CS 378: Natural Language Processing
Lecture 21: Constituency Parsing / QA

TEXAS
The University of Texas at Austin

Eunsol Choi

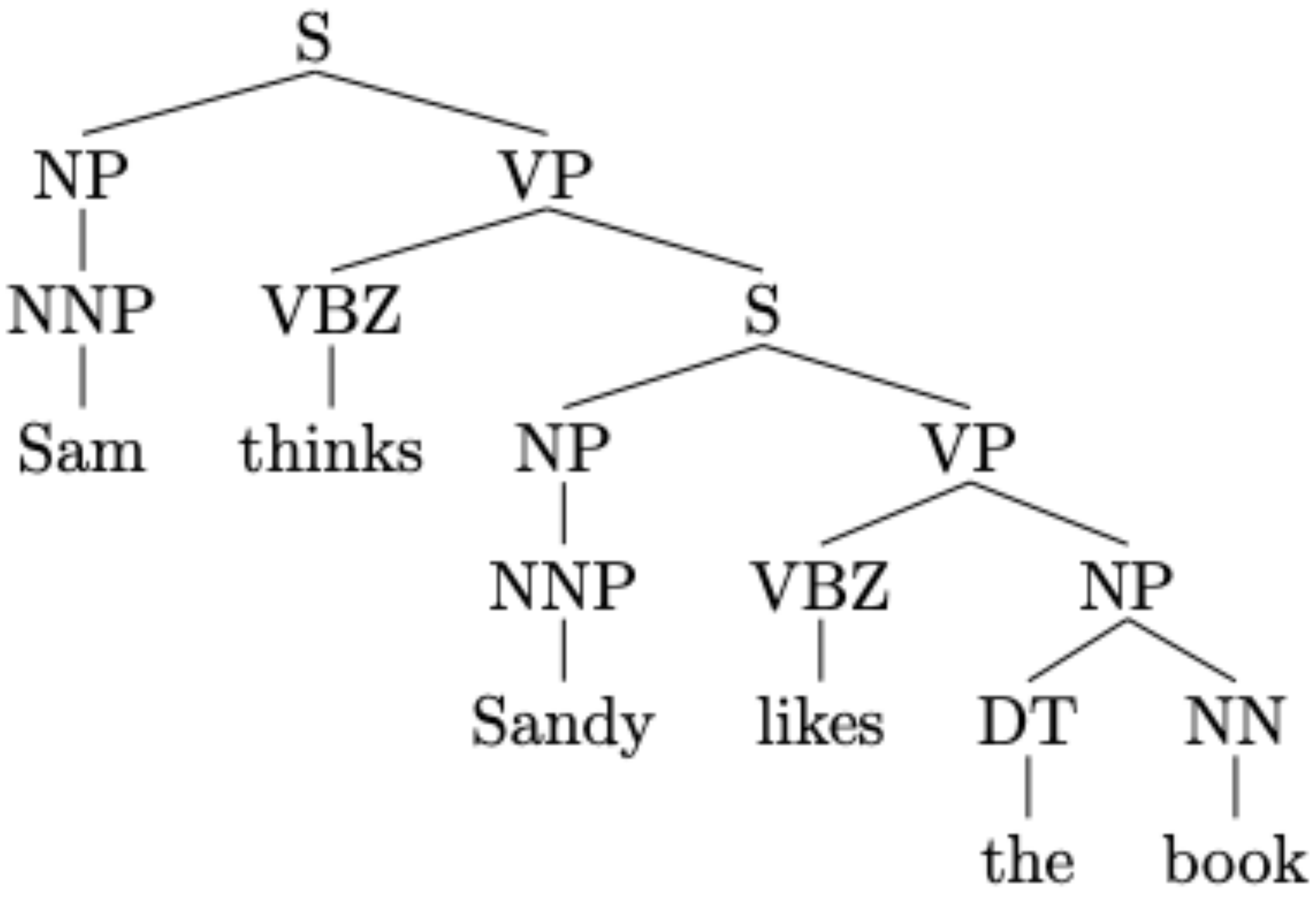
Today

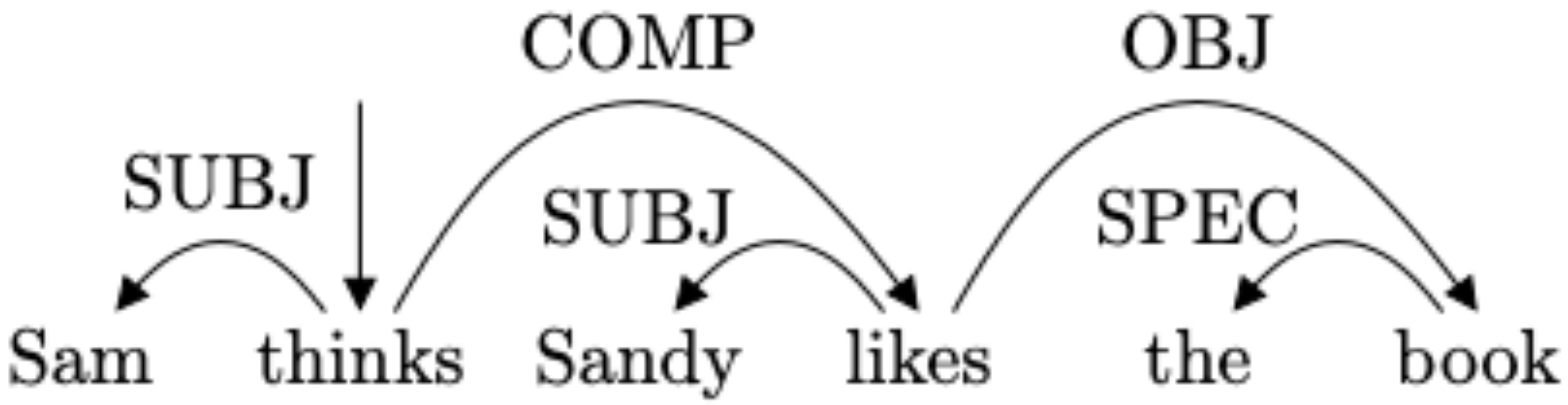
- Trees
    - Constituency Trees
        - How to find best-scoring trees given probabilistic CFG grammar
    - Dependency Trees vs. Constituency Trees

- Applications
    - Question Answering

Recap: Syntax

- Study of word order and how words form sentences
  - Constituency Parsing
  - Dependency Parsing





Words organized as nested structure of constituents

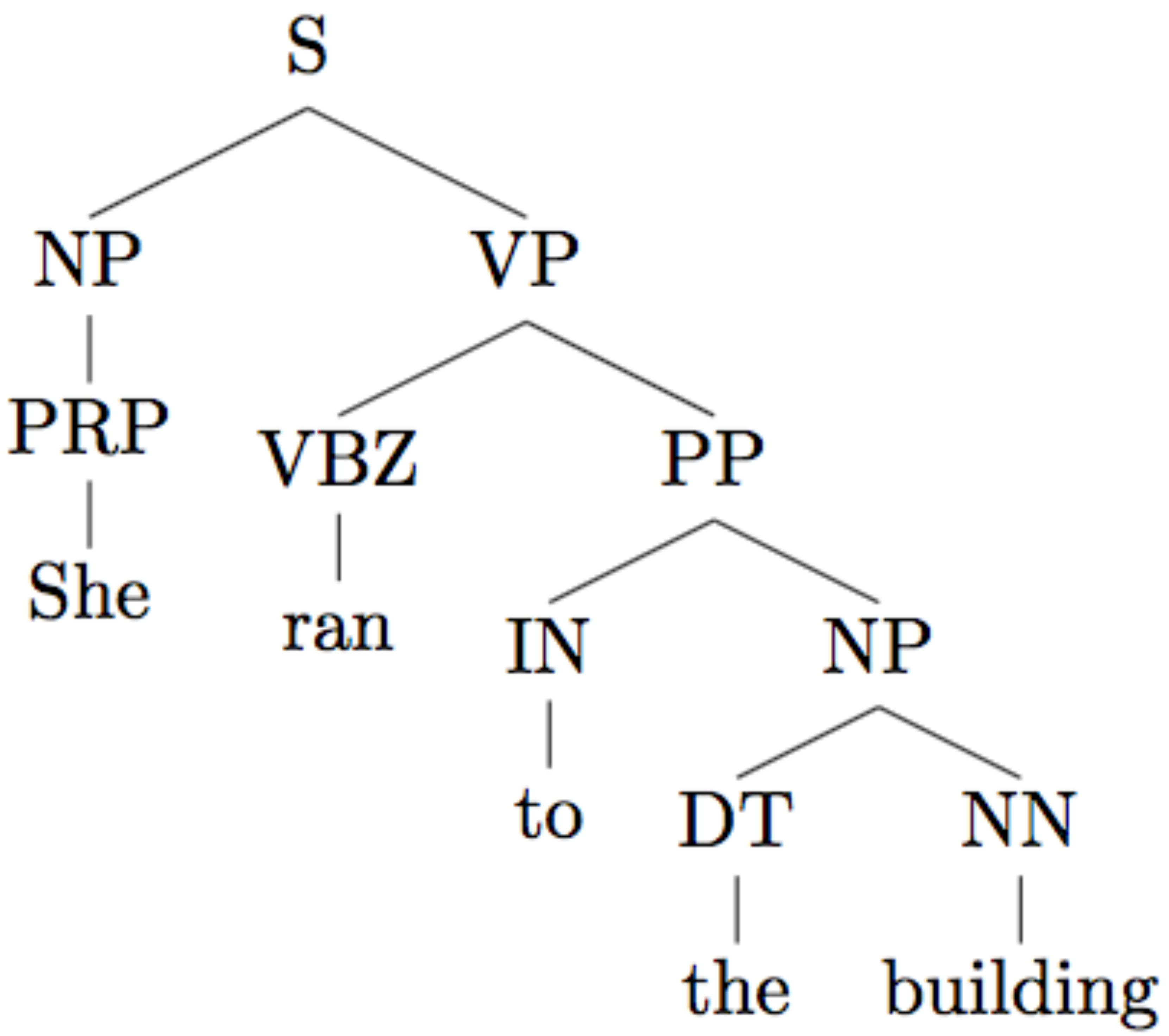
which words depend on (modify or are arguments of) which other words.

Constituency Parsing

- Phrase structure organizes words into nested constituents

- Constituent: a ***unit*** that can appear in different places
  - John talked to the children about drugs.
  - John talked [to the children] [about drugs].
  - John talked [about drugs] [to the children].
  - *John talked drugs to the children about

- Common constituents: noun phrases, verb phrases, prepositional phrases



Context-Free Grammar (CFG)

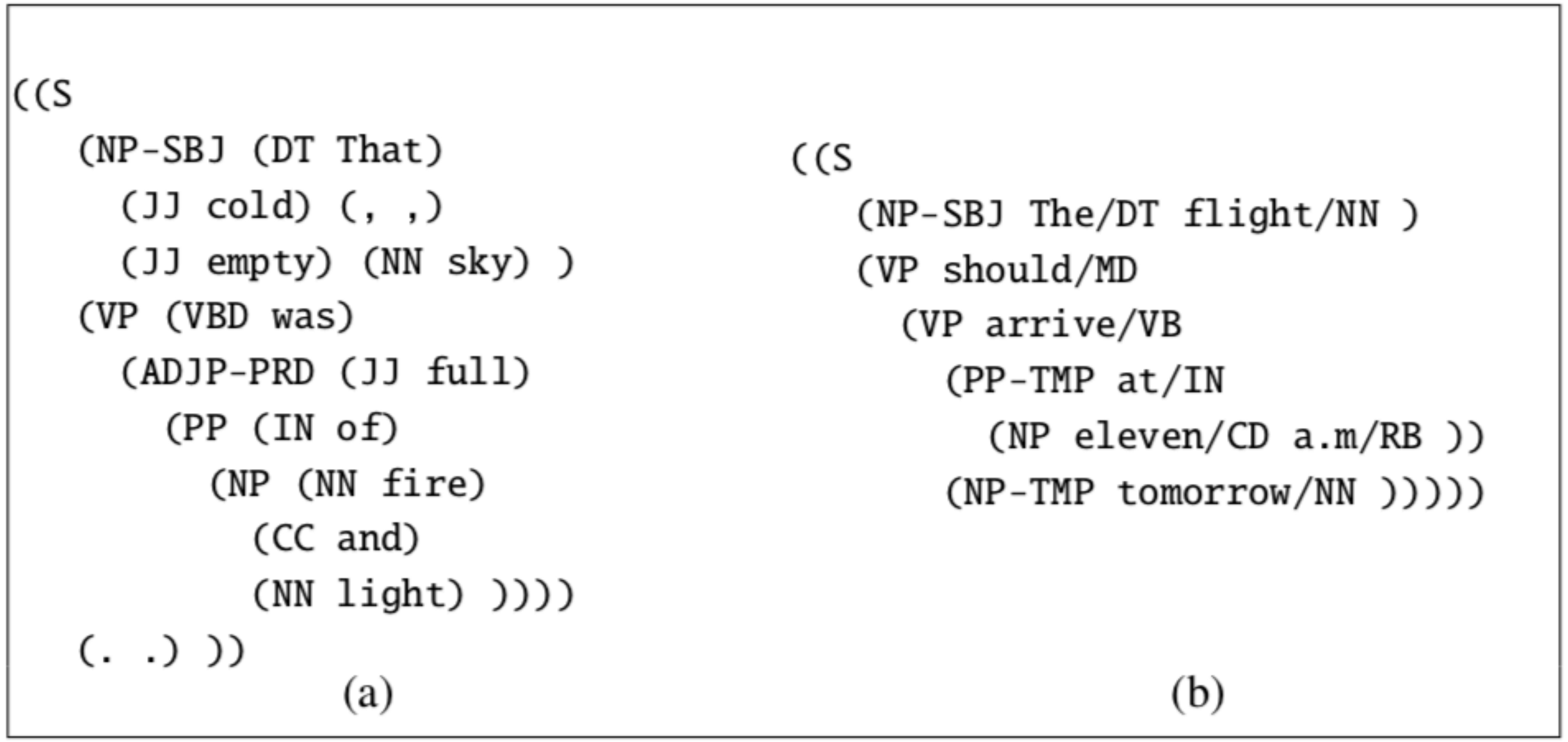
- Formal system for modeling constituency structure in language

- A context-free grammar is a tuple <N, Σ , S, R>
  - N: a set of non terminal symbols
  - Σ: a set of terminal symbols
  - R: set of rules
  - S: a start symbol

Statistical Parsing

- Learning from data: Treebanks
- Adding probabilities to the rules: probabilistic CFGs (PCFGs)

Treebanks: a collection of sentences paired with their parse trees



```
((S
   (NP-SBJ (DT That)
     (JJ cold) (, ,)
     (JJ empty) (NN sky) )
   (VP (VBD was)
     (ADJP-PRD (JJ full)
       (PP (IN of)
         (NP (NN fire)
           (CC and)
           (NN light) ))))
   (. .) ))
```

(a)

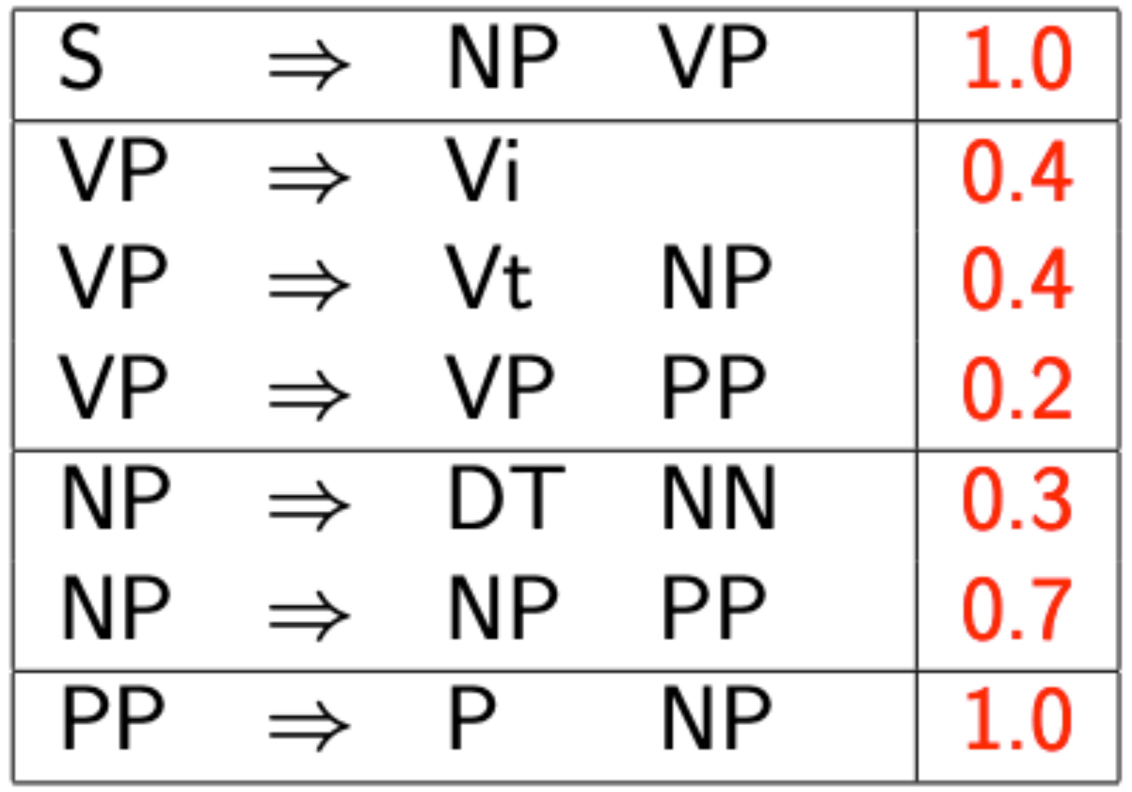
```
((S
   (NP-SBJ The/DT flight/NN )
   (VP should/MD
     (VP arrive/VB
       (PP-TMP at/IN
         (NP eleven/CD a.m/RB ))
       (NP-TMP tomorrow/NN )))))
```

(b)

