SQuAD:100,000+ Questions for Machine Comprehension of Text

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, Percy Liang

Published in EMNLP 2016

Presented by Jiaming Shen
April 17, 2018

SQuAD = Stanford Question Answering Dataset

Online challenge: https://rajpurkar.github.io/SQuAD-explorer/

Overall contribution

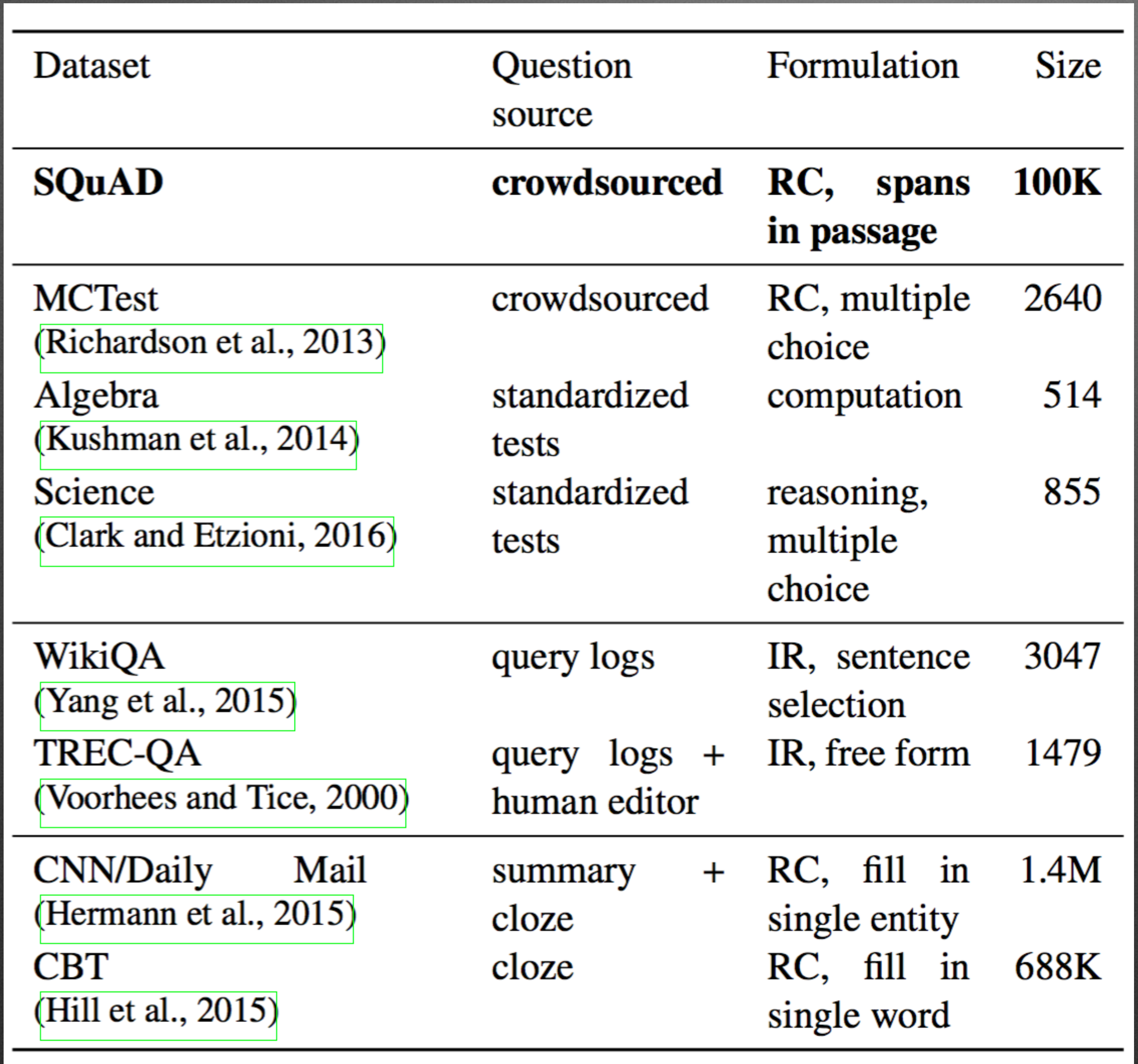
- A benchmark dataset with:
    - Proper difficulty
    - Principled curation process
    - Detailed data analysis

Outline

- What are the QA datasets prior to SQuAD?
- What does SQuAD look like?
- How is SQuAD created?
- What are the properties of SQuAD?
- How well we can do on SQuAD?

What are the QA datasets prior to SQuAD?

Related 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

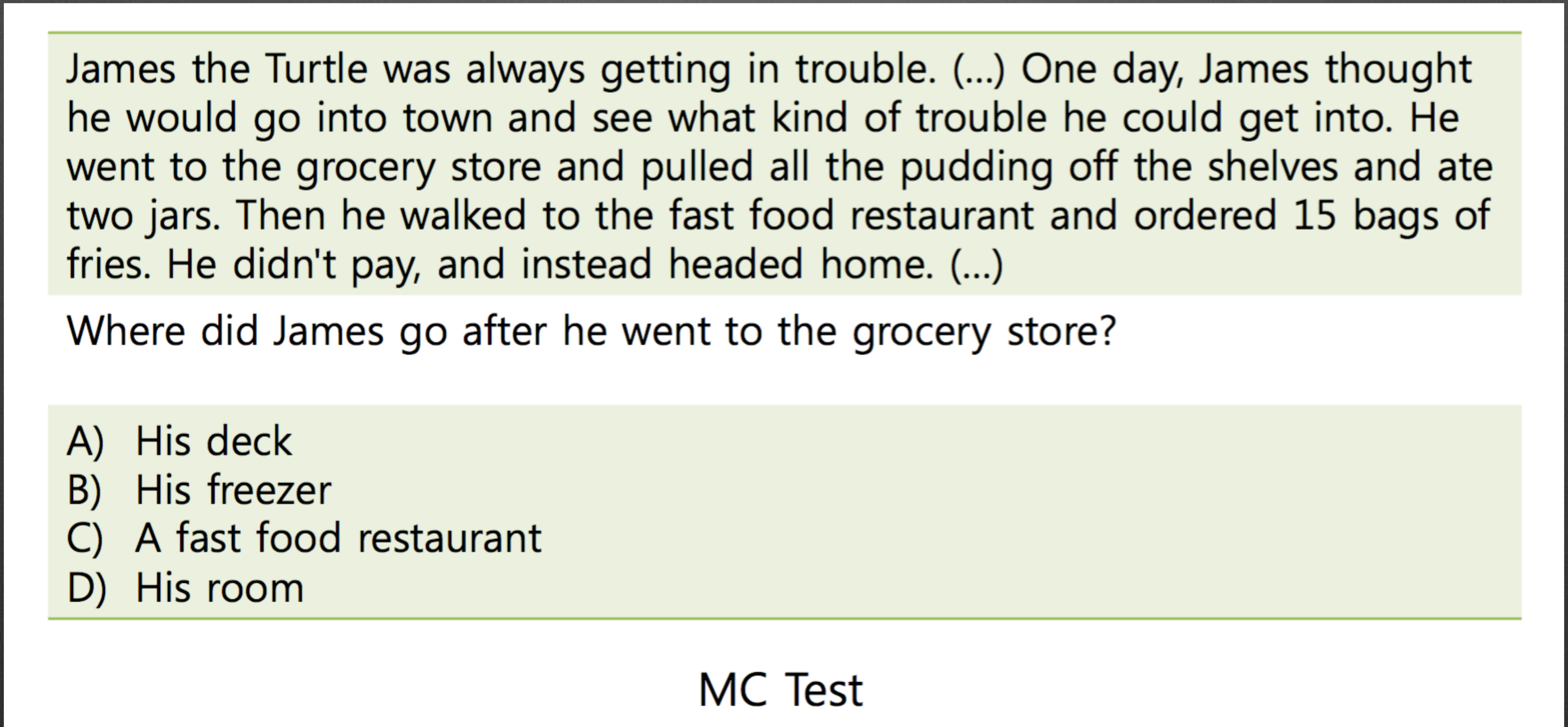
Type I: Complex reading comprehension datasets

Type II: Open-domain QA datasets

Type III: Cloze datasets

Type I: Complex Reading Comprehension Datasets

- Require commonsense knowledge, very challenge
- Dataset size too small



James the Turtle was always getting in trouble. (...) One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home. (...)

Where did James go after he went to the grocery store?

