Question Answering In Context

Eunsol Choi*, He He*, Mohit Iyyer*, Mark Yatskar*
Scott Yih, Yejin Choi, Percy Liang, Luke Zettlemoyer

** : the authors contributed equally*

http://quac.ai

Overview

Question answering is a multi-turn task where you ask further questions based on what you have learned. We present a large-scale, multi-turn question answering dataset, simulating an information seeking dialogue.
Each interaction is between a student and a teacher on a text. A student poses questions to learn as much as possible about a hidden Wikipedia text, and a teacher who answers the questions by providing short excerpts from the text.

Task Setup

	Student	*Teacher*
Given	Article / Section Title/ Article Summary	Section Text
Provides	Ask a question to learn as much as possible about this topic!	Answer the questions by choosing **a span** from the text or return **no answer** Give a feedback to the question! FollowUp ↪ Maybe ⇢ Don't ↛ Yes / No Yes No Not a Yes / No

Dataset Statistics

avg turns per dialogue	7.2
avg tokens per question	6.5
avg tokens per answer	14.1
avg tokens per section	401

	Train	Eval	Total
# unique sections	6,843	2,002	8,854
# of dialogues	11,567	2,002	13,594
# QA pairs	83,568	14,707	98,407

Collected Dialogue Examples

Section: Daffy Duck, Origin & History

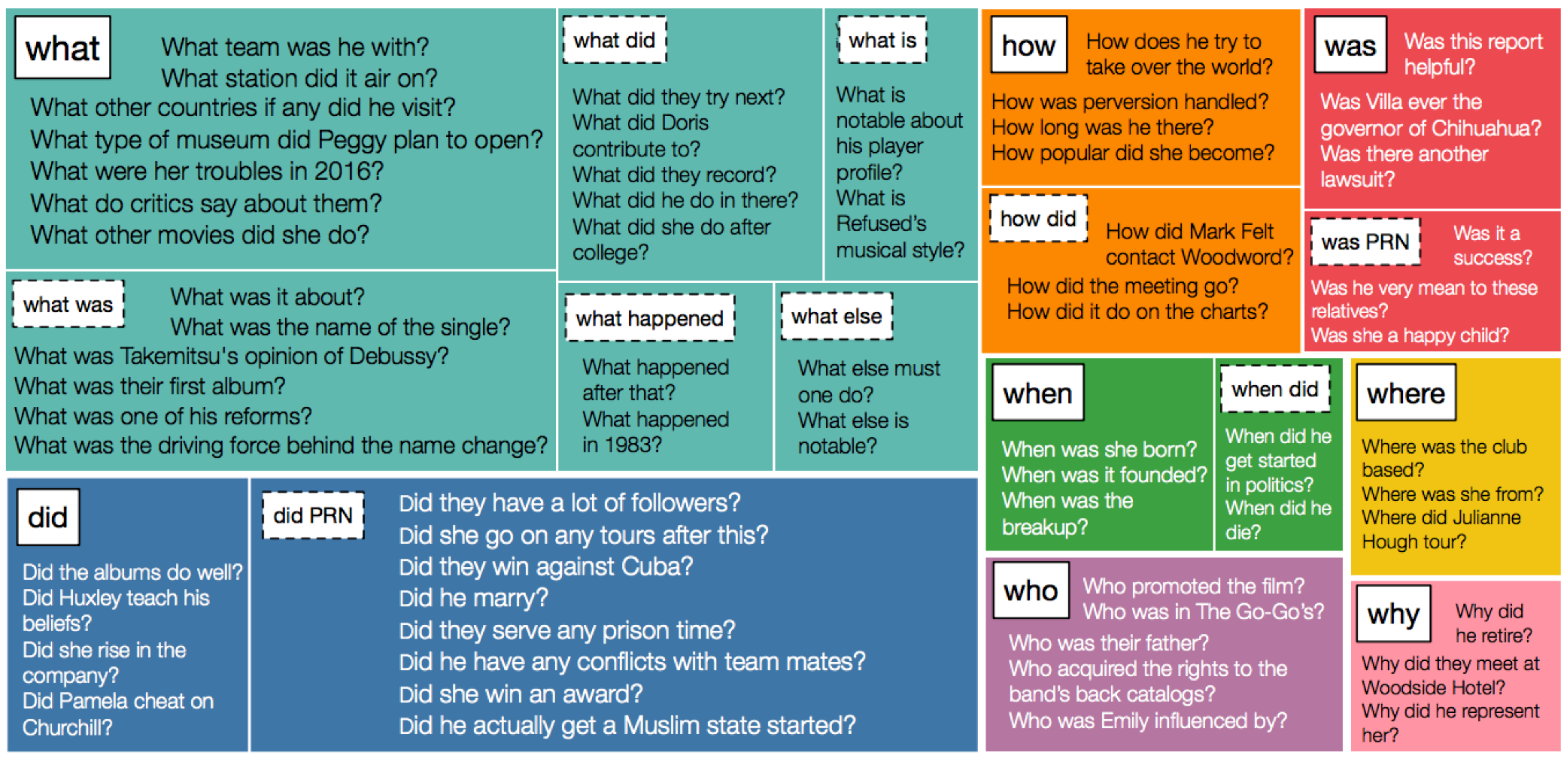
STUDENT: **What is the origin of Daffy Duck?**
TEACHER: ↪ first appeared in Porky's Duck Hunt
STUDENT: **What was he like in that episode?**
TEACHER: ↪ assertive, unrestrained, combative
STUDENT: **Was he the star?**
TEACHER: ⇢ No, barely more than an unnamed bit player in this short
STUDENT: **Who was the star?**
TEACHER: ↛ No answer
STUDENT: **Did he change a lot from that first episode in future episodes?**
TEACHER: ↪ Yes, the only aspects of the character that have remained consistent (...) are his voice characterization by Mel Blanc
STUDENT: **How has he changed?**
TEACHER: ↪ Daffy was less anthropomorphic
STUDENT: **In what other ways did he change?**
TEACHER: ↪ Daffy's slobbery, exaggerated lisp (...) is barely noticeable in the early cartoons.
STUDENT: **Why did they add the lisp?**
TEACHER: ↪ One often-repeated "official" story is that it was modeled after producer Leon Schlesinger's tendency to lisp.
STUDENT: **Is there an "unofficial" story?**
TEACHER: ↪ Yes, Mel Blanc (...) contradicts that conventional belief
. . .