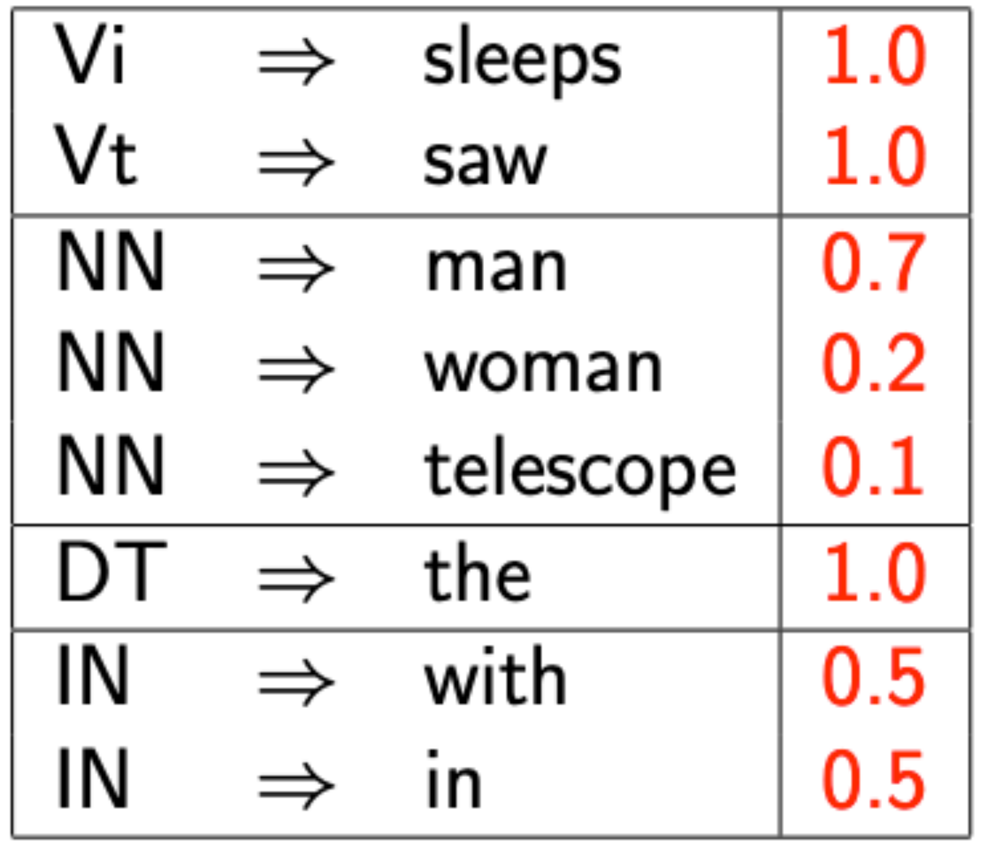
50K annotated sentences

The Penn Treebank Project (Marcus et al, 1993)

Probabilistic context free grammar



S	⇒	NP	VP	1.0
VP	⇒	Vi		0.4
VP	⇒	Vt	NP	0.4
VP	⇒	VP	PP	0.2
NP	⇒	DT	NN	0.3
NP	⇒	NP	PP	0.7
PP	⇒	P	NP	1.0



Vi	⇒	sleeps	1.0
Vt	⇒	saw	1.0
NN	⇒	man	0.7
NN	⇒	woman	0.2
NN	⇒	telescope	0.1
DT	⇒	the	1.0
IN	⇒	with	0.5
IN	⇒	in	0.5

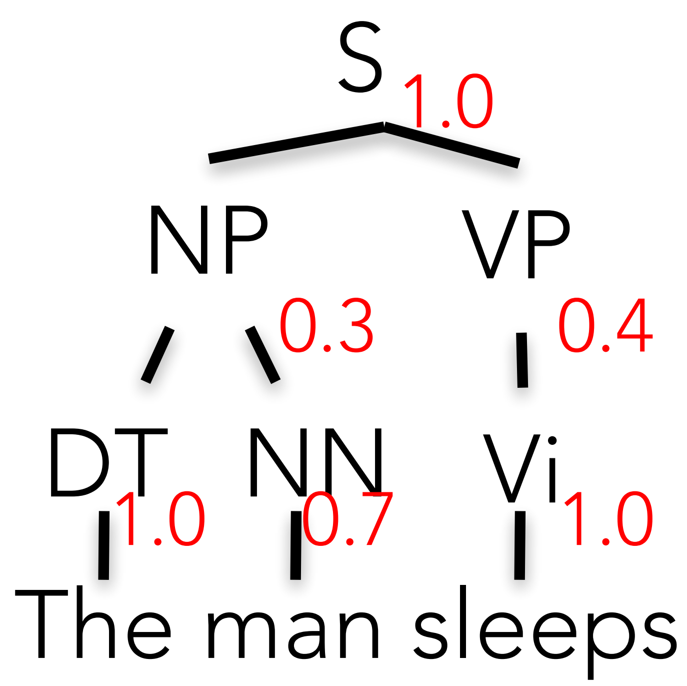
A context-free grammar: $G = (N, \Sigma, R, S)$

For each rule $\alpha \rightarrow \beta \in R$, there is a parameter $q(\alpha \rightarrow \beta) \geq 0$. For any $X \in N$,

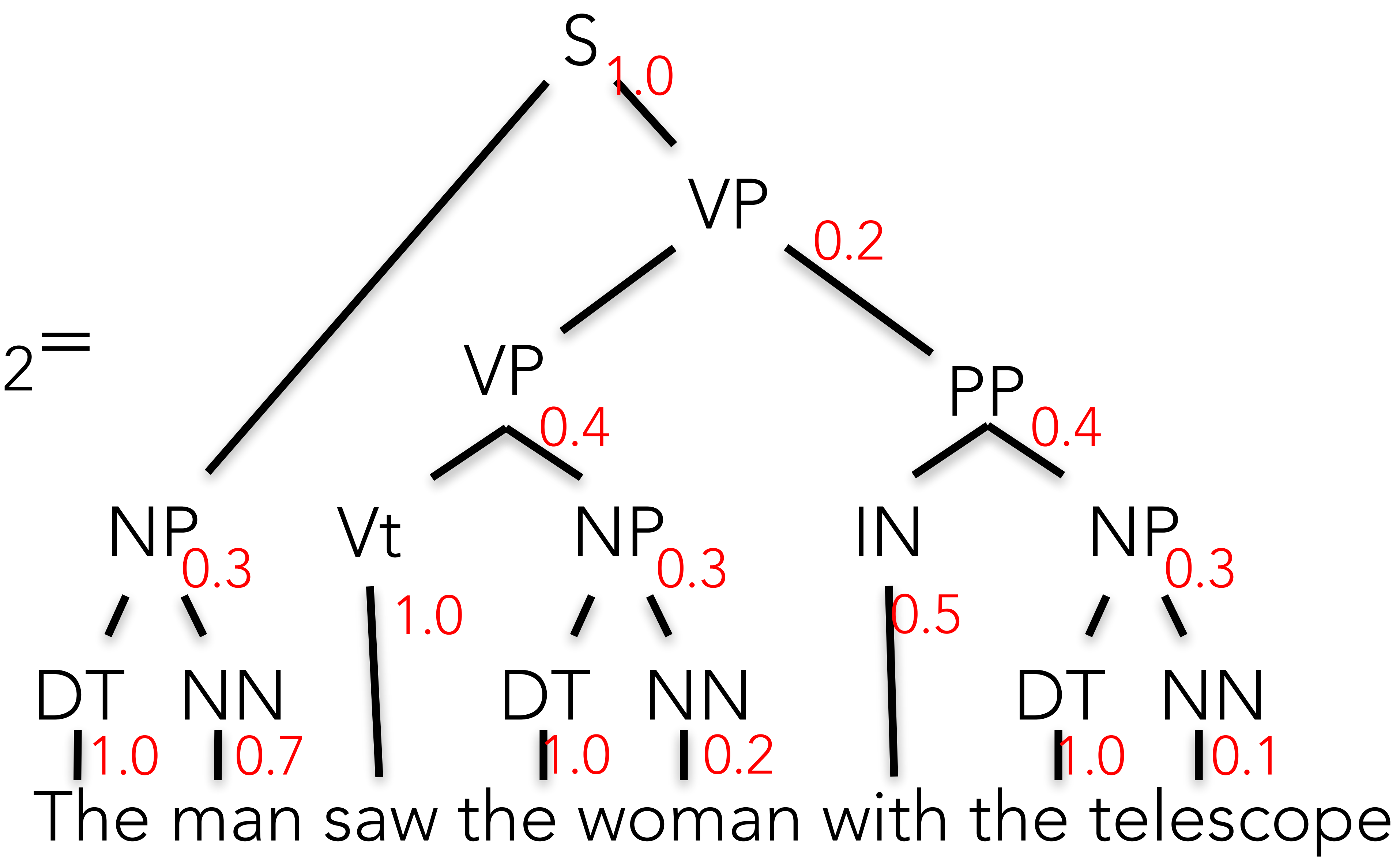
$$\sum_{\alpha \rightarrow \beta : \alpha = X} q(\alpha \rightarrow \beta) = 1$$

Tree Scorer

$t_1 =$



$t_2 =$



$$P(t_1) = q(\mathrm{S} \to \mathrm{NP\ VP}) \times q(\mathrm{NP} \to \mathrm{DT\ NN}) \times q(\mathrm{DT} \to \mathrm{the})$$

$$\times\, q(\mathrm{NN} \to \mathrm{man}) \times q(\mathrm{VP} \to \mathrm{Vi}) \times q(\mathrm{Vi} \to \mathrm{sleeps})$$

$$= 1.0 \times 0.3 \times 1.0 \times 0.7 \times 0.4 \times 1.0 = 0.084$$

$p(t_s)$=1.0*0.3*1.0*0.7*0.2*0.4*1.0*0.3*1.0*0.2*0.4*0.5*0.3*1.0*0.1

Recap: Deriving a PCFG from a treebank

- Training data: a set of parse trees $t_1, t_2, \ldots, t_m$
- A PCFG (N, Σ, S, R, q):
  - N is the set of all non-terminals seen in the trees
  - Σ is the set of all words seen in the trees
  - S is taken to be S.
  - R is taken to be the set of all rules $\alpha \rightarrow \beta$ seen in the trees
- The maximum-likelihood parameter estimates are:

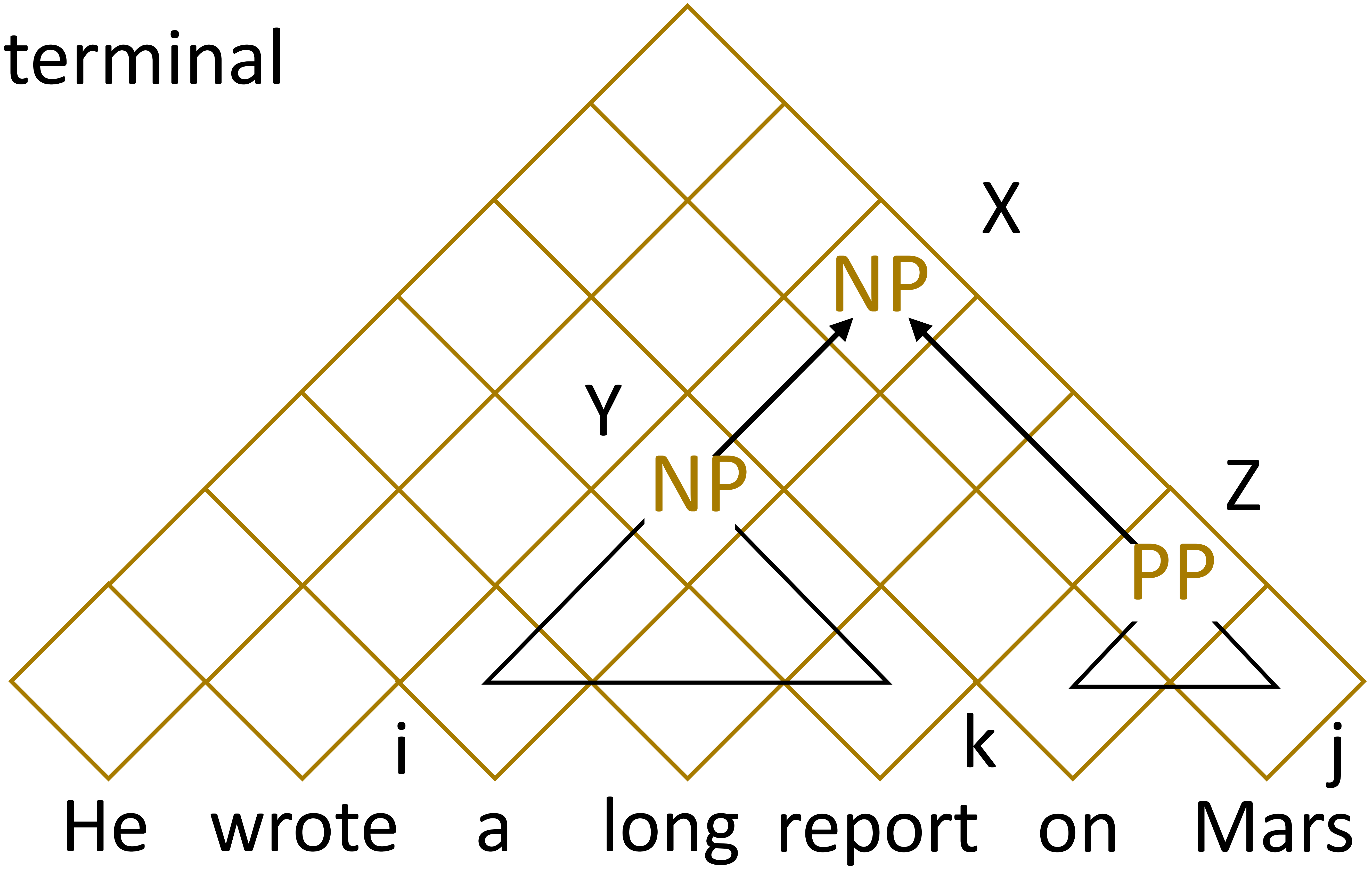
$$q_{ML}(\alpha \rightarrow \beta) = \frac{\text{Count}(\alpha \rightarrow \beta)}{\text{Count}(\alpha)}$$

If we have seen the rule VP → Vt NP 105 times, and the the non-terminal VP 1000 times, $q(\text{VP} \rightarrow \text{Vt NP}) = 0.105$

Inference: Cocke-Kasami-Younger (CKY algorithm)

- Given a sentence $x_1, x_2, \ldots, x_n$, denote T[i,j,X] as the highest score for any parse tree that covers words $x_i, \ldots, x_j$ with non-terminal $X \in N$ as its root.

- Find T[l, n, S]



Cocke-Kasami-Younger (CKY algorithm)

- Chart: T[i,j,X] = best score

- Base: T[i,i,X] = log P(X → w_i)

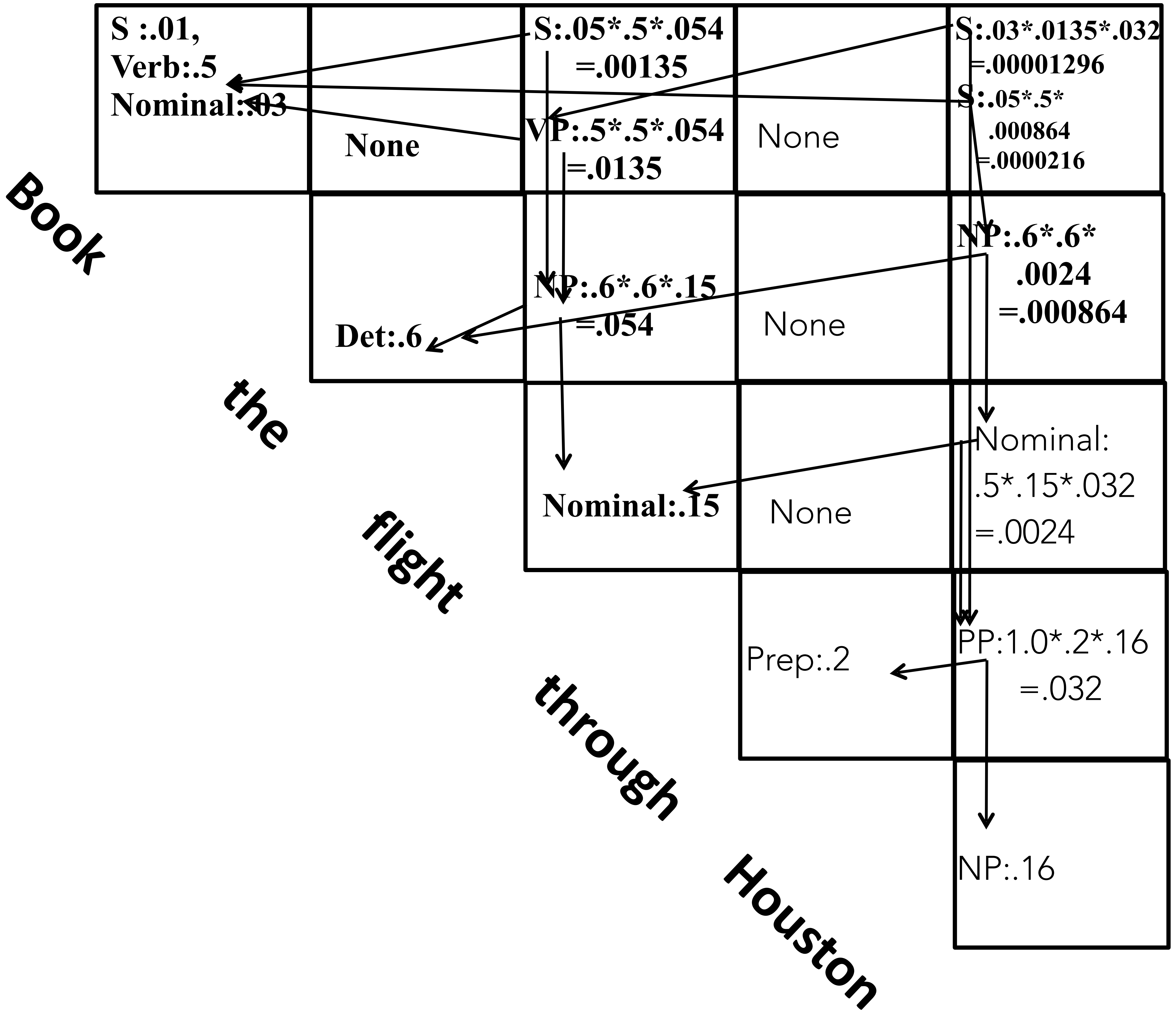
- For each constituent length l
  - For each left end point i
    - For each split point k
      - For each rule X → Y Z
        - Do constant work