A) His deck
B) His freezer
C) A fast food restaurant
D) His room

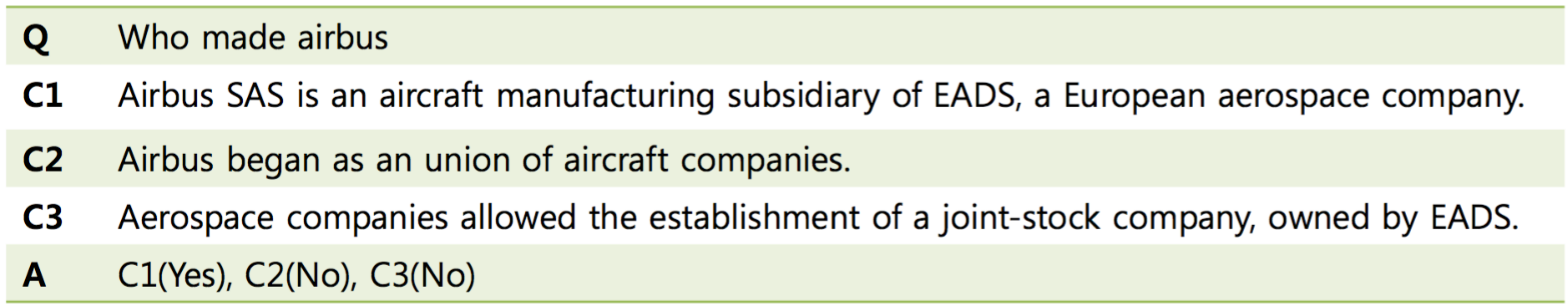
MC Test

Type II: Open-domain QA Datasets

- Open-domain QA: answer a question from a large collection of documents.

- WikiQA: only sentence selection

- TREC-QA: free-form answer -> hard to evaluate

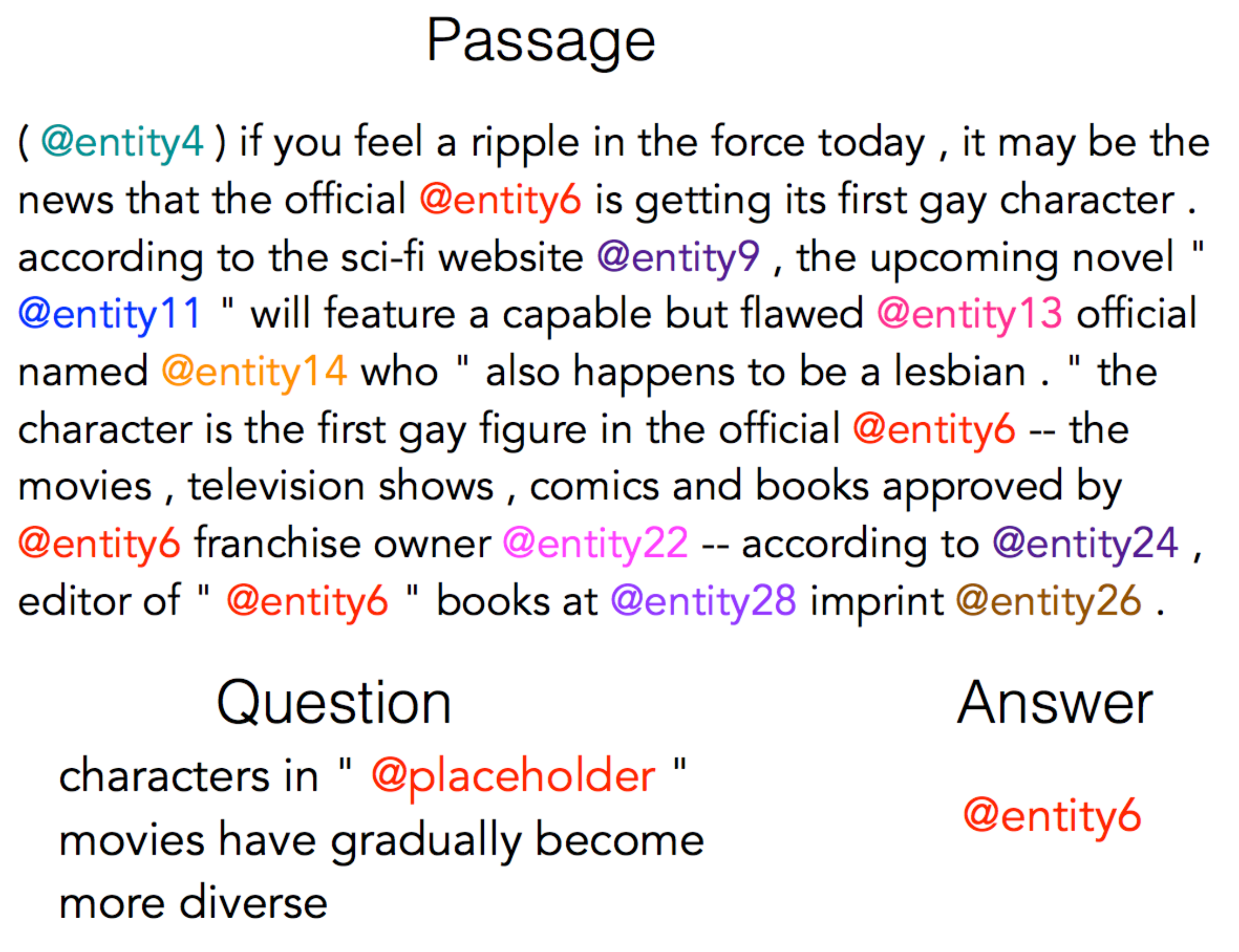


Q	Who made airbus
C1	Airbus SAS is an aircraft manufacturing subsidiary of EADS, a European aerospace company.
C2	Airbus began as an union of aircraft companies.
C3	Aerospace companies allowed the establishment of a joint-stock company, owned by EADS.
A	C1(Yes), C2(No), C3(No)

WikiQA

Type III: Cloze Datasets

- Automatically generated -> large scale
- Limitations are described in ACL 2016 Best Paper.



Passage

(@entity4) if you feel a ripple in the force today , it may be the news that the official @entity6 is getting its first gay character . according to the sci-fi website @entity9 , the upcoming novel " @entity11 " will feature a capable but flawed @entity13 official named @entity14 who " also happens to be a lesbian . " the character is the first gay figure in the official @entity6 -- the movies , television shows , comics and books approved by @entity6 franchise owner @entity22 -- according to @entity24 , editor of " @entity6 " books at @entity28 imprint @entity26 .

Question

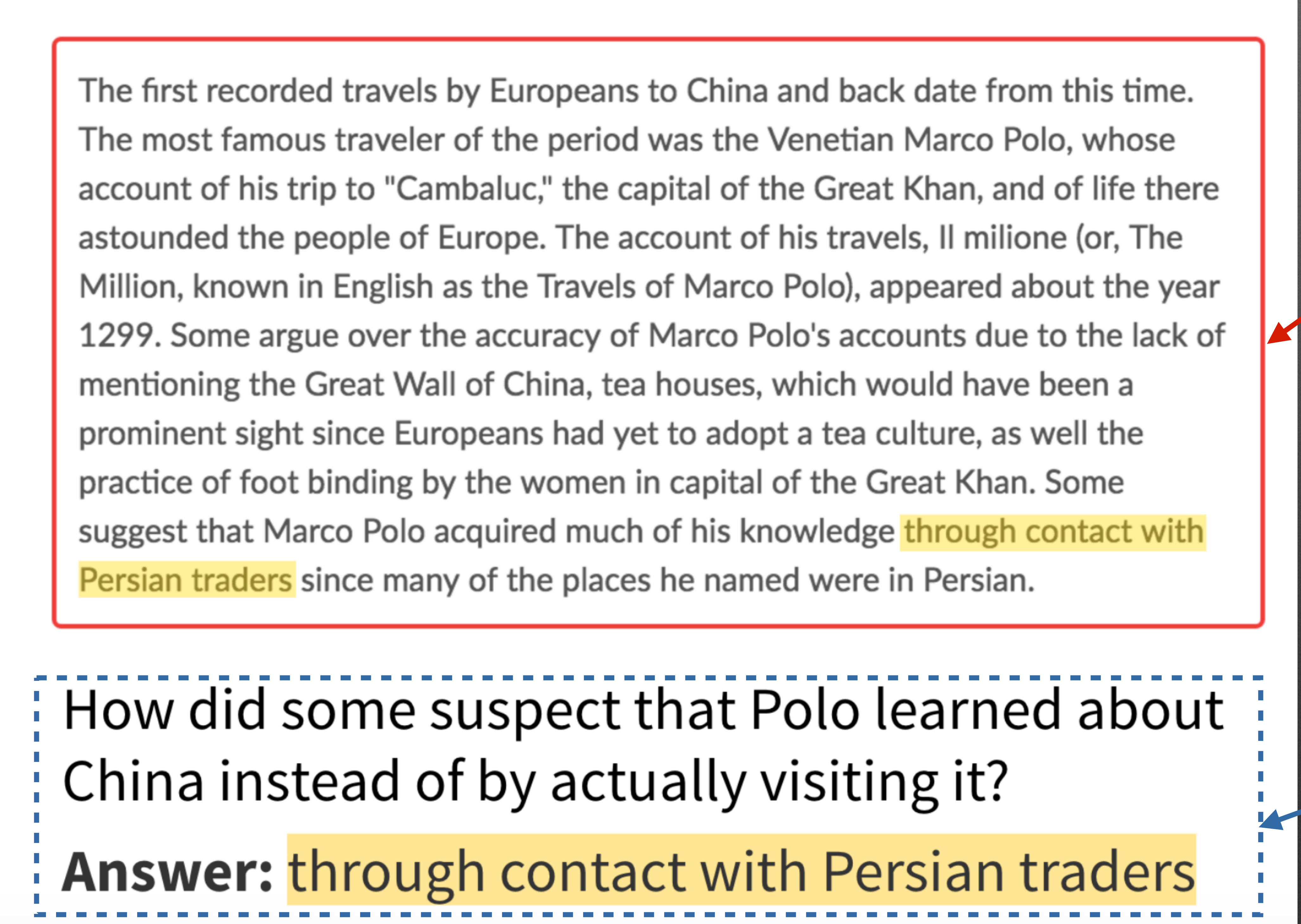
characters in " @placeholder " movies have gradually become more diverse