Section: Augusto Pinochet : Intellectual life...

STUDENT: **Was he known for being intelligent?**
TEACHER: ↪ No, Pinochet was publicly known as a man with a lack of culture.
STUDENT: **why did people feel that way?**
TEACHER: ↪ reinforced by the fact that he also portrayed himself as a common man
STUDENT: **did he have any hobbies?**
TEACHER: ↪ Yes, Before wresting power from Allende, Pinochet had written two books.
STUDENT: **what is the name of a book written by him?**
TEACHER: ↪ Geopolitica (1968) and Campana de Tarapaca (1972).
STUDENT: **what were the books about?**
TEACHER: ↪ Chile's military literature.
STUDENT: **was there anything noteworthy regarding his books?**
TEACHER: ↪ Yes, In Geopolitica Pinochet plagiarized (...) Gregorio Rodriguez Tascon
STUDENT: **did he deny those allegations?**
TEACHER: ↛ No answer
STUDENT: **what did he plagiarize in Geopolitica?**
TEACHER: ⇢ In Geopolitica Pinochet plagiarized (...) paragraphs from a 1949 presentation
. . .

Question Analysis

- Manual analysis showed about half the questions are not factoid. (Asking more open ended question such as "was the show success?"
- Also 80% of the questions are contextual, on the conversation history (40%) and to the document (60%)
- About 20% questions are unanswerable
- 40% questions have multiple possible answer span in the text



what
What team was he with?
What station did it air on?
What other countries if any did he visit?
What type of museum did Peggy plan to open?
What were her troubles in 2016?
What do critics say about them?
What other movies did she do?

what did
What did they try next?
What did Doris contribute to?
What did they record?
What did he do in there?
What did she do after college?

what is
What is notable about his player profile?
What is Refused's musical style?

what was
What was it about?
What was the name of the single?
What was Takemitsu's opinion of Debussy?
What was their first album?
What was one of his reforms?
What was the driving force behind the name change?

what happened
What happened after that?
What happened in 1983?

what else
What else must one do?
What else is notable?

did
Did the albums do well?
Did Huxley teach his beliefs?
Did she rise in the company?
Did Pamela cheat on Churchill?

did PRN
Did they have a lot of followers?
Did she go on any tours after this?
Did they win against Cuba?
Did he marry?
Did they serve any prison time?
Did he have any conflicts with team mates?
Did she win an award?
Did he actually get a Muslim state started?

how
How does he try to take over the world?
How was perversion handled?
How long was he there?
How popular did she become?

how did
How did Mark Felt contact Woodword?
How did the meeting go?
How did it do on the charts?

was
Was this report helpful?
Was Villa ever the governor of Chihuahua?
Was there another lawsuit?

was PRN
Was it a success?
Was he very mean to these relatives?
Was she a happy child?

when
When was she born?
When was it founded?
When was the breakup?

when did
When did he get started in politics?
When did he die?

where
Where was the club based?
Where was she from?
Where did Julianne Hough tour?

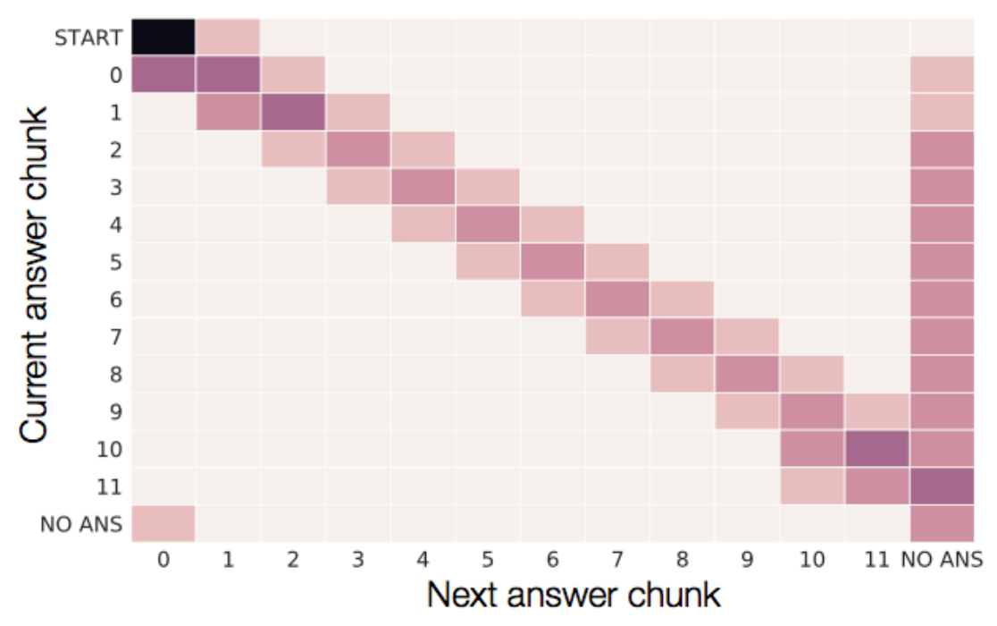
who
Who promoted the film?
Who was in The Go-Go's?
Who was their father?
Who acquired the rights to the band's back catalogs?
Who was Emily influenced by?

why
Why did he retire?
Why did they meet at Woodside Hotel?
Why did he represent her?

