

question answering

CS 585, Fall 2018

Introduction to Natural Language Processing

<http://people.cs.umass.edu/~miyyer/cs585/>

Mohit Iyyer

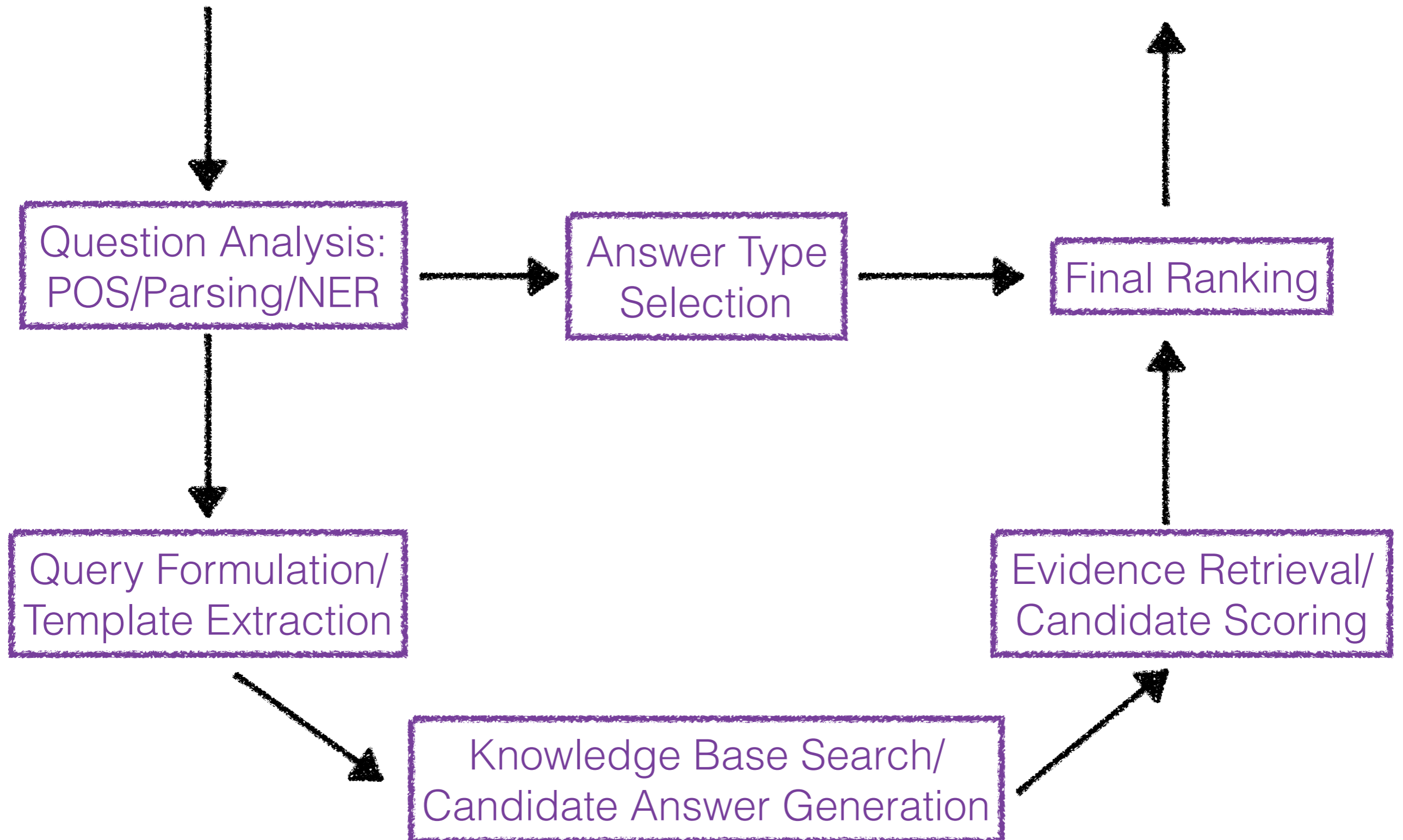
College of Information and Computer Sciences

University of Massachusetts Amherst

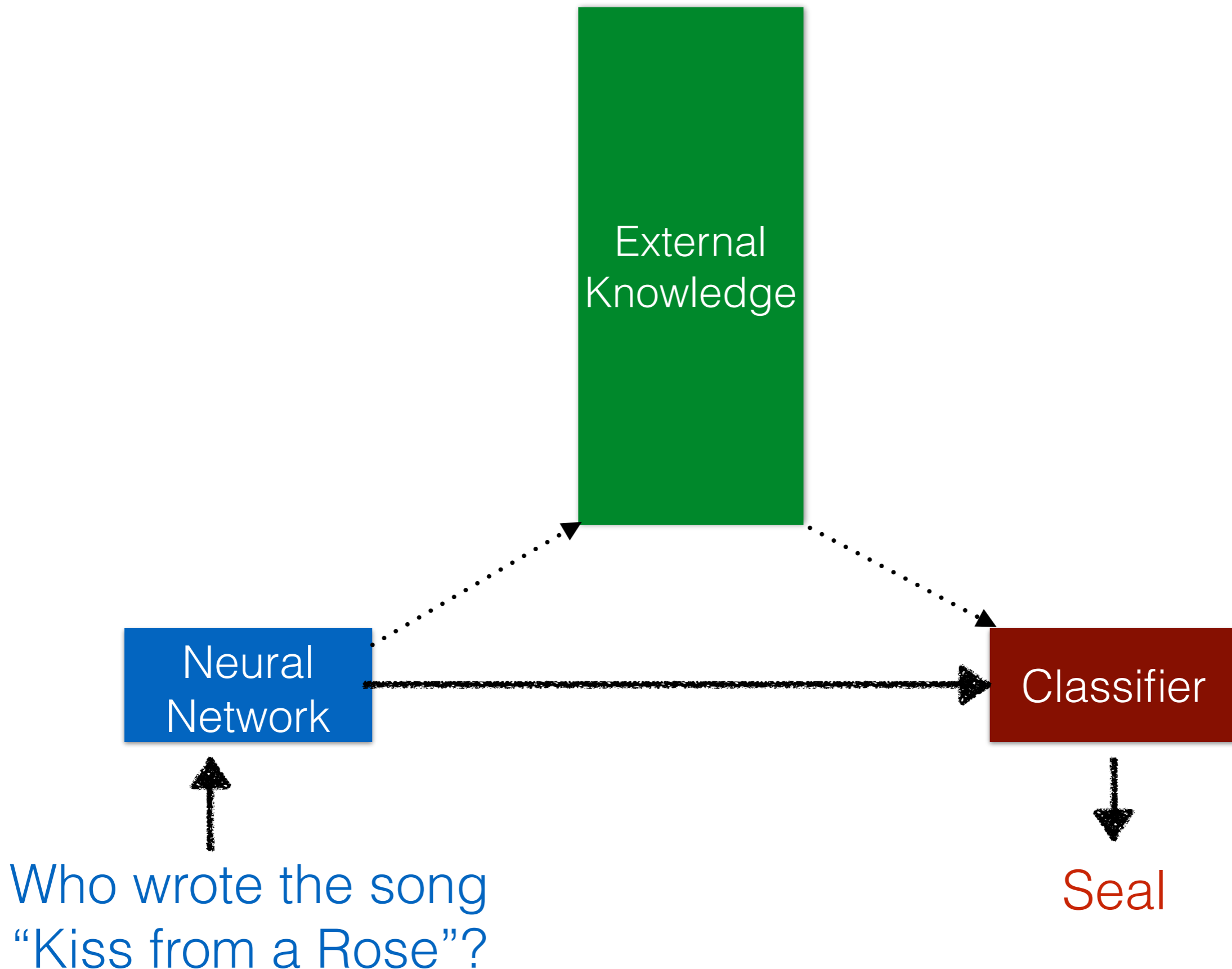
questions from last time

- progress reports?
- topics for future classes?
 - can we have a practical coding class on neural nets?
 - summarization / more neural net lectures?
- extra credit??
 - will announce next class
- HW3
 - to be released this weekend

Who wrote the song
“Kiss from a Rose”?



Can we replace all of these modules with a single neural network?