Answer

@entity6

What does SQuAD look like?

SQuAD Dataset Format



The first recorded travels by Europeans to China and back date from this time. The most famous traveler of the period was the Venetian Marco Polo, whose account of his trip to "Cambaluc," the capital of the Great Khan, and of life there astounded the people of Europe. The account of his travels, Il milione (or, The Million, known in English as the Travels of Marco Polo), appeared about the year 1299. Some argue over the accuracy of Marco Polo's accounts due to the lack of mentioning the Great Wall of China, tea houses, which would have been a prominent sight since Europeans had yet to adopt a tea culture, as well the practice of foot binding by the women in capital of the Great Khan. Some suggest that Marco Polo acquired much of his knowledge through contact with Persian traders since many of the places he named were in Persian.

How did some suspect that Polo learned about China instead of by actually visiting it?

Answer: through contact with Persian traders

A passage

One QA pair

SQuAD Dataset Format

- One passage can have multiple question-answer pairs.
- Totally 100,000+ QA pairs from 23,215 passages.

How is SQuAD created?

SQuAD Dataset Collection

- Consisting three steps:
  - Step1: Passage curation
  - Step2: Question-answer collection
  - Step3: Additional answers collection

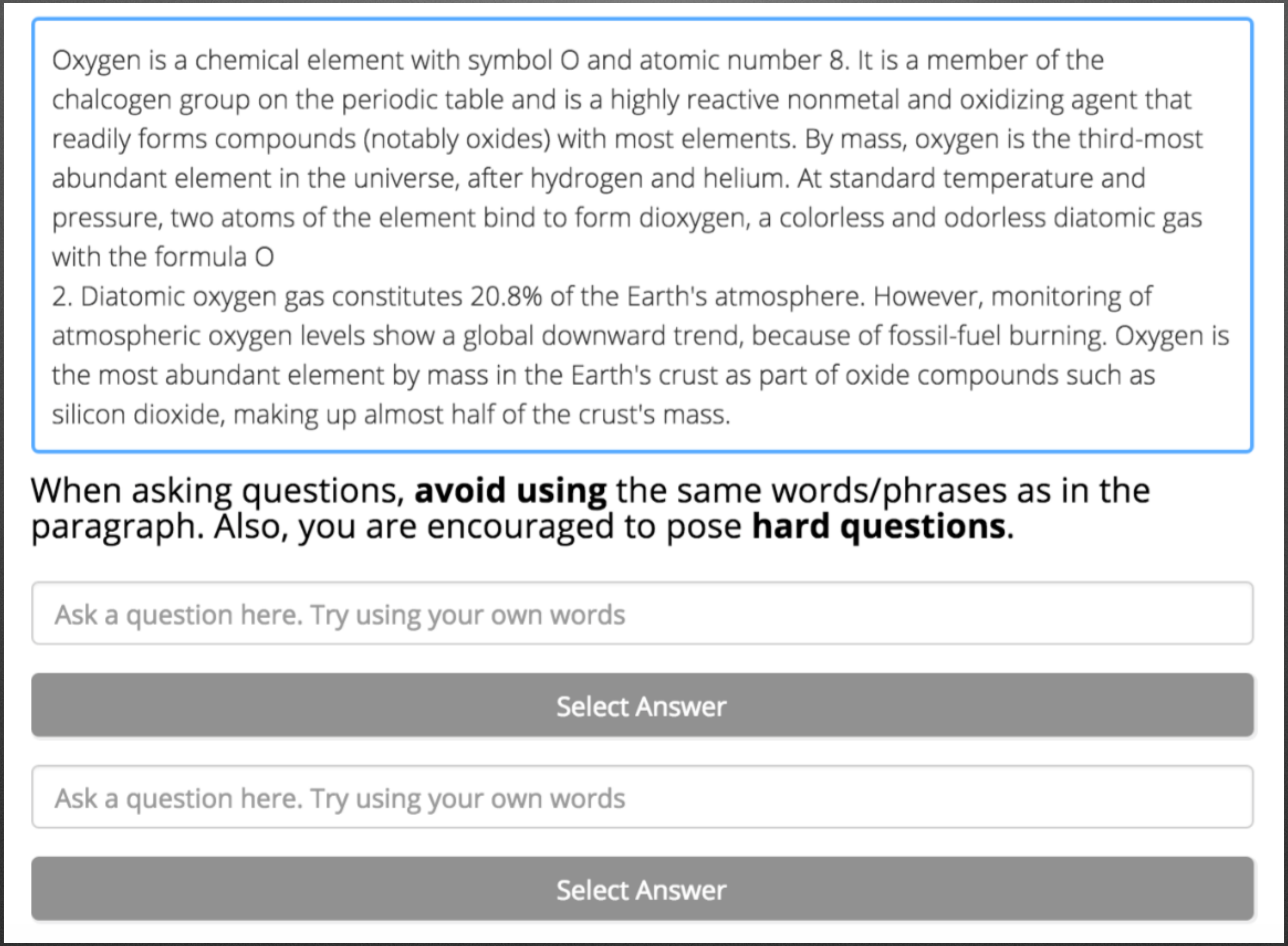
Step 1: Passage Curation

- Select top 10000 articles of English Wikipedia based on Wikipedia’s internal PageRanks scores.
- Random sample 536 articles out of 10000 articles.
- Extract passages longer than 500 characters from all 536 articles -> 23,115 paragraphs.
- Train/dev/test datasets are **split in the article level.**
- Train/dev datasets are released and test dataset is holdout.

Step 2: Question-Answer Collection

- Using crowdsourcing technique
    - Crowd-workers with 97% HIT acceptance rate, larger than 1000 HITs, and located in USA/Canada.
    - Spend 4 minutes on each paragraph and asking up to 5 questions with answer highlighted in the text.

Step 2: Question-Answer Collection



Oxygen is a chemical element with symbol O and atomic number 8. It is a member of the chalcogen group on the periodic table and is a highly reactive nonmetal and oxidizing agent that readily forms compounds (notably oxides) with most elements. By mass, oxygen is the third-most abundant element in the universe, after hydrogen and helium. At standard temperature and pressure, two atoms of the element bind to form dioxygen, a colorless and odorless diatomic gas with the formula O
2. Diatomic oxygen gas constitutes 20.8% of the Earth's atmosphere. However, monitoring of atmospheric oxygen levels show a global downward trend, because of fossil-fuel burning. Oxygen is the most abundant element by mass in the Earth's crust as part of oxide compounds such as silicon dioxide, making up almost half of the crust's mass.

When asking questions, **avoid using** the same words/phrases as in the paragraph. Also, you are encouraged to pose **hard questions**.

Ask a question here. Try using your own words
Select Answer
Ask a question here. Try using your own words
Select Answer

Step 3: Additional Answers Collection

- For each question in dev/test datasets, get at least two additional answers.
- Why we do this?
  - Make evaluation more robust.
  - Assess human performance.