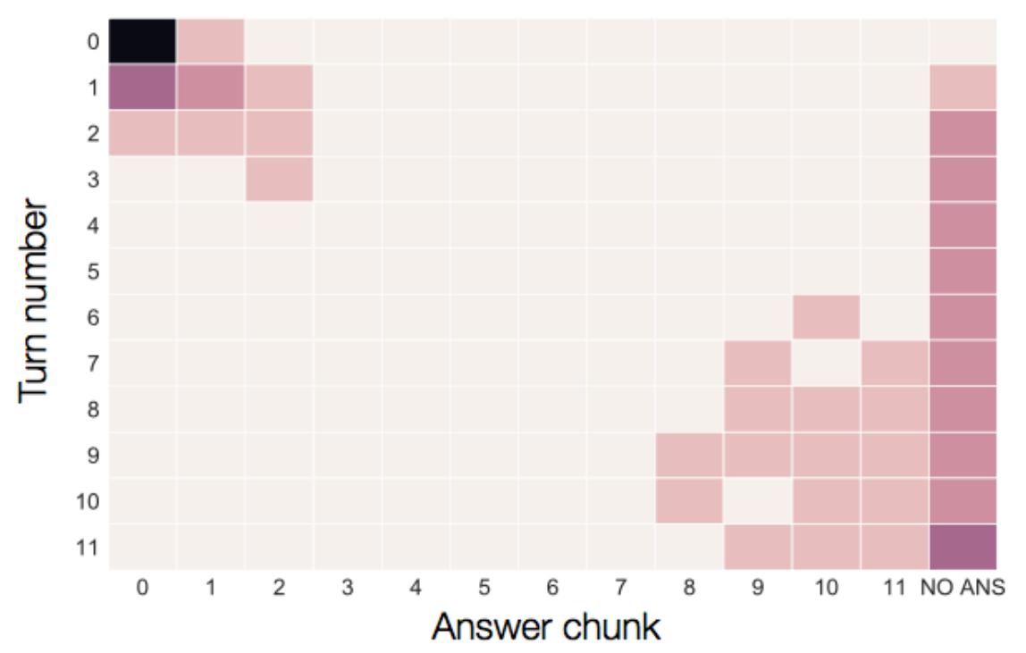
Treemap visualization of the eight most frequent "Wh" word in the dataset, where the box area is proportional to number of occurrences

The Role of Context

We show the location of answer within the text (where each text is divided into 12 chunks of equal size) in the following figures as a heatmap.



Location of next answer given the location of current answer: Answer to the next question is near the current answer.



Answer location by the position in the dialogue: Even without the access to the document, student's question roughly follows text in a linear order.