- **factoid QA:** the answer is a single entity / numeric
 - “who wrote the book “Dracula”?”
- **non-factoid QA:** answer is free text
 - “why is Dracula so evil?”
- **QA subtypes (could be factoid or non-factoid):**
 - **semantic parsing:** question is mapped to a logical form which is then executed over some database
 - “how many people did Dracula bite?”
 - **reading comprehension:** answer is a span of text within a document (could be factoid or non-factoid)
 - **community-based QA:** question is answered by multiple web users (e.g., Yahoo! Answers)
 - **visual QA:** questions about images

my QA research

- *2014-2015*: answering factoid questions within the context of a trivia game called **quiz bowl**
- *2016-2017*: semantic parsing within **conversational contexts**
- *2018*: reading comprehension within the context of **information-seeking dialog**

what is quiz bowl?

- a trivia game that contains questions about famous entities (e.g., novels, battles, countries)
- developed a deep learning system, **QANTA**, to play quiz bowl
- one of the first applications of deep learning to question answering

In one novel, one of these figures antagonizes an impoverished family before leaping into an active volcano.

Another of these figures titles a novella in which General Spielsdorf describes the circumstances of his niece Bertha Reinfeldt's death to the narrator, Laura.

In addition to Varney and Carmilla, another of these creatures sails to London on the Russian ship Demeter.

That figure bites Lucy Westenra before being killed by a coalition including Jonathan Harker and Van Helsing.

For 10 points, identify these bloodsucking beings most famously exemplified by Bram Stoker's Dracula.

A: Vampire

... identify these bloodsucking beings most famously exemplified by Dracula ...



QANTA

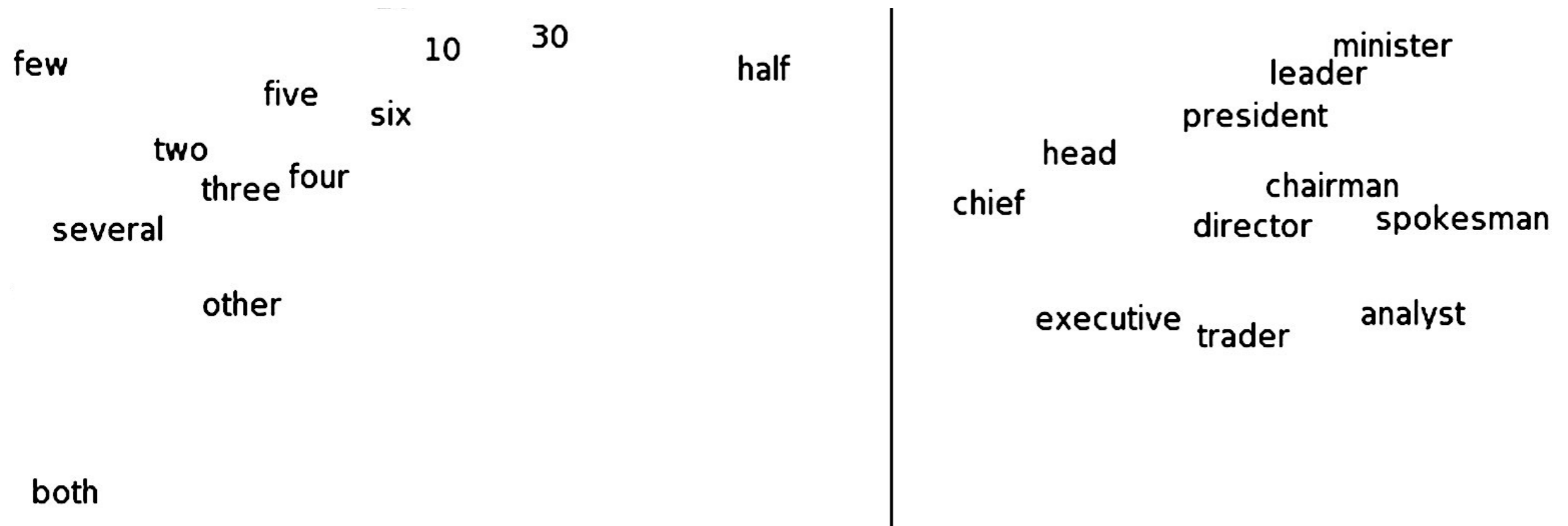


Vampires

words as basic building blocks

- represent words with low-dimensional vectors called **embeddings** (Mikolov et al., NIPS 2013)

president =
[0.23, 1.3, -0.3, 0.43]



composing embeddings

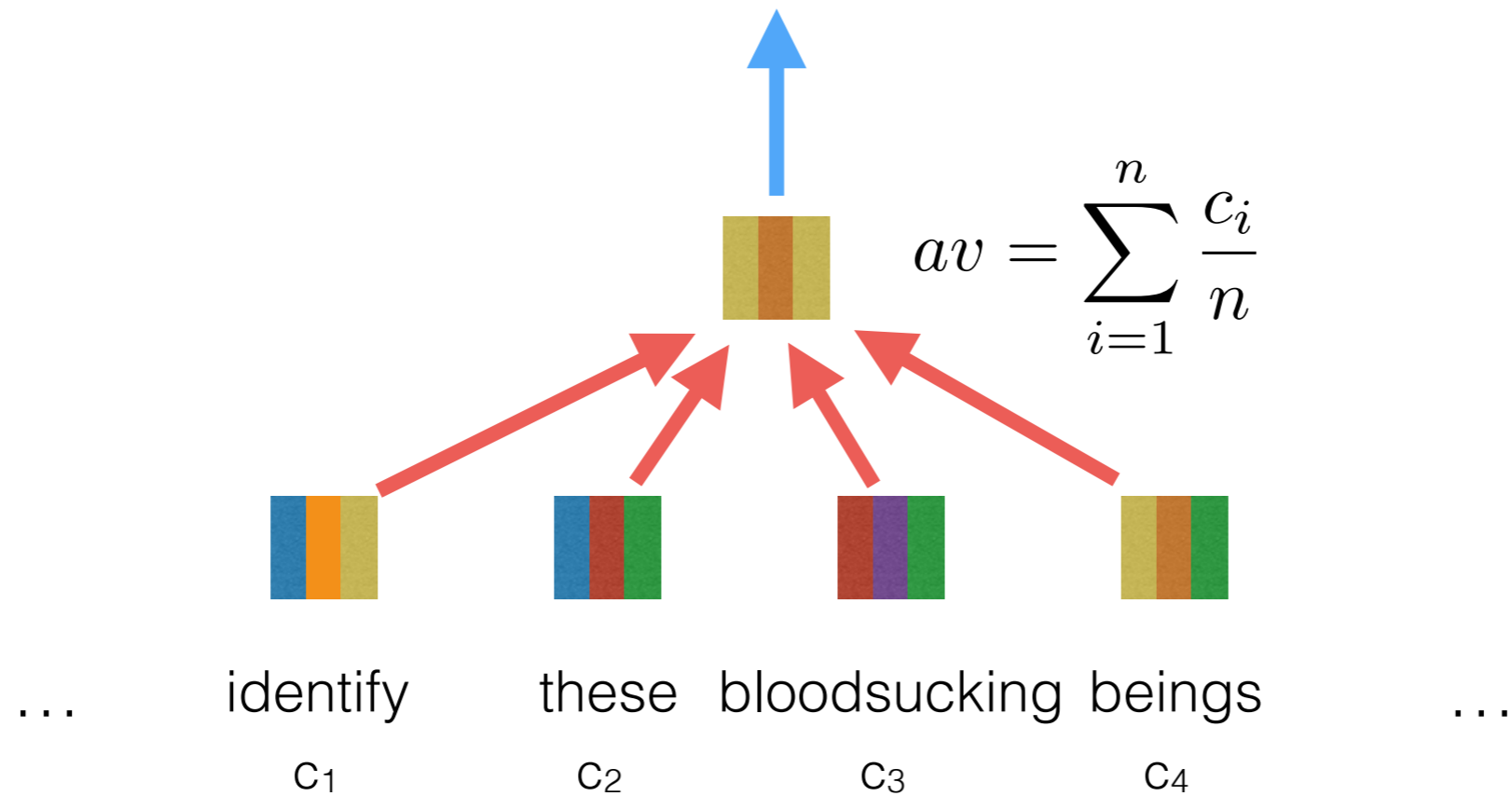
- neural networks **compose** word embeddings into vectors for phrases, sentences, and documents