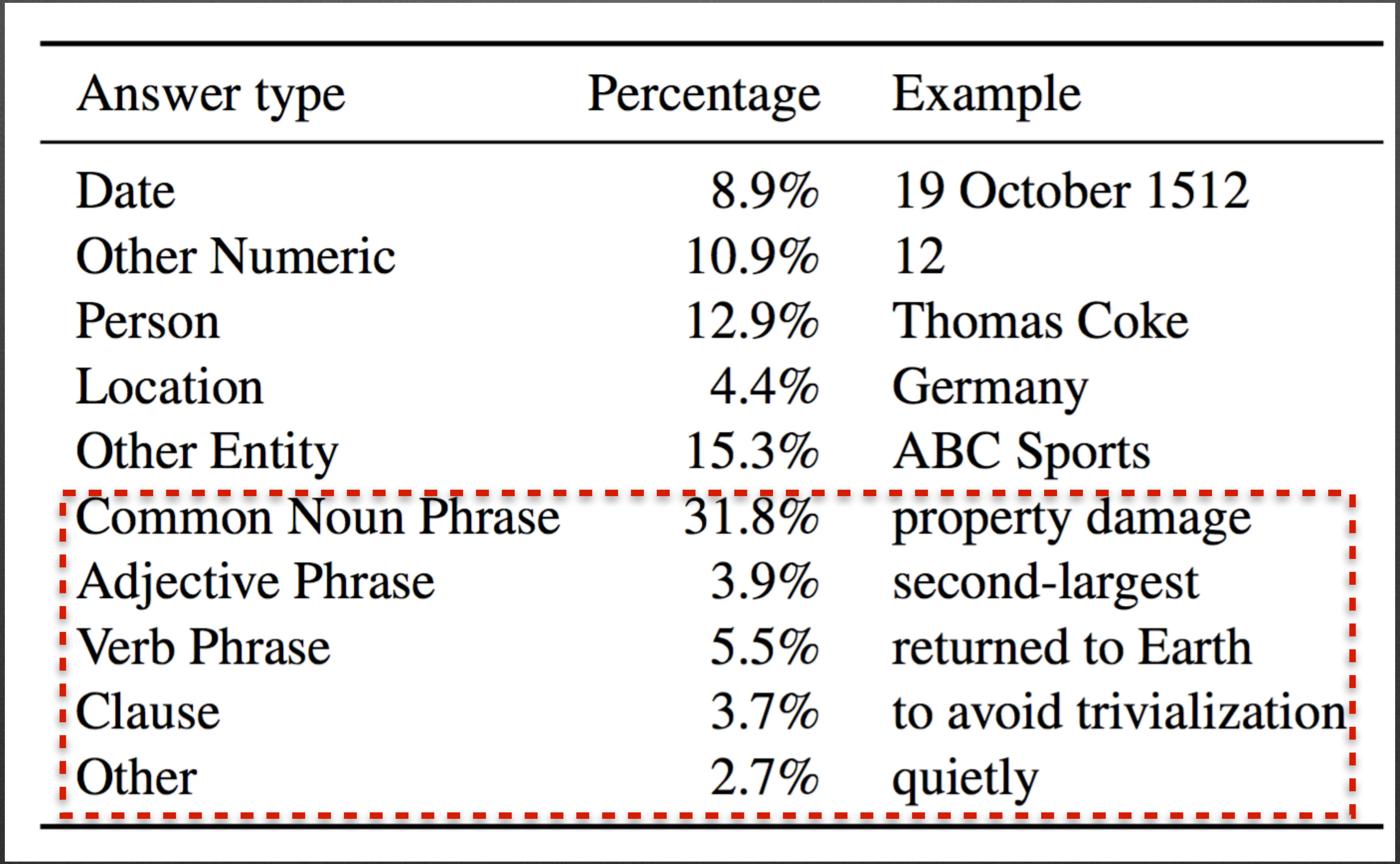
What are the properties of SQuAD?

Data Analysis

- Diversity in answers
- Reasoning for answering questions
- Syntactic divergence

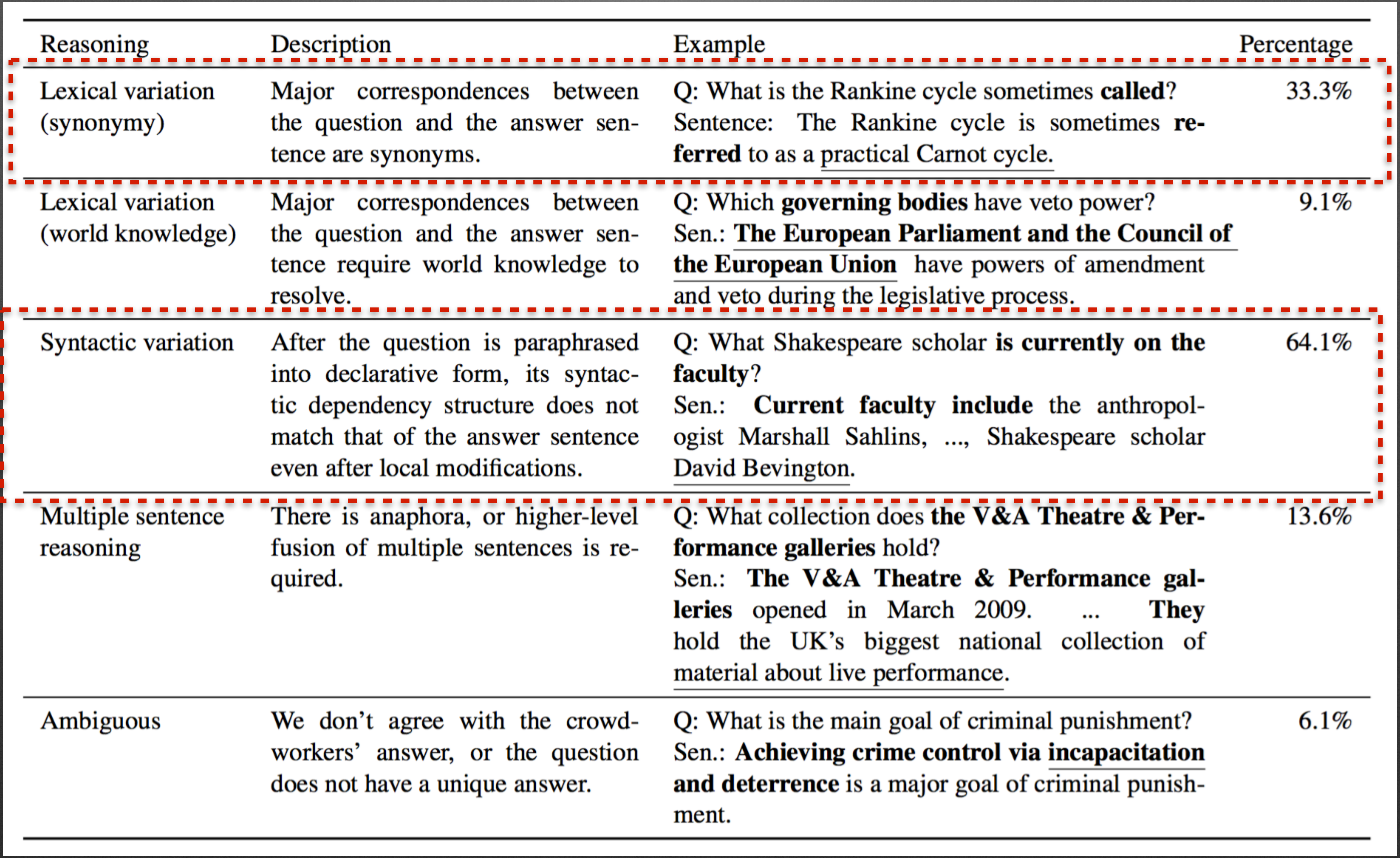
Diversity in Answers

- 67.4% non-entity answers, and many answers are not even noun -> Can be challenging.



