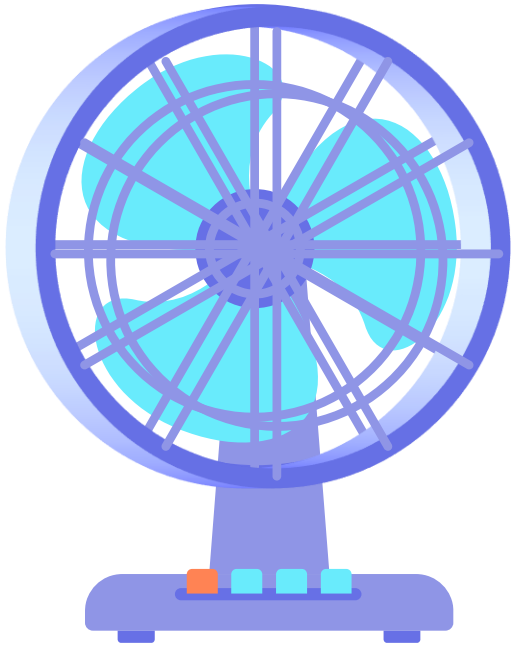


SQuAD

한양대학교 인공지능연구실 석사 1기
조충현

Index

Introduction

What is SQuAD?

- SQuAD = Stanford Question Answering Dataset
- 100,000+ questions by crowdworkers using Wikipedia articles
- Answer to each question is a segment of text from the corresponding reading 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".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

Figure 1: Question-answer pairs for a sample passage in the SQuAD dataset. Each of the answers is a segment of text from the passage.

EX) 첫번째 질문에 답하기 위해 먼저 **"precipitation... falls under gravity"**절의 관련 부분을 찾은 다음 **"under"**이 원인이라고 판단하고 **"gravity"**라는 정답을 결정한다.

Introduction

Why create the SQuAD?

- 존재하는 데이터셋들은 높은 품질의 데이터를 만들기에는 데이터가 적다
- 큰 규모의 데이터셋들은 **QA**의 특성을 가지고 있지 않다는 문제가 있었다.

그래서!!!

데이터 규모도 크고 높은 품질의 데이터인 **SQuAD** 데이터셋을 만들었다.

Introduction

Comparison with Existing datasets

1. 이전에 있는 데이터셋보다 두 자리수나 더 많다
2. **SQuAD**는 주관식이다~!!
3. **Span-based answers**여서 평가하기가 더 쉽다

Existing Datasets

Dataset	Question source	Formulation	Size
SQuAD	**crowdsourced**	**RC, spans in passage**	**100K**
MCTest (Richardson et al., 2013)	crowdsourced	RC, multiple choice	2640
Algebra (Kushman et al., 2014)	standardized tests	computation	514
Science (Clark and Etzioni, 2016)	standardized tests	reasoning, multiple choice	855
WikiQA (Yang et al., 2015)	query logs	IR, sentence selection	3047
TREC-QA (Voorhees and Tice, 2000)	query logs + human editor	IR, free form	1479
CNN/Daily Mail (Hermann et al., 2015)	summary + cloze	RC, fill in single entity	1.4M
CBT (Hill et al., 2015)	cloze	RC, fill in single word	688K

Table 1: A survey of several reading comprehension and question answering datasets. SQuAD is much larger than all datasets except the semi-synthetic cloze-style datasets, and it is similar to TREC-QA in the open-endedness of the answers.

Reading comprehension

- **MCTest**
  - **crowdsourcing**으로 **660**개의 스토리를 포함하고 각 스토리마다 **4**개의 질문이 있고 각 질문마다 **4**개의 대답을 고르는 데이터셋
- **Algebra** 수식을 찾는 데이터셋
- **Science 4**학년 수준의 과학 시험문제를 푸는 데이터셋

단점**:** 너무 데이터 크기가 작다

Open-domain question answering

- 큰 **documents**들로부터 나오는 질문에 답하는 데이터셋

차이**: sentence selection VS. select a specific span**

Cloze datasets

- **passage**에서 빈 단어를 예측하는 데이터셋
- 이러한 데이터셋은 자동으로 생성될 수 있어서 상당히 크다

차이**:** 단순한 하나의 단어를 찾는거라면 **SQuAD**는 더 큰 구절을 찾을 수도 있다**.**

Dataset Collection

총 세가지의 단계로 **dataset**을 모음

1. **Curating passages**
2. **Crowdsourcing question-answers on those passages**
3. **Additional answers collections**