neural network (identify these bloodsucking beings) =



neural bag-of-words

softmax: predict **Vampire** out of a set of ~5000 answers



neural bag-of-words

softmax: predict **Vampire** out of a set of ~5000 answers



train network on
30k quiz bowl QA pairs



... identify

C₁



these

C₂



bloodsucking beings

C₃



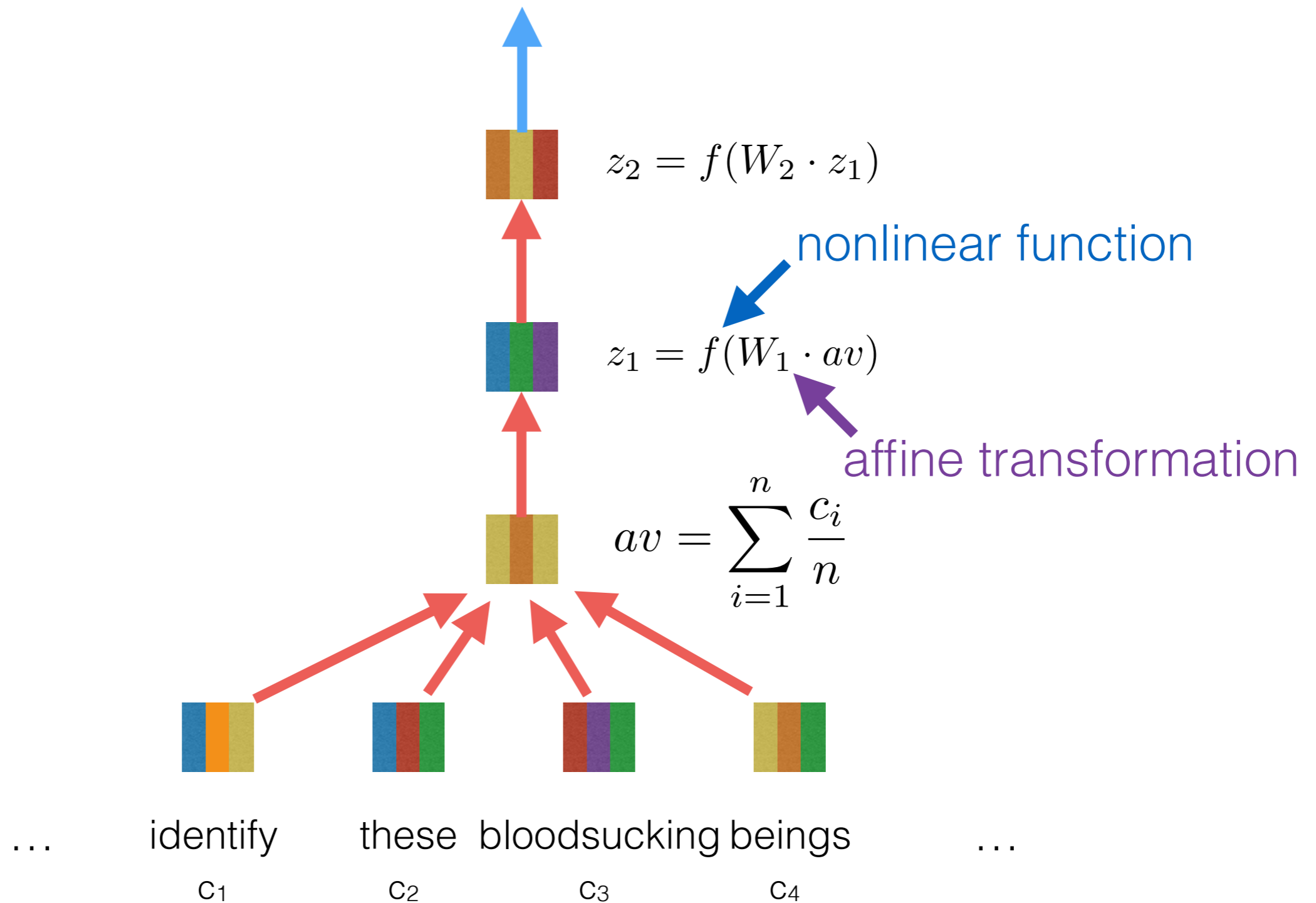
...

