italian-sounding name

Native language of French-speaking actors according to BERT

Person Name	BERT
Jean Marais	French
Daniel Ceccaldi	Italian
Orane Demazis	Albanian
Sylvia Lopez	Spanish
Annick Alane	English

03 Evaluating knowledge in LMs

Developing better prompts to query knowledge in LMs

- **prompt and performance**
  - prompt 형식에 따라 성능이 많이 달라짐
  - LM은 input query에 extremely sensitive함

ID	Modifications	Acc. Gain
P413	x plays in→at y position	+23.2
P495	x was created→made in y	+10.8
P495	x was→is created in y	+10.0
P361	x is a part of y	+2.7
P413	x plays ~~in~~ y position	+2.2

감사합니다