Dataset Collection

Passage curation

- **10000 articles of English Wikipedia**를 가져오고
- **Random**하게 **536 articles**를 뽑는다
- 뽑은 **articles**들을 개별적인 **paragrap**로 추출하고 **images, figures, tables, 500**자 미만의 **paragrap**들은 버림
- 그 결과 **23,215**개의 **paragrap**가 나오게 된다**.**

Dataset Collection

Question-answer collection

crowdworker들에게 **question**들과 **aswers**를 만들라고 이용한다**.**

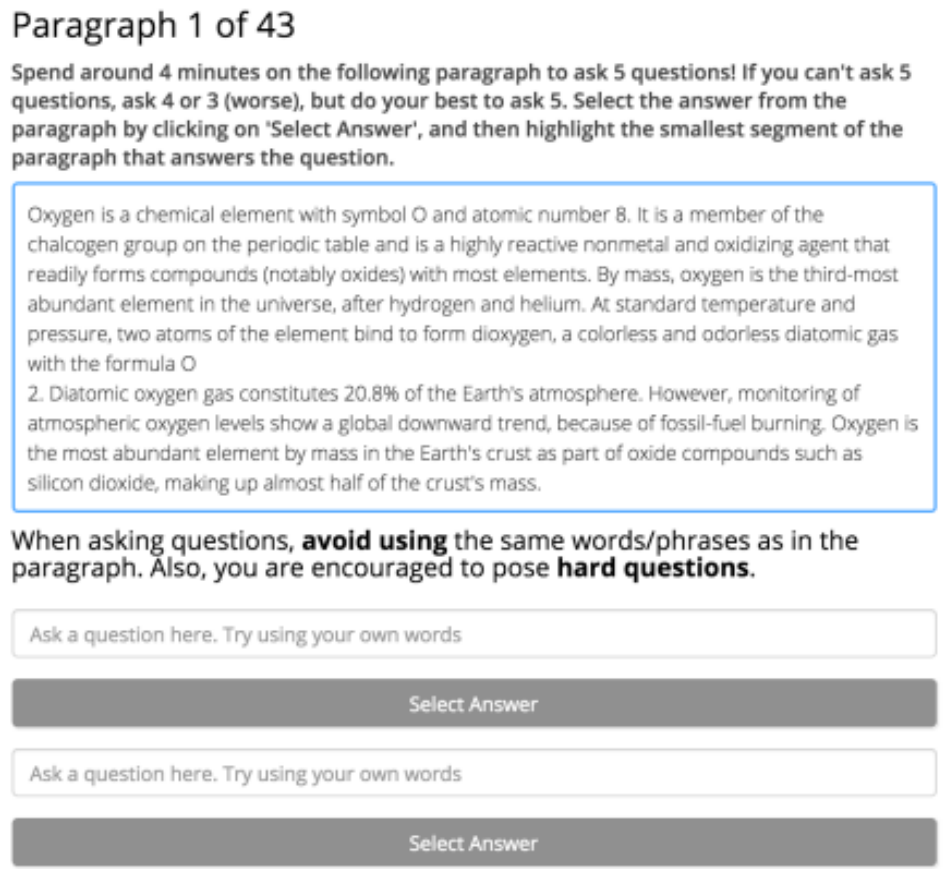


Figure 2: The crowd-facing web interface used to collect the dataset encourages crowdworkers to use their own words while asking questions.