C₄

n
 c_i

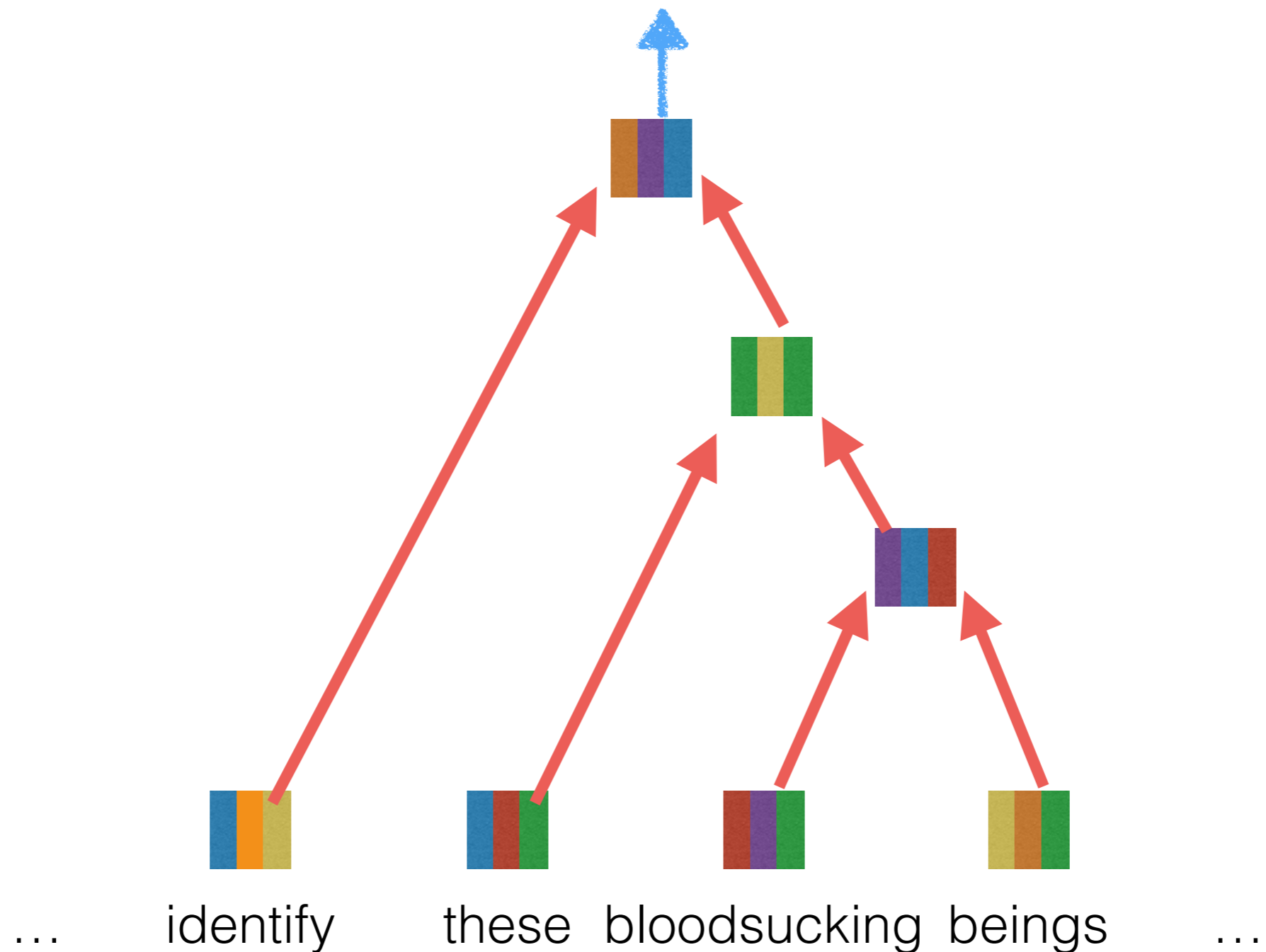
deep averaging networks

softmax: predict **Vampire** out of a set of ~5000 answers



dependency-tree NNs

softmax: predict **Vampire** out of a set of ~5000 answers



simple discourse-level representations by averaging

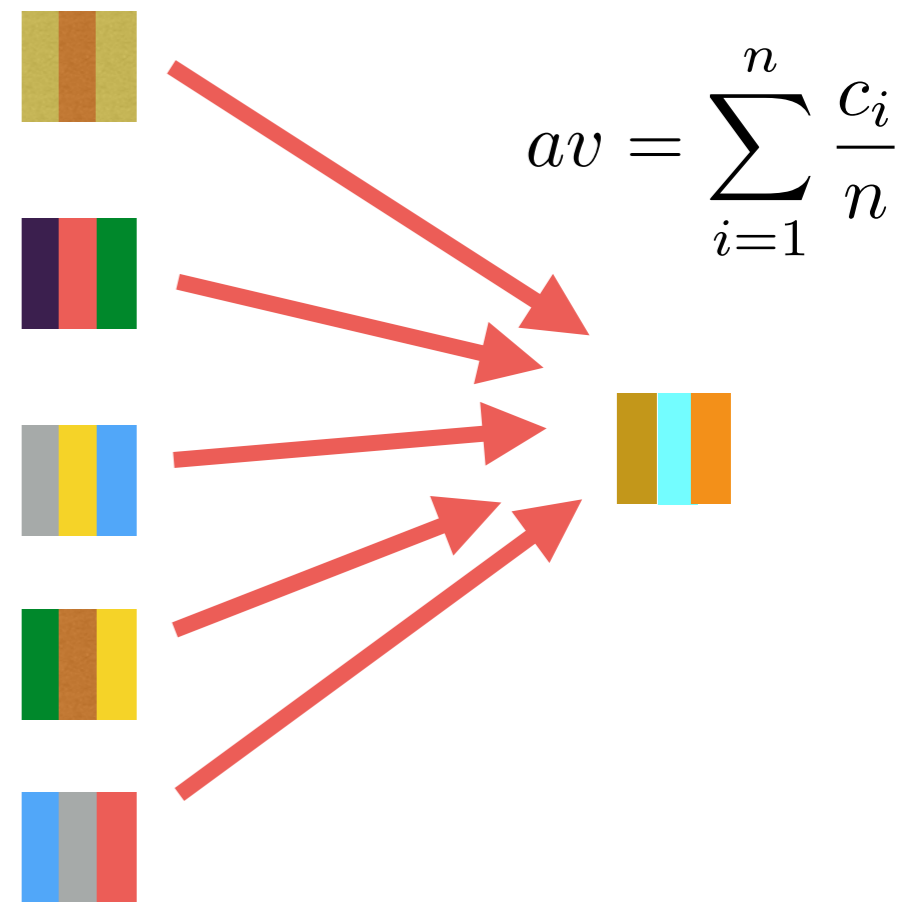
In one novel, one of these figures antagonizes an impoverished family before leaping into an active volcano.

Another of these figures titles a novella in which General Spielsdorf describes the circumstances of his niece Bertha Reinfeldt's death to the narrator, Laura.

In addition to Varney and Carmilla, another of these figures sails on the Russian ship Demeter in order to reach London.

That figure bites Lucy Westenra before being killed by a coalition including Jonathan Harker and Van Helsing.

For 10 points, identify these bloodsucking beings most famously exemplified by Bram Stoker's Dracula.

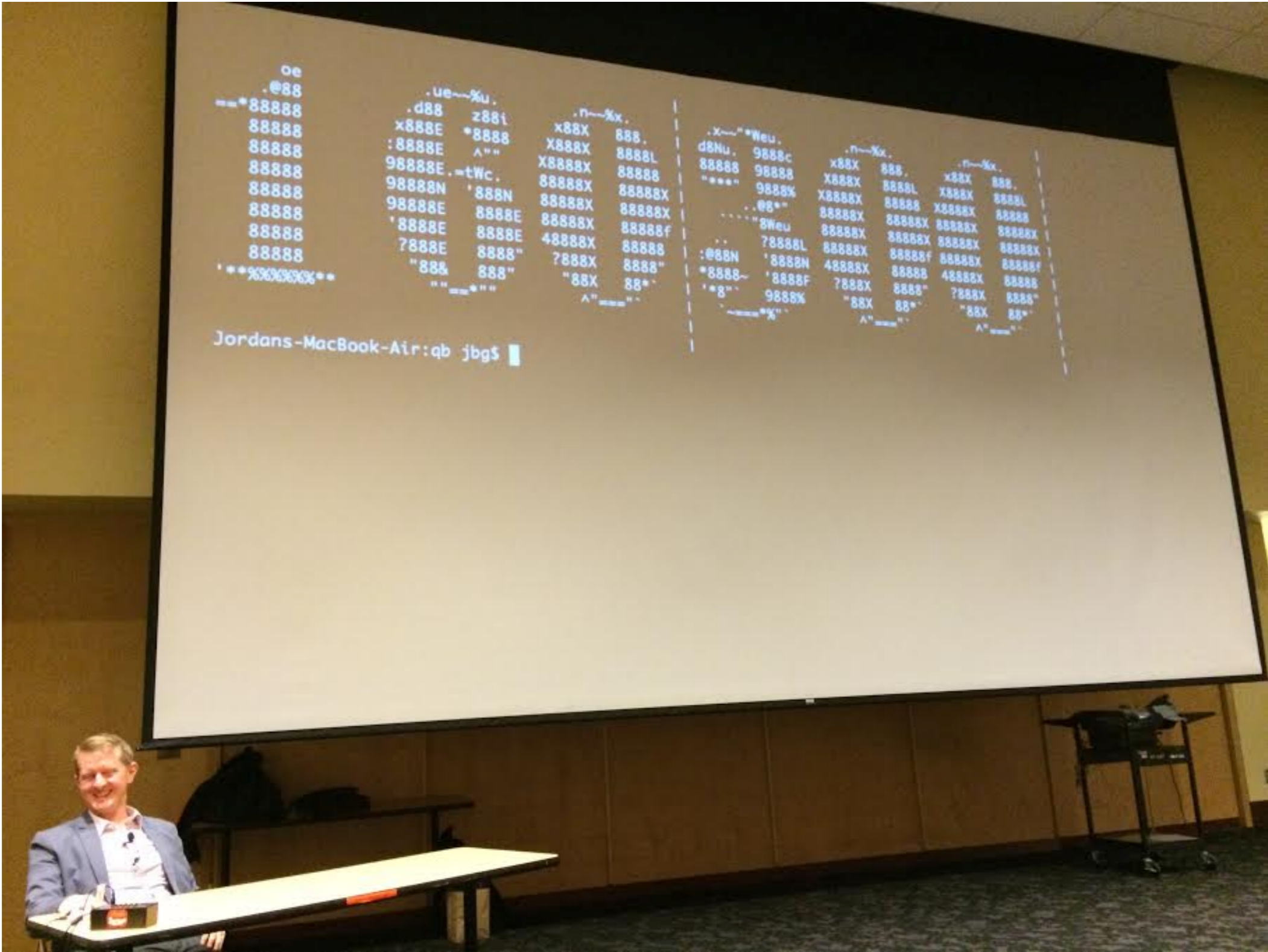


comparison of architectures

Model	Quizbowl Accuracy	Time / Epoch (s)
NBOW	66.3	11
DAN	70.8	18
Tree-NN	72.1	314

similar results have since been shown for other tasks such as entailment and sentence similarity (Wieting et al., ICLR 2016, Hill et al., NAACL 2016)

