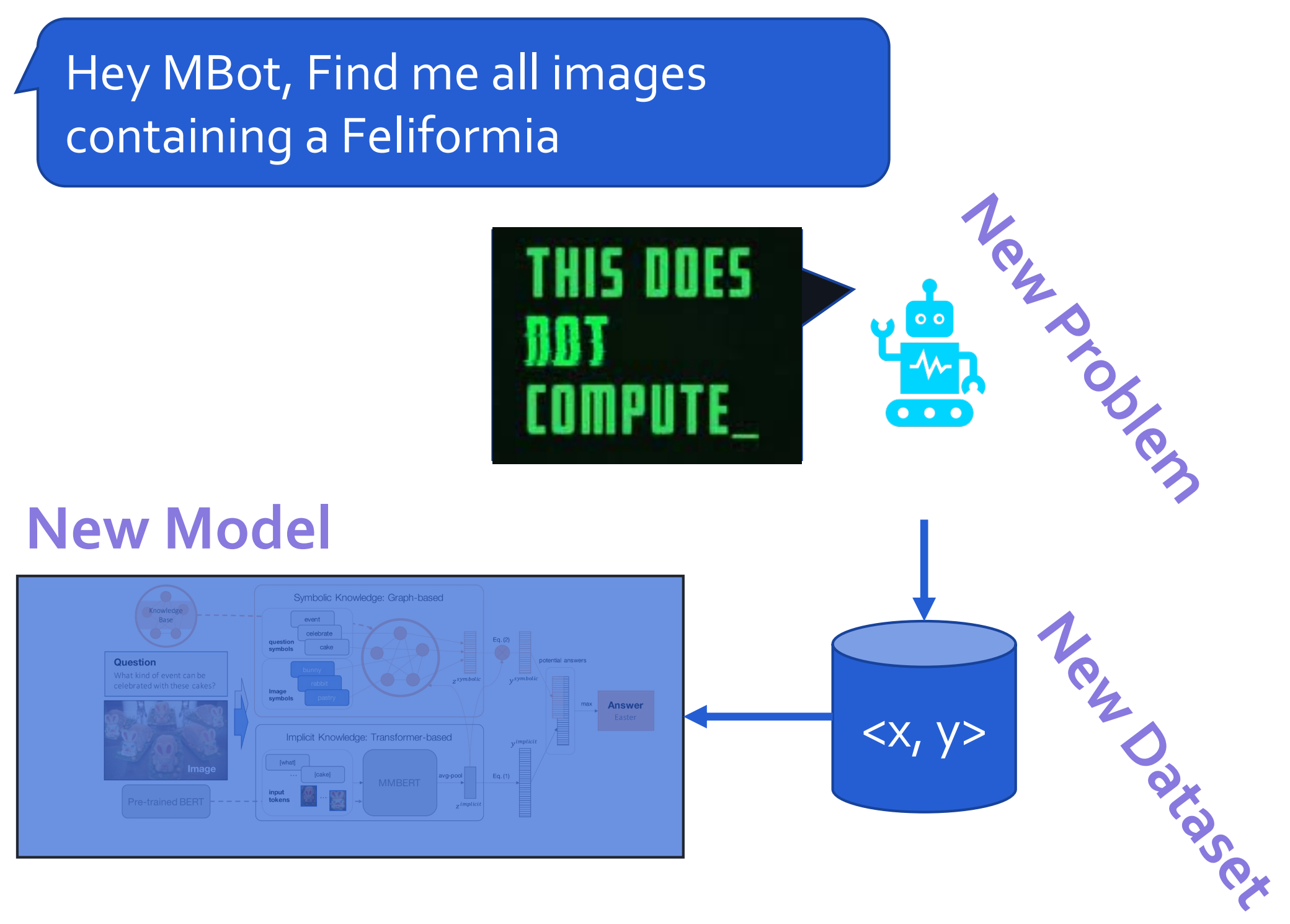
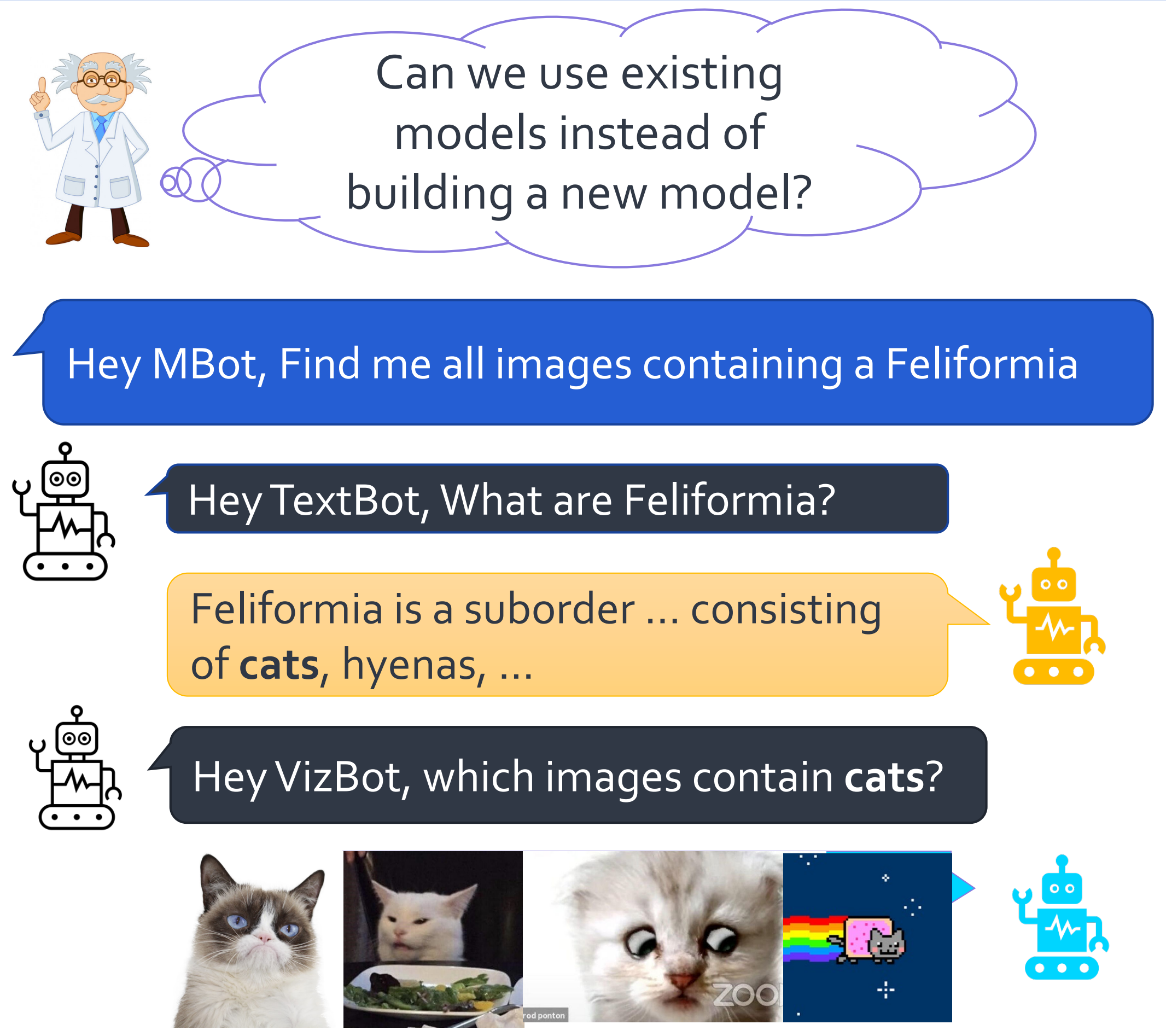
Text Modular Networks: Learning to Decompose Tasks in the Language of Existing Models