- Recurrence:

$$T[i,j,X] = \max_{\substack{k \\ l<=k<j}} \quad \max_{r\,:\,X \to X1\ X2} \quad T[i,k,X1] + T[k+1,j,X2] + \log P(X \to X1\ X2)$$

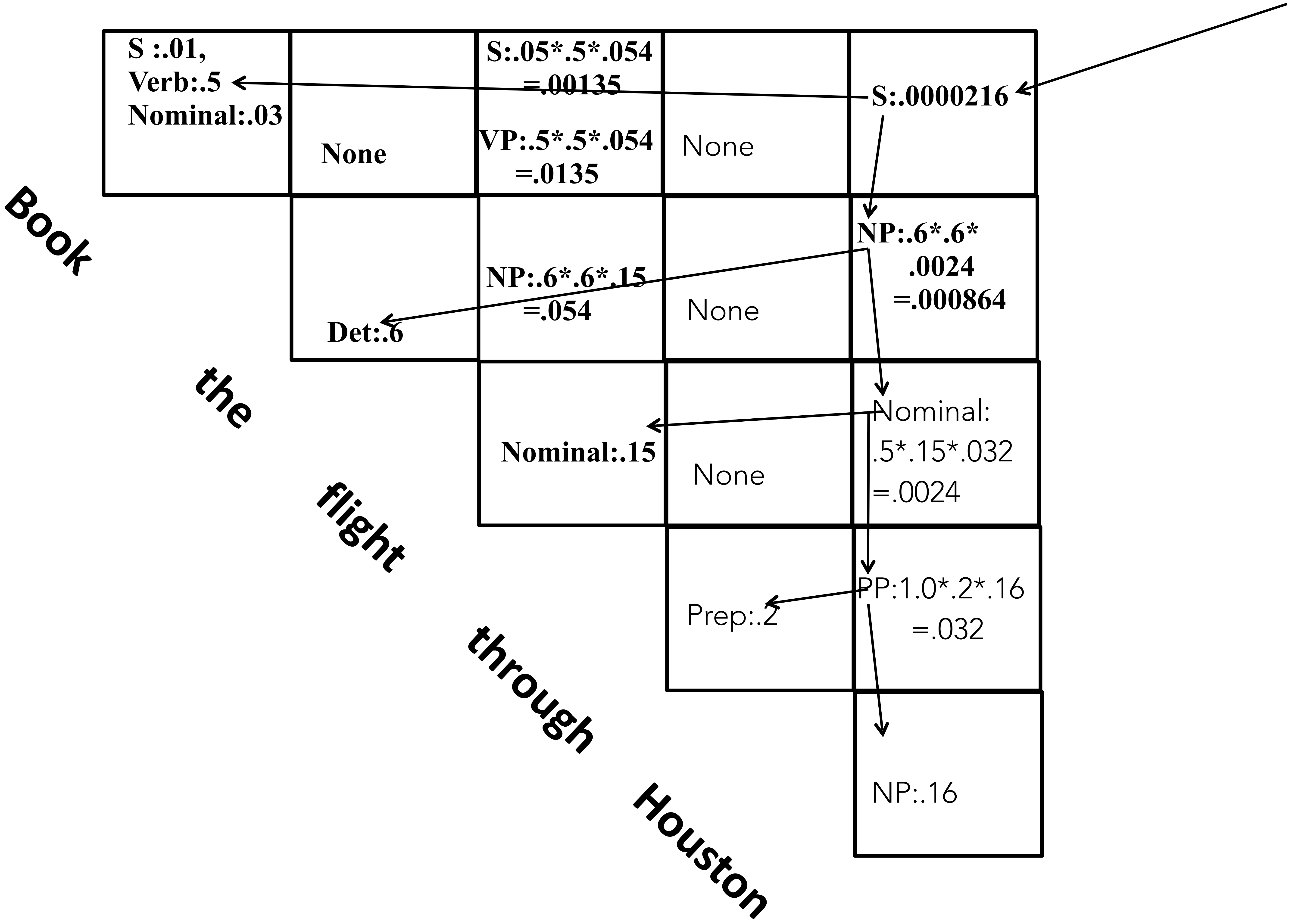
CKY [Example]

S → NP VP 0.8
S → X1 VP 0.1
X1 → Aux NP 1.0
S → book | include | prefer
0.01 0.004 0.006
S → Verb NP 0.05
S → VP PP 0.03
NP → I | he | she | me
0.1 0.02 0.02 0.06
NP → Houston | NWA
0.16 .04
Det→ the | a | an
0.6 0.1 0.05
NP → Det Nominal 0.6
Nominal → book | flight | meal | money
0.03 0.15 0.06 0.06
Nominal → Nominal Nominal 0.2
Nominal → Nominal PP 0.5
Verb→ book | include | prefer
0.5 0.04 0.06
VP → Verb NP 0.5
VP → VP PP 0.3
Prep → through | to | from
0.2 0.3 0.3
PP → Prep NP 1.0



Book	S :.01, Verb:.5 Nominal:.03	None	S:.05*.5*.054 =.00135 VP:.5*.5*.054 =.0135	None	S:.03*.0135*.032 =.00001296 S:.05*.5* .000864 =.0000216
the		Det:.6	NP:.6*.6*.15 =.054	None	NP:.6*.6* .0024 =.000864
flight			Nominal:.15	None	Nominal: .5*.15*.032 =.0024
through				Prep:.2	PP:1.0*.2*.16 =.032
Houston					NP:.16

CKY [Example]

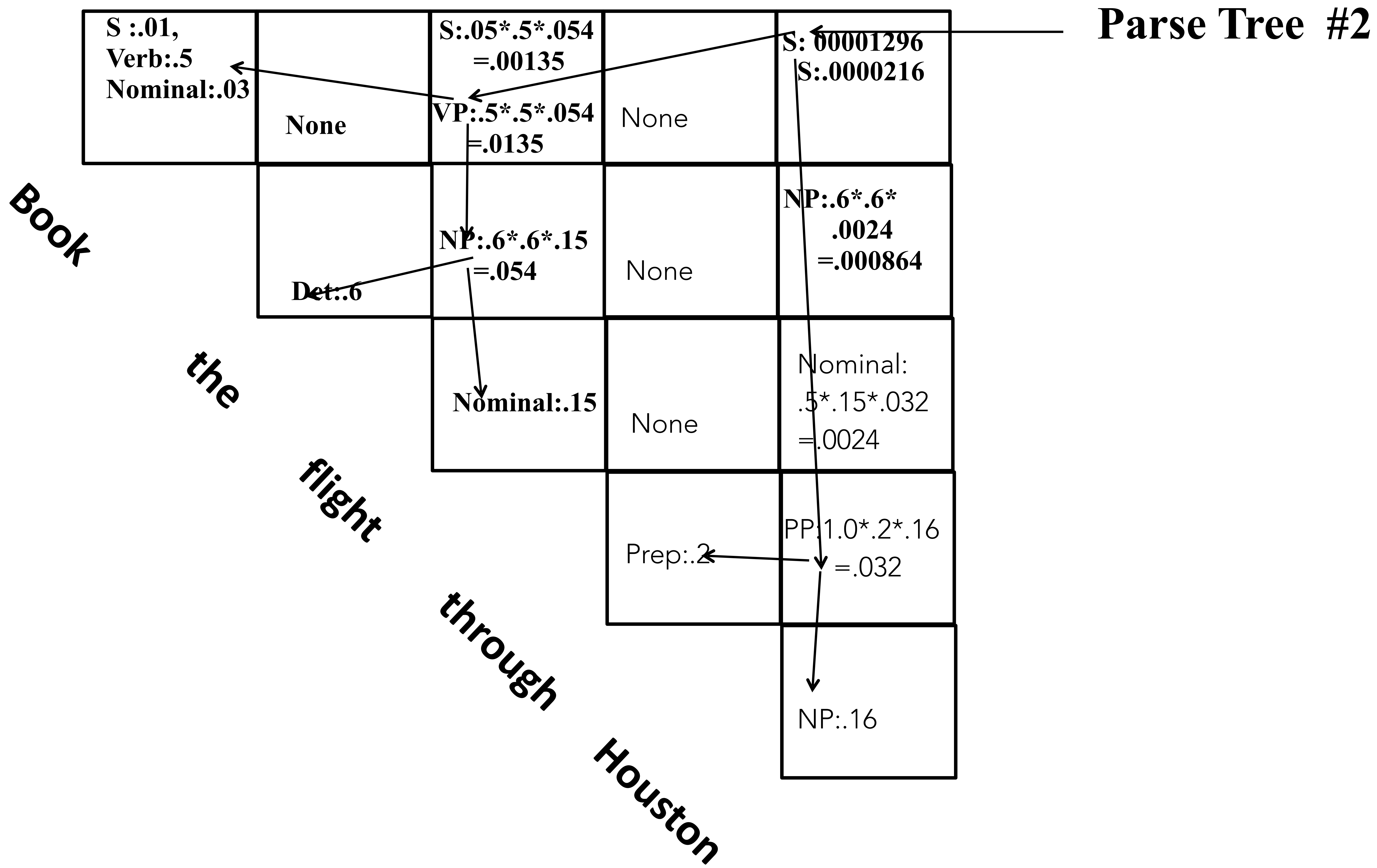


Parse Tree #1

Book	S :.01, Verb:.5 Nominal:.03	None	S:.05*.5*.054 =.00135 VP:.5*.5*.054 =.0135	None	S:.0000216
the		Det:.6	NP:.6*.6*.15 =.054	None	NP:.6*.6* .0024 =.000864
flight			Nominal:.15	None	Nominal: .5*.15*.032 =.0024
through				Prep:.2	PP:1.0*.2*.16 =.032
Houston					NP:.16

Pick most probable parse, i.e. take max to combine probabilities of multiple derivations of each constituent in each cell.

CKY [Example]



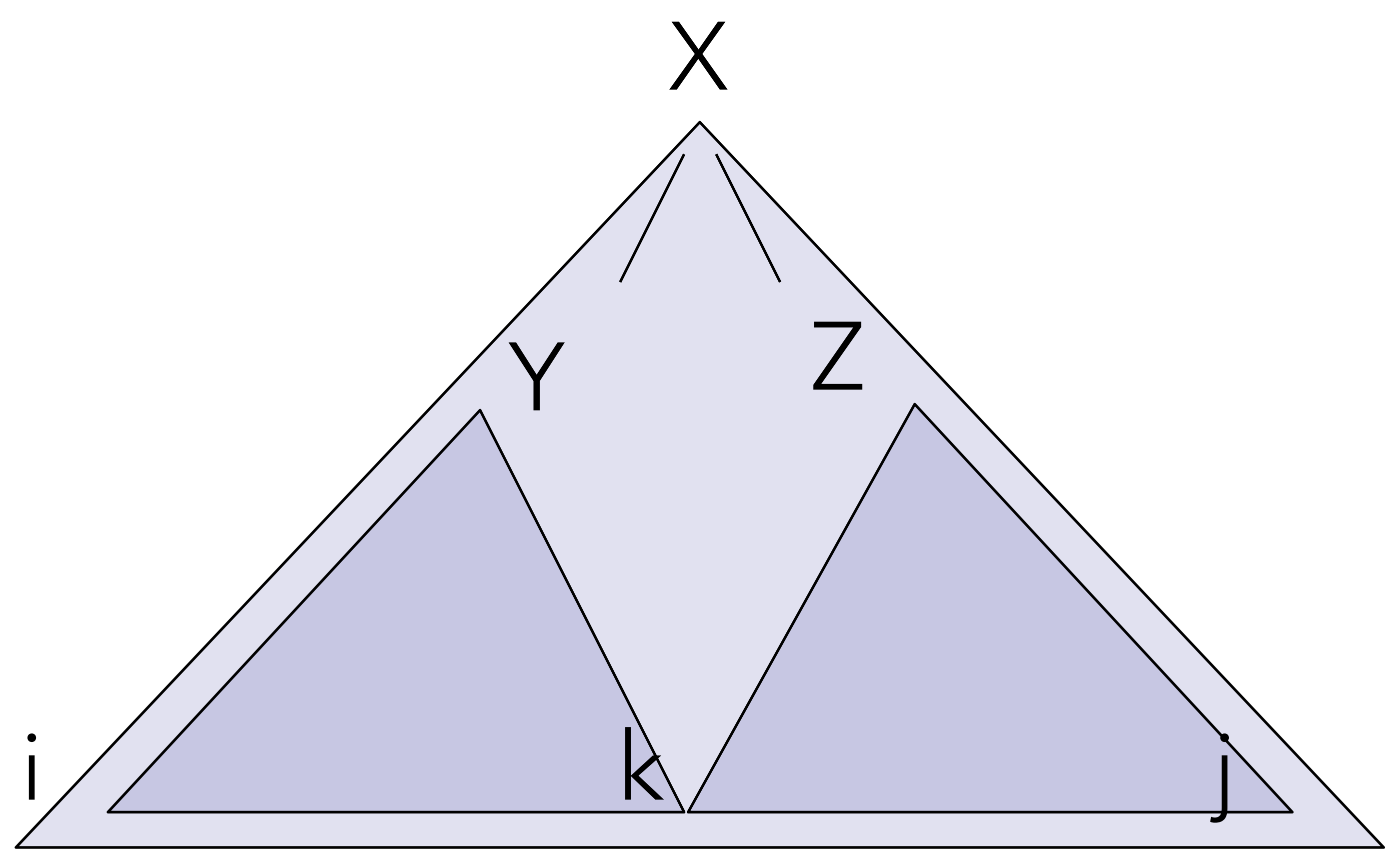
Parse Tree #2

Book the flight through Houston

S :.01, Verb:.5 Nominal:.03	None	S:.05*.5*.054 =.00135 VP:.5*.5*.054 =.0135	None	S: 00001296 S:.0000216
	Det:.6	NP:.6*.6*.15 =.054	None	NP:.6*.6* .0024 =.000864
		Nominal:.15	None	Nominal: .5*.15*.032 =.0024
			Prep:.2	PP:1.0*.2*.16 =.032
				NP:.16

Efficiency?

- For each constituent length l (<= n)
  - For each left end point i (<= n)
    - For each split point k
      - For each rule X → Y Z
        - Do constant work



- Runtime: O(n^3*|R|) R = grammar constant
- Memory:
  - Need to store score caches
  - Cache size: |Symbols| * n^2

Efficiency?

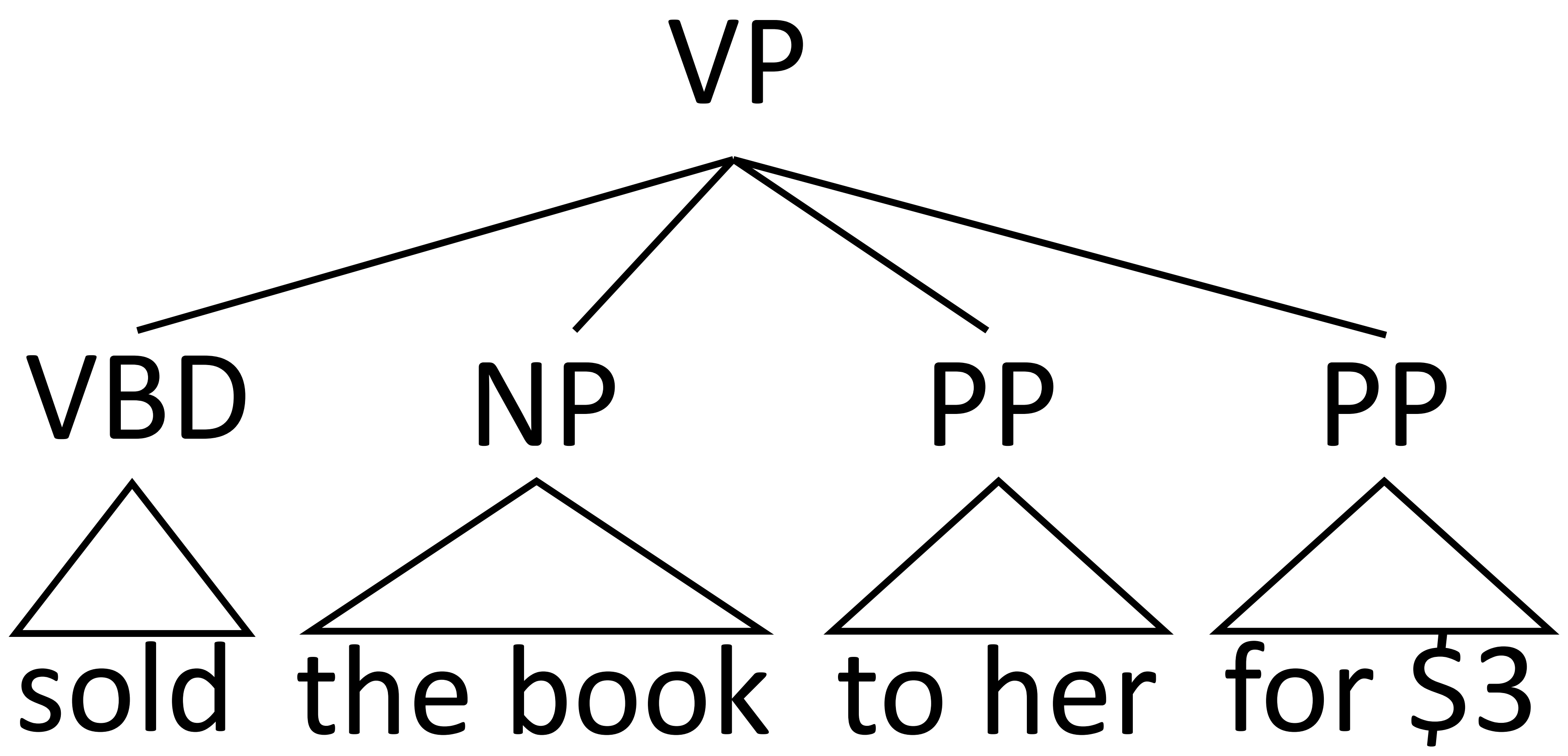
- Can we keep N-ary (N>2) rules and still do dynamic programming?

- Can we keep unary rules and still do dynamic programming?

