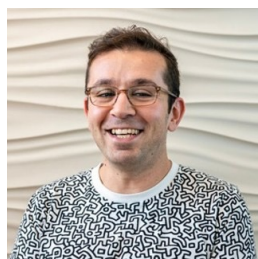


2021 Annual Conference of the North American Chapter of the  
Association for Computational Linguistics

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

Tushar Khot, Daniel Khashabi, Kyle Richardson, Peter Clark & Ashish Sabharwal



# Motivation

- Let's say you are researcher studying *Feliformia*



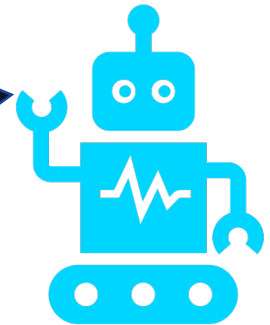
Hey VizBot, Find me all images containing a Feliformia

...



Feliformia

THIS DOES  
BOT  
COMPUTE\_

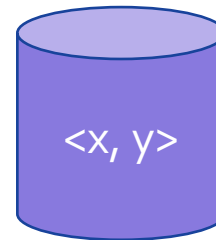


# Motivation

- Let's build a dataset

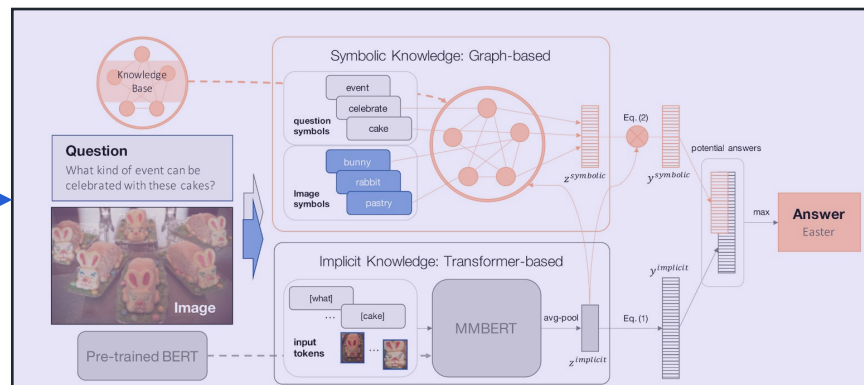


Hey that's an interesting unsolved problem. Let's build a dataset



Hey that's an interesting unsolved dataset. Let's build a *novel model* for it

$x$



$y$



**Q:** What phylum does this animal belong to?

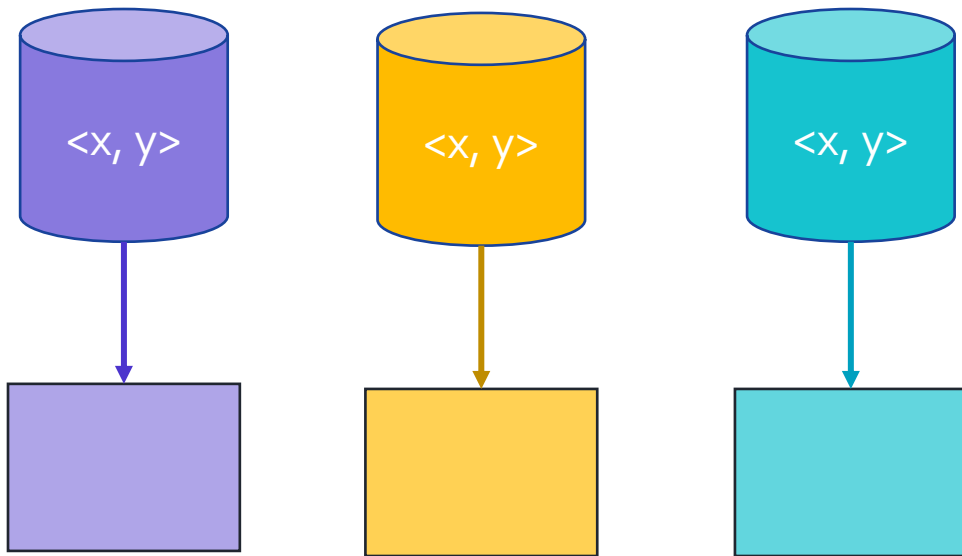
**A:** chordate, chordata

Marino et al' 19

Marino et al' 20

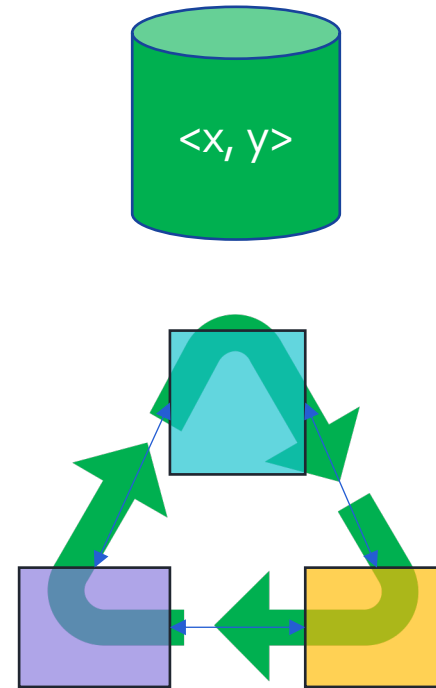
# Motivation

New Problem => New Dataset => New Model



Develop and train new (and often large) models  
for each task **from scratch each time**

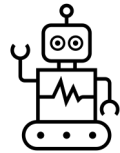
Can we reuse existing models (*stand on the  
shoulder of giants*) to solve new tasks?



# Motivation

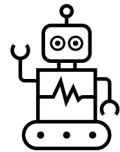
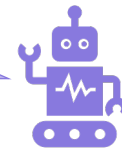