Tushar Khot, Daniel Khashabi, Kyle Richardson, Peter Clark & Ashish Sabharwal
{tushark, danielk, kyler, peterc, ashishs} @allenai.org

Motivation

Standard Approach:
New Dataset ➔ New Model




Can we use existing models instead of building a new model?
Hey MBot, Find me all images containing a Feliformia
Hey TextBot, What are Feliformia?
Feliformia is a suborder … consisting of **cats**, hyenas, …
Hey VizBot, which images contain **cats**?

Given a set of existing QA models, can one leverage them to answer complex questions by communicating with these existing models?

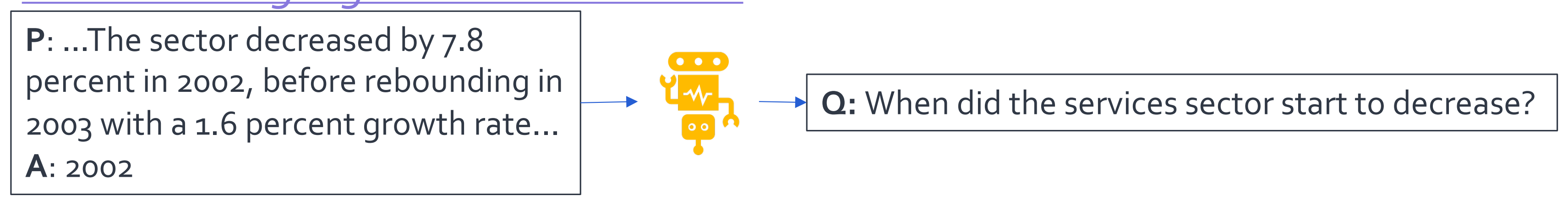
Text Modular Networks

Contribution 1:
Text Modular Networks (TMNs): A general framework that can *leverage existing simpler models -- neural and symbolic --* as blackboxes for answering complex questions.

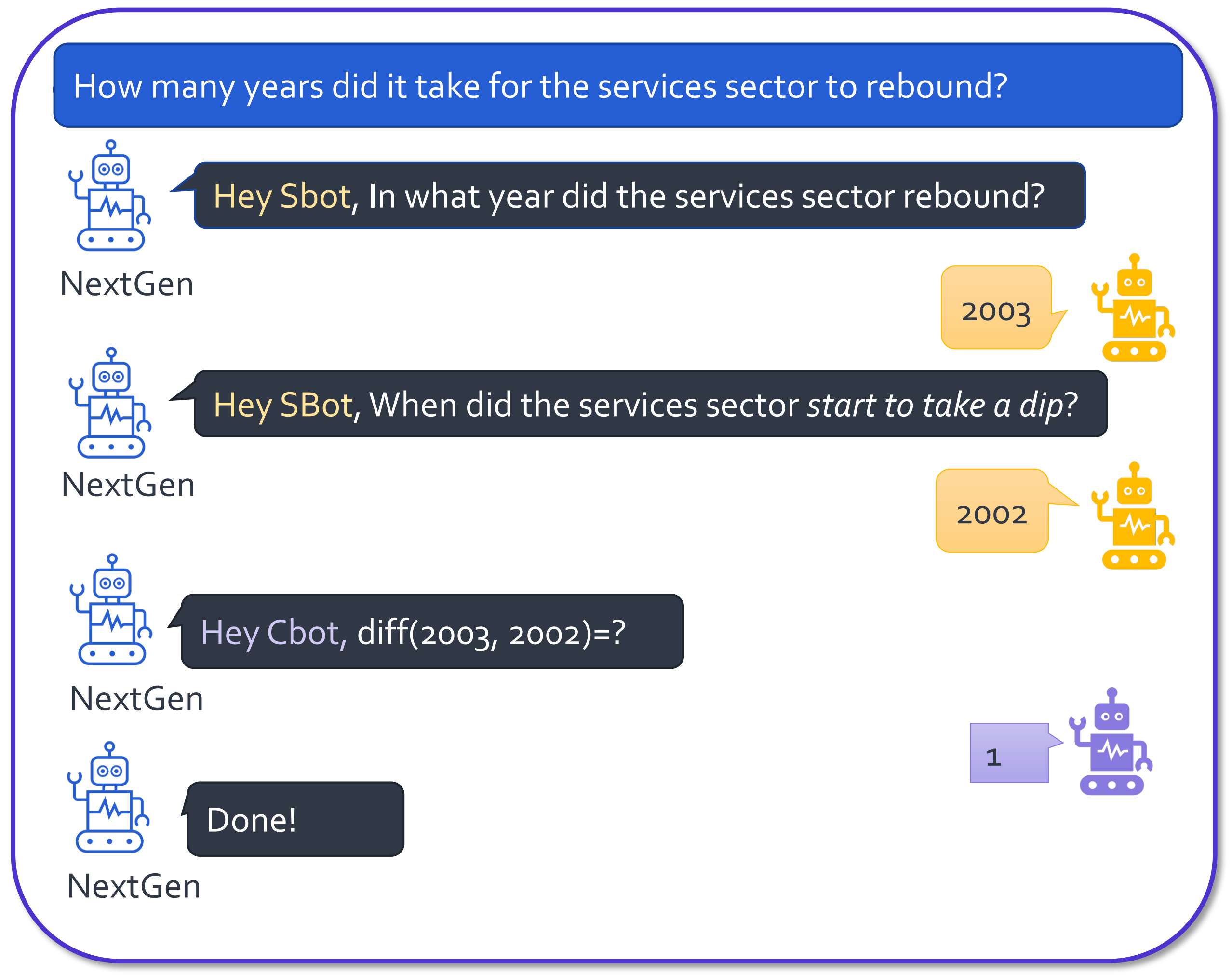
Key problem: Training NextGen to ask the right questions to the appropriate models in the model's language

Solution:

1. Learn the *language* of the sub-models


P: …The sector decreased by 7.8 percent in 2002, before rebounding in 2003 with a 1.6 percent growth rate…
A: 2002
Q: When did the services sector start to decrease?

