Overall

현존하는 대부분의 QA dataset들이 거의 몇 가지 뿌리에서 나옴.

기존 pre-trained model + 우리가 풀고자 하는 목적에 부합하는 dataset(에서 기인한 branch)

transfer learning(TANDA method 등) + fine-tuning을 위한 추가 dataset(소량).

MRC models from pre-trained model

https://github.com/yuanbit/FinBERT-QA (BERT)

https://github.com/BioMedBERT/biomedbert (BERT)

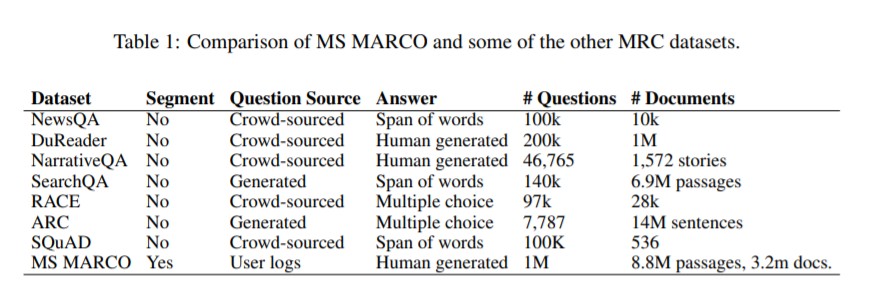
https://github.com/HLTCHKUST/HLTC-MRQA (XLNET)

~ https://github.com/zihangdai/xlnet

~ <https://arxiv.org/abs/1906.08237>

영어 QA dataset

1. 영문 QA dataset 비교 (2016)



+ Hotpot QA (Crowd-sourced, multiple choice); wiki based

2. 요구되는 QA 데이터셋 및 MRC 모델의 특징

**현실적인 데이터셋 (이하 현실성):** 대부분의 기존 데이터셋에는 종종 위키백과 같은 출처의 고품질 스토리나 텍스트 범위가 포함되어 있다. 실제 텍스트는 messy하다. 오타 또는 약어, 음성 인터페이스의 경우 transformation error가 포함될 수 있음. 우리가 사용하는 회의록 MRC 모델은 노이즈가 많고 문제가 많은 현실적인 데이터에서 만든 데이터셋을 사용해야 한다.

**복수 구절 및 문서에서 답변 추출 가능 (이하 multihop):** API 시간 제한 문제 이슈가 해결되지 않는다면, 우리가 사용하는 MRC 모델은 복수의 구절과 문서에서 최종 답변에 대한 정보와 지원 능력이 있는지 확인하는 것이 중요하다.

3. 주요 QA 데이터셋

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 종류 | 예시 | 현실성 | multi-hop | 특징 |
| Extractive MRC | SQuAD | △ | X | 가장 많이 언급되고 사용되는 dataset |
| Generated MRC | **MS MARCO**  **(2016)** | O | △ | Bing 검색데이터. 질문이 실제 사용자 검색 쿼리에서 파생됨. 🡪 사용자가 충족하고자 할 수 있는 "자연스러운" question 분포를 더 잘 나타낼 수 있음. (답변 또한 웹페이지에서 추출) |
| Multi hop MRC | Hotpot QA  (2018) | △ | O | 여러 문서 + 복잡한 추론을 거친 답변 도출 가능 (using hyperlink graph network) |
| Merged form | **Multi QA**  **(2019)** | △ | O | BERT-based RC model: 10개의 데이터셋을 병합한 모델. |

SQuAD

1) wiki based dataset 🡪 답변이 잘 정제된 문어체이므로 현실을 잘 반영하지 못함

2) 질문을 하나의 단락만이 문맥으로 주어지도록 설계 🡪 실제로 대부분의 질문은 해당 단락의 단일 문장과 일치하는 양상을 보임. (Zhilin et al. 2018)

3) 따라서 더 큰 맥락에서 추론할 수 있는 능력은 상대적으로 떨어짐.

MS MARCO (A Human Generated MAchine Reading COmprehension Dataset)

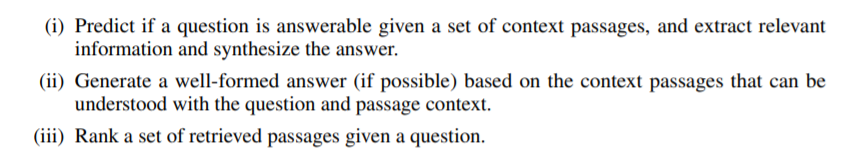
1) Bing의 검색 질의 로그에서 샘플링

2) 1,010,916개의 익명화된 question **+** 3,563,535개의 웹 문서에서 추출한 8,841,823개의 answer **+** 사람이 제대로 생성한 182,669개의 human generated & curated answer

3) MSMARCO 데이터 세트의 질문에는 답이 여러 개 있거나 답이 전혀 없을 수 있음.

4) the MS-MARCO dataset defines the task as answering a question from multiple passages and the words in the answer are not necessary in the passages (Chuanqi Tan, 2018)

+ 논문에서 권장하는 MRC task



<https://arxiv.org/pdf/1611.09268.pdf>

HotpotQA

1) 위키백과 기사

2) 여러 문서에 대한 추론이 필요한 MRC 모델에 필요한 QA 데이터 세트. 🡪 후속 연구 등으로 발전

<https://github.com/hotpotqa/hotpot>

<https://arxiv.org/pdf/1809.09600.pdf>

multi-QA

<https://www.aclweb.org/anthology/P19-1485.pdf>

<https://github.com/alontalmor/multiqa>

영어 자연어 처리 모델

<https://github.com/yuanbit/FinBERT-QA> (BERT)

<https://github.com/BioMedBERT/biomedbert> (BERT)

<https://github.com/HLTCHKUST/HLTC-MRQA> (XLNET)

~ <https://github.com/zihangdai/xlnet>

~ <https://arxiv.org/abs/1906.08237>

<https://arxiv.org/pdf/1911.04118.pdf> (TANDA)

