



# A Survey on MRC : Tasks, Evaluation Metrics and Benchmark Datasets

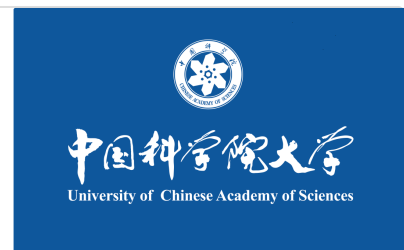
paper

[https://s3-us-west-2.amazonaws.com/secure.notion-static.com/67e34382-18e3-4d69-8150-2b60c7f7e6f1/A\\_Survey\\_on\\_Machine\\_Reading\\_Comprehension\\_Tasks.pdf](https://s3-us-west-2.amazonaws.com/secure.notion-static.com/67e34382-18e3-4d69-8150-2b60c7f7e6f1/A_Survey_on_Machine_Reading_Comprehension_Tasks.pdf)

github

## Machine Reading Comprehension Tasks, Metrics, and Datasets

In this section, we introduce the computation methods of MRC evaluation metrics. Typical evaluation metrics of MRC datasets are: Accuracy, Exact Match, F1 score, ROUGE, BLEU, HEQ and Meteor.  
<https://mrc-datasets.github.io/>



## Abstract

- 57개의 MRC task & dataset 분석 후 4개의 속성(attribute)를 가진 분류 방법 제안
- MRC를 위한 9개의 metric 분석, 7개의 속성과 10개의 특징
- 현안과 앞으로의 연구 방향 제안