- Binary trees over n words have at most n-1 nodes, but you can have unlimited numbers of nodes with unaries (S → SBAR → NP → S → …)

- We introduce Chomsky Normal Form (CNF):
  - All rules are either X → Y Z or X → w
  - Rewriting PCFG grammar into equivalent CNF is possible

Binarization

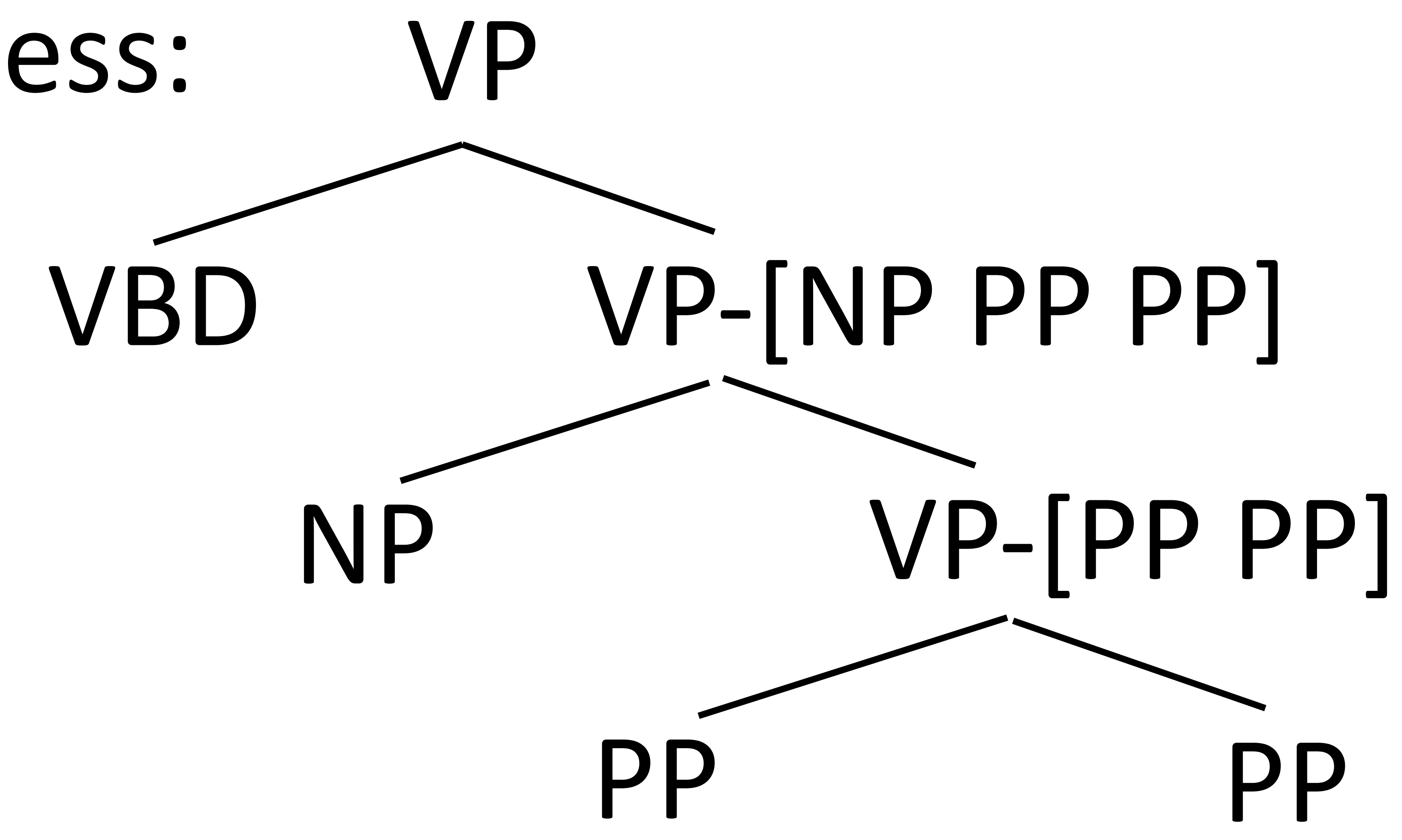


P(VP → VBD NP PP PP) = 0.2

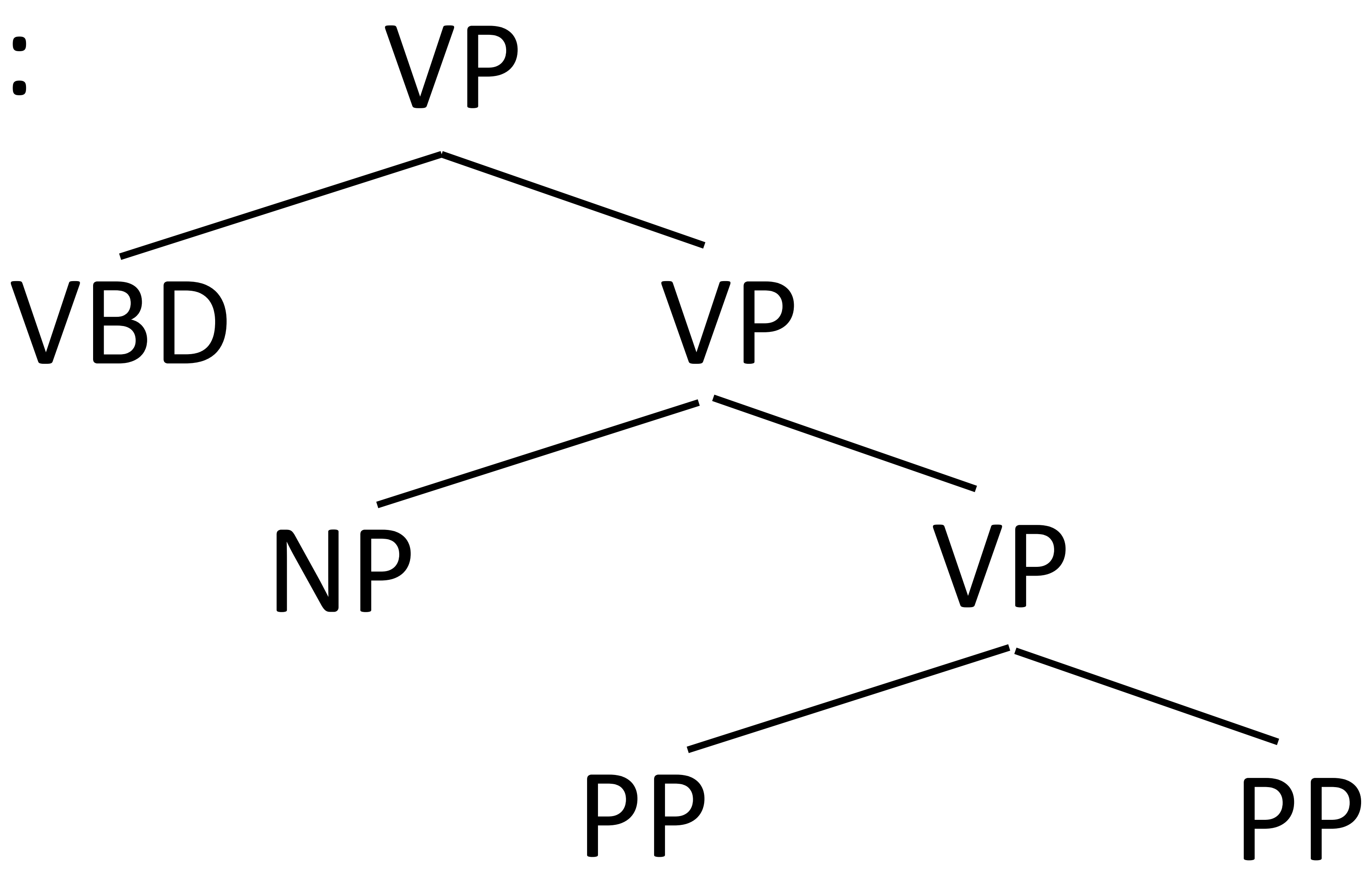
P(VP → VBZ PP) = 0.1

…

- Lossless:

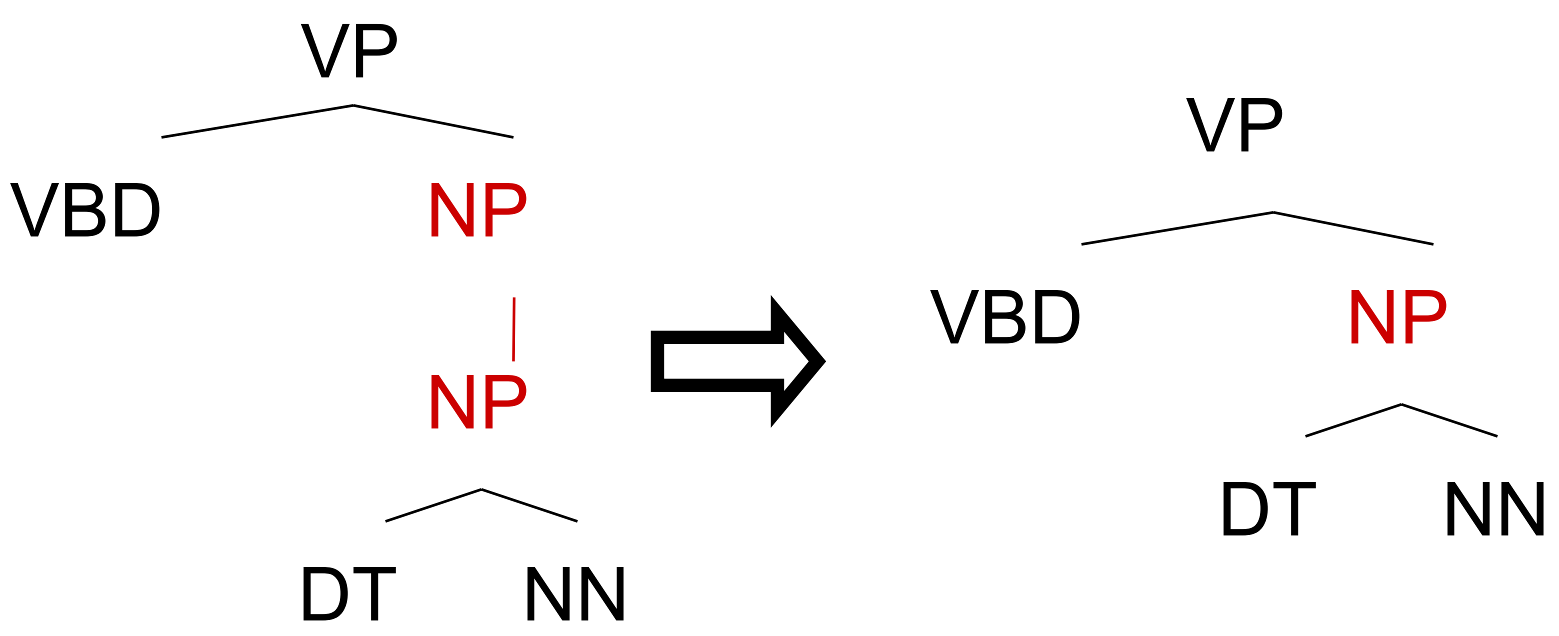


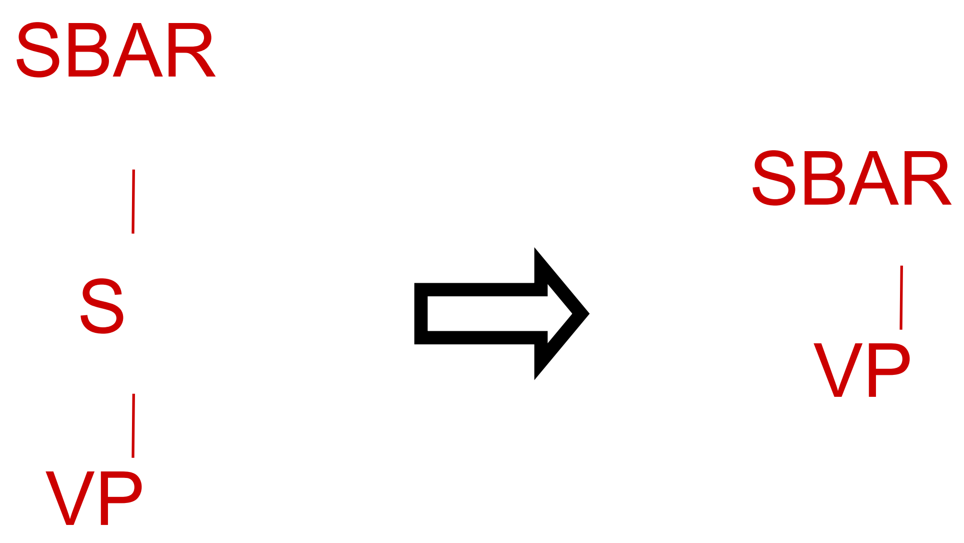
- Lossy:



Unary Rules

- Enforce at most one unary over each span by modifying grammar

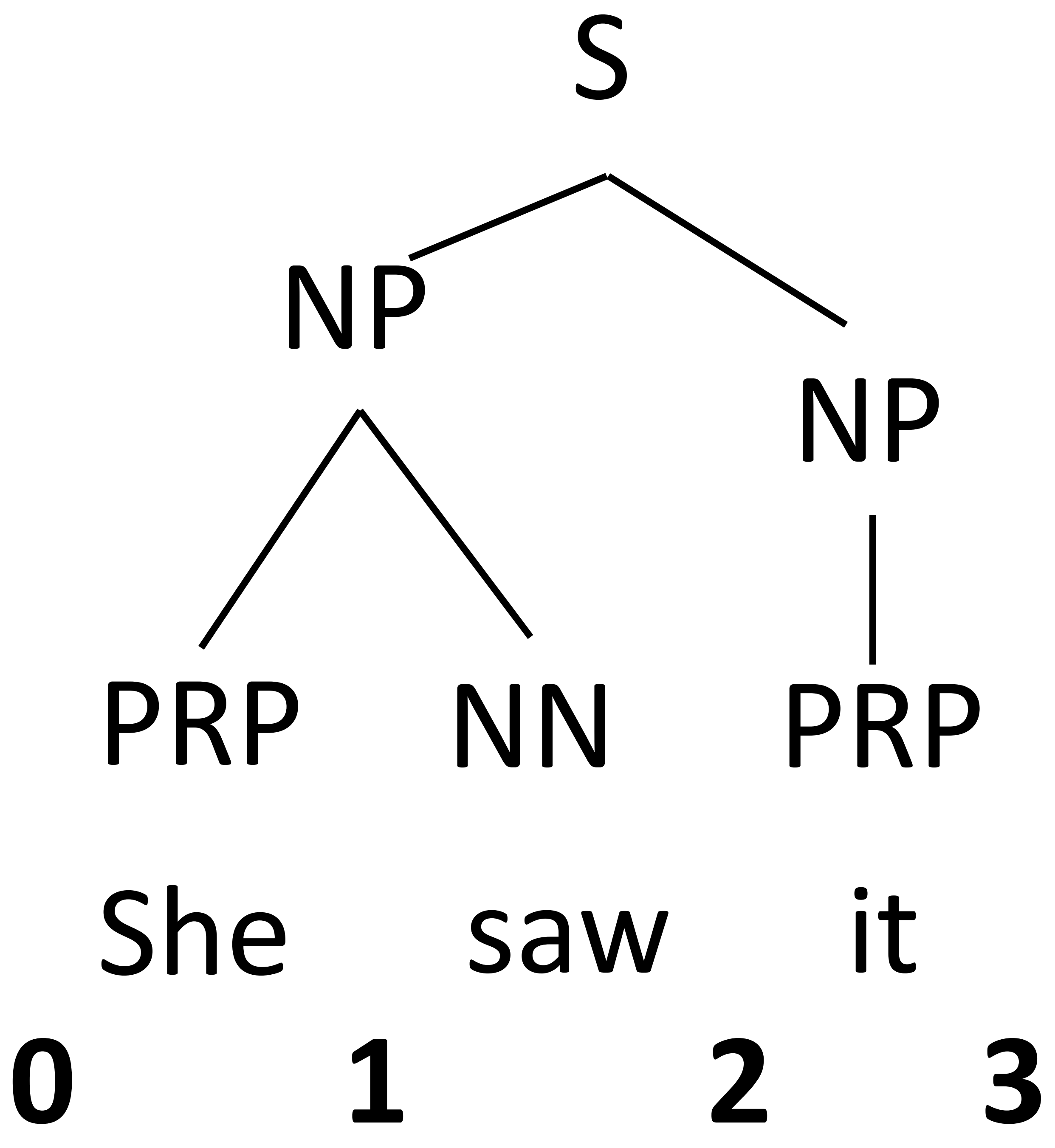




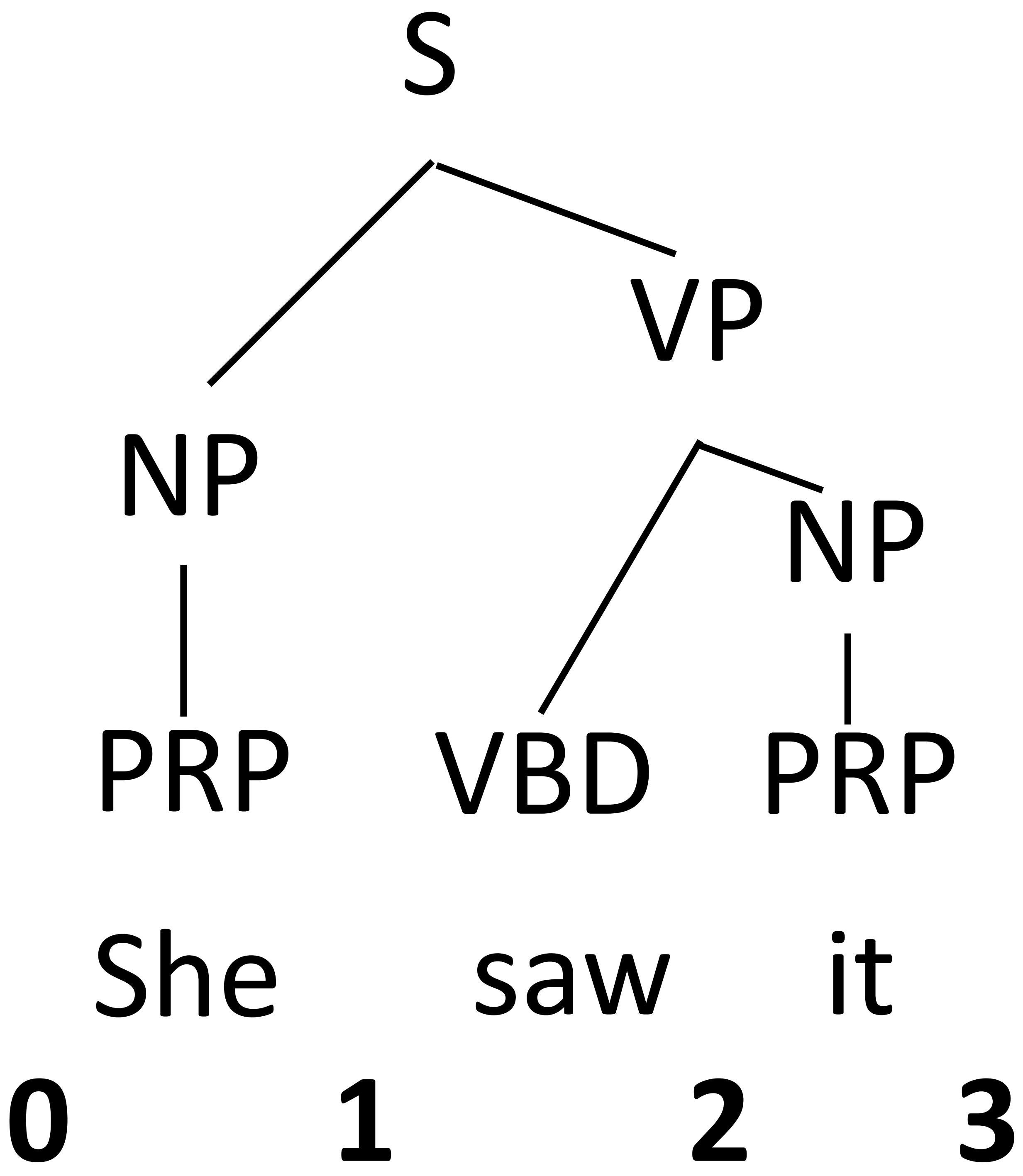
Compute unary closure: if there is a rule chain $X \to Y_1, Y_1 \to Y_2, \ldots, Y_k \to Y$, add $q(X \to Y) = q(X \to Y_1) \times \cdots \times q(Y_k \to Y)$