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

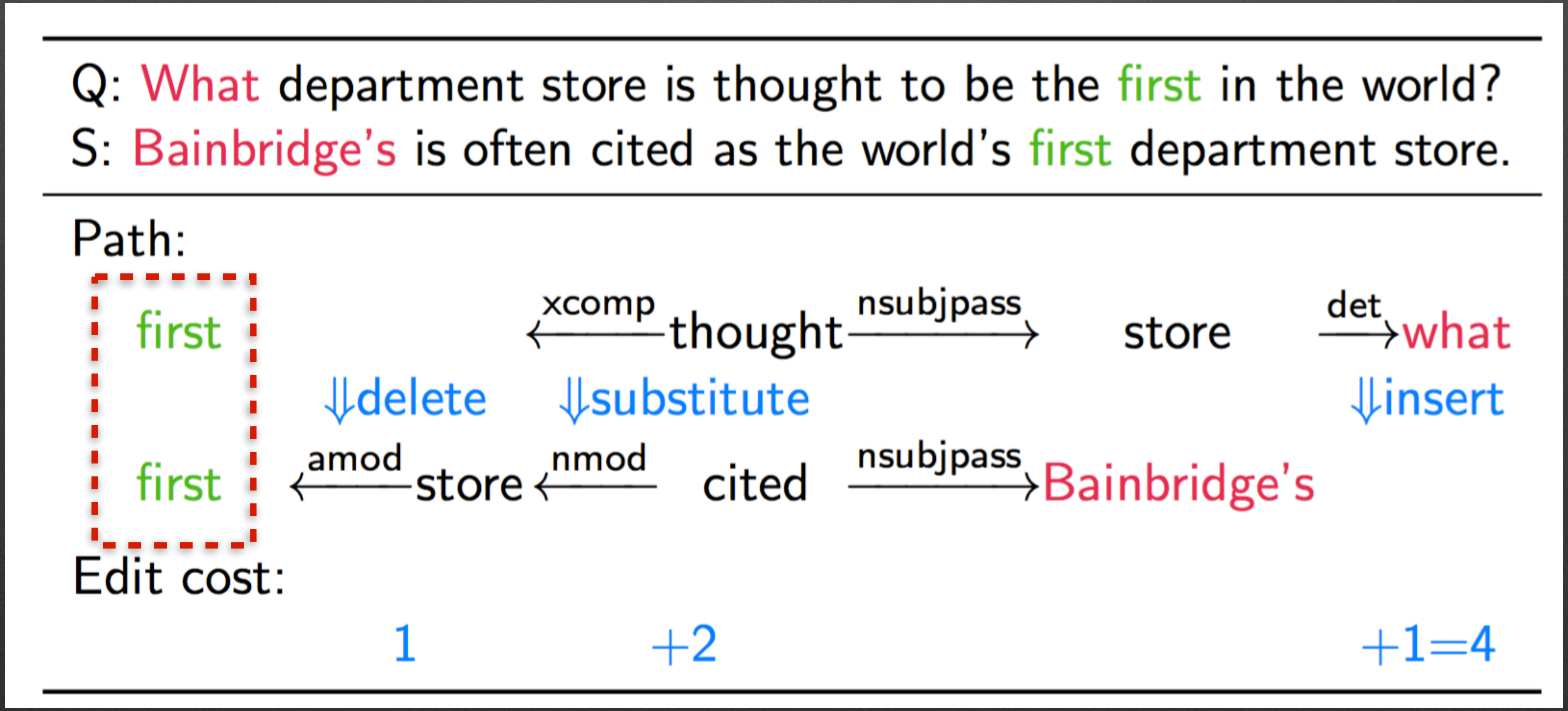
Reasoning for answering questions



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 crowd-workers' 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%

Syntactic divergence

- Syntactic divergence is the minimum edit distance (belong all unlexicalized dependency path) over all possible anchors (word-lemma pairs).



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 $\xleftarrow{\text{xcomp}}$ thought $\xrightarrow{\text{nsubjpass}}$ store $\xrightarrow{\text{det}}$ what

$\Downarrow$delete $\Downarrow$substitute $\Downarrow$insert

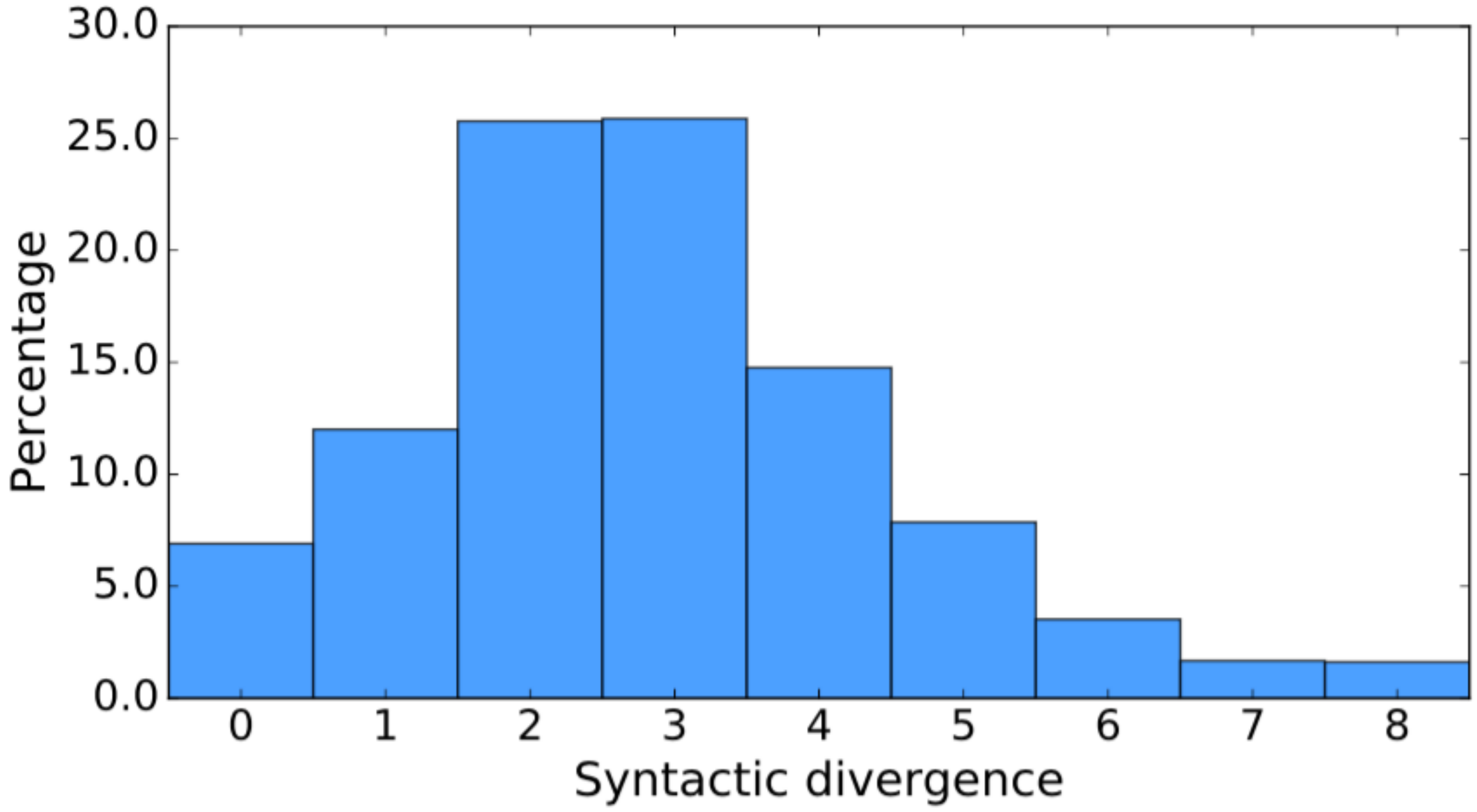
first $\xleftarrow{\text{amod}}$ store $\xleftarrow{\text{nmod}}$ cited $\xrightarrow{\text{nsubjpass}}$ Bainbridge's

Edit cost:

1 +2 +1=4

Syntactic divergence

- Histogram of syntactic divergence

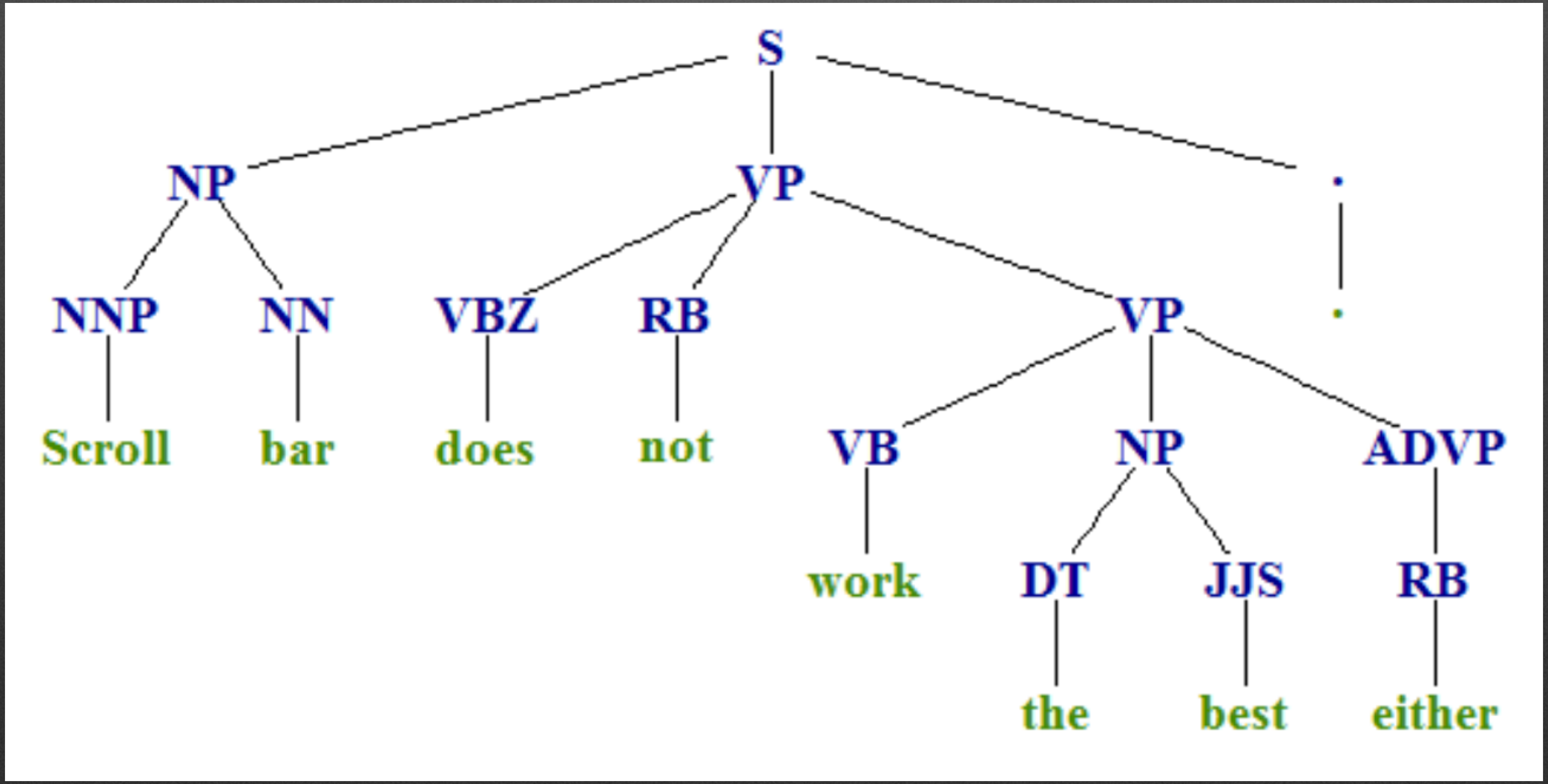


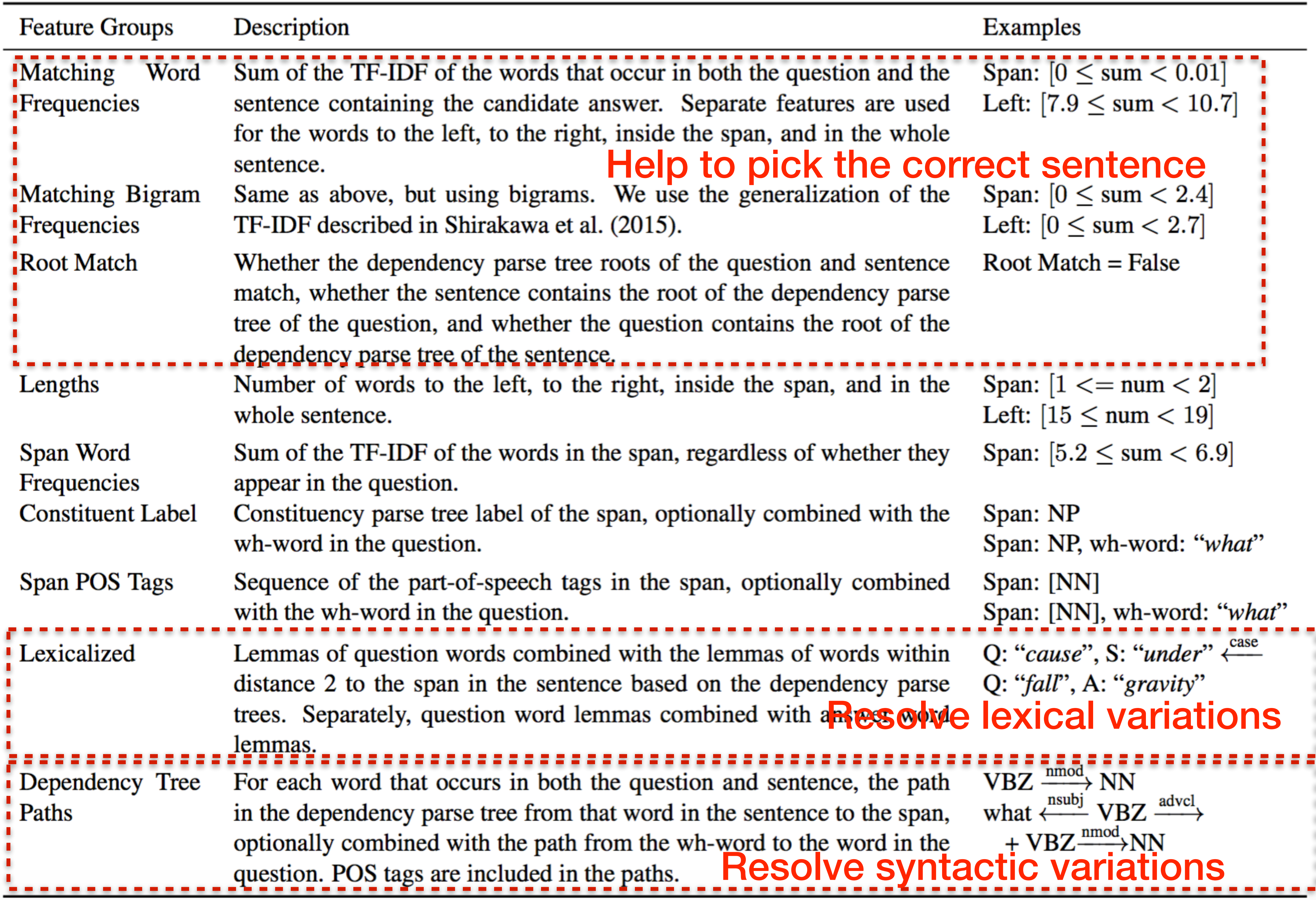
(a) Histogram of syntactic divergence.

How well we can do on SQuAD?

“Baseline” method

- Candidate answer generation: use constituency parser.
- Feature extraction
- Train a Logistic Regression Model





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

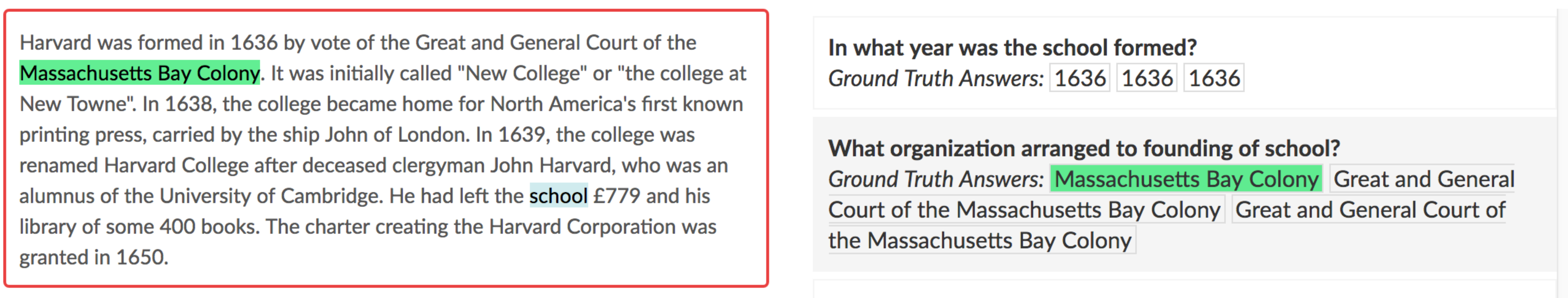
Help to pick the correct sentence

Resolve lexical variations

Resolve syntactic variations

Evaluation

- After ignoring punctuations and articles, using the following two metrics:
  - Exact Match
  - Macro-averaged F1 score: maximum F1 over all of the ground truth answers