2. Decompose complex tasks into this language: Use to generate questions for each reasoning step
3. Train NextGen to generate decompositions: Given previous history of questions, train to generate the next question (and model)


How many years did it take for the services sector to rebound?
NextGen
Hey Sbot, In what year did the services sector rebound?
2003
NextGen
Hey SBot, When did the services sector start to take a dip?
2002
NextGen
Hey Cbot, diff(2003, 2002)=?
1
NextGen
Done!

ModularQA

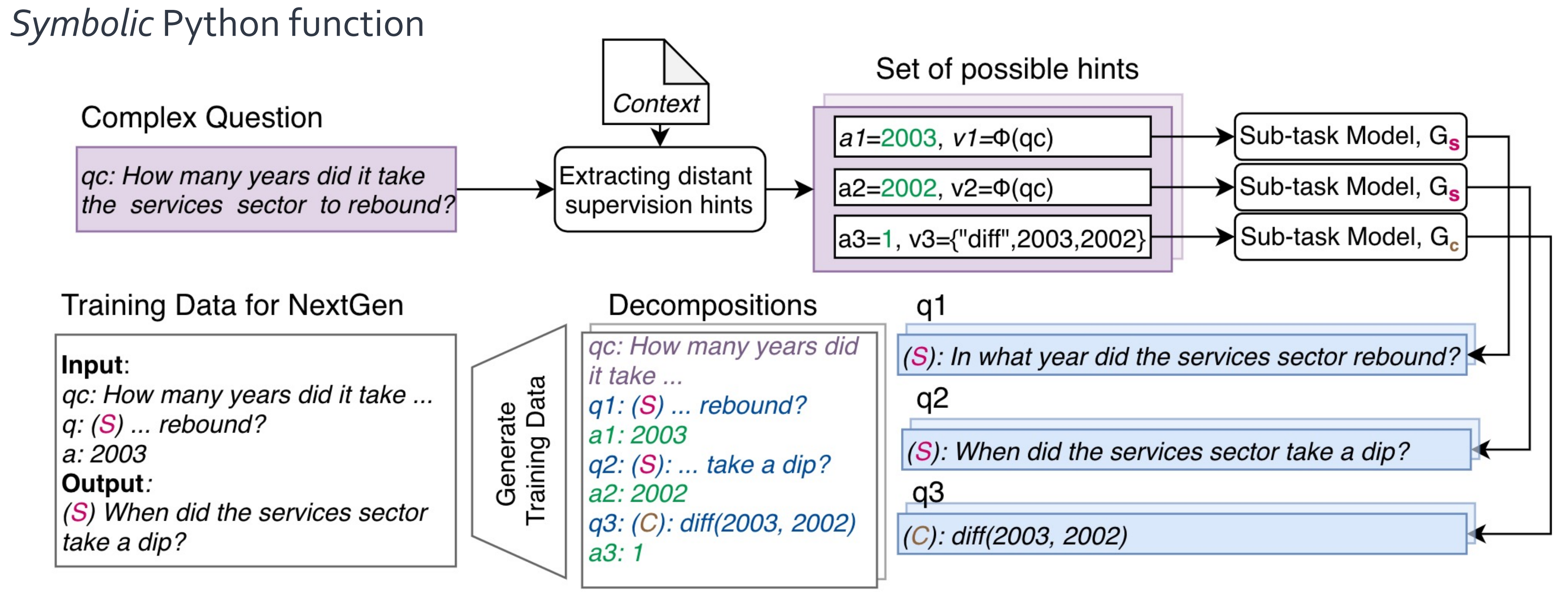
Contribution 2:
ModularQA: An implementation of the TMN framework that learns to decompose *multi-hop and discrete reasoning* questions.