## 목차

### Abstract

### 1.Introduction

#### 1.1 Overview

#### 1.2 History

#### 1.3 Motivation

### 2. Tasks

#### 2.1 Definition of Typical MRC task

#### 2.2 Discussion on MRC task

#### 2.3 Classification of MRC tasks

#### 2.4. Definition of each category in the new classification method

#### 2.5. Statistics of MRC Tasks

### 3. Evaluation Metrics

#### 3.2 Accuracy

#### 3.3 Exact Match

#### 3.4 Precision

#### 3.5 Recall

#### 3.6 F1

#### 3.7 ROUGE

#### 3.8 BLEU

### 4. Benchmark Datasets

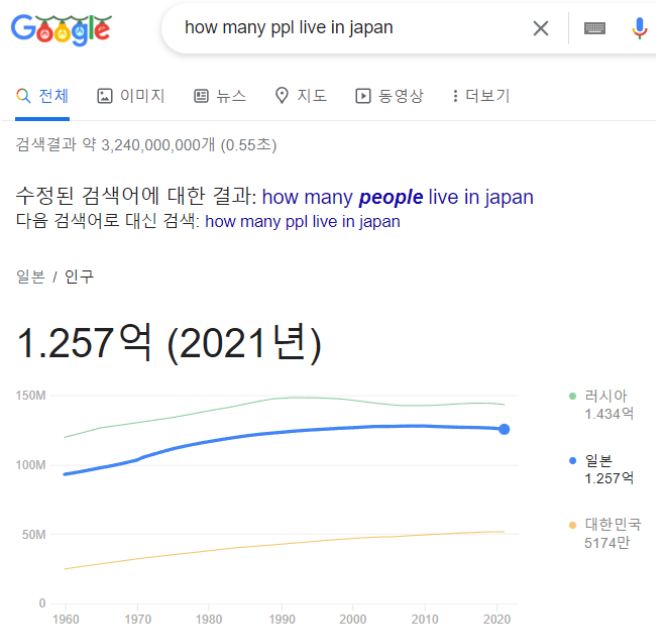
#### 4.10 introducing all datasets

### 5. Open issues

## 1.Introduction

### 1.1 Overview

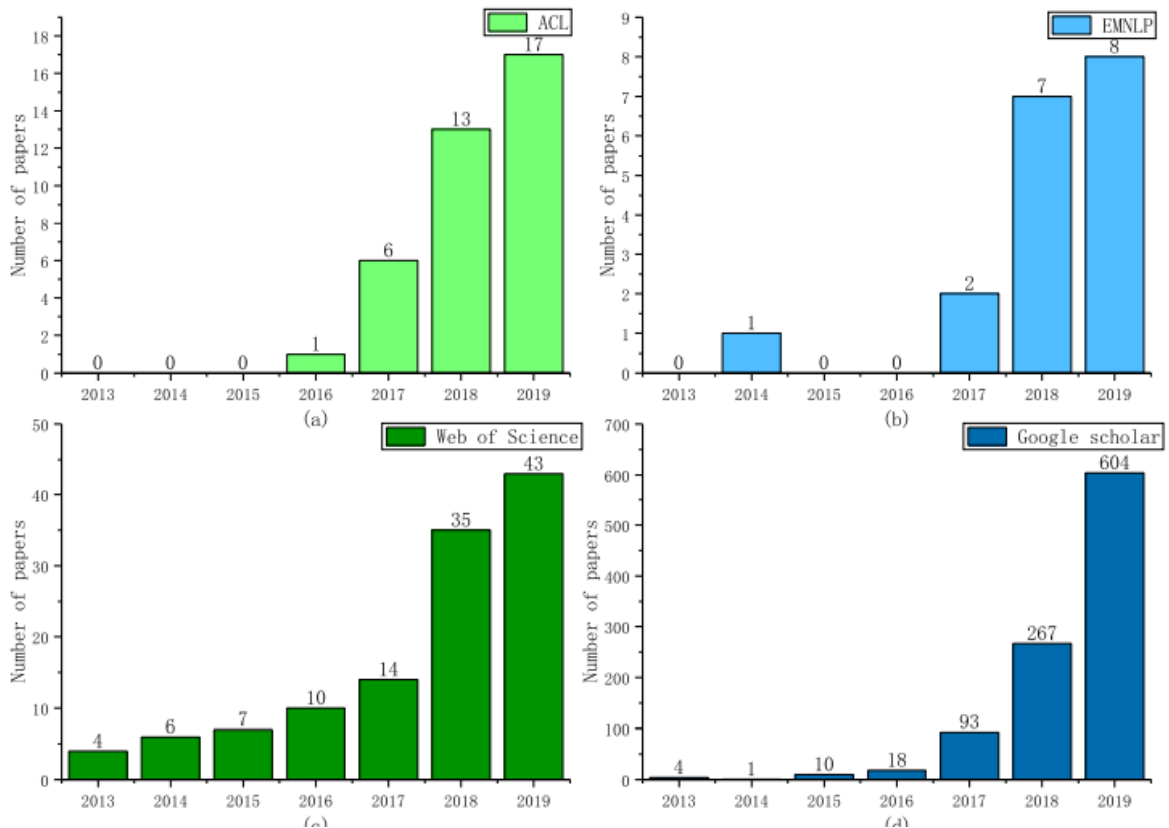
- MRC : 컴퓨터가 텍스트를 읽고 내용을 이해하도록 하는 것
- 사람이 텍스트를 읽고 이해하는 방식을 모사함
  - 읽고 토익 RC문제 푸는 거 생각하면 될듯
  - 이 RC가 그 RC였네...
- 따라서 MRC는 검색엔진(search engine), 대화시스템(dialogue system) 에 사용될 수 있다.
  - Google 검색에 how ~로 시작되는 문장을 넣어도 검색되는 게 이것 때문인 것 같다



## 1.2 History

- 1977년에 QUALM이 가장 초기형이다.
  - 2 story understanding systems
- 1999 Hirschman : corpus w/ 60 dev, 60 test story. 수준은 초등학교 3-6학년
  - 주로 rule-base or statistic 모델, 11 subtask에 30-40%의 정확도를 보임
  - 좋은 데이터셋이 없어서 한동안 발전이 없었음
- 2013년 MCTest dataset
  - 이것도 베이스라인은 rule-base
  - 500 story, 2000 question
  - 이지만 사람들이 ml을 적용하기도
- 2015년 Hermann et al. Teaching machines to read and comprehend
  - dataset 생성하는 방법
  - attention based DNN 적용 모델
  - 이때를 기점으로 많이 연구되기 시작함
- Figure 2 :

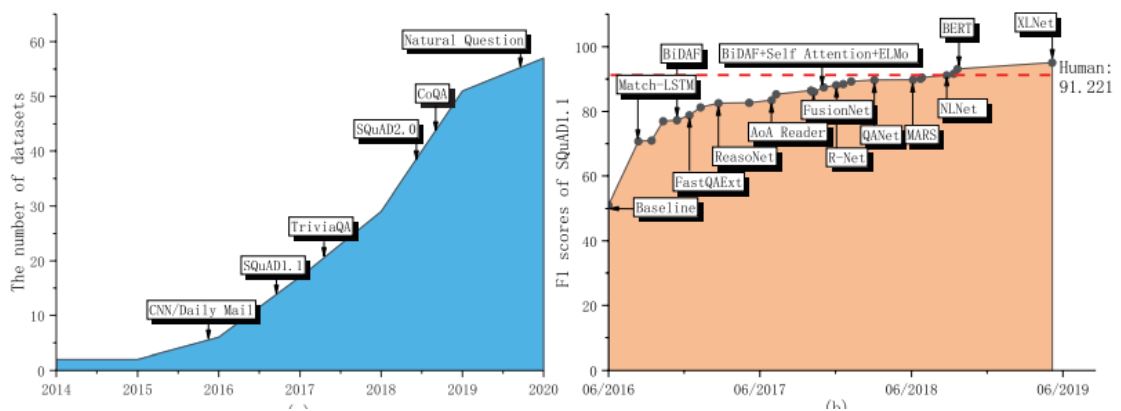
a) ACL, b) EMNLP, c) Web of Science, d) Google scholar 매체별로 발표된 연도별 MRC페이퍼개수



- 2018년부터 페이퍼 갯수 급증(을 위해서는 15-16부터 연구가 활발히 되었겠지?)

### 1.3 Motivation

- Figure 3



a. MRC dataset의 누적 갯수 + 및 대표적인 것들의 발표 시기

- b. SQuAD 1.1에서 SOTA model history
  - 다른 MRC review들과의 차이점
    - 대부분 모델, 성능에 관한 내용인데 우리는 데이터셋 중심
    - task type이 체계적으로 정리되어 있지 않음
- 