- In CKY: Update unary rule once after the binary rules

Parser Evaluation



S(0,3),
NP(0,2),
NP(2,3),
~~PRP(0,1),~~
~~NN(1,2),~~
~~PRP(2,3)~~



S(0,3),
NP(0,1),
VP(1,3),
NP(2,3),
~~PRP(0,1),~~
~~VBD(1,2),~~
~~PRP(2,3)~~

- Precision: number of correct brackets / num pred brackets = 2/3
- Recall: number of correct brackets / num of gold brackets = 2/4
- F1: harmonic mean of precision and recall = $(1/2 * ((2/4)^{-1} + (2/3)^{-1}))^{-1}$

  $= 0.57$

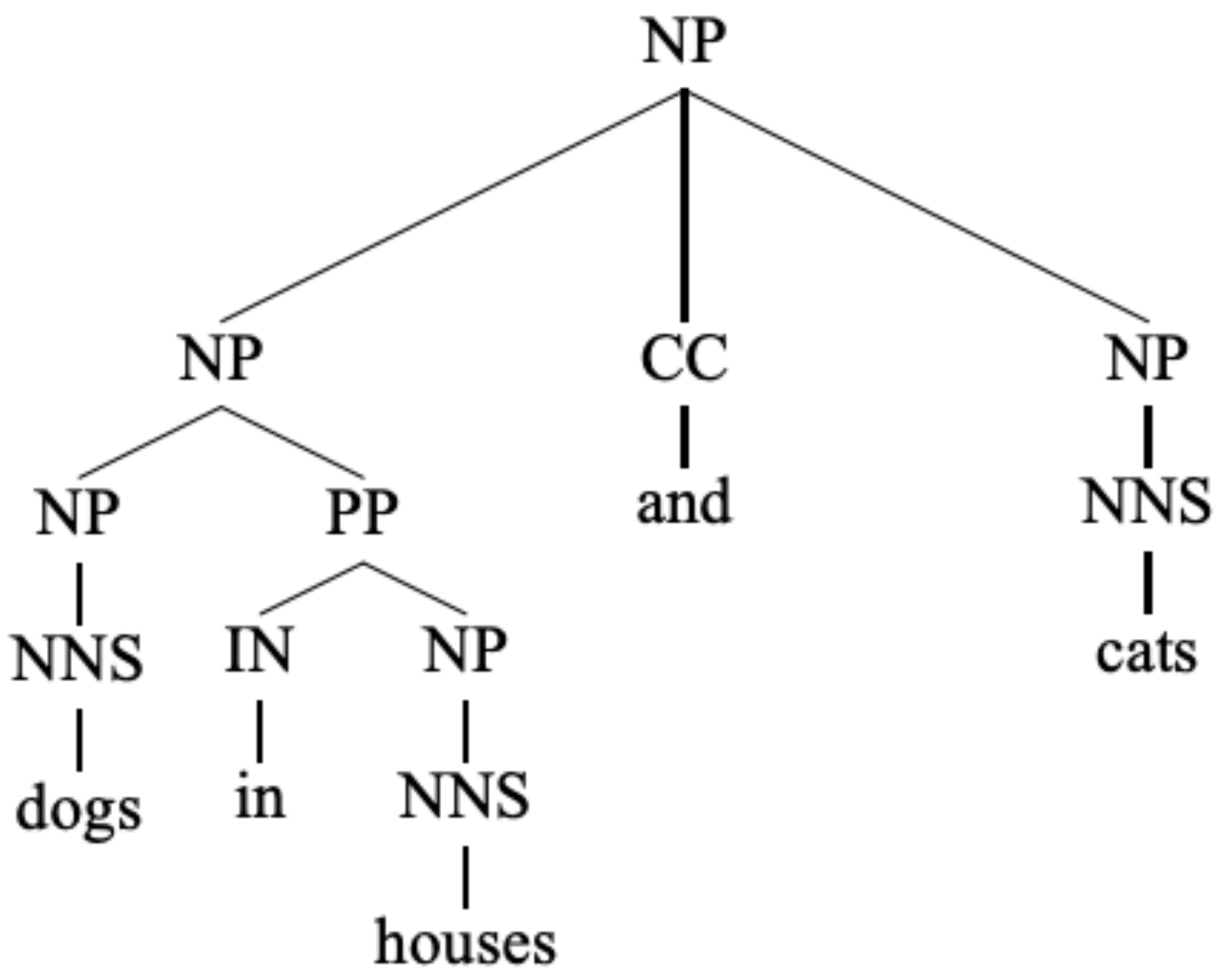
Results

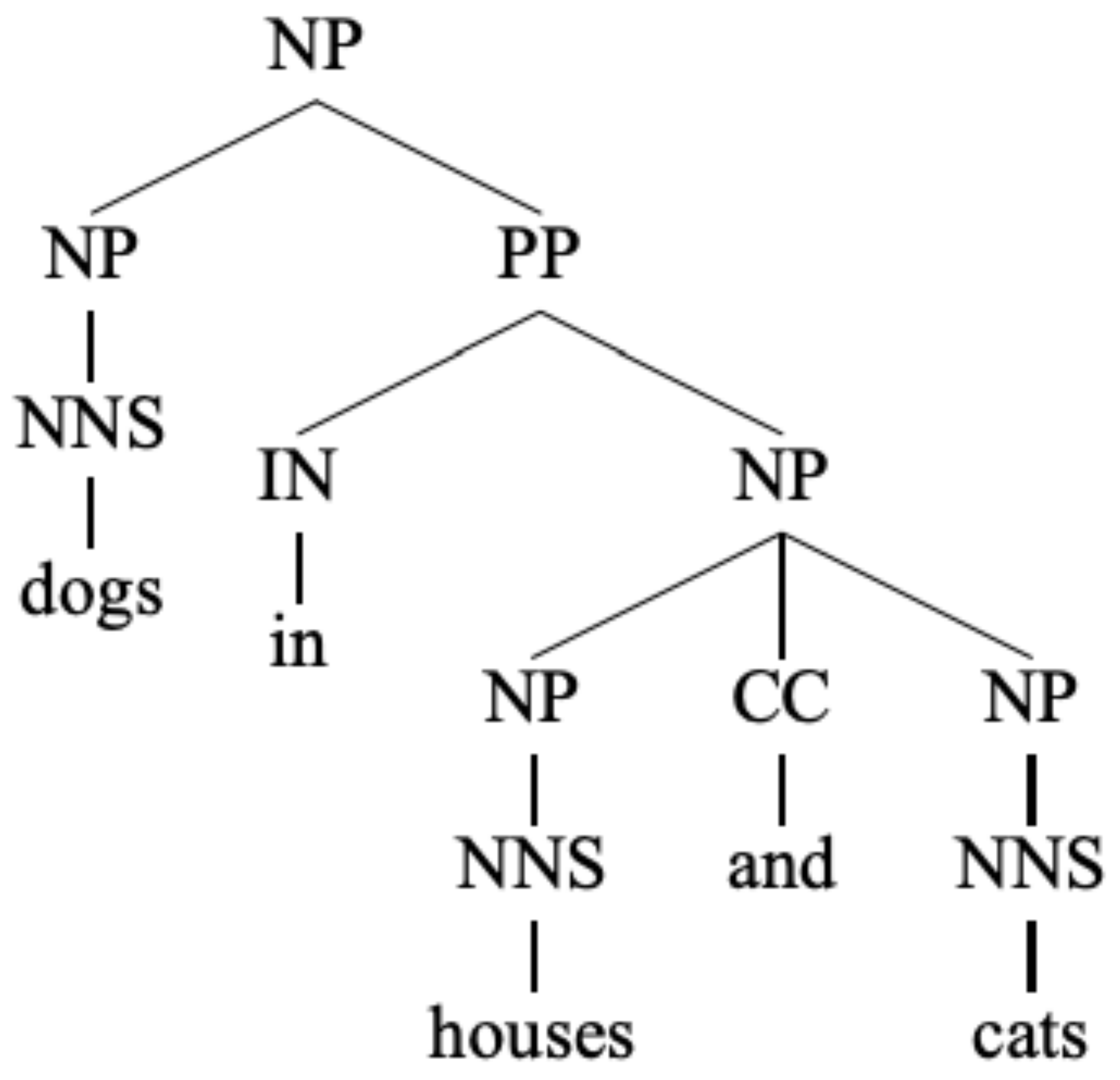
- Standard dataset for English: Penn Treebank (Marcus et al., 1993)
  - Evaluation: F1 over labeled constituents of the sentence
- Vanilla PCFG: ~75 F1
- Best PCFGs for English: ~90 F1
- SOTA (discriminative models): 95 F1
- Other languages: results vary widely depending on annotation + complexity of the grammar

Klein and Manning (2003)

Weaknesses of PCFG

- Lexical information (words) is lost!





Two trees have the exact same set of PCFG rules!

Lexicalized Parsers

- Annotate each grammar symbol with its “head w
  word of t

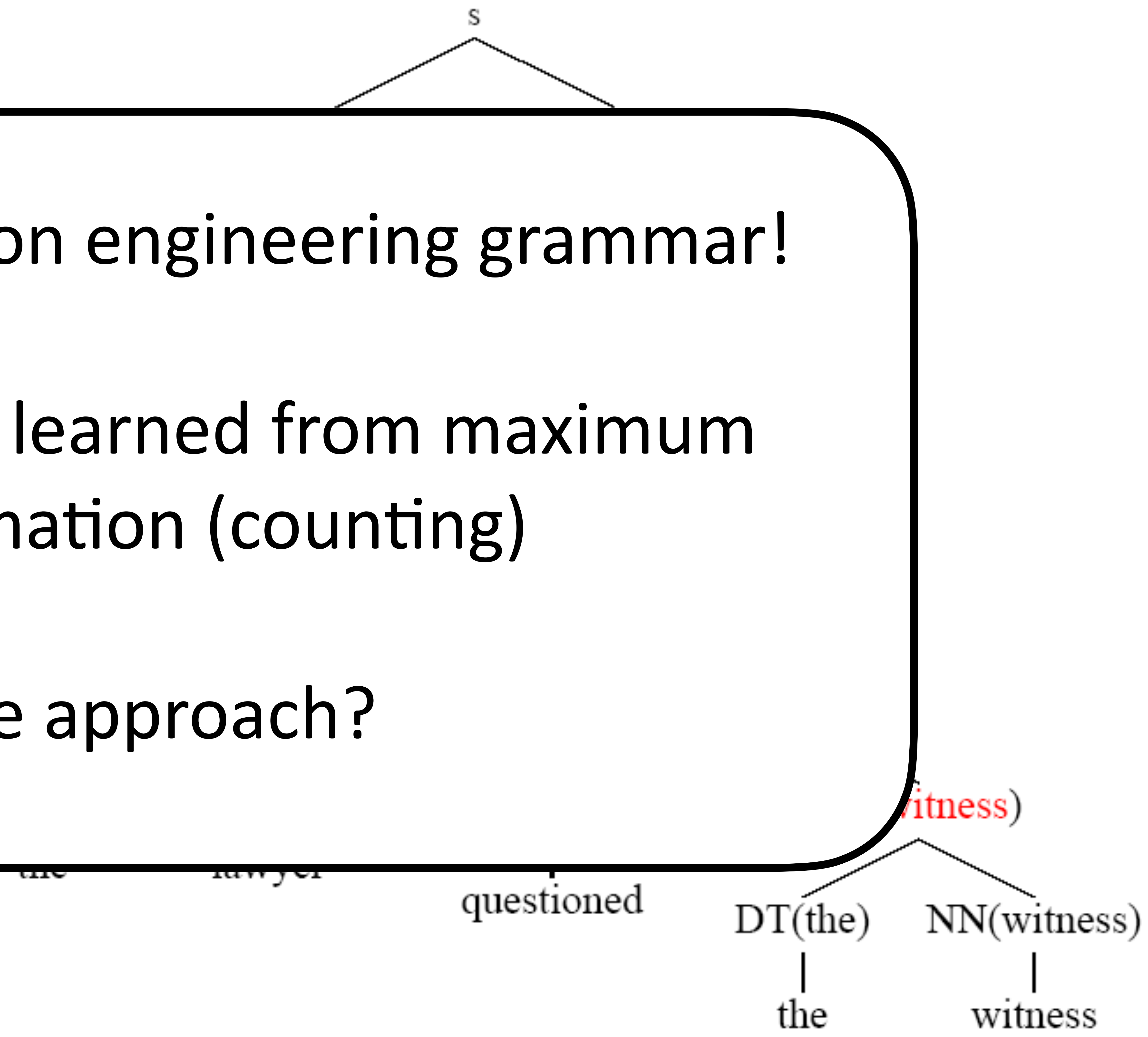
- Rules for
  the last w
  prepositi

- Collins ar
  ~89 F1 with these

All these work focuses on engineering grammar!

Constituency parser is learned from maximum likelihood estimation (counting)

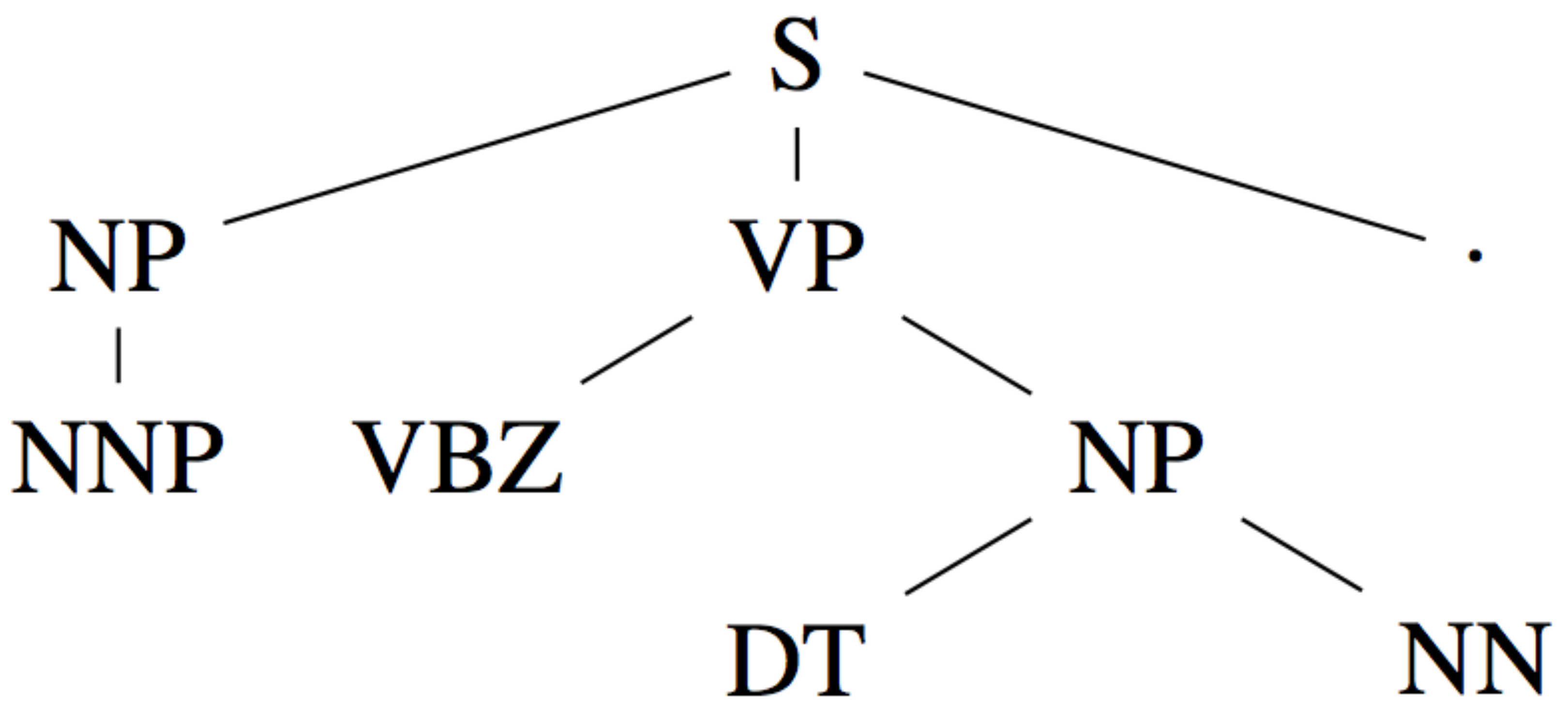
Alternative approach?