2015: defeated Ken Jennings 300-160



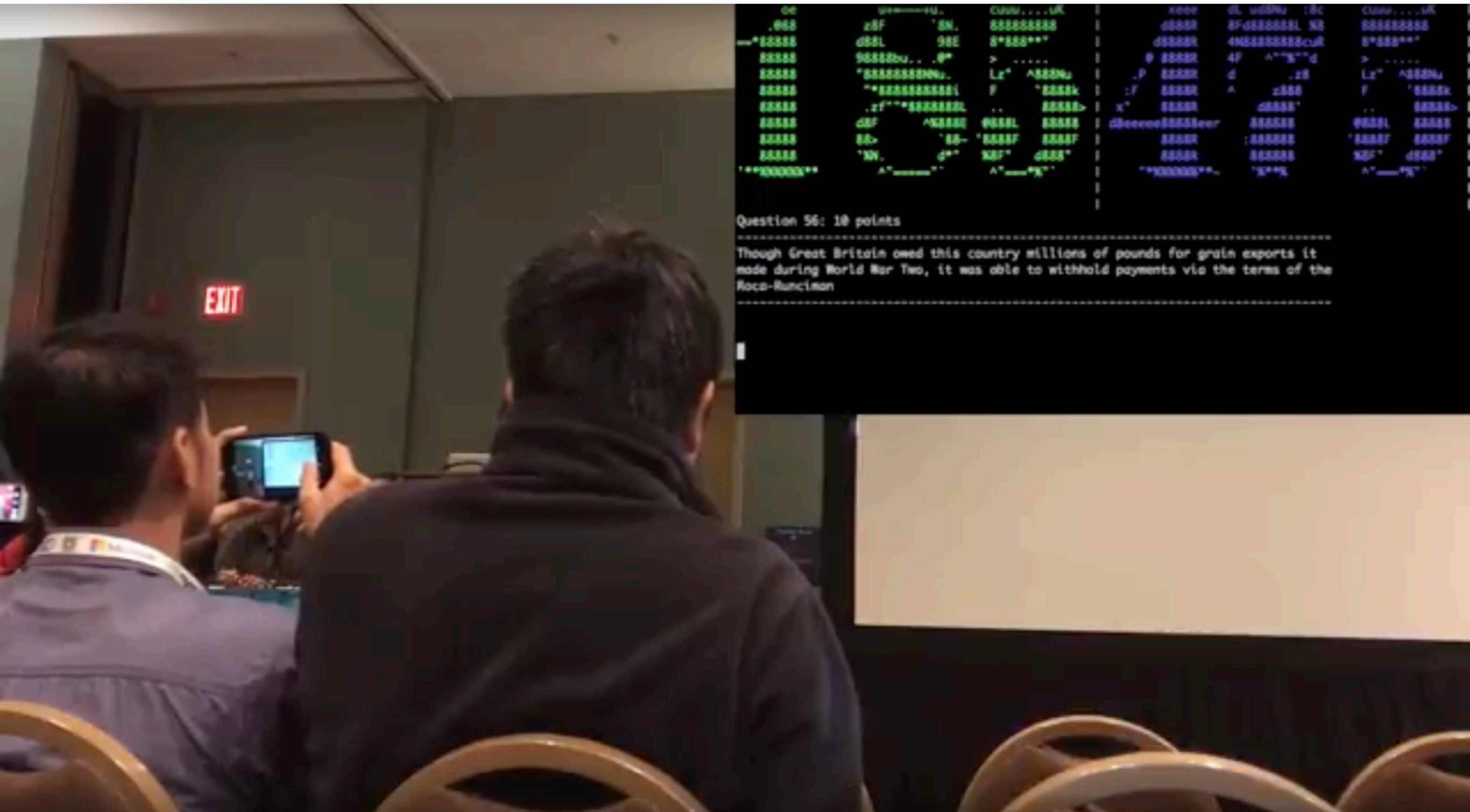
2016: lost to top quiz bowlers 345-145



2017: beat top quiz bowlers 260-215



late 2017: crushed top team 475-185



late 2018: adversarial QB?

questions written to fool an NN
competition on Dec. 15!
submit your systems!
qanta.org

deep learning ~ memorization

during training, QANTA becomes very good at associating **named entities** in questions with answers...

That figure bites **Lucy Westenra** before being killed by a coalition including **Jonathan Harker** and **Van Helsing**.

Vampire

The diagram illustrates the concept of named entities in a sentence. A blue-bordered box contains the sentence: "That figure bites **Lucy Westenra** before being killed by a coalition including **Jonathan Harker** and **Van Helsing**." Three green arrows point from the named entities "Lucy Westenra", "Jonathan Harker", and "Van Helsing" to the word "Vampire" below the box. This indicates that the model has learned to associate these specific names with the concept of a vampire.

deep learning ~ memorization

during training, QANTA becomes very good at associating **named entities** in questions with answers...

In one novel, one of these figures antagonizes an impoverished family before leaping into an active volcano.

???

let's look at **sequential semantic parsing**:
a QA task where memorization doesn't work!

conversational contexts are hard!

How much protein is in an egg?

And how many carbohydrates?

Are eggs on my shopping list?

What about butter?

Do I need an umbrella today?

Where can I buy one?

What's 42 plus 8 minus 13?

Is the answer divisible by 4?

conversational contexts are hard!

How much protein is in an egg?

And how many carbohydrates?

Are eggs on my shopping list?

What about butter?

Do I need an umbrella today?

Where can I buy one?

What's 42 plus 8 minus 13?

Is the answer divisible by 4?

the follow-up question can only be answered by resolving either an *explicit* or *implied* reference to the previous question

FINA Women's Water Polo World Cup

Rank	Nation	Gold	Silver
1	Netherlands	8	3
2	Australia	3	3
3	USA	2	5
4	Hungary	1	1
5	Canada	0	0

FINA Women's Water Polo World Cup

Rank	Nation	Gold	Silver
1	Netherlands	8	3
2	Australia	3	3
3	USA	2	5
4	Hungary	1	1
5	Canada	0	0

1. Which nations competed in the FINA women's water polo cup?

FINA Women's Water Polo World Cup

Rank	Nation	Gold	Silver
1	Netherlands	8	3
2	Australia	3	3
3	USA	2	5
4	Hungary	1	1
5	Canada	0	0

1. Which nations competed in the FINA women's water polo cup?

SELECT Nation

semantic parse:
a logical form
executed on table
to yield answer

FINA Women's Water Polo World Cup

Rank	Nation	Gold	Silver
1	Netherlands	8	3
2	Australia	3	3
3	USA	2	5
4	Hungary	1	1
5	Canada	0	0

1. Which nations competed in the FINA women's water polo cup?

```
SELECT Nation
```

2. Of these nations, which ones took home at least one gold medal?

```
SUBSEQUENT WHERE Gold != 0
```

FINA Women's Water Polo World Cup

Rank	Nation	Gold	Silver
1	Netherlands	8	3
2	Australia	3	3
3	USA	2	5
4	Hungary	1	1
5	Canada	0	0

1. Which nations competed in the FINA women's water polo cup?

SELECT Nation

2. Of these nations, which ones took home at least one gold medal?

SUBSEQUENT **WHERE** Gold **!=** 0

SUBSEQUENT:
handles references
between questions

FINA Women's Water Polo World Cup

Rank	Nation	Gold	Silver
1	Netherlands	8	3
2	Australia	3	3
3	USA	2	5
4	Hungary	1	1
5	Canada	0	0