## 2. Tasks

### 2.1 Definition of Typical MRC task

- 다양한 형태(multimodal MRC, textual MRC 등) 가 있지만
- 표준은 textual QA-based MRC
  - 정의 : 지도학습 형태

$$\{(p_i, q_i, a_i)\}_{i=1}^n$$

- p: 텍스트 문단, q: p에서 나온 질문, a: p에서 추론한 다른 정보를 포함하지 않는 q의 답

$$a = f(p, q)$$

- 목표는 적절한 f(predictor) 를 찾는 것

### 2.2 Discussion on MRC task

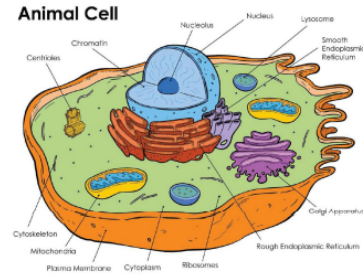
#### 1. Multimodal MRC vs Textual MRC

- Multimodal MRC
  - 예시

**Passage with illustration:**

This diagram shows the anatomy of an Animal cell. Animal Cells have an outer boundary known as the plasma membrane. The nucleus and the organelles of the cell are bound by this membrane. The cell organelles have a vast range of functions to perform like hormone and enzyme production to providing energy for the cells. They are of various sizes and have irregular shapes. Most of the cells size range between 1 and 100 micrometers and are visible only with help of microscope.

**Animal Cell**

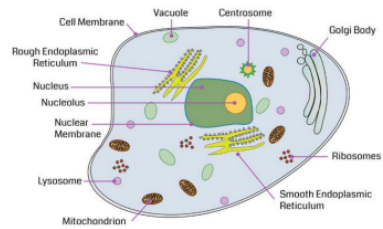


**Question with illustration:**

What is the outer surrounding part of the Nucleus?

**Choices:**

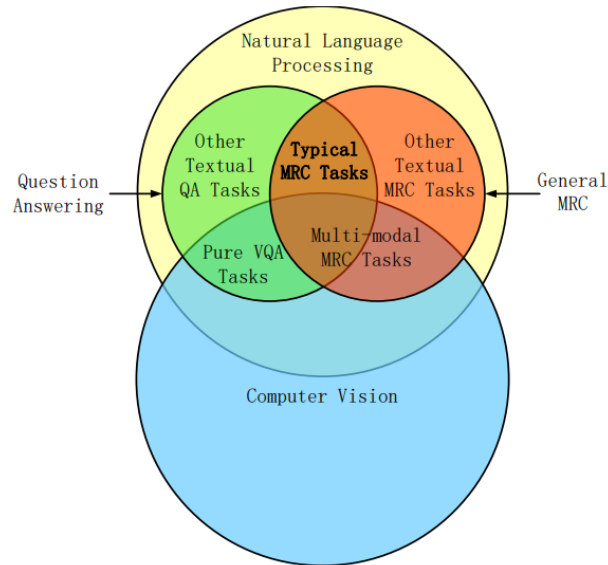
- (1) Nuclear Membrane ✓
- (2) Golgi Body
- (3) Cell Membrane
- (4) Nucleolus



- 사람 대상 연구에서 예시처럼 그림과 글을 같이 읽게 했을 때 상상력이 높은 아이가 상상력이 낮은 아이보다 더 좋은 이해력을 보임
- 따라서 MRC에서도 그림을 이해할 수 있게 하는 게 전체 내용에 대한 이해력을 올리기 위한 중요한 task임.
- dataset list : TQA, MovieQA, COMICS, RecipeQA (위 예시는 TQA)

## 2. MRC vs QA

- MRC는 QA task를 풀기 위한 방법론
  - 물론 반대로 MRC가 QA의 한 장르라고 말하는 사람도 있음
  - 그러나  $MRC \subset QA$  는 아니다
  - QA가 아닌 MRC도 있고, MRC가 아닌 다른 방법으로 QA를 풀 수도 있따
  - **what is difference btw VQA and MMRC?**
  - MRC와 CV의 포함관계 그림



- Chen[36] 은 wikipedia에서 document retrieval을 이용해 범위를 지정해준 다음, MRC로 정답 범위를 찾는 식으로 QA에 접근함
- Hu[37] 은 QA에 접근하는 4가지 방법론(룰베이스, info-retrival, knowledge-based, MRC) 중 하나로 MRC를 생각
- MRC하면 text-QA 형태만 생각하지만, 사실 다양함. Lucy[38] 은 오히려 질문을 잘 하는 게 이해의 척도라고 언급하기도.
- ShARC : 대화형 MRC dataset
  - 다른 데이터셋과 달리, 텍스트에서 직접적으로, 충분히 언급되지 않은 질문을 함. 따라서 모델은 배경지식을 활용해서 원래 질문에 대답하기 위한 두번째 질문을 생성해야 함.

ex> 아이유는 2022 서울가요대상 본상을 수상했다. 아이유는 가수인가?  
 → 서가대는 누구에게 주는 상인가? 그 해 발매한 곡에 대한 상  
 ⇒ 아이유는 (서가대를 타려면) 그 해 곡을 발매했을 것이다 == 아이유는 가수가 맞다
- RecipeQA[35] : 음식 레시피에 대한 멀티모달 MRC 데이터셋
  - ordering task : 제목과 레시피 텍스트가 주어졌을 때, 요리 과정 사진 순서를 배열하는 task
  - 레시피의 순서에 따라 일어나는 이벤트이므로 시퀀스간의 관계를 추론해야 함