“Grammar as Foreign Language” (deep learning)

Vinyals et al., 2015

John has a dog ➔



John has a dog ➔

$(\text{S } (\text{NP NNP })_{\text{NP}} \ (\text{VP VBZ } (\text{NP DT NN })_{\text{NP}} \ )_{\text{VP}} \ . \ )_{\text{S}}$

- Linearize parse trees into a sequence
- Then parsing becomes similar to machine translation, takes a sentence as an input sequence and output a parse tree sequence
- Data augmentation tricks, gets up to 92 F1

Today

- Trees
  - Constituency Trees
    - How to find best-scoring trees given probabilistic CFG grammar
  - Dependency Trees vs. Constituency Trees

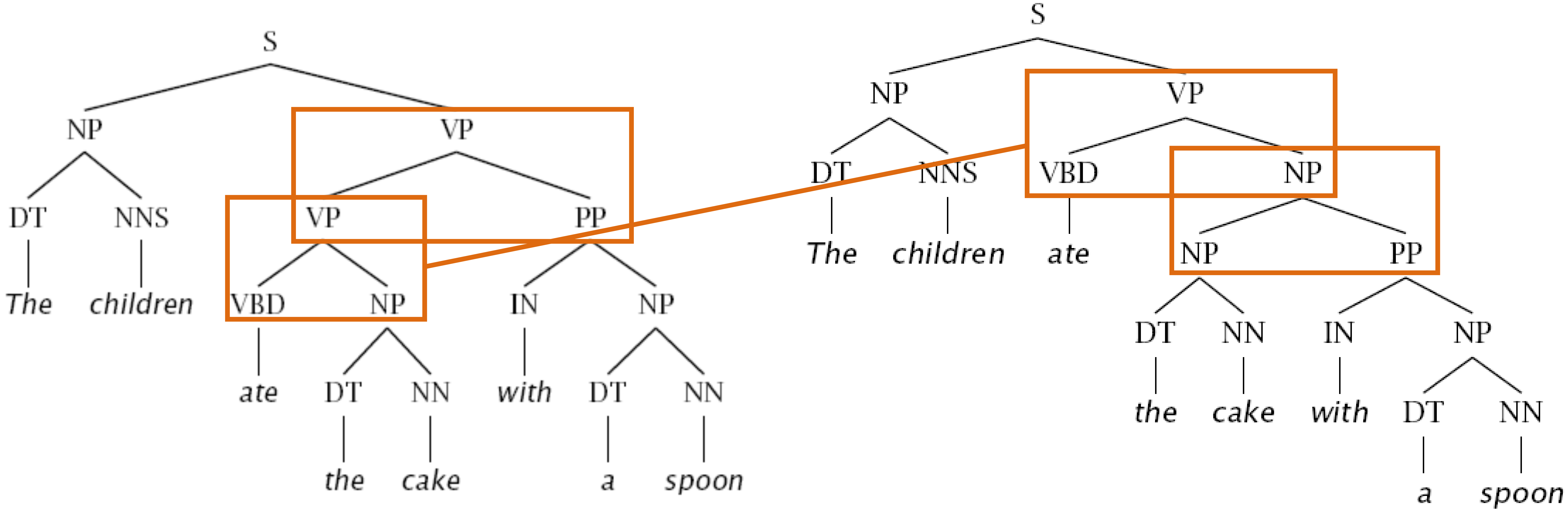
- Applications
  - Question Answering

Dependency parsing vs. Constituency parsing

- Can we get transform dependency parse into constituency parse? How about the other way around?
  - Mostly yes, with some caveats
    - Dependency parse can capture non-projective dependencies, while constituency parse cannot
    - Mapping from constituency to dependency edges and head is heuristic, somewhat lossy.

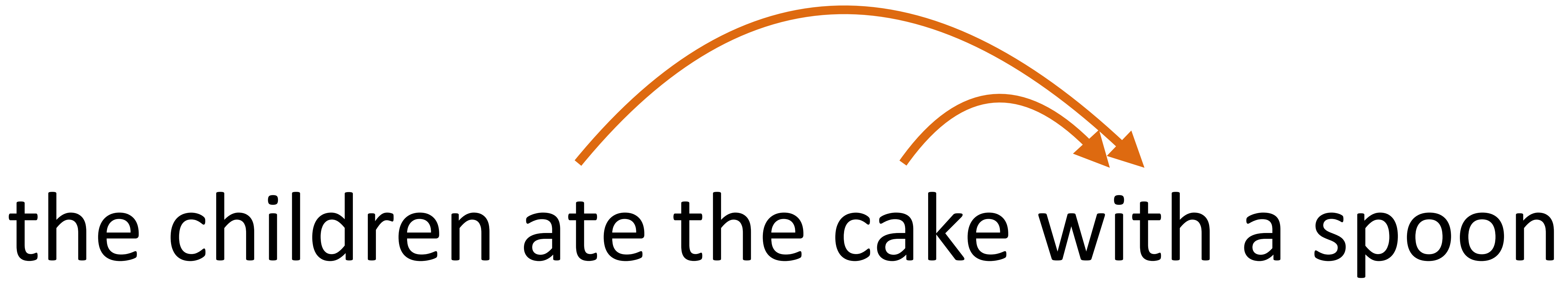
Dependency vs. Constituency: PP Attachment

- Constituency: several rule productions need to change



Dependency vs. Constituency: PP Attachment

- Dependency: one word (with) assigned a different parent



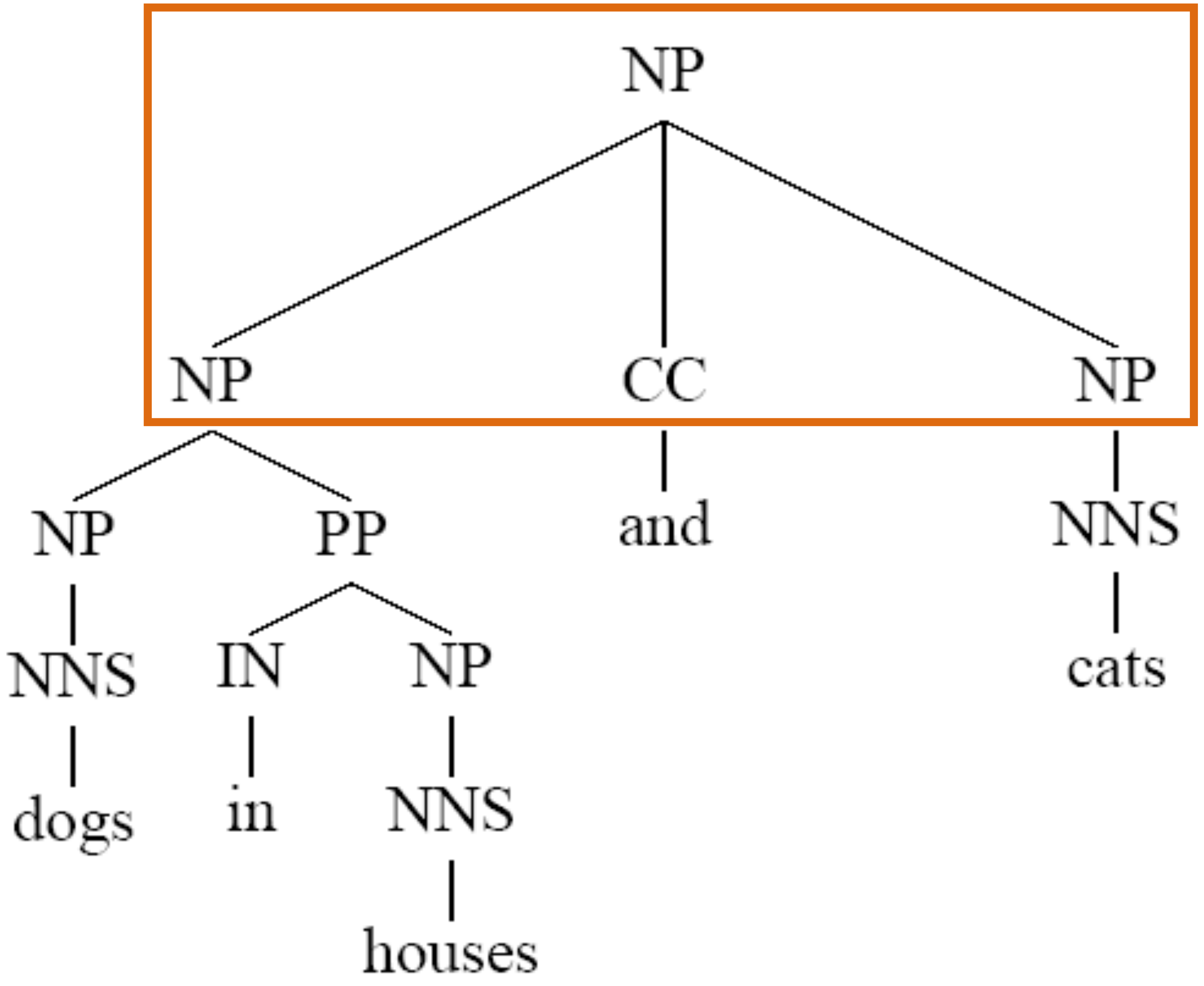
the children ate the cake with a spoon

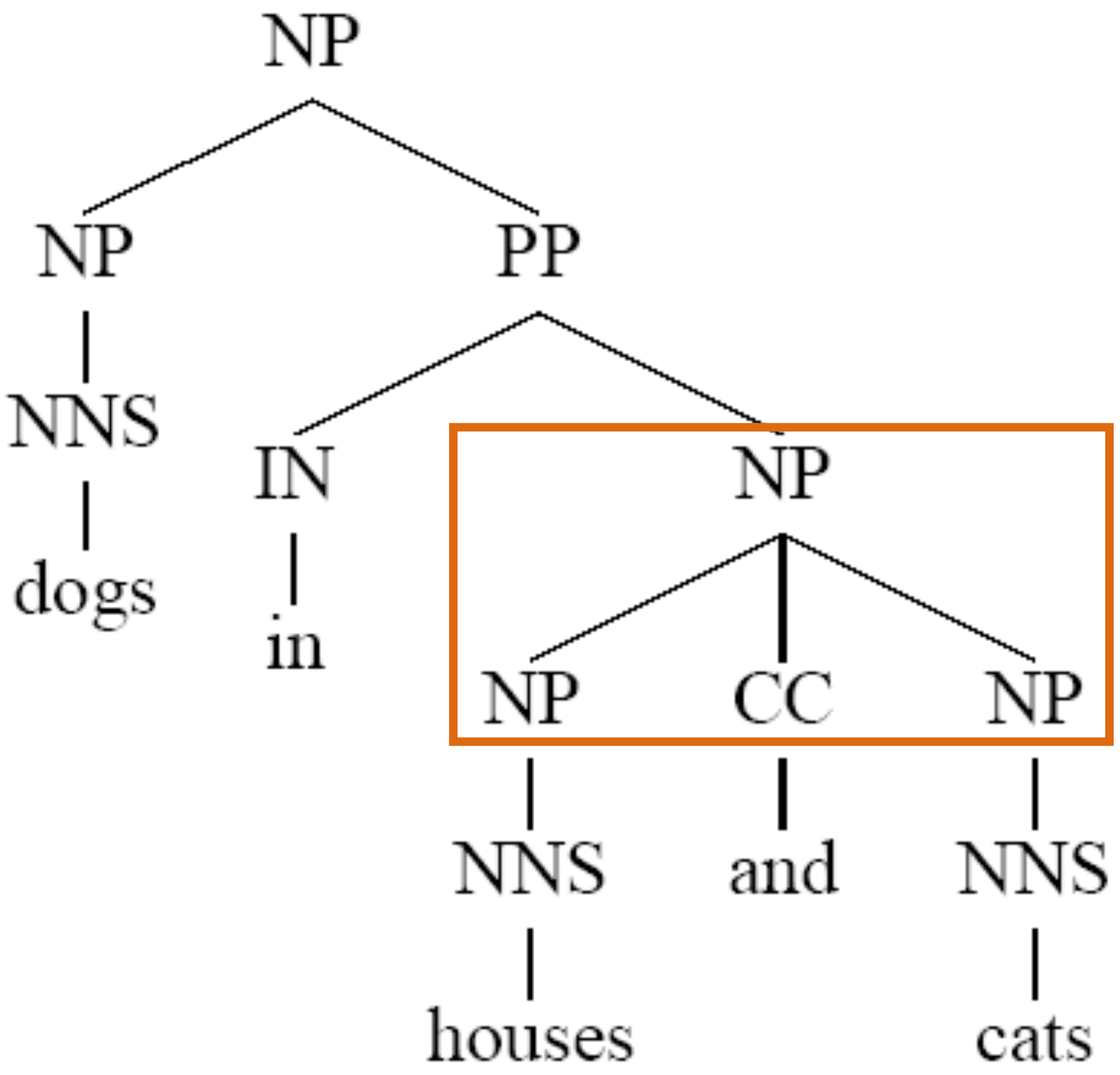
- More predicate-argument focused view of syntax

- “What’s the main verb of the sentence? What is its subject and object?” — easier to answer under dependency parsing

Dependency vs. Constituency: Coordination

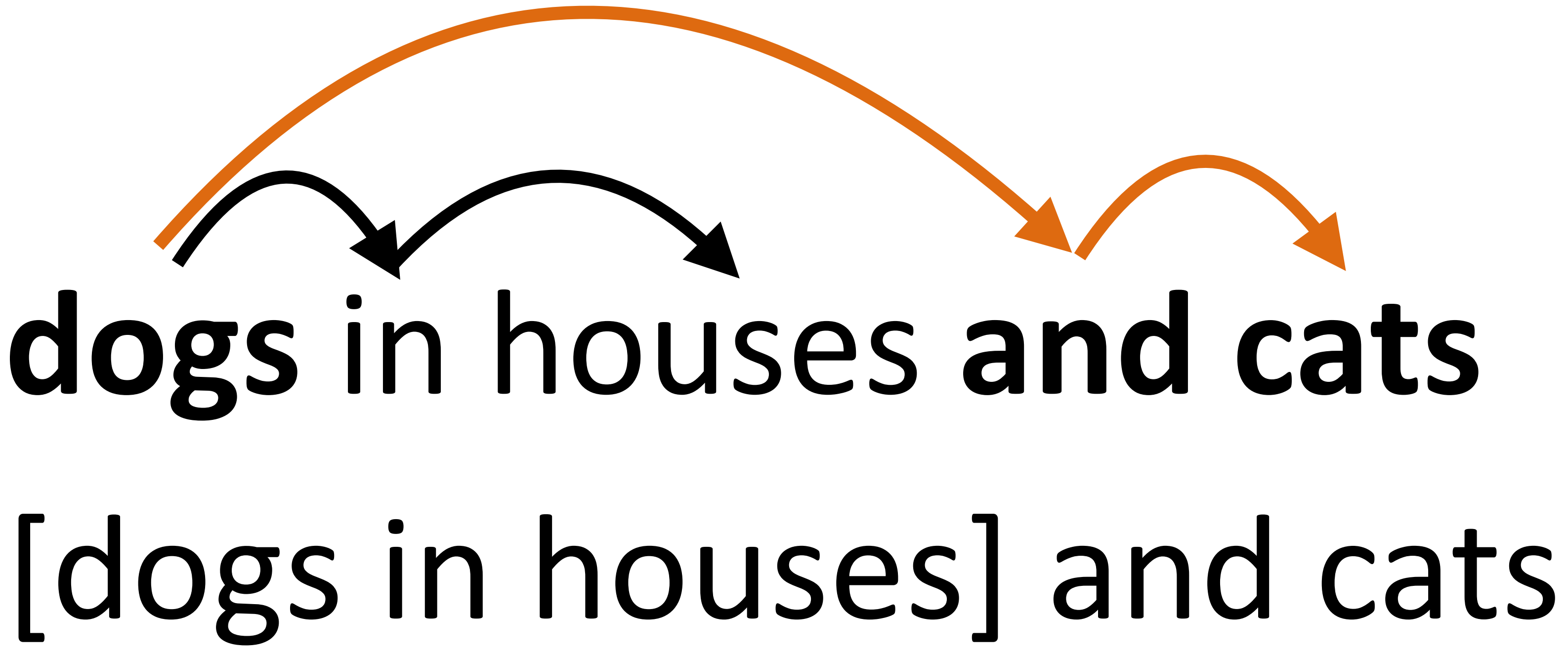
- Constituency: ternary rule NP -> NP CC NP





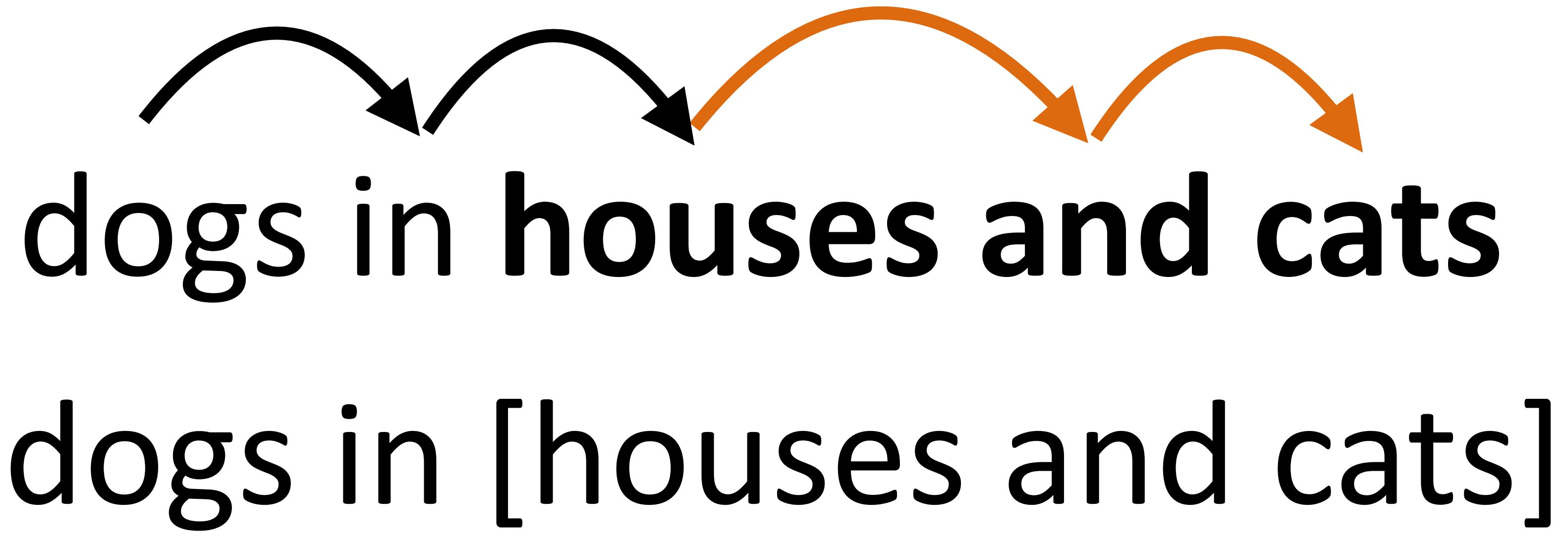
Dependency vs. Constituency: Coordination