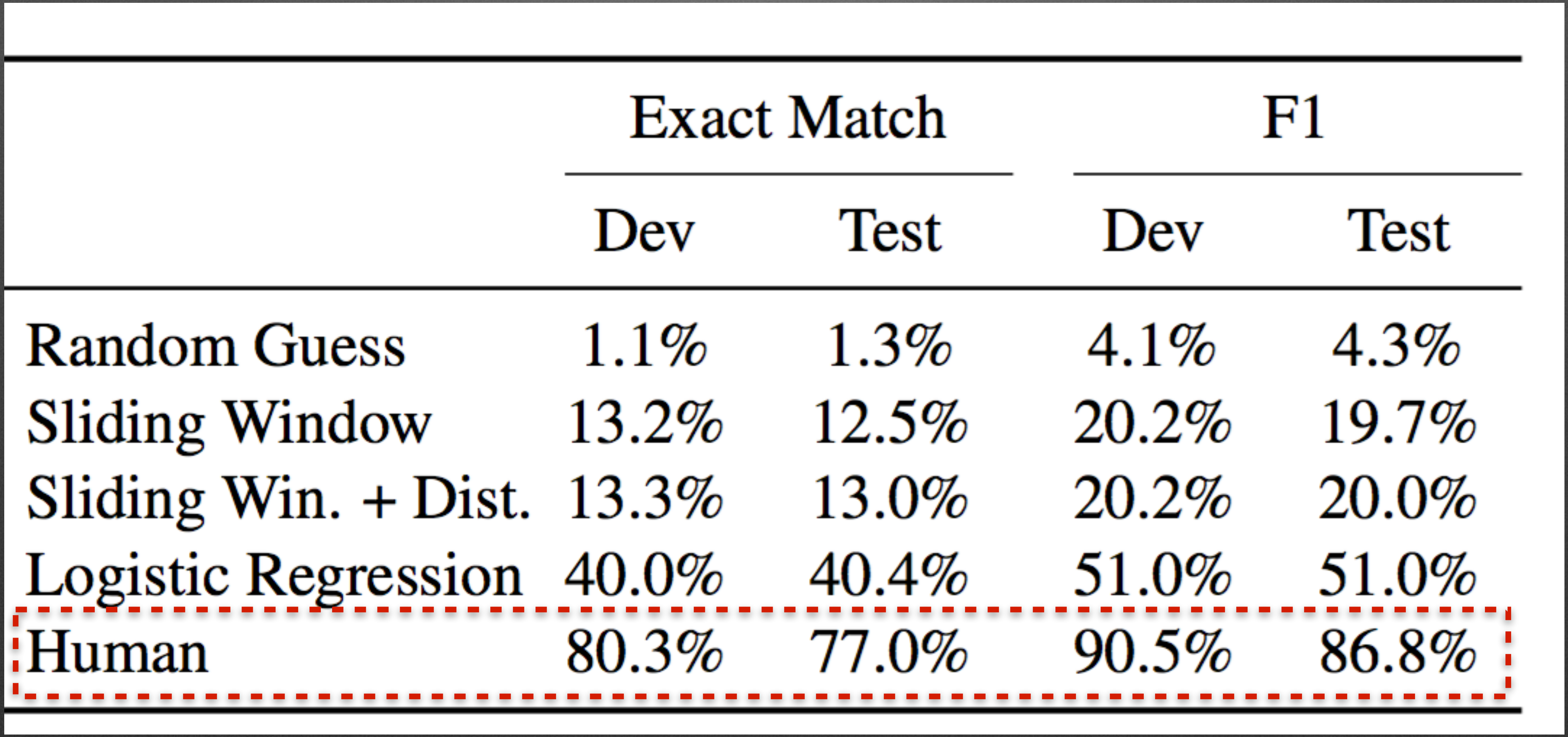
Harvard was formed in 1636 by vote of the Great and General Court of the Massachusetts Bay Colony. It was initially called "New College" or "the college at New Towne". In 1638, the college became home for North America's first known printing press, carried by the ship John of London. In 1639, the college was renamed Harvard College after deceased clergyman John Harvard, who was an alumnus of the University of Cambridge. He had left the school £779 and his library of some 400 books. The charter creating the Harvard Corporation was granted in 1650.

In what year was the school formed?
Ground Truth Answers: 1636 1636 1636

What organization arranged to founding of school?
Ground Truth Answers: Massachusetts Bay Colony Great and General Court of the Massachusetts Bay Colony Great and General Court of the Massachusetts Bay Colony

Experiment results

- Overall results

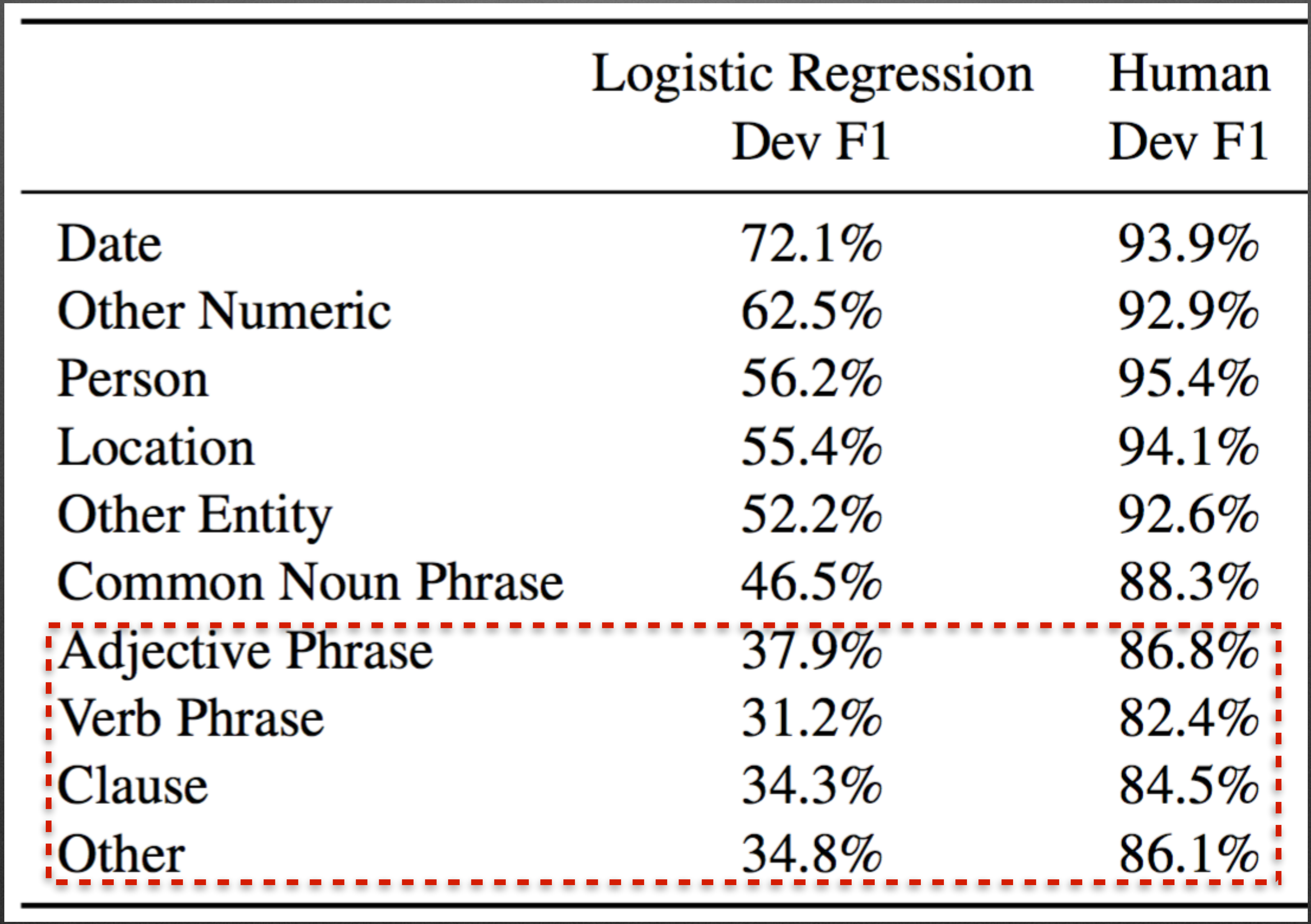


	Exact Match		F1	
	Dev	Test	Dev	Test
Random Guess	1.1%	1.3%	4.1%	4.3%
Sliding Window	13.2%	12.5%	20.2%	19.7%
Sliding Win. + Dist.	13.3%	13.0%	20.2%	20.0%
Logistic Regression	40.0%	40.4%	51.0%	51.0%
Human	80.3%	77.0%	90.5%	86.8%