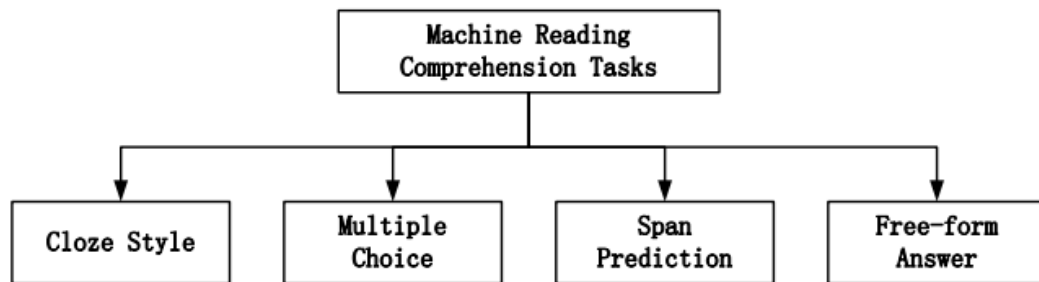
ex>냄비에 불켄때, 물을 붓고 라면을 넣는다 vs 라면을 넣고 물을 붓는다

- MSMARCO에도 순서 맞추기 task가 있음

### 3. MRC vs other NLP tasks

## 2.3 Classification of MRC tasks

- 현재 사용되고 있는 MRC task 분류법



**Figure 6.** Existing classification method of machine reading comprehension tasks.

- cloze-style: 빈칸채우기
- Multiple Choice: 객관식
- Span prediction: paragraph에서 답이 되는 부분 찾아서 highlighting
- Free from : and others...
- 그러나 중복 해당되는 경우가 많음

**Passage:** Tottenham won 2-0 at Hapoel Tel Aviv in UEFA Cup action on Thursday night in a defensive display which impressed Spurs skipper Robbie Keane. ... Keane scored the first goal at the Bloomfield Stadium with **Dimitar Berbatov**, who insisted earlier on Thursday he was happy at the London club, heading a second. The 26-year-old Berbatov admitted the reports linking him with a move had affected his performances ... Spurs manager Juande Ramos has won the UEFA Cup in the last two seasons ...

**Question:** Tottenham manager Juande Ramos has hinted he will allow \_\_\_\_\_ to leave if the Bulgaria striker makes it clear he is unhappy.

**Choices:** (A) Robbie Keane (B) **Dimitar Berbatov** ✓

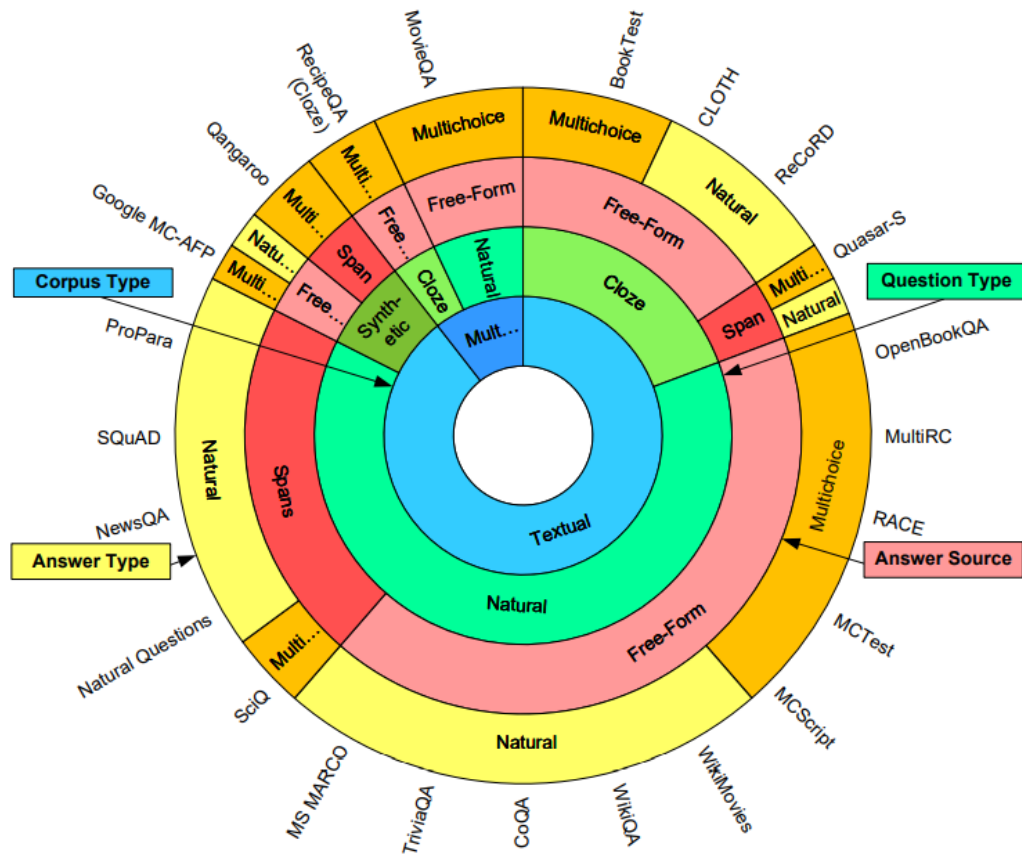
- 이 경우, 질문이 cloze-syle이면서 동시에 MC이기도 함

## 2.4. Definition of each category in the new classification method

- 따라서 새로운 분류법 제시
  - corpus type : textual vs multimodal
  - question type : natural vs cloze vs synthesis



- answer source: span vs freeform
- answer type : natural vs MC



## 1. corpus type

- multimodal(위에 세포그림 예시 참고)
- text

**Passage:** In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under *gravity*. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called "showers".