Hey MBot, Find me all images containing a Feliformia

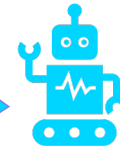
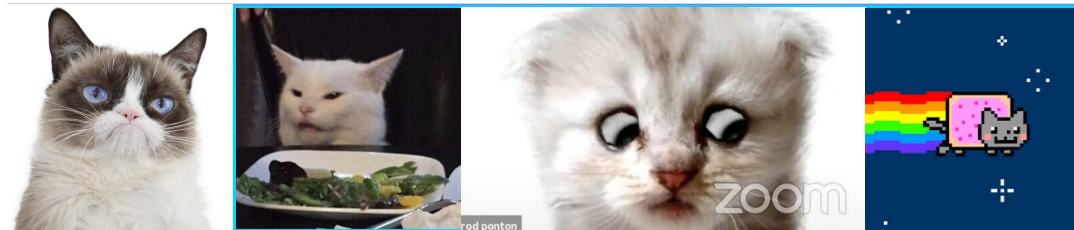


Hey TextBot, What are Feliformia?

Feliformia is a suborder ... consisting of **cats**, hyenas, mongooses, and viverrids



Hey VizBot, which images contain **cats**?



...

# Our Contribution

Research Question:

*Can we learn to decompose complex tasks into sub-tasks solvable by existing models?*



- **Text Modular Networks** (TMNs): A general framework that can *leverage existing simpler models -- neural and symbolic --* as blackboxes for answering complex questions.
- **ModularQA**: An implementation of this framework that learns to decompose *multi-hop and discrete reasoning* questions.
- A model that is more *robust, versatile, sample-efficient and interpretable*

# Setup

Research Question:

*Can we learn to decompose the complex tasks into sub-tasks solvable by the existing models?*

- Sub-Tasks & Models:



SBot

- Task: Reading Comprehension
- Model: *RoBERTa* model trained on SQuAD



CBot

- Task: Basic Math Calculation
- Model: *Symbolic* Python function

- Complex Tasks: Multi-hop + discrete reasoning

- HotpotQA (Yang et al' 18)
- DROP Subset (Dua et al' 19)

P: ...The sector decreased by 7.8 percent in 2002 ...

Q: When did the services sector start to decrease?



