CS224N : Lecture 15 - Add Knowledge to Language Models (1)

What does a LM know?

- Predictions generally make sense, but are not all factually correct

    - Why might this happen?

        - Unseen facts : some facts may not have occured 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

- Factually correct한 정보를 정확히 담기위해 LM을 다양한 방법으로 개선할 수 있음

    - Querying traditional knowledge bases

        - Manual annotation needed

        - Complex NLP pipeline needed

    - Querying LMs as knowledge bases

        - More flexible

        - Hard to interpret, trust, modify

Techniques to add knowledge to LMs

Add pretrained entity embeddings

- ERNIE : Enhanced Language Representation with informative Entities

- Contribution

    - Structured Knowledge Encoding

    - Heterogeneous Information Fusion

- model architecture

    - Textual Encoder (T-Encoder)

    - Knowledgeable Encoder (K-Encoder)

(a) Model Achitecture

(b) Aggregator
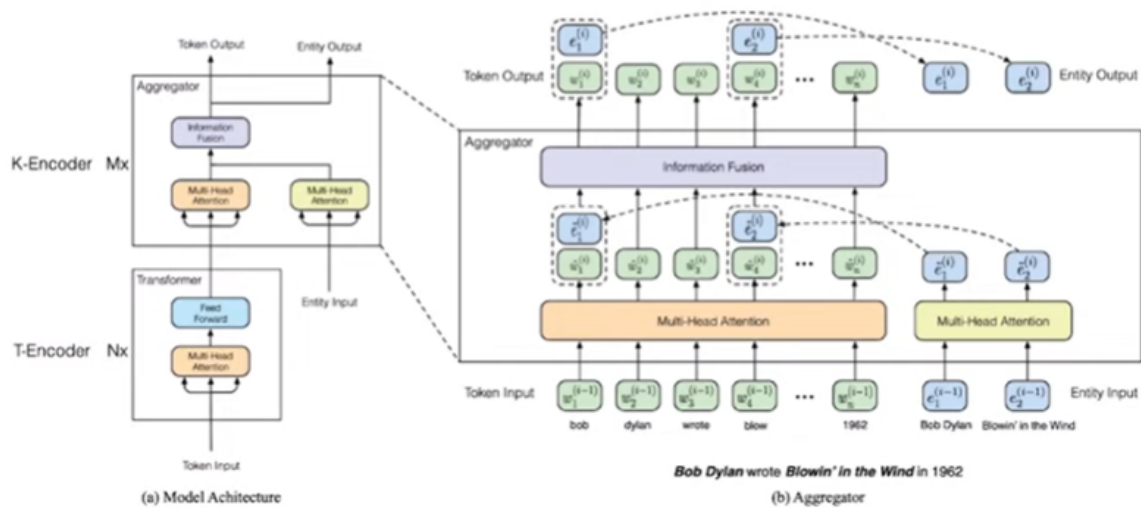
- KnowBERT

Use an external memory

- KGLM

    - KG의 형태에 의존적

- kNN-LM

    - 의미가 유사한 text를 KNN을 통해 탐색 → 단어 예측 : LM 학습

Modify the training data

- WKLM

- ERNIE, salient span masking : Enhanced Representation through Knowledge Integration

- 칭화대 ERNIE vs. 바이두 ERNIE

- Knowledge masking strategy로 사전학습 수행

- basic level, entity level, phrase level masking

- 사전학습 과정에서 highlevel knowledge 미리 학습 → 기존 LM보다 good representation

- Salient span masking

  - Named entity를 사용하여 salient spans 생성

Evaluating knowledge in LMs

- LAMA Probe

  - 동일한 setting에서 학습하여 어떤 LM이 가장 많은 정보를 포함하는지?

  - 하나의 benchmark로서 factual and commonsense knowledge를 probe

- LAMA-UHN

  - LAMA에서 relational knowledge 없이 답변 가능한 예제 모두 제거 → 새로운 데이터 생성 → entity가 있어야 답변 가능한 데이터만 → LAMA 대비 성능 저하

  - BERT가 entity name의 surface form에 too 의존

- promt and performance

  - prompt 형식에 따라 성능 varies

  - LM is extremely sensitive to input query