Sub-Tasks & Models:

- Task: Reading Comprehension
- Model: *Neural* model trained on SQuAD
- Task: Basic Math Calculation
- Model: *Symbolic* Python function

Complex Tasks: Multi-hop + discrete reasoning

- HotpotQA (conjunction, composition, comparison)
- DROP Subset (comparison, difference, complementation)


Complex Question
qc: How many years did it take the services sector to rebound?
Context
Extracting distant supervision hints
Set of possible hints
a1=2003, v1=Φ(qc)
a2=2002, v2=Φ(qc)
a3=1, v3={"diff",2003,2002}
Sub-task Model, G_S
Sub-task Model, G_S
Sub-task Model, G_C
Training Data for NextGen
Input:
qc: How many years did it take …
q: (S) … rebound?
a: 2003
Output:
(S) When did the services sector take a dip?
Generate Training Data
Decompositions
qc: How many years did it take …
q1: (S) … rebound?
a1: 2003
q2: (S): … take a dip?
a2: 2002
q3: (C): diff(2003, 2002)
a3: 1
q1
(S): In what year did the services sector rebound?
q2
(S): When did the services sector take a dip?
q3
(C): diff(2003, 2002)

Results

No decomposition annotations needed!

Sample Decompositions

12 Years a Slave starred what British actor born 10 July 1977)
Q: Who stars in 12 Years a Slave? A: Chiwetel Ejiofor
Q: Who is the British actor born 10 July 1977? A: Chiwetel Umeadi Ejiofor

How many children's books has the writer of the sitcom Maid Marian and her Merry Men written ?
Q: What writer was on Maid Marian and her Merry Men? A: Tony Robinson
Q: How many children's books has Tony Robinson written? A: sixteen

Did Holland's Magazine and Moondance both begin in 1996?
Q: When did Holland's Magazine begin? A:1876
Q: When did Moondance begin? A:1996
Q: `if_then`(1876=1996, no, yes) A: no

How many days passed between the Sendling Christmas Day Massacre and the Battle of Aidenbach?
Q: When was the Battle of Aidenbach? A: 8 January 1706
Q: When was the Sendling Christmas Massacre? A: 25 December 1705
Q: diff(8 January 1706, 25 December 1705, days) A: 14

Contribution 3:
A system that is more *robust, versatile, sample-efficient and interpretable*

More versatile

Modular

	DROP F1	HotpotQA
ModularQA	87.9	61.8
NMN-D	79.1*	🟥
SNMN	🟥	63.1

Black-box

	DROP F1	HotpotQA
ModularQA	87.9	61.8
NumNet+V2	91.6	🟥
Quark	🟥	75.5

More robust

Contrast Test	DROP EM	DROP F1
ModularQA	<u>55.7</u>	<u>63.3</u>
NumNet+V2	45.2	56.2

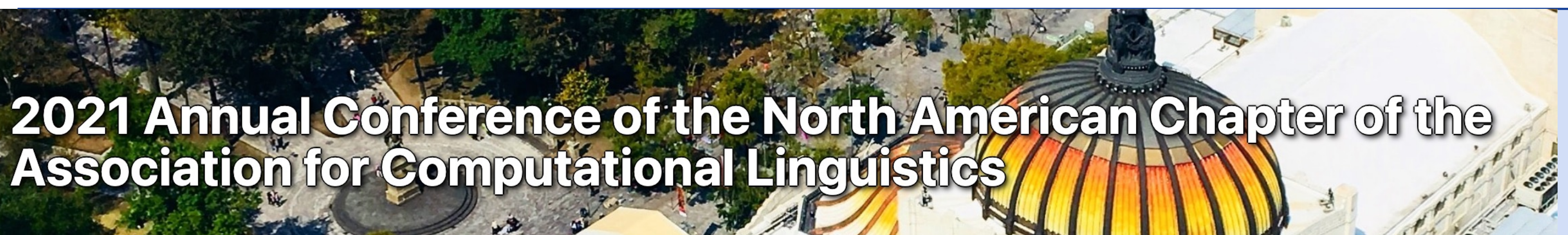
Sample Efficient

Training Set %	100%	60%	20%
ModularQA	87.8	<u>89.3</u>	<u>87.0</u>
NumNet+V2	<u>91.6</u>	88.3	85.4

Interpretable

Human Eval	Trust	Understand	Prefer
ModularQA	<u>67%</u>	<u>78%</u>	<u>68%</u>
DecompRC	33%	22%	32%

* Evaluated on an overlapping test set



2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics

Data, Code and Models: https://github.com/allenai/modularqa/
Demo: https://modularqa-demo.apps.allenai.org/