**Question:** What causes precipitation to fall?

**Answer:** gravity


## 2. type of question

- cloze style


- 적절한 단어/어구/문장/이미지를 삽입후보(context) 에서 골라 빈칸에 삽입하는 것
- 후보 제시 형식이 객관식/주관식인지는 상관 없음 → 이건 answer type에서 고려됨

**Passage**  
Last-Minute Lasagna:  
1. Heat oven to 375 degrees F. Spoon a thin layer of sauce over the bottom of a 9-by-13-inch baking dish.  
2. Cover with a single layer of ravioli.  
3. Top with half the spinach half the mozzarella and a third of the remaining sauce.  
4. Repeat with another layer of ravioli and the remaining spinach mozzarella and half the remaining sauce.  
5. Top with another layer of ravioli and the remaining sauce not all the ravioli may be needed. Sprinkle with the Parmesan.  
6. Cover with foil and bake for 30 minutes. Uncover and bake until bubbly, 5 to 10 minutes.  
7. Let cool 5 minutes before spooning onto individual plates.

**Question:** Choose the best image for the missing blank to correctly complete the recipe.



**Choices:**



(A) ✓ (B) (C) (D)

- natural style
  - 질문: 대부분 문장의 형식을 갖춘 의문형 문장. 물음표로 끝남. 명령문 형식일 수도 있음
  - ex> 손흥민이 소속된 EPL 팀은?
  - ex> 손흥민이 소속된 EPL 팀을 찾으시오

**Passage:** In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under *gravity*. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called "showers".

**Question:** What causes precipitation to fall?

**Answer:** gravity

예시에서는 답이 passage안에 있긴 한데, 없어도 됨

- synthesis form
  - 질문 : 문장의 형식이 완전하지 않고 단어의 나열
  - 왜 하는지는 모르겠음. 모델이 이해하기 더 어렵게 하려고 그러나?

**Passage:** The hanging Gardens, in [Mumbai], also known as Pheroze Shah Mehta Gardens, are terraced gardens ... They provide sunset views over the [Arabian Sea] ... Mumbai (also known as Bombay, the official name until 1995) is the capital city of the Indian state of Maharashtra. It is the most populous city in *India* ... The Arabian Sea is a region of the northern Indian Ocean bounded on the north by Pakistan and Iran, on the west by northeastern Somalia and the Arabian Peninsula, and on the east by India ...

**Synthetic Question:** (Hanging gardens of Mumbai, country, ?)

**Choices:** (A)Iran, (B)India✓, (C)Pakistan, (D) Somalia

### 3. answer type

- multiple choice : 나도알고 너도알고 우리모두아는 객관식
- freeform: 나머지 전부
  - 주어진 passage내에 답이 있을수도, 없을수도
  - 이런식으로 계산이 필요할 수도

**Passage:** That year, his Untitled (1981), a painting of a haloed, black-headed man with a bright red skeletal body, depicted amid the artist's signature scrawls, was sold by Robert Lehrman for \$16.3 million, well above its \$12 million high estimate.