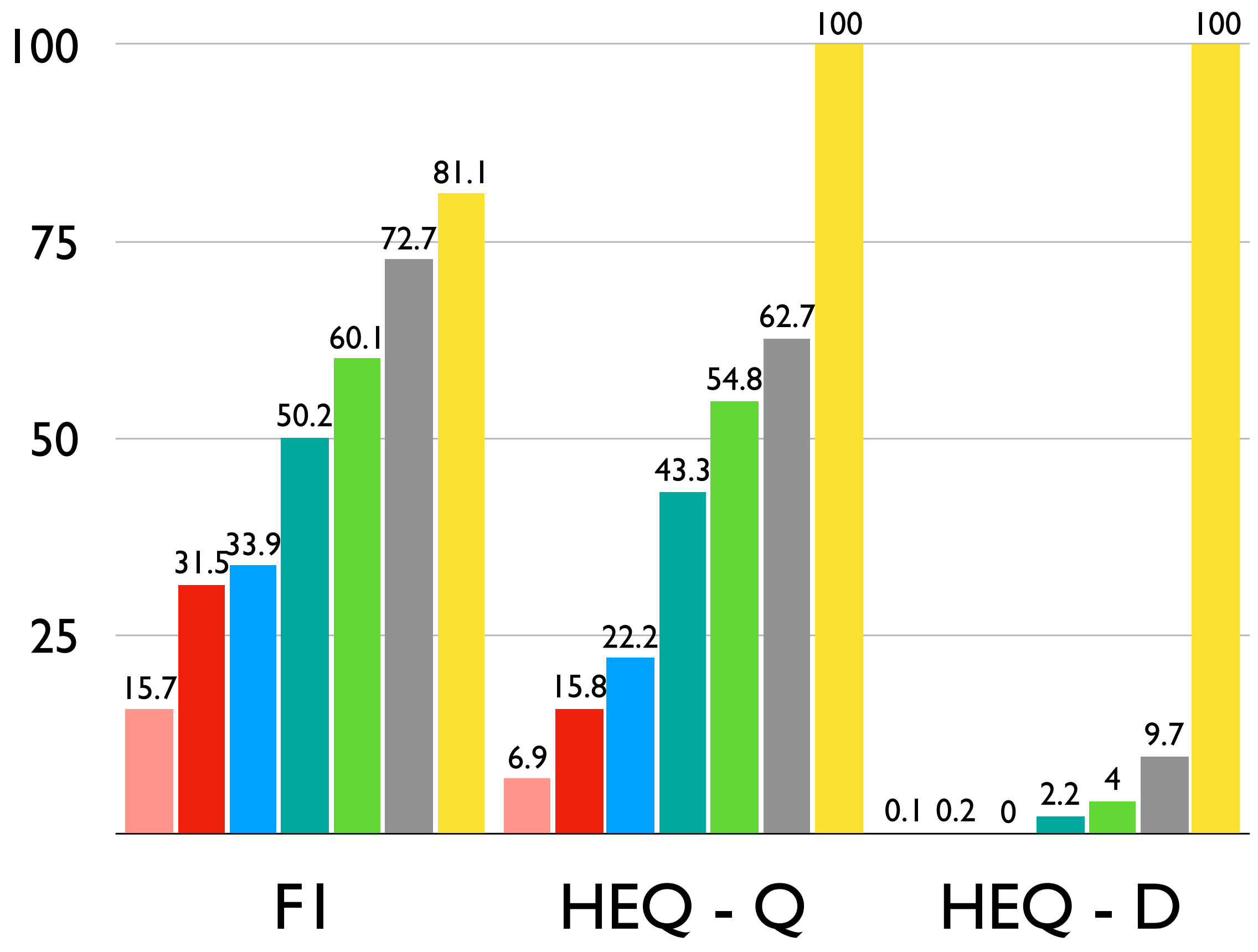
Experiments

Evaluation Measures

Word-level F1: Precision and recall computed by the words in predictions / words in the reference
Human Equivalence Score (HEQ): Whether system's output is as good as that of an average human (we collected 5 annotated answer spans per questions in evaluation data)