1. Which nations competed in the FINA women's water polo cup?

```
SELECT Nation
```

2. Of these nations, which ones took home at least one gold medal?

```
SUBSEQUENT WHERE Gold != 0
```

3. Of those, which ranked in the top 2 positions?

```
SUBSEQUENT WHERE Rank <= 2
```

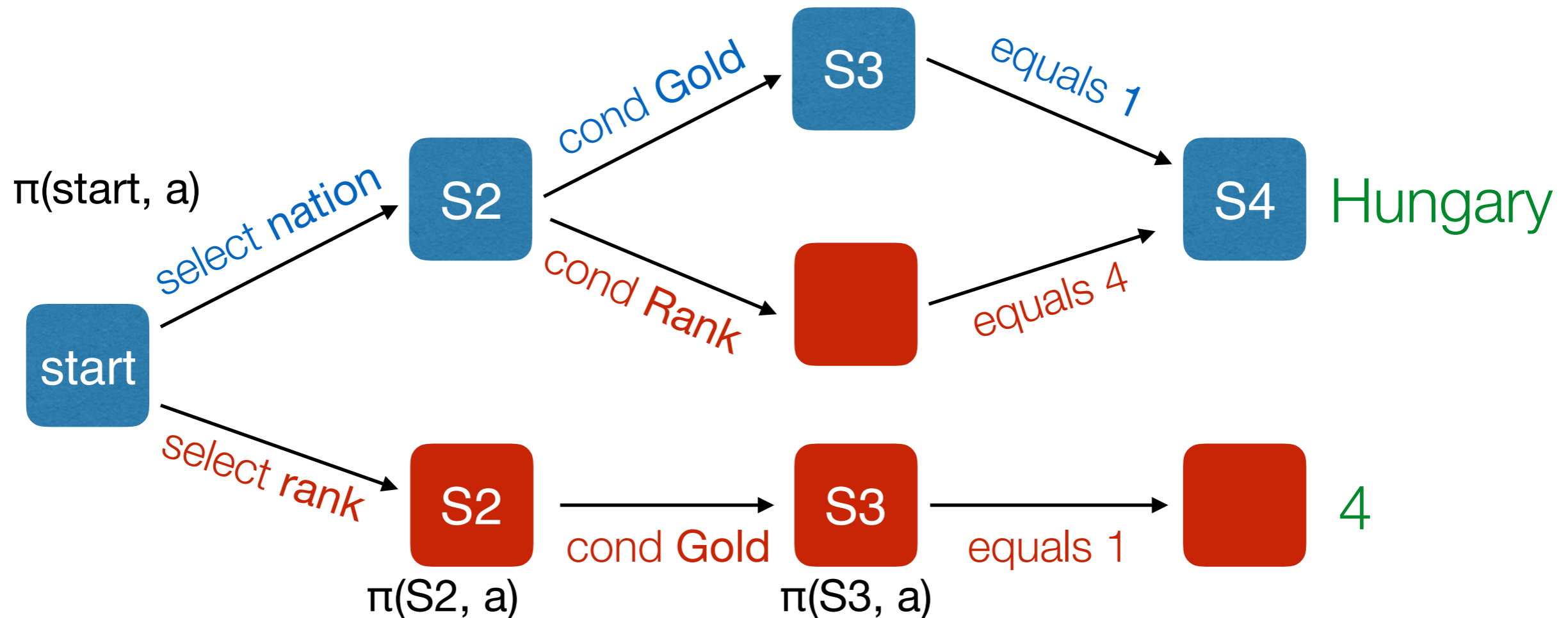
dynamic semantic parsing

- We collect **SQA**, a dataset of ~6000 question/answer sequences
- Since we only know the answer to a question and *not* its ground-truth logical form, this problem is only weakly supervised.
- To solve it, we use **reward-guided structured-output learning**

dynamic semantic parsing

Q: which nations won exactly one gold medal? A: Hungary

1. select-column Nation
2. cond-column Gold
3. op-equal 1

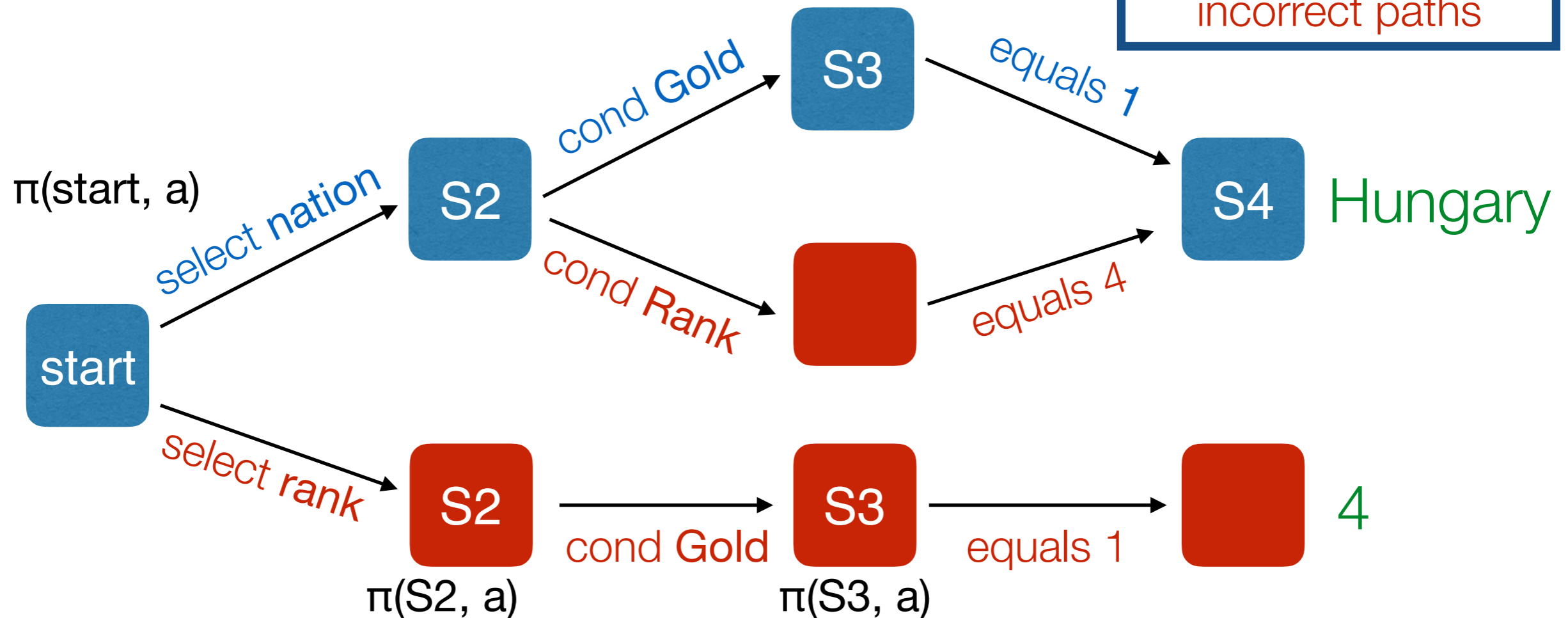


dynamic semantic parsing

Q: which nations won exactly one gold medal? A: Hungary

- 1. select-column Nation
- 2. cond-column Gold
- 3. op-equal 1

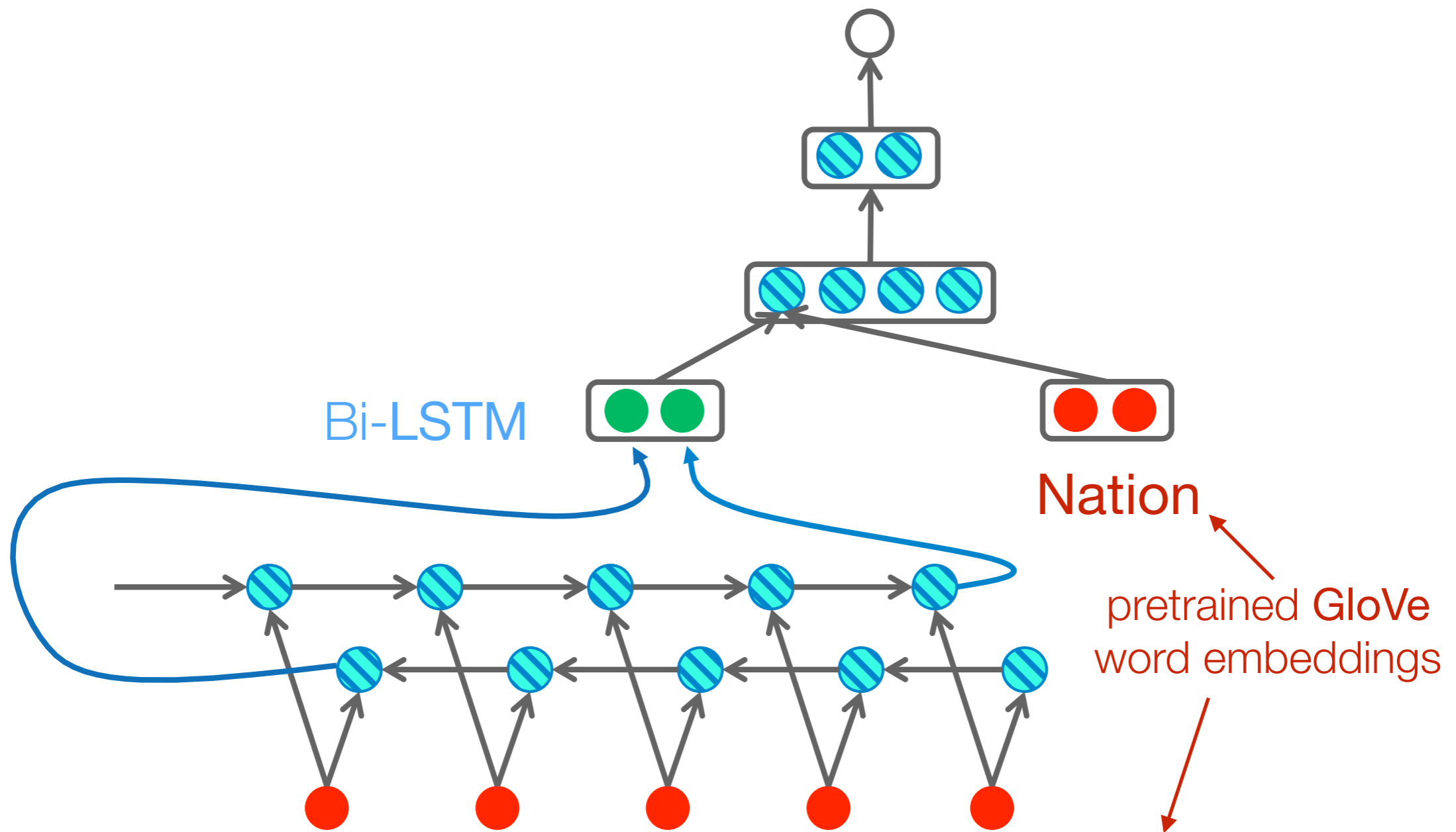
maximize score of (approx) correct path
minimize score of incorrect paths