- Dependency: first item is the head



dogs in houses **and cats**

[dogs in houses] and cats

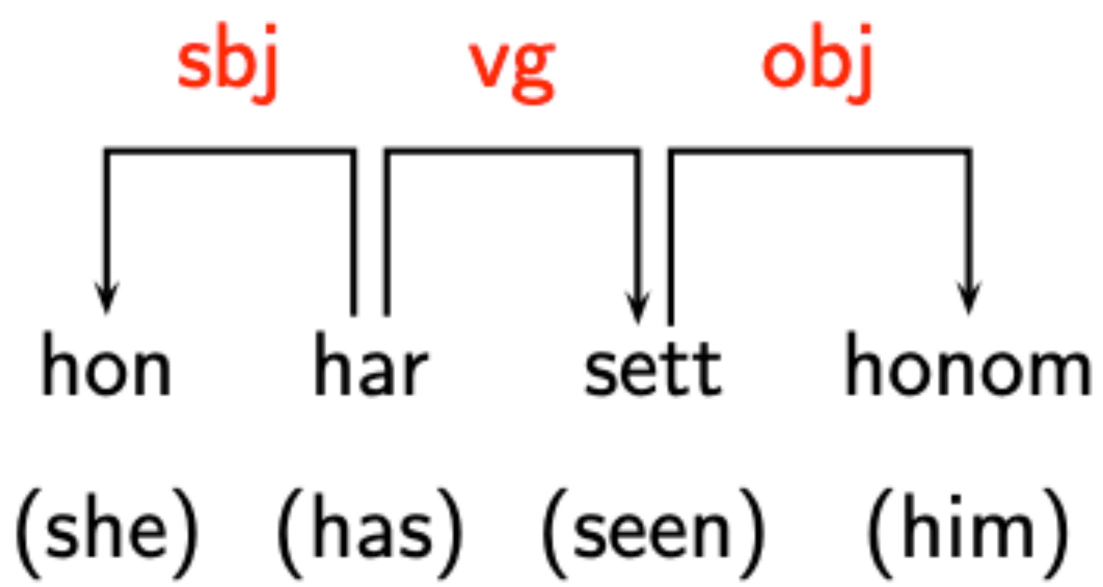


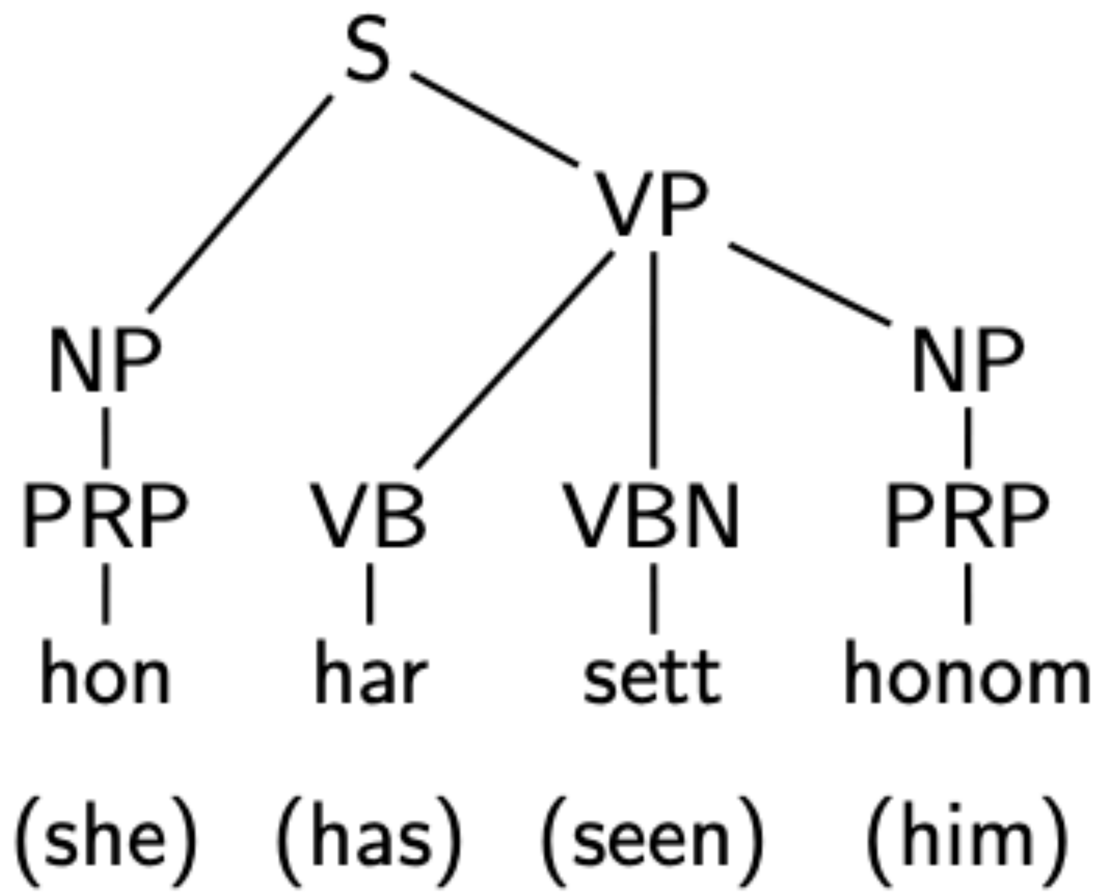
dogs in **houses and cats**

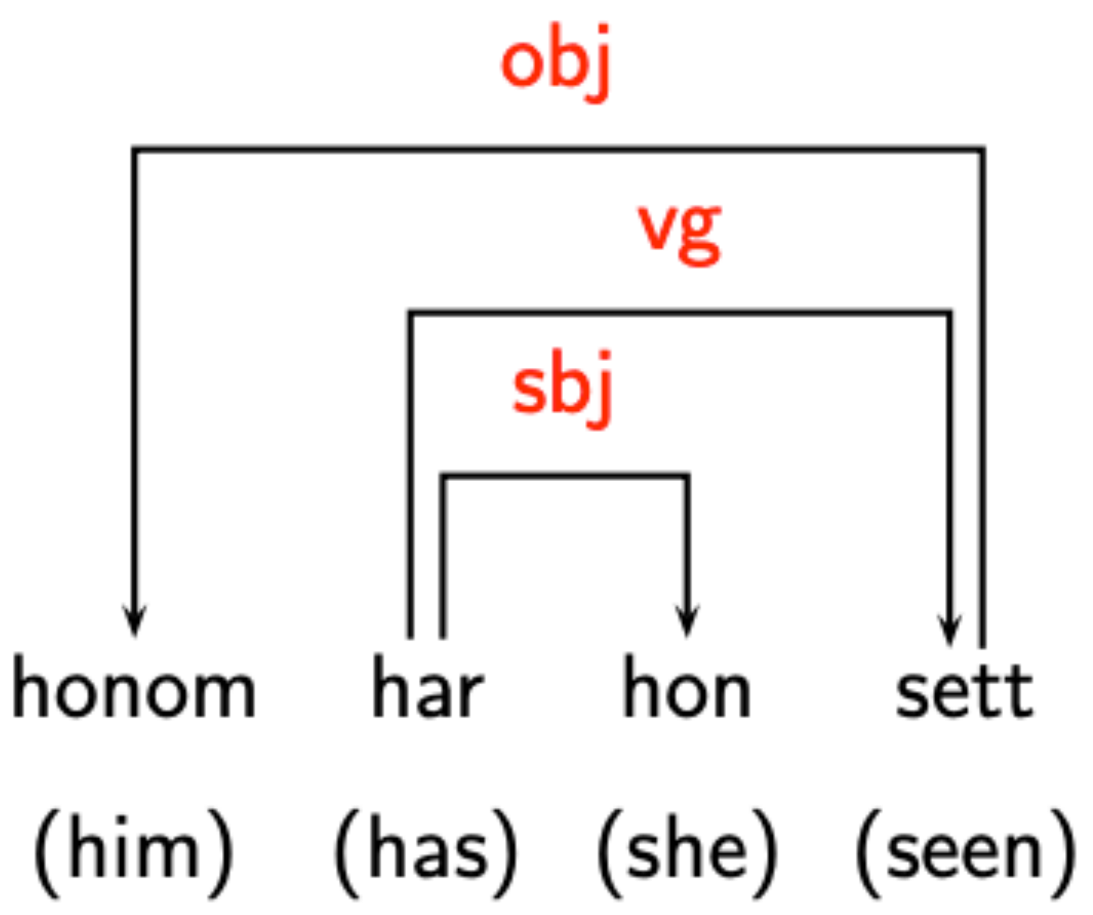
dogs in [houses and cats]

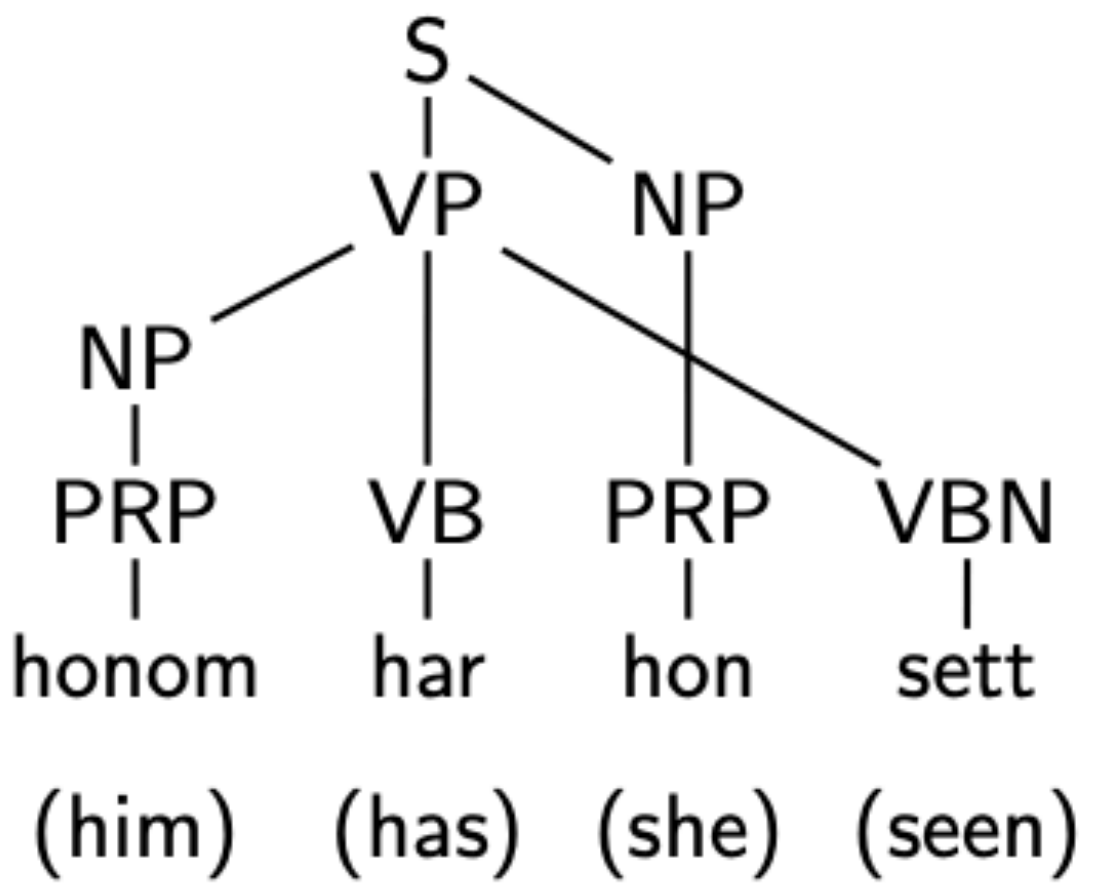
- Coordination is decomposed across a few arcs as opposed to being a single rule production as in constituency
- Can also choose *and* to be the head
- In both cases, headword doesn’t really represent the phrase — constituency representation makes more sense

Advantages of dependency tree









Dependency vs. Constituency

- Dependency is often more useful in practice (models predicate argument structure)

- Dependency parsers are easier to build: no "grammar engineering", no unaries, easier to get structured discriminative models working well

- Dependency parsers are usually faster

- Dependencies are more universal cross-lingually

Dependency vs. Constituency

- Constituency includes non-terminals, and their edges are not typed.

- Dependency types encode “grammatical roles”.

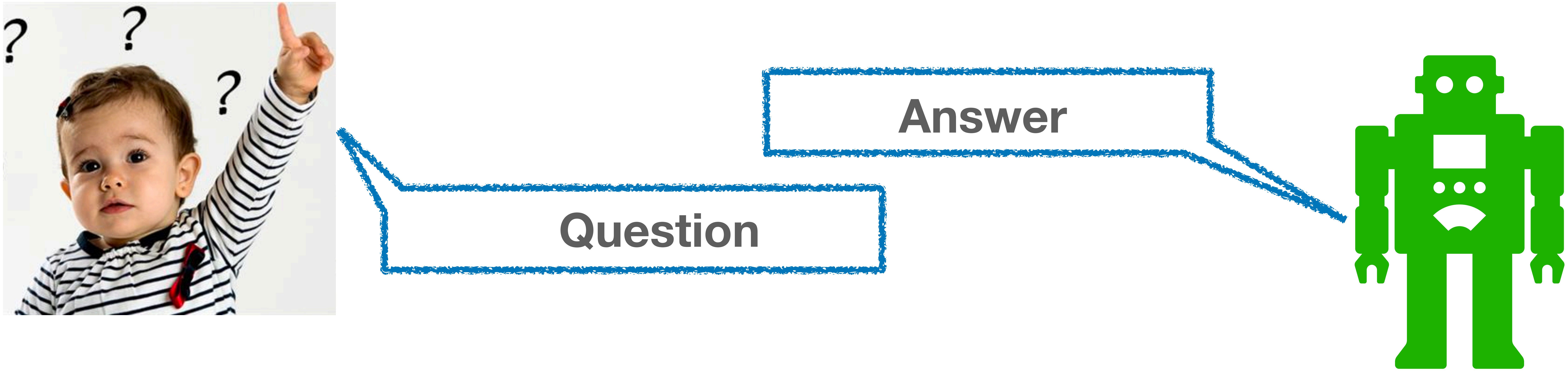
CS 378: Natural Language Processing
Lecture 22: Question Answering

TEXAS
The University of Texas at Austin

Eunsol Choi

This Lecture

- Introduction to question answering task in NLP
- More on how the task is framed and formalized, less on modeling and learning

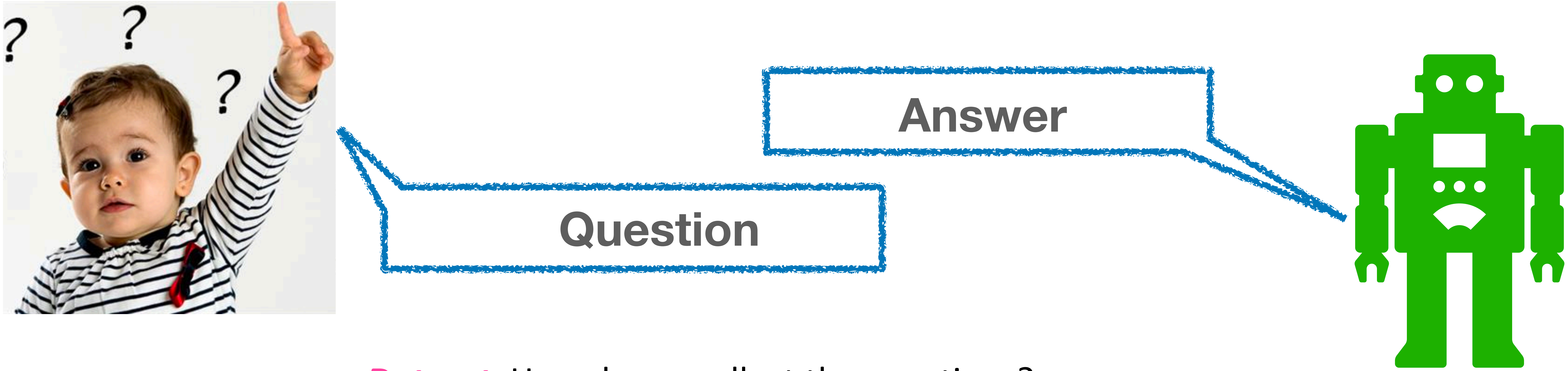


Question Answering

Dataset: How do we collect the answers?

Model: How should we find the answers?

Presentation: How should we present the answers?



Dataset: How do we collect the questions?