**Question:** How many more dollars was the Untitled (1981) painting sold for than the 12 million dollar estimation?

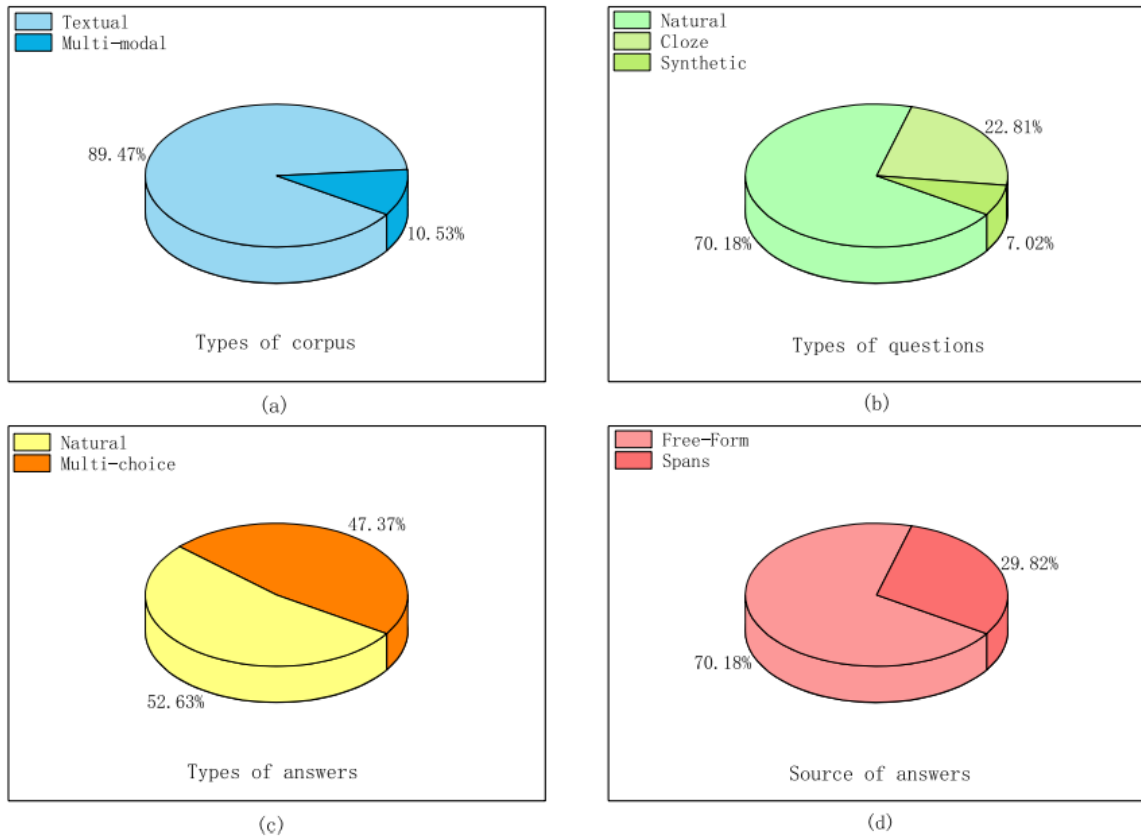
**Answer:** 4300000

### 4. answer source

- span : passage내에 답이 있고, 그걸 highlighting하면 됨
- freeform : 나머지 전부 222

## 2.5. Statistics of MRC Tasks

- corpus type : text >> multimodal
- question type : natural >> cloze > synthetic
- answer type : natural ~ MC
- answer source : freeform >> span



- 소개된 dataset을 이 기준으로 표기한 table이 있는데, 생략. Table 1 in 19-20p

### 3. Evaluation Metrics

**Table 3.** Statistics on the usage of different evaluation metrics in 57 machine reading comprehension tasks.

Metrics	Accuracy	F1	EM	BLEU	Recall	Precision	ROUGE-L	HEQ-D	Meteor
Usage	61.40%	36.84%	22.81%	7.02%	5.26%	5.26%	3.51%	1.75%	1.75%

#### 3.2 Accuracy

$$\frac{\text{맞힌 토큰 갯수}}{\text{총 토큰 갯수}}$$

$$\text{Accuracy} = \frac{M}{N}$$

- 위치 정보가 틀려도 맞았다고 치나? 위치 상관없이 맞았다고 치나?
- 중간에 단어가 들어간 경우 인덱스가 하나 밀리니까 → 뭔가 해결하는 방법이 있음

### 3.3 Exact Match

$$\frac{\text{정확하게 맞힌 문제 갯수}}{\text{총 문제 갯수}}$$

ex> GT for Ground Truth, Model for Model result

GT1) Biden is president of the United States

Model1) Biden is president of United States

GT2) I am Korean

Model2) I am Korean

의 경우에 2문제에 대한

$$\text{accuracy} = 9/10$$

$$\text{EM} = 1/2$$

### 3.4 Precision

$$\frac{\text{정확하게 맞힌 } x \text{ 갯수}}{TP \text{ } x \text{ 개수} + FP \text{ } x \text{ 개수}}$$

x 는 token, question 모두 가능

TS : Token level score for Single question

T : Token lv score

Q : Question lv score

$$Precision_{TS} = \frac{Num(TP_T)}{Num(TP_T) + Num(FP_T)}$$

$$Precision_Q = \frac{Num(TP_Q)}{Num(TP_Q) + Num(FP_Q)}$$

### 3.5 Recall

$$\frac{\text{모델이 양성이라고 예측 and 진짜 양성인 } x \text{ 갯수}}{\text{진짜 양성인 } x \text{ 갯수}}$$

$$Recall_{TS} = \frac{Num(TP_T)}{Num(TP_T) + Num(FN_T)}$$

$$Recall_Q = \frac{Num(TP_Q)}{Num(TP_Q) + Num(FN_Q)}$$

### 3.6 F1

- token lv F1 for single question

$$F1_{TS} = \frac{2 \times Precision_{TS} \times Recall_{TS}}{Precision_{TS} + Recall_{TS}}$$

- token lv F1

$$F1_T = \frac{\sum Max(Precision_{TS})}{Num(Questions)}$$

- question lv F1

$$F1_Q = \frac{2 \times Precision_Q \times Recall_Q}{Precision_Q + Recall_Q}$$

### 3.7 ROUGE

- n-gram과 recall 을 사용한 평가 시스템
- RS : ReferenceSummary = Ground Truth
- S : Summary = 모델이 만든 문장

$$ROUGE-N = \frac{\sum_{S \in \{RS\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{RS\}} \sum_{gram_n \in S} Count(gram_n)}$$

$$\frac{S의\ n - gram중\ RS와\ 일치하는\ n - gram\ 갯수}{RS의\ n - gram\ 갯수}$$

ex> RS: 애플은 2022년 iphone 14 pro max 발매했다

S : 애플 주식회사는 iphone 14 를 발매했다

Rouge-1 : 3/7

### 3.8 BLEU

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

$$BLEU = BP \cdot \exp \left( \sum_{n=1}^N w_n \log p_n \right)$$

- BP(Brevity Penalty) : 원래 주어진 문장 대비 문장이 짧아지는 것에 대한 페널티

ex>

GT) I am super duper fancy and pretty

M) I am

이때  $BP = e^{((1-7)/2)} = e^{-3} = 0.04$

- $p_n$  = modified precision

$$\frac{\text{생성 길이의 } n\text{-gram 중 Ground Truth에 있는 } n\text{-gram (단, GT의 갯수를 초과할 수 없음)}}{\text{생성 길이의 총 } n\text{-gram 수}}$$