dynamic semantic parsing

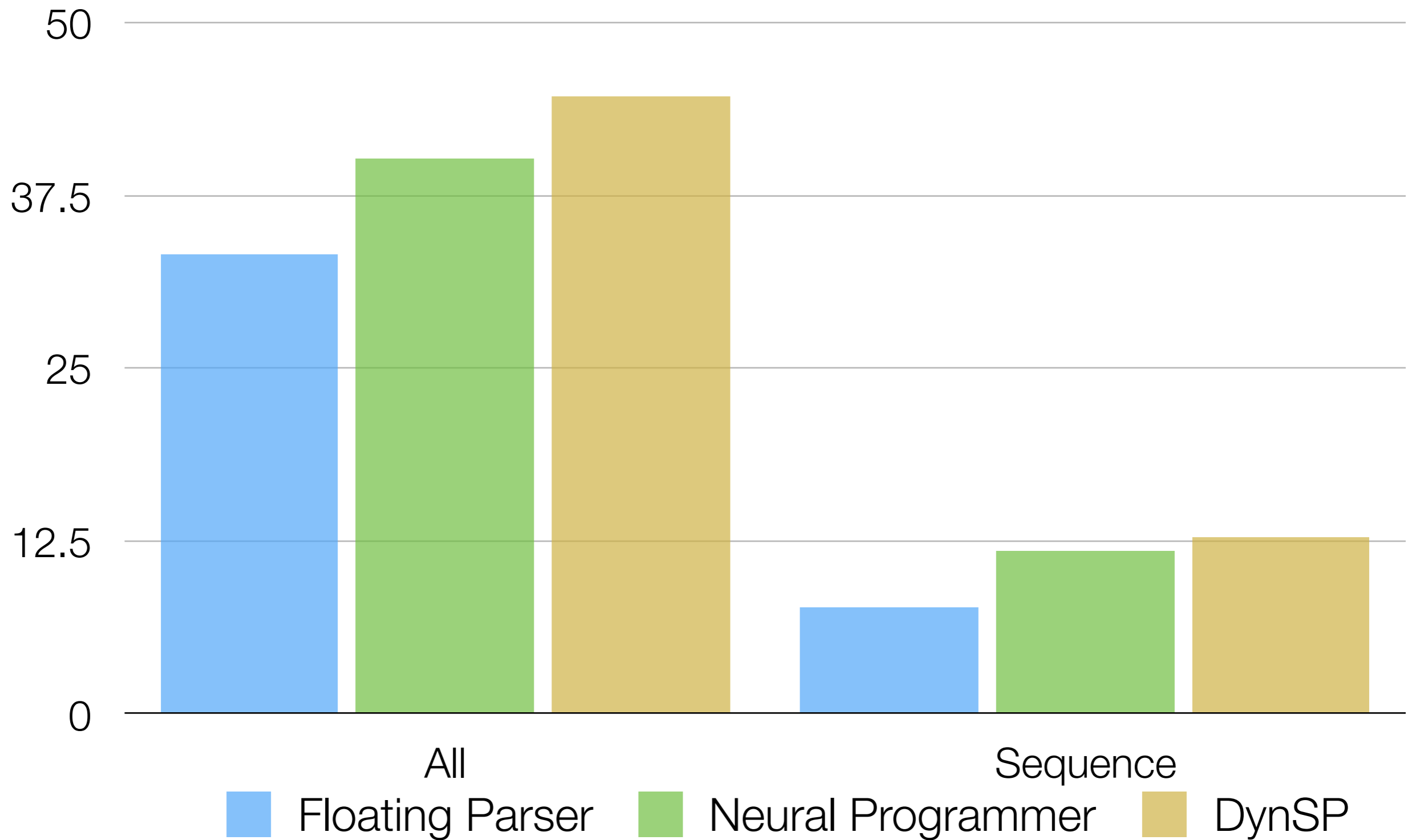
- neural network modules output scalar values which we use in the value function
 $\pi(\text{current parse, next operation})$
- end-to-end training algorithm: approximate a reference parse and train the value function to favor that parse
- discourse-level information incorporated with **SUBSEQUENT** statements, which have their own action semantics

ex: module implementation



Which nations won one gold medal?

Answer Accuracy on SQA



this problem is still very difficult!

- model lacks world knowledge due to small training data size

this problem is still very difficult!

- model lacks world knowledge due to small training data size

Call sign	City	Genre
KJMS	Topeka	Bluegrass
WAJS	Washington	Rock, pop
WAOY	Boston	Hip-hop
KQMP	Omaha	R&B
WETX	Seattle	Synthwave

what radio station plays
synthwave music?

composing embeddings

- neural networks **compose** word embeddings into vectors for phrases, sentences, and documents

neural network (           ) =   

embeddings often initialized with “pretrained” vectors and then fine-tuned on a specific dataset.

how do we handle infrequent words/phrases such as *synthwave* and *call sign*?

learn *contextualized representations* with
large scale neural language models

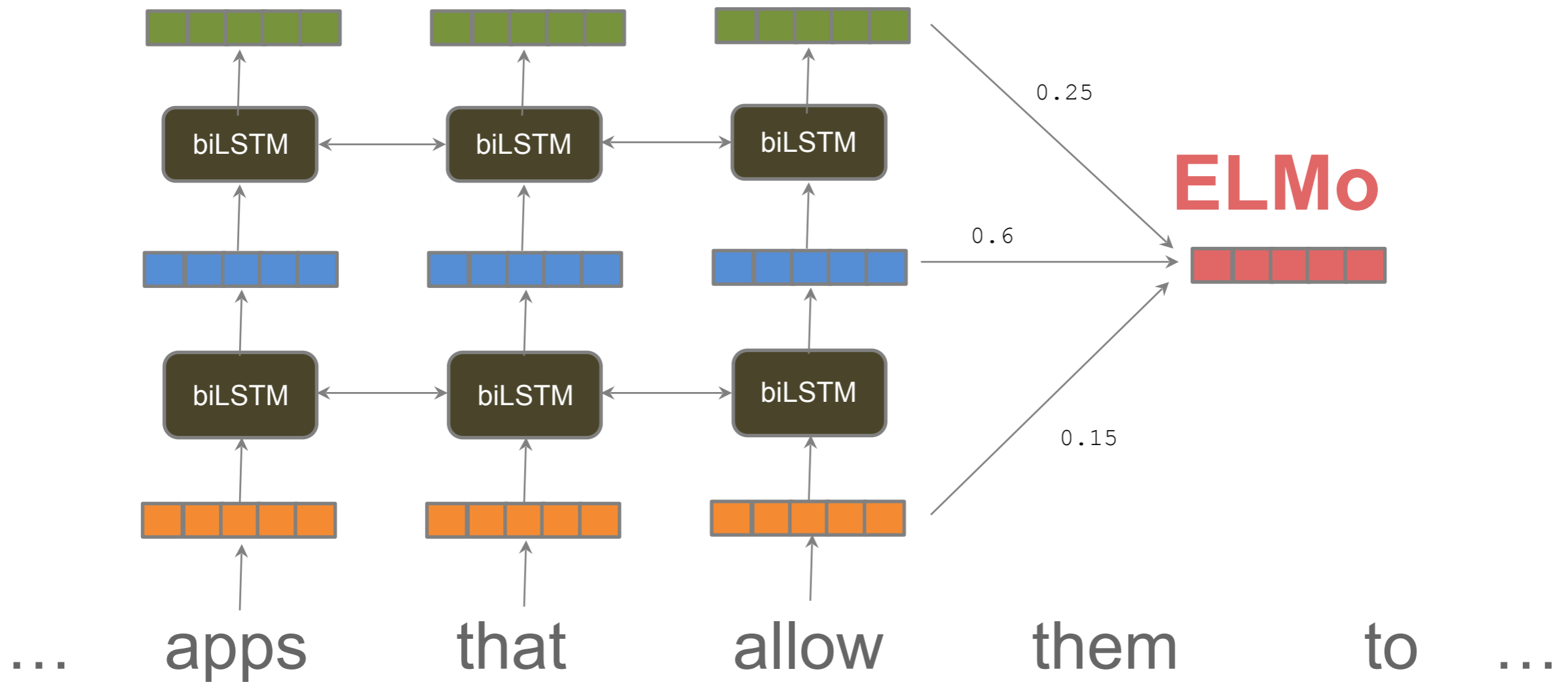
Peters et al., NAACL 2018

predicting the next word

With smartphones, users can now listen to radio stations online through apps that allow them to...

choose call signs from a variety of different locations and genres.