Dataset Collection

Additional answers collection

평가를 더욱 **robust**하게 하기 위해 **dev**와 **test sets**에 최소 각각의 **question**에 **2**개의 추가적인 **answer**를 얻는다**.**

2번째 **answer**를 만들때 **crowdworker**는 한 **article**의 **paragraphs**와 **questions**를 본다**.**

그리고 질문에 대답하는 **paragraph**의 **shortest span**을 찾는다**.**

만약 질문의 답을 찾을 수 없다면 **unaswerable**로 마크가 된다**.**

Dataset Analysis

1. Diversity in answers
2. Reasoning required to answer questions
3. Stratification by syntactic divergence

Dataset Analysis

Diversity in answers

- 먼저 **numerical** 과 **non-numerical answer**로 나눈다.
- **Non-numerical**한 **answer**는 **Stanford CoreNLP**를 이용해 구분한다.
- 고유 명수구는 **NER**태그를 이용하여 사람, 위치, 기타항목으로 더 나뉜다.

Answer type	Percentage	Example
Date	8.9%	19 October 1512
Other Numeric	10.9%	12
Person	12.9%	Thomas Coke
Location	4.4%	Germany
Other Entity	15.3%	ABC Sports
Common Noun Phrase	31.8%	property damage
Adjective Phrase	3.9%	second-largest
Verb Phrase	5.5%	returned to Earth
Clause	3.7%	to avoid trivialization
Other	2.7%	quietly

Table 2: We automatically partition our answers into the following categories. Our dataset consists of large number of answers beyond proper noun entities.

Dataset Analysis

Reasoning required to answer questions

Reasoning	Description	Example	Percentage
Lexical variation (synonymy)	Major correspondences between the question and the answer sentence are synonyms.	Q: What is the Rankine cycle sometimes **called**? Sentence: The Rankine cycle is sometimes **referred** to as a practical Carnot cycle.	33.3%
Lexical variation (world knowledge)	Major correspondences between the question and the answer sentence require world knowledge to resolve.	Q: Which **governing bodies** have veto power? Sen.: **The European Parliament and the Council of the European Union** have powers of amendment and veto during the legislative process.	9.1%
Syntactic variation	After the question is paraphrased into declarative form, its syntactic dependency structure does not match that of the answer sentence even after local modifications.	Q: What Shakespeare scholar **is currently on the faculty**? Sen.: **Current faculty include** the anthropologist Marshall Sahlins, ..., Shakespeare scholar David Bevington.	64.1%
Multiple sentence reasoning	There is anaphora, or higher-level fusion of multiple sentences is required.	Q: What collection does **the V&A Theatre & Performance galleries** hold? Sen.: **The V&A Theatre & Performance galleries** opened in March 2009. ... **They** hold the UK's biggest national collection of material about live performance.	13.6%
Ambiguous	We don't agree with the crowdworkers' answer, or the question does not have a unique answer.	Q: What is the main goal of criminal punishment? Sen.: **Achieving crime control via incapacitation and deterrence** is a major goal of criminal punishment.	6.1%

Table 3: We manually labeled 192 examples into one or more of the above categories. Words relevant to the corresponding reasoning type are bolded, and the crowdsourced answer is underlined.

Dataset Analysis

Stratification by syntactic divergence

Q: What department store is thought to be the first in the world?
S: Bainbridge's is often cited as the world's first department store.

Path:

first ←xcomp— thought —nsubjpass→ store —det→ what

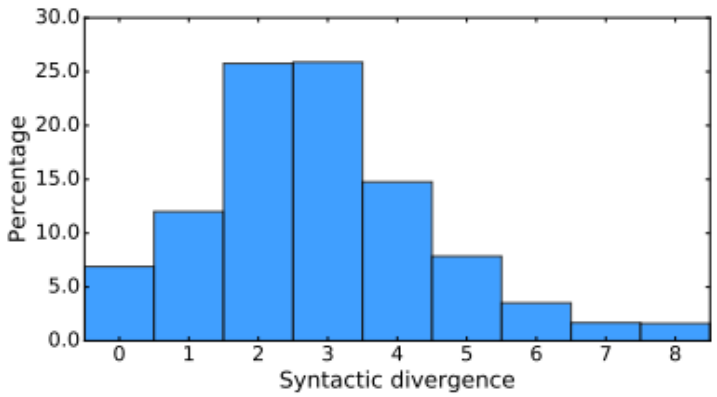
⇓delete ⇓substitute ⇓insert

first ←amod— store ←nmod— cited —nsubjpass→ Bainbridge's

Edit cost:

1 +2 +1=4

Figure 3: An example walking through the computation of the syntactic divergence between the question Q and answer sentence S.