ex> modified unigram precision

GT) for Ground Truth, M) for Model result

GT) my feet sucks

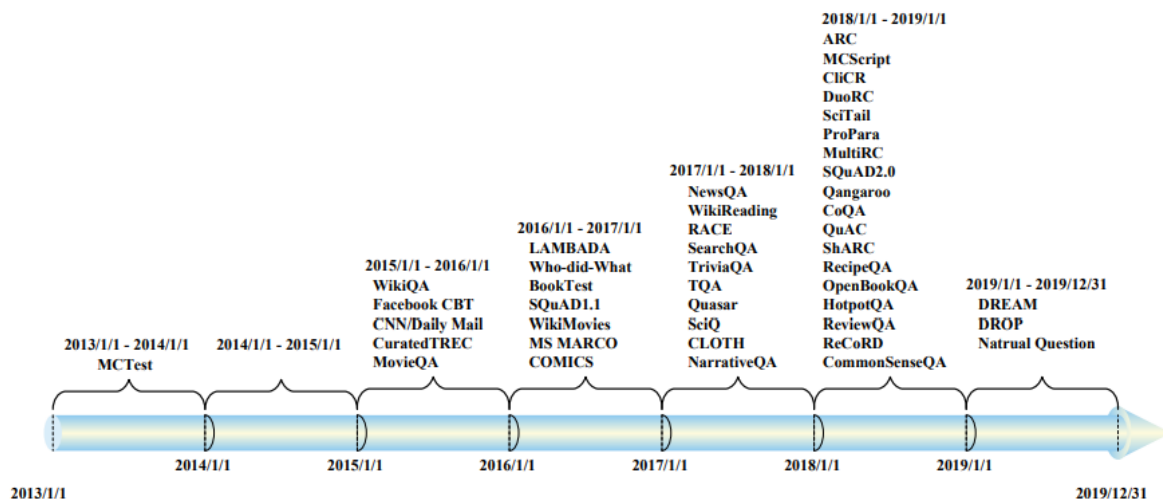
M) my feet is over your feet, under his feet next to her feet

no modified precision = 5/13

modified precision = 2/13

## 4. Benchmark Datasets

- timeline



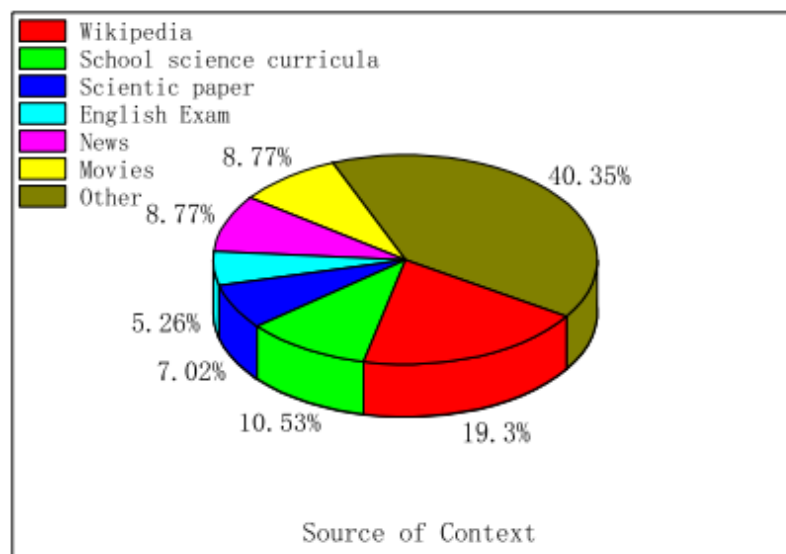
- question 사이즈별 나열

논문 30-32p table 4 참고

- maxsize : 18.87M (WikiReading, 2016)
- minsize : 488(ProPara, 2018)



- 어쨌든 사이즈는 커지는 추세
- corpus 사이즈별 나열  
논문 33-34p table 5 참고
  - 다양한 corpus unit : article, book, dialogue, document, lesson, movie, passage, sentence, story, etc.
  - maxsize : 4.7M (WikiReading, 2016)
  - minsize : 108(Facebook CBT, 2016)
- generation method 별 나열  
논문 35-37p table 6 참고
  - automated(모델 생성), 클라우드소싱, 전문가
  - 클라우드소싱 >>모델생성 > 전문가
- 다양한 corpus source
  - 시험문제, 위키피디아, 뉴스기사, 이공계논문 초록, 이야기(소설/동화), 기술문서, 교과서, 영화 줄거리, 레시피, 정부 웹사이트, 검색엔진 질문기록, 호텔 리뷰 등



- 기타 특성
  - unanswerable questions
    - unanswerable question을 데이터셋에 포함시키면 모델을 더 강건하게 할 수 있음
  - multi-hop MRC

- 질문에 대답하기 위해 2개 이상의 정보가 필요한 질문
  - ex> SSG랜더스와 키움히어로즈 중 2022 한국시리즈 우승팀은?
  - SSG의 우승 여부, 키움히어로즈의 우승 여부 2개의 정보가 필요
- paraphrased paragraph
  - 같은 내용의 paragraph를 paraphrase해서 2개 적어놓고 둘의 정보를 비교함으로써 comprehension test
  - DuoRC
- commonsense(knowledge)
  - 문제를 풀기 위해 다의어 및 대명사 이해, 간단한 연산 등을 필요로 함
  - ex> this book can't be put into a school bag. **it's** too small
  - 여기에서 it의 의미는?
- complex reasoning
  - 상대적으로 복잡한 논리적 추론을 포함
- multimodal
  - RecipeQA
- domain specific
  - 영화, 요리 레시피, 의학 등
- open-domain QA
  - (범위를 특정하는게 의미 없을 만큼) 대량의 문서로부터 주어진 질문에 대한 답변 찾기
  - BookTest, SearchQA
- 인기투표 결과(월간 citation 갯수 순위)
  1. SQuAD 1.1
  2. CNN/Daily Mail
  3. SQuAD 2.0
  4. Natural Questions
  5. TriviaQA
  6. CoQA

7. WikiMovies
8. CBT
9. MSMARCO
10. WikiQA

## 4.10 introducing all datasets

지금까지 발표된 47개 dataset 전부에 대한 한문단씩 설명이 있는데 일단 인기투표 결과 1-10위 만 요약.