Use all layers of language model



Contextual representations

ELMo representations are *contextual* – they depend on the entire sentence in which a word is used.

No issues representing *synthwave* and *call sign*

back to QA:

a dataset for *information-seeking dialog*

why do we need *another* QA dataset?

- questions in existing datasets are often unnatural
of countries that won at least one bronze and three gold medals,
were there any in Europe with populations under four million?
- ... or uninteresting (too easy)
in what year was Emma Stone born?
- our goal is to collect challenging, interesting
questions that incorporate both *conversational* and
external context

information-seeking dialog

“The purpose is for you to form an impression of an area that previously you knew only very little about, guided by your interests and by what you learn”

Stede & Shlangen, 2004

- more structured than “chatbots”
- less structured than goal-oriented agents
- actually has practical applications!

student: how big are grizzly bears?

teacher: they're huge! adults can grow to 9 ft tall at 900 lbs

student: do they attack humans?

teacher: rarely, most often when they smell humans' food

student: since they're so big, how can I protect myself?

teacher: bear spray is usually effective

student: what do you mean "usually"?

teacher: sometimes grizzlies just walk through it!

student: ... what do i do then?

teacher: play dead

student: that sounds risky... will they get me if i climb a tree?

teacher: yes, they can climb trees too (or knock them down)

student: in a fight, would a grizzly bear defeat a polar bear?

teacher: ...

crowdsourced setting with two workers

teacher

The Anti-Hazing Act of 1995, also known as Republic Act No. 8049 is a national legislation in the Philippines regulating the act of hazing and other initiation rites in fraternities and sororities in the country. It prohibits and penalizes physical harm and violence in such practices.... *(rest of Wiki article)*

student's current question:
your selected answer:

student

title: Anti-Hazing Act of 1995

short summary:

The Anti-Hazing Act of 1995 is a national legislation in the Philippines

your last question: what is the purpose of the act?

teacher's answer: regulating the act of hazing

... type your next question here

crowdsourcing a large-scale reading comprehension dataset

Q: why was the act introduced?

A: a response to the death of Leonardo Villa in 1991

The Anti-Hazing Act of 1995, also known as Republic Act No. 8049 is a national legislation in the Philippines regulating the act of hazing and other initiation rites in fraternities and sororities in the country. It prohibits and penalizes physical harm and violence in such practices.... *(rest of Wiki article)*

crowdsourcing a large-scale reading comprehension dataset

The Anti-Hazing Act of 1995, also known as Republic Act No. 8049 is a national legislation in the Philippines regulating the act of hazing and other initiation rites in fraternities and sororities in the country. It prohibits and penalizes physical harm and violence in such practices.... *(rest of Wiki article)*

Q: why was the act introduced?

A: a response to the death of Leonardo Villa in 1991

Q: How did Leonardo Villa die?

A: multiple injuries when he underwent hazing rites

conversational context:
the student asked about information revealed in the previous answer

crowdsourcing a large-scale reading comprehension dataset

The Anti-Hazing Act of 1995, also known as Republic Act No. 8049 is a national legislation in the Philippines regulating the act of hazing and other initiation rites in fraternities and sororities in the country. It prohibits and penalizes physical harm and violence in such practices.... *(rest of Wiki article)*

Q: why was the act introduced?

A: a response to the death of Leonardo Villa in 1991

Q: How did Leonardo Villa die?

A: multiple injuries when he underwent hazing rites

Q: Was the general public in favor of this act?

A: cannot answer!

models have to decide if the question is **answerable** from the text before producing an answer!

crowdsourcing a large-scale reading comprehension dataset

The Anti-Hazing Act of 1995, also known as Republic Act No. 8049 is a national legislation in the Philippines regulating the act of hazing and other initiation rites in fraternities and sororities in the country. It prohibits and penalizes physical harm and violence in such practices.... *(rest of Wiki article)*

sometimes a question is *indirectly* answered by text from the article!

Q: why was the act introduced?

A: a response to the death of Leonardo Villa in 1991

Q: How did Leonardo Villa die?

A: multiple injuries when he underwent hazing rites

Q: Was the general public in favor of this act?

A: cannot answer!

Q: Did the situation improve after the Act was passed?

A: No, By 2017, only 1 conviction has been made under the Anti-Hazing Act



QuAC: QUestion Answering in Context

# of dialogs	13,594
# of questions	98,407
# of unanswerable questions	16,950
# of articles	3,611
# of unique sections	8,854
<hr/>	
avg questions per dialog	7.2
avg tokens per question	6.5
avg tokens per answer	15.1
avg tokens per section	401.0

first large-scale dataset of dialog-style QA!

all dialogs are about well-known “people” entities

what can we do with these dialogs?

- build teacher models that answer a user's questions about a topic / document
- build student models that *generate* information-seeking questions about a topic

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

...

↪ *ask a follow-up!*

↪ *you can follow up, but there are more interesting things to learn*

↧ *do not ask a follow-up!*



question analysis

question type	examples
non-factoid (47%)	Q: Were the peace talks a success? Q: What was her childhood like?
coref to article (72%)	Title: Paul Cezanne - Early life and family Q: When did he start painting?
contextual (27%)	Q: Did [Ian McKellen] do any charity work? A: Yes, McKellen is a member of Only Make Believe (...) Q: what does he do there?
“anything else” (8%)	Q: What other acting did she do? Q: What else did he research?

different people ask different questions!

Hattie McDaniel — Whereabouts of the Missing Oscar

1. What are the whereabouts of Hattie's oscar?
2. Was it stolen?
3. What was the Oscar for?
4. What did the article say about the investigation?
5. When was it noticed missing?
6. Is it on display somewhere today?

1. Did the McDaniel Oscar go missing?
2. When did it go missing?
3. When did McDaniel win the Oscar?
4. Has anyone searched for the missing Oscar?
5. What were the results?
6. If it was returned, why can't it be found?
7. Was there any truth to that?

1. What happened to the Oscar?
2. How did it go missing?
3. Who had the Oscar after she died?
4. Why did the University have it?
5. What did she win the Oscar for?
6. What else can you tell me about this time period?
7. What were the results she published?
8. What did she think happened to it?
9. What was she known for in her community?

1. Where was the McDaniel Oscar?
2. Does anyone think they know where it is?
3. What is the McDaniel Oscar?
4. Is it an award that she won?
5. Are there any other interesting aspects about this article?
6. Do they think someone stole it?
7. Did they ever search the Potomac River for it?
8. Did Carter have an alternate theory?

evaluating teacher models

- many questions have multiple valid answers, so for our dev / test sets we acquire five total answers per question

Q: Was Robey successful in his career?

A1: Robey's turn was a great success, and as a result he secured a permanent theatrical residency at the venue.

A2: By the early months of 1891, Robey was much in demand, and he decided to change his stage name.

- we evaluate by both **F1** (like SQuAD) and a new dialog-level metric, human equivalence (**HEQ**), that measures whether a system is as good at answering questions as an average human

humans >>> our best model

- Humans:
 - 81.0 F1, 100 HEQ
- BiDAF + self-attention + ELMo (no dialog context)
 - 50.2 F1, 2.2 HEQ
- BiDAF + self-attention + ELMo + prev 3 QA pairs
 - 60.1 F1, 4.0 HEQ
- **> 10 F1 when incorporating dialog context!**
- FlowQA (Huang et al., 2018 in submission)
 - 64.1 F1, 5.8 HEQ



quizbowl:
qanta.org

SQA:
aka.ms/sqa

QuAC:
quac.ai