(a) Histogram of syntactic divergence.

Q: Who went to Wittenberg to hear Luther speak?
S: Students thronged to Wittenberg to hear Luther speak.
Path:

Wittenberg ←nmod— went —nsubj→ Who

Wittenberg ←nmod— thronged —nsubj→ Students

(b) An example of a question-answer pair with edit distance 0 between the dependency paths (note that lexical variation is ignored in the computation of edit distance).

Q: What impact did the high school education movement have on the presence of skilled workers?
S: During the mass high school education movement from 1910 – 1940 , there was an increase in skilled workers.
Path:

school ←compound— movement ←nsubj— have —dobj→ impact —det→ What

school ←compound— movement —nmod→ 1910 —acl→ was —nsubj→ increase

(c) An example of a question-answer pair with edit distance 6.

Figure 4: We use the edit distance between the unlexicalized dependency paths in the question and the sentence containing the answer to measure *syntactic divergence*.

Methods

logistic Regression

- 여러가지 **features**를 뽑아내고 그 **features**로 **multiclass log-likelihood loss**를 이용하여 학습

Feature Groups	Description	Examples
Matching Word Frequencies	Sum of the TF-IDF of the words that occur in both the question and the sentence containing the candidate answer. Separate features are used for the words to the left, to the right, inside the span, and in the whole sentence.	Span: $[0 \leq \text{sum} < 0.01]$ Left: $[7.9 \leq \text{sum} < 10.7]$
Matching Bigram Frequencies	Same as above, but using bigrams. We use the generalization of the TF-IDF described in Shirakawa et al. (2015).	Span: $[0 \leq \text{sum} < 2.4]$ Left: $[0 \leq \text{sum} < 2.7]$
Root Match	Whether the dependency parse tree roots of the question and sentence match, whether the sentence contains the root of the dependency parse tree of the question, and whether the question contains the root of the dependency parse tree of the sentence.	Root Match = False
Lengths	Number of words to the left, to the right, inside the span, and in the whole sentence.	Span: $[1 <= \text{num} < 2]$ Left: $[15 \leq \text{num} < 19]$
Span Word Frequencies	Sum of the TF-IDF of the words in the span, regardless of whether they appear in the question.	Span: $[5.2 \leq \text{sum} < 6.9]$
Constituent Label	Constituency parse tree label of the span, optionally combined with the wh-word in the question.	Span: NP Span: NP, wh-word: "*what*"
Span POS Tags	Sequence of the part-of-speech tags in the span, optionally combined with the wh-word in the question.	Span: [NN] Span: [NN], wh-word: "*what*"
Lexicalized	Lemmas of question words combined with the lemmas of words within distance 2 to the span in the sentence based on the dependency parse trees. Separately, question word lemmas combined with answer word lemmas.	Q: "*cause*", S: "*under*" $\xleftarrow{\text{case}}$ Q: "*fall*", A: "*gravity*"
Dependency Tree Paths	For each word that occurs in both the question and sentence, the path in the dependency parse tree from that word in the sentence to the span, optionally combined with the path from the wh-word to the word in the question. POS tags are included in the paths.	VBZ $\xrightarrow{\text{nmod}}$ NN what $\xleftarrow{\text{nsubj}}$ VBZ $\xrightarrow{\text{advcl}}$ + VBZ $\xrightarrow{\text{nmod}}$ NN

Table 4: Features used in the logistic regression model with examples for the question "*What causes precipitation to fall?*", sentence "*In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity.*" and answer "*gravity*". Q denotes question, A denotes candidate answer, and S denotes sentence containing the candidate answer.

Experiments

	Exact Match		F1	
	Dev	Test	Dev	Test
Random Guess	1.1%	1.3%	4.1%	4.3%
Sliding Window	13.2%	12.5%	20.2%	19.7%
Sliding Win. + Dist.	13.3%	13.0%	20.2%	20.0%
Logistic Regression	40.0%	40.4%	51.0%	51.0%
Human	80.3%	77.0%	90.5%	86.8%

Table 5: Performance of various methods and humans. Logistic regression outperforms the baselines, while there is still a significant gap between humans.

