Integrating knowledge in Language Model

Language Model
- Masked LM — ex. I went [MASK] to the store [MASK].
- Standard LM — The students opened their books.

- predictions generally make sense, but are not all factually correct.
  Due to:
  1) Unseen facts
  2) Rare facts
  3) Model sensitivity

개선 방향
1) Querying traditional knowledge bases
   - 강화된 dataset
   - manual annotation 필요
   - Complex NLP pipeline 필요
2) Querying LMs as knowledge bases
   - More flexible 하지만 hard to interpret, trust, modify
   ⇒ T5

Knowledge를 LM에 주입하는 방법
1) Add pretrained entity embeddings
2) Use an external memory
3) Modify the training data

ERNIE - entity embedding
- 사전 학습된 BERT를 확장
ex. Bob Dylan wrote Blowin' in the Wind in 1962.
에 대한 wiki data 사전학습의 기존 것과 합침
T-Encoder ⇒ K-Encoder

KGLM - KG 형태에 의존적

kNN-LM
- 이미 유사한 text를 kNN통해 탐색하고 단어 예측하는 방법통해 LM학습
- kNN과 LM prob 을 함께 결합해 사용 가능

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

- pretraining 과정에서 자연스럽게 knowledge 주입

WKLM - 모델이 true knowledge와 false knowledge 구별하게끔
- Weakly supervised Knowledge Pretrained LM

BERT 이후 masking strategy 변경한 다양한 모델 제안
그 중에서도 Span 단위로 masking 수행하는 모델 등장

Salient span masking
- Named entity를 사용해 salient span 생성
- 간단하지만 entity 정보 효과적 학습
- QA task에서 두각

REALM : Retrieval-Augmented LM-Training
- masking 방법을 학습해 QA task 성능 개선

LAMA Probe
- 동일 setting에서 학습해 어떤 LM이 가장 많은 정보 포함하는지 비교