### 1. SQuAD 1.1(2016)

- crowdworker, 100000 문제
- 표준으로 받아들여짐

▼ example

A prime number (or a prime) is a natural number greater than 1 that has no positive divisors other than 1 and itself. A natural number greater than 1 that is not a prime number is called a composite number. For example, 5 is prime because 1 and 5 are its only positive integer factors, whereas 6 is composite because it has the divisors 2 and 3 in addition to 1 and 6. The fundamental theorem of arithmetic establishes the central role of primes in number theory: any integer greater than 1 can be expressed as a product of primes that is unique up to ordering. The uniqueness in this theorem requires excluding 1 as a prime because one can include arbitrarily many instances of 1 in any factorization, e.g.,  $3$ ,  $1 \cdot 3$ ,  $1 \cdot 1 \cdot 3$ , etc. are all valid factorizations of 3.

**What is the only divisor besides 1 that a prime number can have?**

Ground Truth Answers: itself itself itself itself itself

**What are numbers greater than 1 that can be divided by 3 or more numbers called?**

Ground Truth Answers: composite number composite number composite number primes

**What theorem defines the main role of primes in number theory?**

Ground Truth Answers: The fundamental theorem of arithmetic fundamental theorem of arithmetic arithmetic arithmetic fundamental theorem of arithmetic fundamental theorem of arithmetic

### 2. CNN/Daily Mail

- super large
- source from CNN/Dailymail

### 3. SQuAD 2.0

- SQuAD 1.1에 50,000개의 unanswerable q 추가
- SQuAD 1.1보다 human accuracy - model SOTA의 차이가 큼

- SQuAD 1.1보다 모델 입장에서 어렵다는 뜻
4. Natural Questions
- 구글 서치엔진에 들어온 질문 답변으로 만들어진 데이터셋
  - 한 질문에 대해 긴 답변 짧은 답변 둘다있음
5. TriviaQA
- 650k qa pair-evidence
  - 상대적으로 어려운 추론 난이도
  - 질문-답변 간 문법구조 및 사용된 단어 상이
  - 따라서 cross-sentence reasoning이 더 필요
6. CoQA
- 8k dialogues, 127k Qs, 7 fields
  - coreference(공통 참조), pragmatic reasoning(단어가 내포하는 의미) 를 찾을 수 있어야 함
    - 다른 MRC에는 없는 task
7. WikiMovies
- raw text + 전처리된 KB(Knowledge Base)
- ▼ example

**Doc: Wikipedia Article for Blade Runner (partially shown)**

Blade Runner is a 1982 American neo-noir dystopian science fiction film directed by Ridley Scott and starring Harrison Ford, Rutger Hauer, Sean Young, and Edward James Olmos. The screenplay, written by Hampton Fancher and David Peoples, is a modified film adaptation of the 1968 novel “Do Androids Dream of Electric Sheep?” by Philip K. Dick. The film depicts a dystopian Los Angeles in November 2019 in which genetically engineered replicants, which are visually indistinguishable from adult humans, are manufactured by the powerful Tyrell Corporation as well as by other “mega-corporations” around the world. Their use on Earth is banned and replicants are exclusively used for dangerous, menial, or leisure work on off-world colonies. Replicants who defy the ban and return to Earth are hunted down and “retired” by special police operatives known as “Blade Runners”. ...

**KB entries for Blade Runner (subset)**

*Blade Runner directed\_by* Ridley Scott  
*Blade Runner written\_by* Philip K. Dick, Hampton Fancher  
*Blade Runner starred\_actors* Harrison Ford, Sean Young, ...  
*Blade Runner release\_year* 1982  
*Blade Runner has\_tags* dystopian, noir, police, androids, ...

**Questions for Blade Runner (subset)**

Ridley Scott directed which films?  
What year was the movie Blade Runner released?  
Who is the writer of the film Blade Runner?  
Which films can be described by dystopian?  
Which movies was Philip K. Dick the writer of?  
Can you describe movie Blade Runner in a few words?

8. Facebook CBT

- 아이 동화책 corpus
- 1개 문제당 21문장인데, 마지막 문장 cloze
- 10개 객관식

9. MSMARCO

- unanswerable question 포함
- 3 subtasks
  - unanswerable q 여부 판단

- if answerable : 대답
- 답변과 관계있는 페이지 링크 가져오기

#### 10. WikiQA

- 위키피디아 paragraph + bing 검색된 질문
- unanswerable question 포함

## 5. Open issues

- Multimodal dataset 및 관련 연구의 부족
- Commonsense, word knowledge를 잘 이해하지 못함
  - ex> this book can't be put into a school bag. **it's** too small
  - 여기에서 it의 의미를 잘 캐치하지 못함
- complex reasoning 부족 (multisentence reasoning을 요구하는 dataset이 적음)
- 단어의 sequence에 지나치게 의존 → 강건함 부족
- interpretability : 왜 이런 결과가 나왔는지 설명할 수 없음
- MRC 문제가 얼마나 어려운지 평가 지표 부족
  - 현 지표 synthetic complexity는 문장의 길이를 지표로만 사용