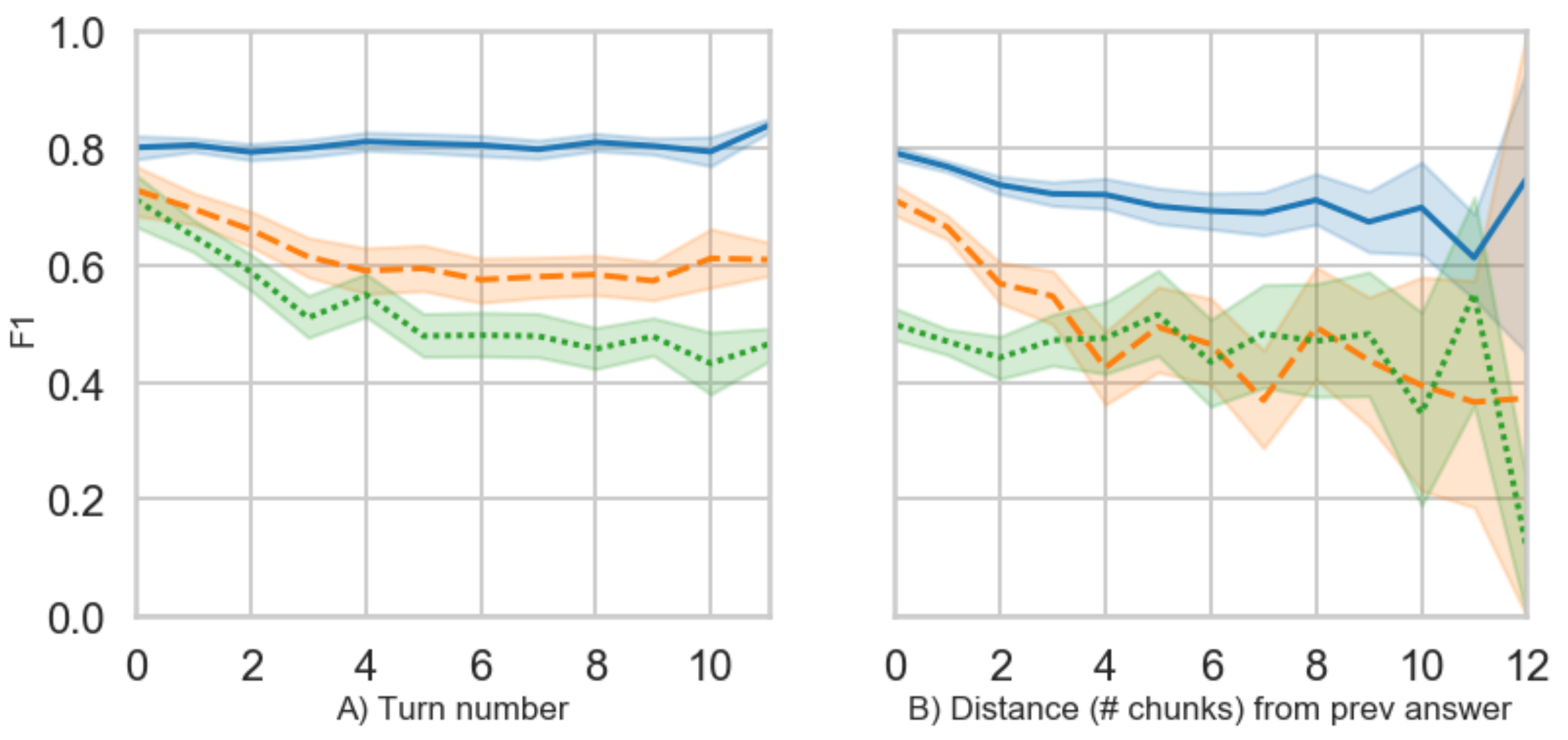
- HEQ-Q: percentage of questions,
- HEQ-D: the percentage of dialogs for which it is true for every Q

Models	🟥	Sanity check	Random Sent	Randomly selects from sentences or no answer option.
	🟥		Transition Matrix	Select the most likely chunk given the previous chunk.
	🟦	Base line	Logistic Regression	Feature-rich model for answer sentence selection (n-gram overlap, bias, context feature)
	🟦		BiDAF ++	A model with a strong performance on existing QA data. Token for 'no answer' appended.
	🟩		with N context	Question turn number is marked, as well as the previous answers in the tcontext
	⬛	Upper Bounds	Oracle Sentence	Selects a sentence with highest F1 or No answer.
	🟨		Humans	We pick one reference as a system output and compute F1 against remaining references.



- There is a large gap (> 20%) between human's and the best model's performance
- High human scores demonstrates high inter-annotator agreement.

Performance Analysis



- Better modeling of dialog history will help.

QuAC Home
QuAC
Question Answering in Context
What is QuAC?
Leaderboard

Data / Model / Leaderboard at http://quac.ai