A: 2002

Q: diff(August 1922, 30 March 1922, months)



A: 5

**HotpotQA Question:** Little Big Girl was a Simpsons episode directed by the animator and artist of what nationality? (answer: American)

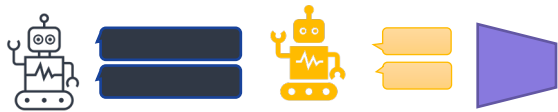
**DROP Question:** How many years did it take for the services sector to rebound? (answer: 1)



# Text Modular Networks

**P:** ...The sector decreased by 7.8 percent in 2002, before rebounding in 2003 with a 1.6 percent growth rate...

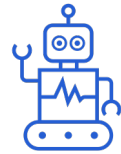
**Question Decomposition** (Talmor and Berant, 2018; Min et al., 2019; Perez et al., 2020): Often need annotations; decompositions independent of the sub-model, target one dataset + sub-model



**Neural Module Networks** (Andreas et al., 2016; Jiang and Bansal, 2019; Gupta et al., 2020): Modules need to be trained on end-task, vector-based communication

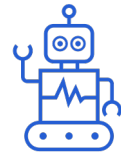


How many years did it take for the services sector to rebound?



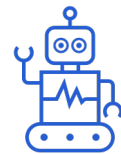
NextGen

Hey Sbot, In what year did the services sector rebound?



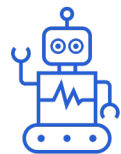
NextGen

Hey SBot, When did the services sector *start to take a dip*?



NextGen

Hey Cbot,  $\text{diff}(2003, 2002)=?$



NextGen

Done!

2003



2002



1



8


Ai2



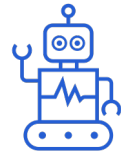
# Text Modular Networks

**P:** ...The sector decreased by 7.8 percent in 2002, before rebounding in 2003 with a 1.6 percent growth rate...

Big Question:

How do we train  to ask the right questions to the appropriate models in the model's language?

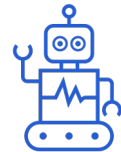
How many years did it take for the services sector to rebound?



NextGen

Hey Sbot, In what year did the services sector rebound?

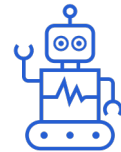
2003



NextGen

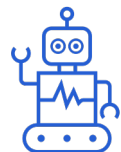
Hey SBot, When did the services sector *start to take a dip*?

2002



NextGen

Hey Cbot,  $\text{diff}(2003, 2002)=?$



NextGen

Done!

1



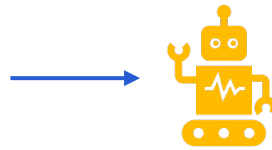
# Training Text Modular Networks (TMNs)

1. Learn the language of the sub-models
2. Decompose complex tasks into questions in this language
3. Train NextGen to generate the decomposition

# Learning the language of models

- Informally, language of a model:  
“What kind of questions can we ask this model?”
- Learn this language: Use the model’s original training data

P: ...The sector decreased by 7.8 percent in 2002, before rebounding in 2003 with a 1.6 percent growth rate...



A: 2002

V: “services”, “sector”, “start”

Q: When did the services sector start to decrease?

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

# Decomposing Complex Tasks

How many years did it take for the services sector to rebound?

Hey **SBot**, In what year did the services sector rebound?



P: ...The sector decreased by 7.8 percent in 2002, before rebounding in 2003 with a 1.6 percent growth rate...

A: 2003

Hey **SBot**, When did the services sector *start to take a dip*?