For SQuAD v1.1 Test: EM: 82.304 F1: 91.221

Experiment results

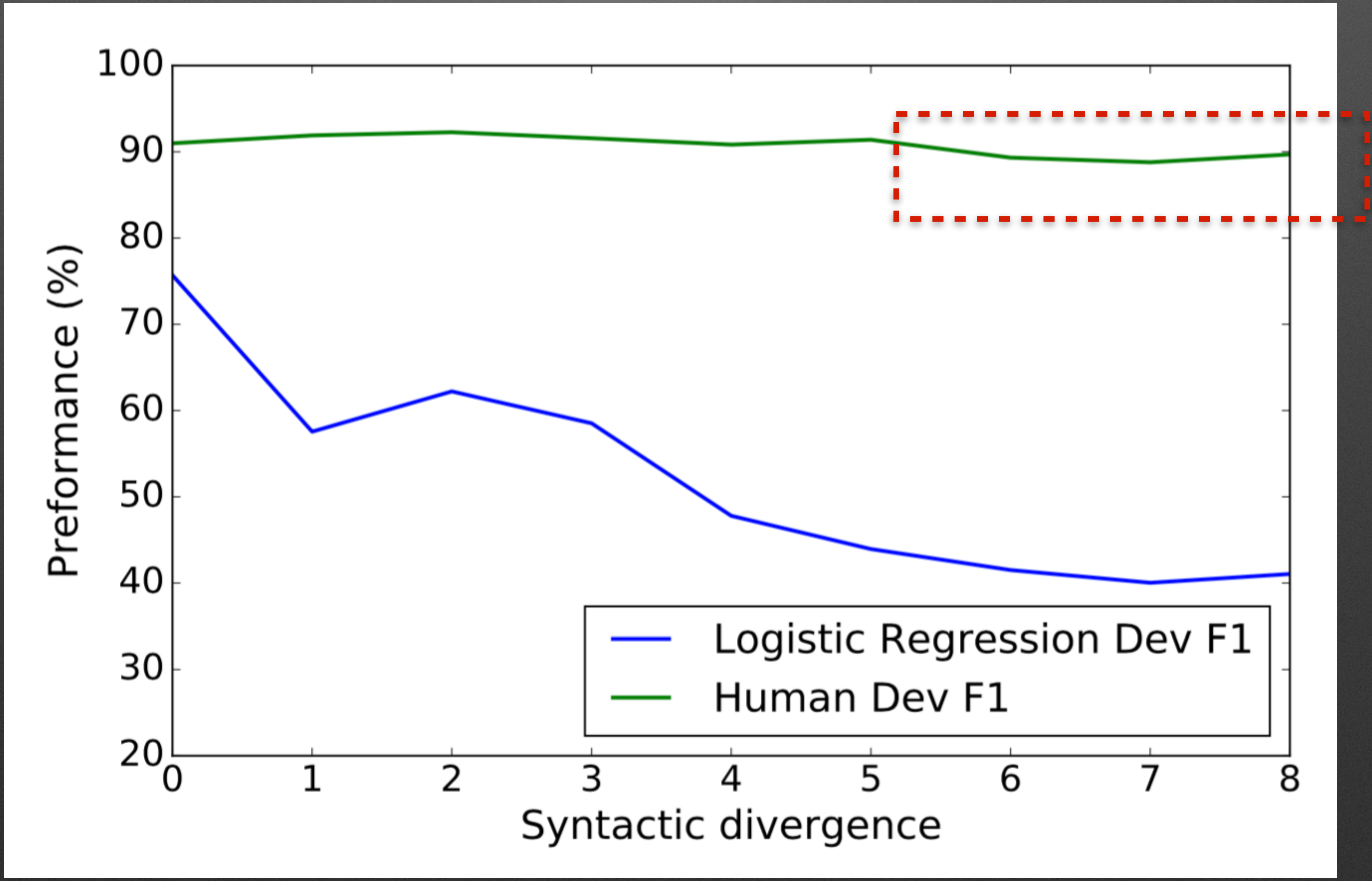
- Performance stratified by answer type



	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%

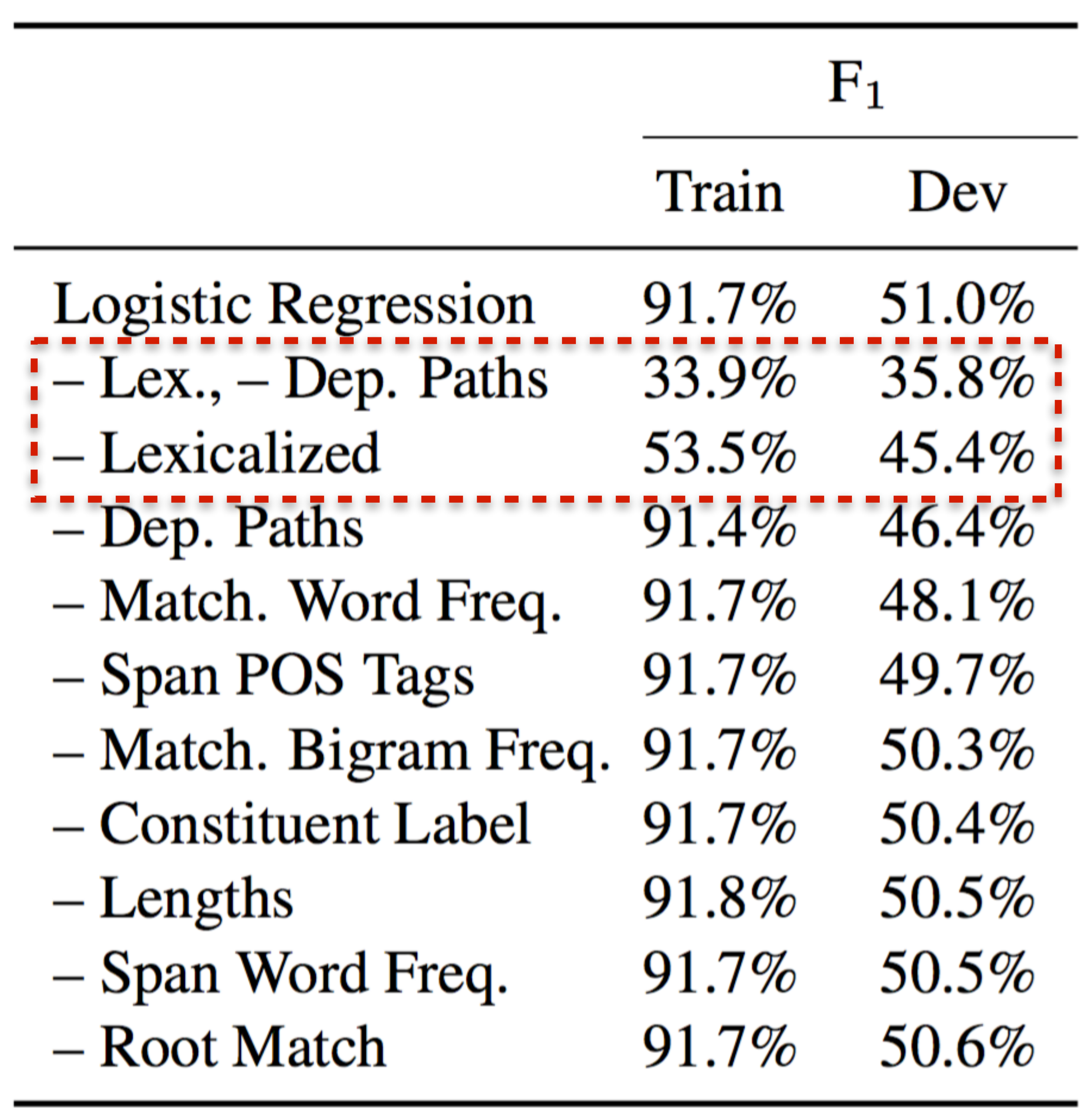
Experiment results

- Performance stratified by syntactic divergence



Experiment results

- Performance with feature ablations



	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%

Summary

- SQuAD is a machine reading style QA dataset.
- SQuAD consists of 100,000+ QA pairs.
- SQuAD is constructed based on crowdsourcing.
- SQuAD drives the field forward.

Thanks

Q & A