	F_1	
	Train	Dev
Logistic Regression	91.7%	51.0%
– Lex., – Dep. Paths	33.9%	35.8%
– Lexicalized	53.5%	45.4%
– Dep. Paths	91.4%	46.4%
– Match. Word Freq.	91.7%	48.1%
– Span POS Tags	91.7%	49.7%
– Match. Bigram Freq.	91.7%	50.3%
– Constituent Label	91.7%	50.4%
– Lengths	91.8%	50.5%
– Span Word Freq.	91.7%	50.5%
– Root Match	91.7%	50.6%

Table 6: Performance with feature ablations. We find that lexicalized and dependency tree path features are most important.

	Logistic Regression Dev F1	Human Dev F1
Date	72.1%	93.9%
Other Numeric	62.5%	92.9%
Person	56.2%	95.4%
Location	55.4%	94.1%
Other Entity	52.2%	92.6%
Common Noun Phrase	46.5%	88.3%
Adjective Phrase	37.9%	86.8%
Verb Phrase	31.2%	82.4%
Clause	34.3%	84.5%
Other	34.8%	86.1%

Table 7: Performance stratified by answer types. Logistic regression performs better on certain types of answers, namely numbers and entities. On the other hand, human performance is more uniform.

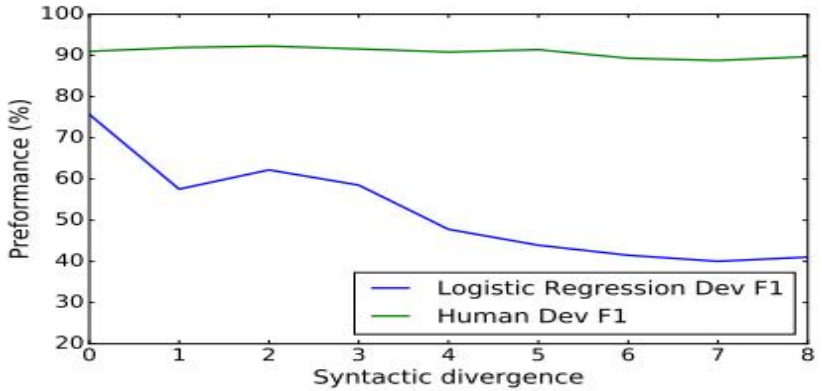


Figure 5: Performance stratified by syntactic divergence of questions and sentences. The performance of logistic regression degrades with increasing divergence. In contrast, human performance is stable across the full range of divergence.

Conclusion

- **SQuAD**는 **Wikipedia articles**를 **crowdsourcing**을 통한 **question-answer** 쌍을 만드는 큰 규모의 **QA** 데이터셋이다
- **SQuAD**는 다양한 범위의 **question**과 **answer type**들이 있다.
- ~~하지만 아직까지는 기계가 인간의 성능을 따라오지는 못하고 있다~~**. (BERT**가 나온 후 뒤바뀜**)**

Conclusion

SQuAD1.1 Leaderboard

Here are the ExactMatch (EM) and F1 scores evaluated on the test set of SQuAD v1.1.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 May 21, 2019	XLNet (single model) *Google Brain & CMU*	**89.898**	**95.080**
2 Jul 21, 2019	SpanBERT (single model) *FAIR & UW*	88.839	94.635
3 Jul 03, 2019	BERT+WWM+MT (single model) *Xiaoi Research*	88.650	94.393
4 Jul 21, 2019	Tuned BERT-1seq Large Cased (single model) *FAIR & UW*	87.465	93.294
5 Oct 05, 2018	BERT (ensemble) *Google AI Language* https://arxiv.org/abs/1810.04805	87.433	93.160
6 May 14, 2019	ATB (single model) *Anonymous*	86.940	92.641
7 Jul 21, 2019	Tuned BERT Large Cased (single model) *FAIR & UW*	86.521	92.617
7 Jul 04, 2019	BERT+MT (single model) *Xiaoi Research*	86.458	92.645
8 Feb 14, 2019	KT-NET (single model) *Baidu NLP*	85.944	92.425
8 Sep 26, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.954	91.677

이미 인간의 성능을 뛰어 넘었다!!

Q&A

질문!?