A: 2002

Hey **CBot**,  $\text{diff}(2003, 2002)=?$



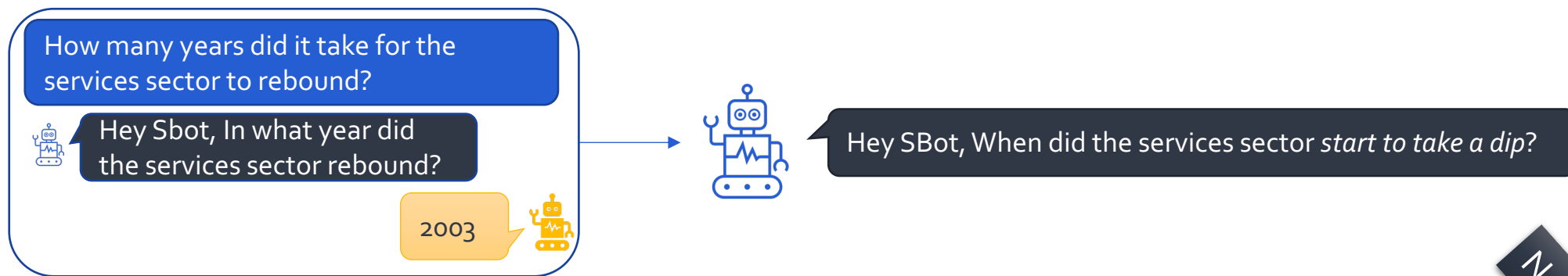
A: 1

Answer: 1

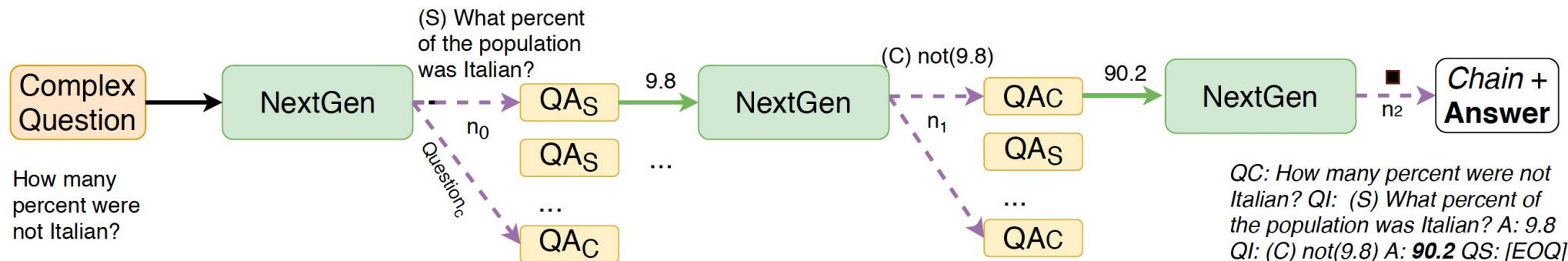
3

# Training & Inference

- Train  on these decompositions to generate the next question



- Inference: Use  to ask questions and the sub-models to answer



No need for answer hints!

# ModularQA: An implementation of TMNs

Checkout demo at:  
<https://modularqa-demo.apps.allenai.org/>

Complex Question

*qc: How many years did it take the services sector to rebound?*

Extracting distant supervision hints

Context

Set of possible hints

$a1=2003, v1=\Phi(qc)$

$a2=2002, v2=\Phi(qc)$

$a3=1, v3=\{"diff", 2003, 2002\}$

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



# Sample Decompositions

No decomposition annotations needed!

## 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

# Results

## More versatile

Modular

	DROP F <sub>1</sub>	HotpotQA
ModularQA	87.9	61.8
NMN-D	79.1*	
SNMN		63.1

Black-box

	DROP F <sub>1</sub>	HotpotQA
ModularQA	87.9	61.8
NumNet+V2	91.6	
Quark		75.5

\* Evaluated on an overlapping test set

## More robust

Contrast Test	DROP EM	DROP F <sub>1</sub>
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%

# Conclusion & Future Work

- We propose Text Modular Networks, a framework to compose existing models – symbolic or neural – to solve more complex tasks
- We develop ModularQA, an implementation of TMNs that can answer multi-hop and numeric reasoning questions
- Resulting model is more robust, versatile, sample-efficient and interpretable

## Future Work

- Dealing with many spurious decompositions (due to distant supervision)  
e.g., “Q: How many FGs were scored? A:3”
- Dealing with sub-optimal models due to unexplored sub-tasks

<https://github.com/allenai/modularqa>

{tushark, danielk, kyler, peterc, ashishs} @allenai.org



# Thanks