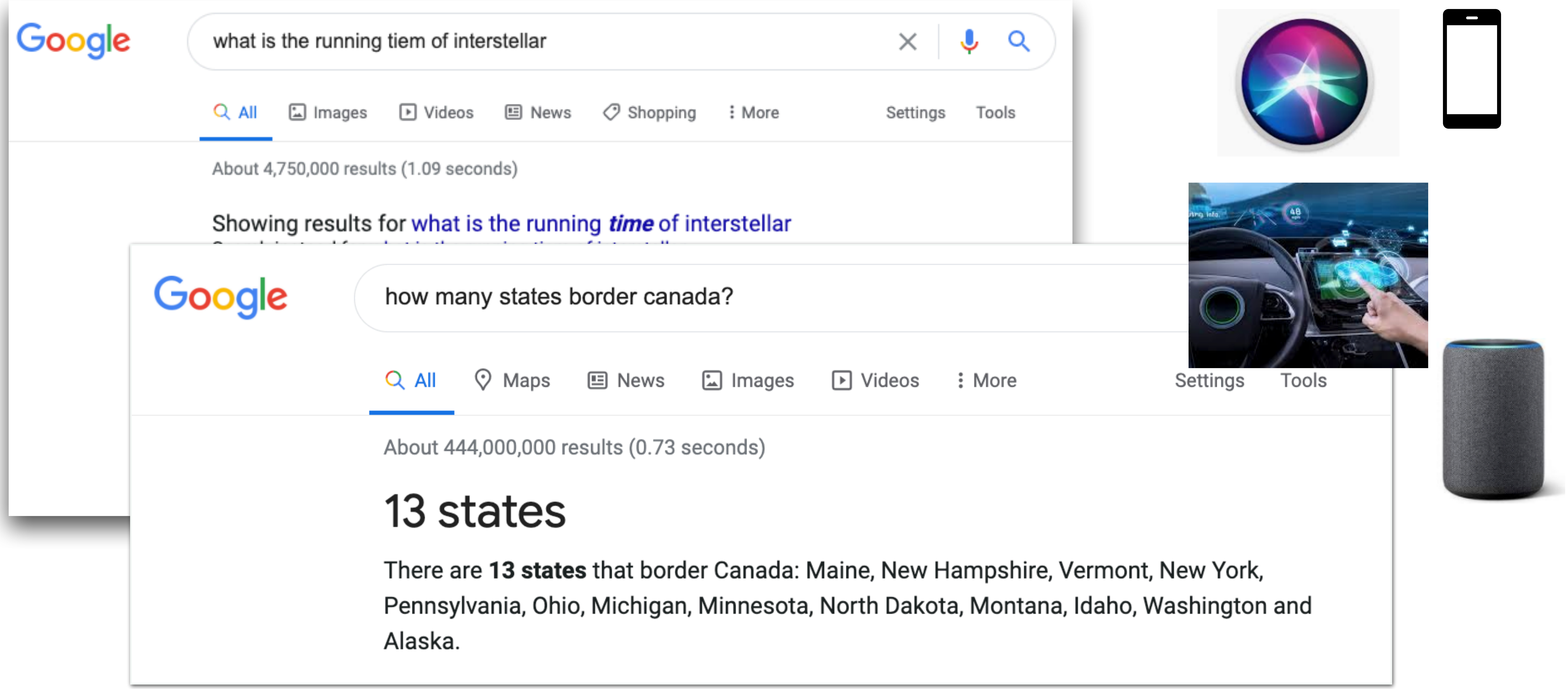
Overview

- Why QA?
- Properties of QA datasets / formalism
  - Output (answer) space
  - Input (evidence) space
  - Required reasoning
  - Further variations
- Presentation of answers
- Remaining challenges

Question Answering

IBM Watson defeated two of Jeopardy's greatest champions in 2011

Question Answering



what is the running tiem of interstellar

About 4,750,000 results (1.09 seconds)

Showing results for what is the running *time* of interstellar

how many states border canada?

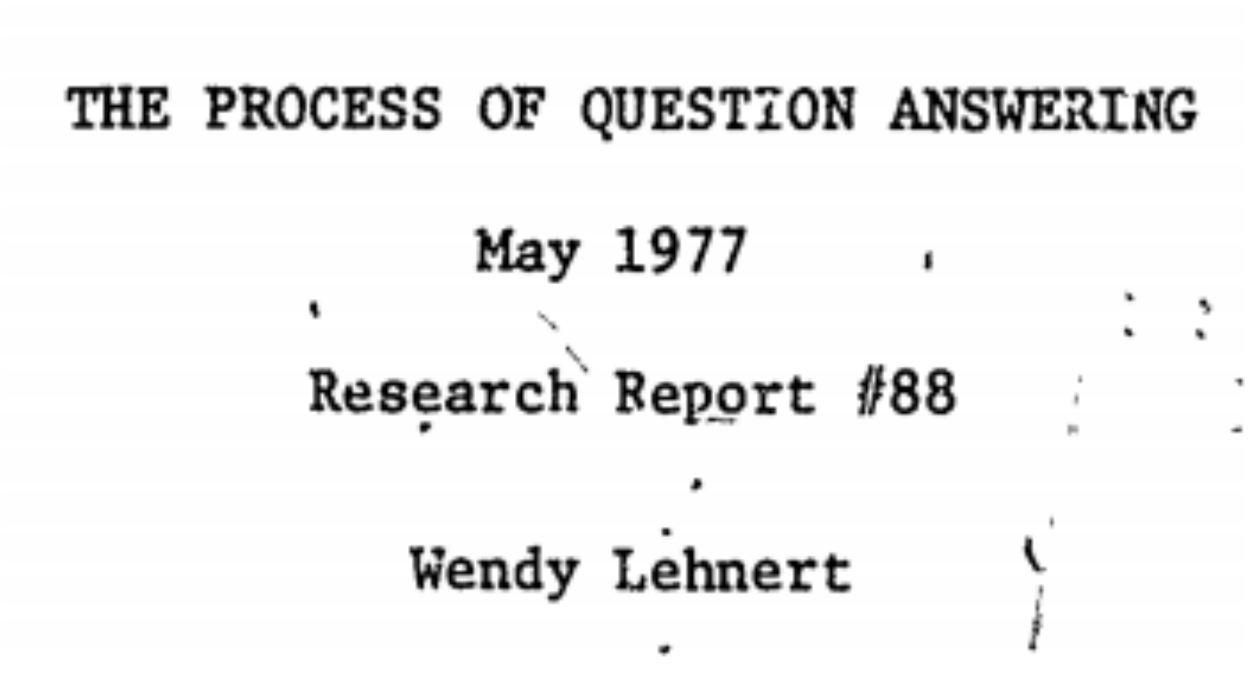
About 444,000,000 results (0.73 seconds)

13 states

There are **13 states** that border Canada: Maine, New Hampshire, Vermont, New York, Pennsylvania, Ohio, Michigan, Minnesota, North Dakota, Montana, Idaho, Washington and Alaska.

Why QA?

- As a testbed to evaluate how machines understand text

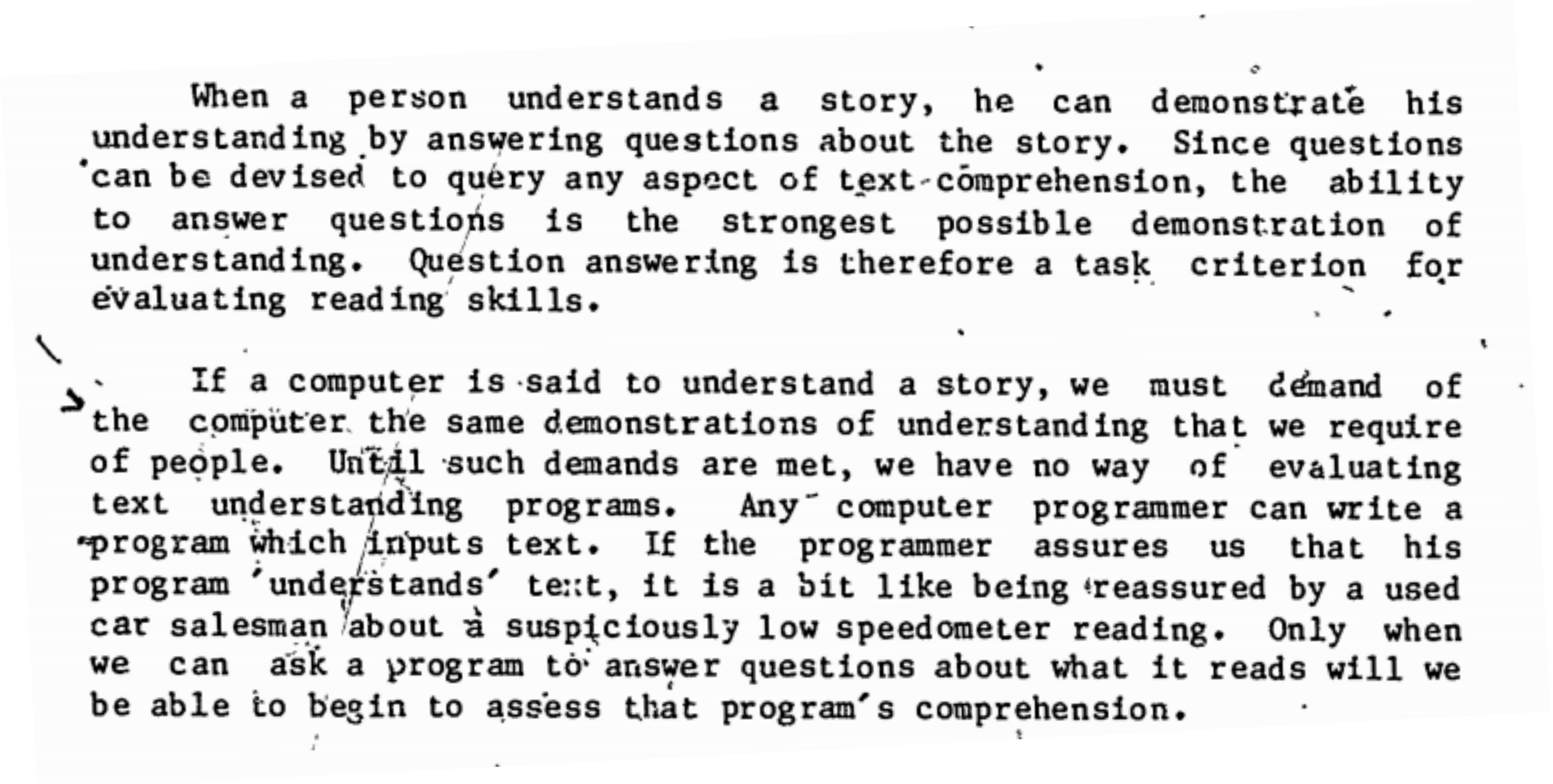


THE PROCESS OF QUESTION ANSWERING

May 1977

Research Report #88

Wendy Lehnert



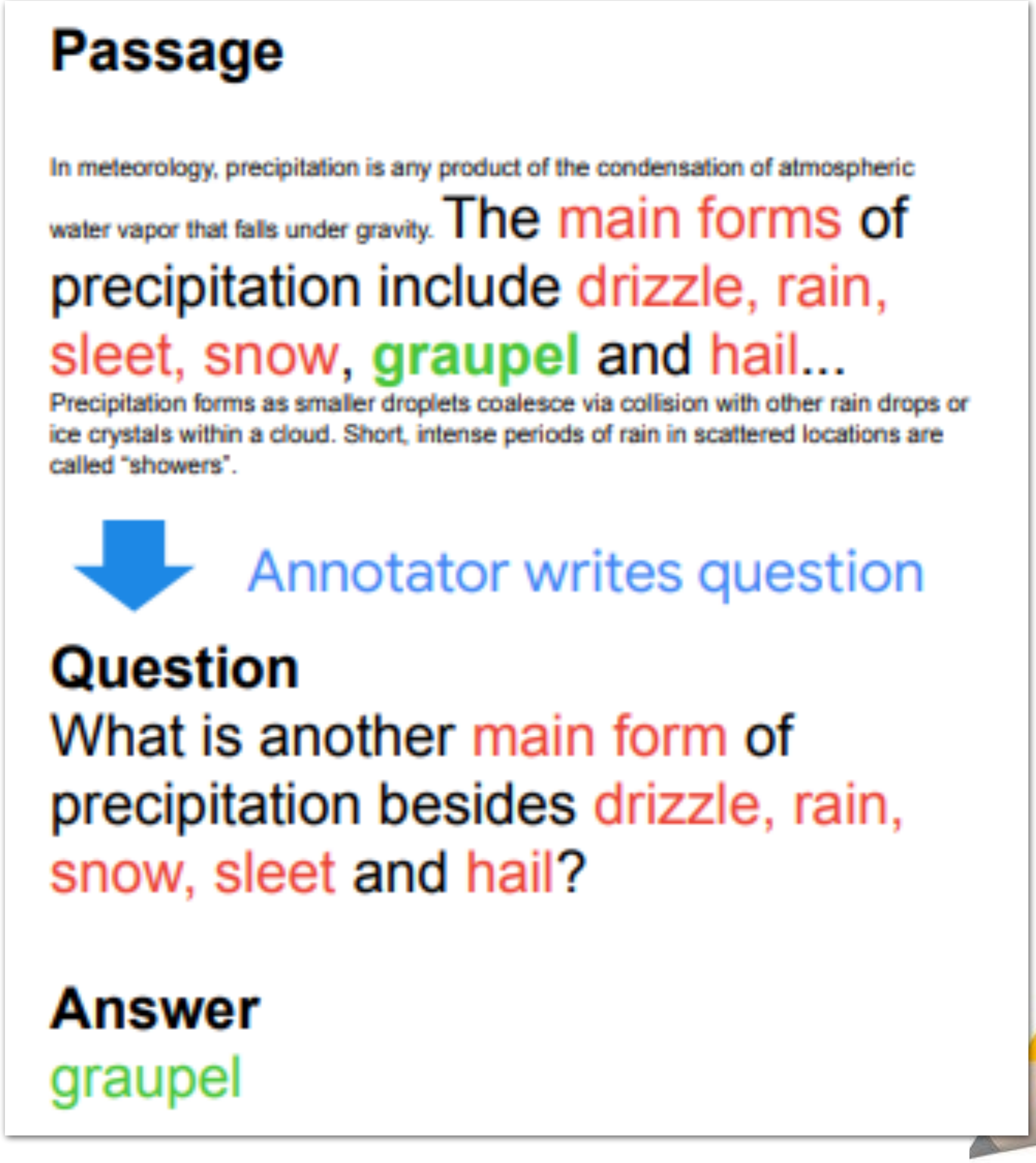
When a person understands a story, he can demonstrate his understanding by answering questions about the story. Since questions can be devised to query any aspect of text-comprehension, the ability to answer questions is the strongest possible demonstration of understanding. Question answering is therefore a task criterion for evaluating reading skills.

If a computer is said to understand a story, we must demand of the computer the same demonstrations of understanding that we require of people. Until such demands are met, we have no way of evaluating text understanding programs. Any computer programmer can write a program which inputs text. If the programmer assures us that his program 'understands' text, it is a bit like being reassured by a used car salesman about a suspiciously low speedometer reading. Only when we can ask a program to answer questions about what it reads will we be able to begin to assess that program's comprehension.

"Since questions can be devised to query **any aspect** of text comprehension, the ability to answer questions is the **strongest possible demonstration of understanding**."

Model-testing Queries

Questioner already knows the answer, aiming to test model's understanding



Passage

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called "showers".

Annotator writes question

Question

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?

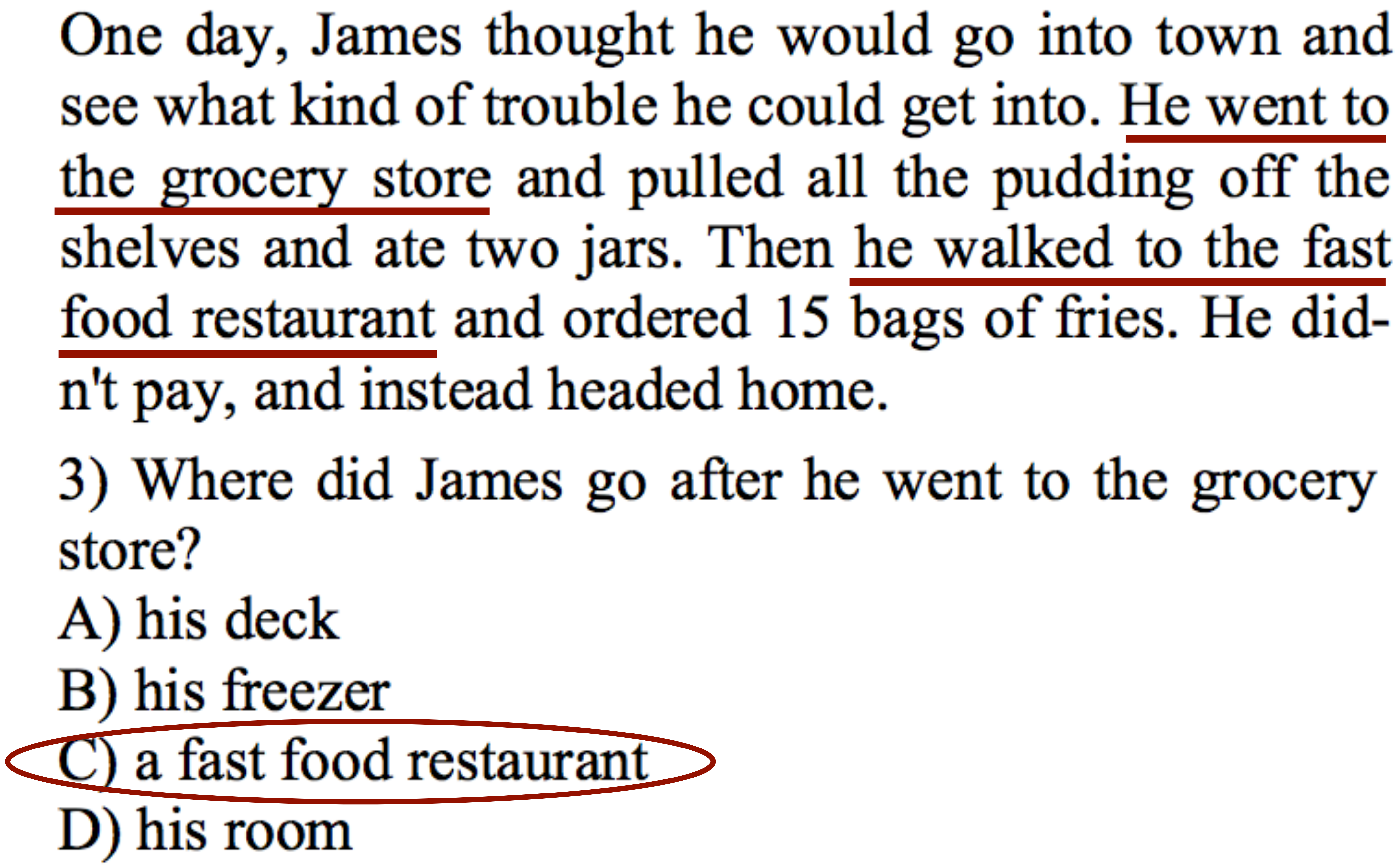
Answer

graupel

[SQuAD, MCTest, RACE, …]

Reading Comprehension

- MCTest (2013): 500 passages, 4 questions per passage
- Two questions per passage explicitly require cross-sentence reasoning



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.

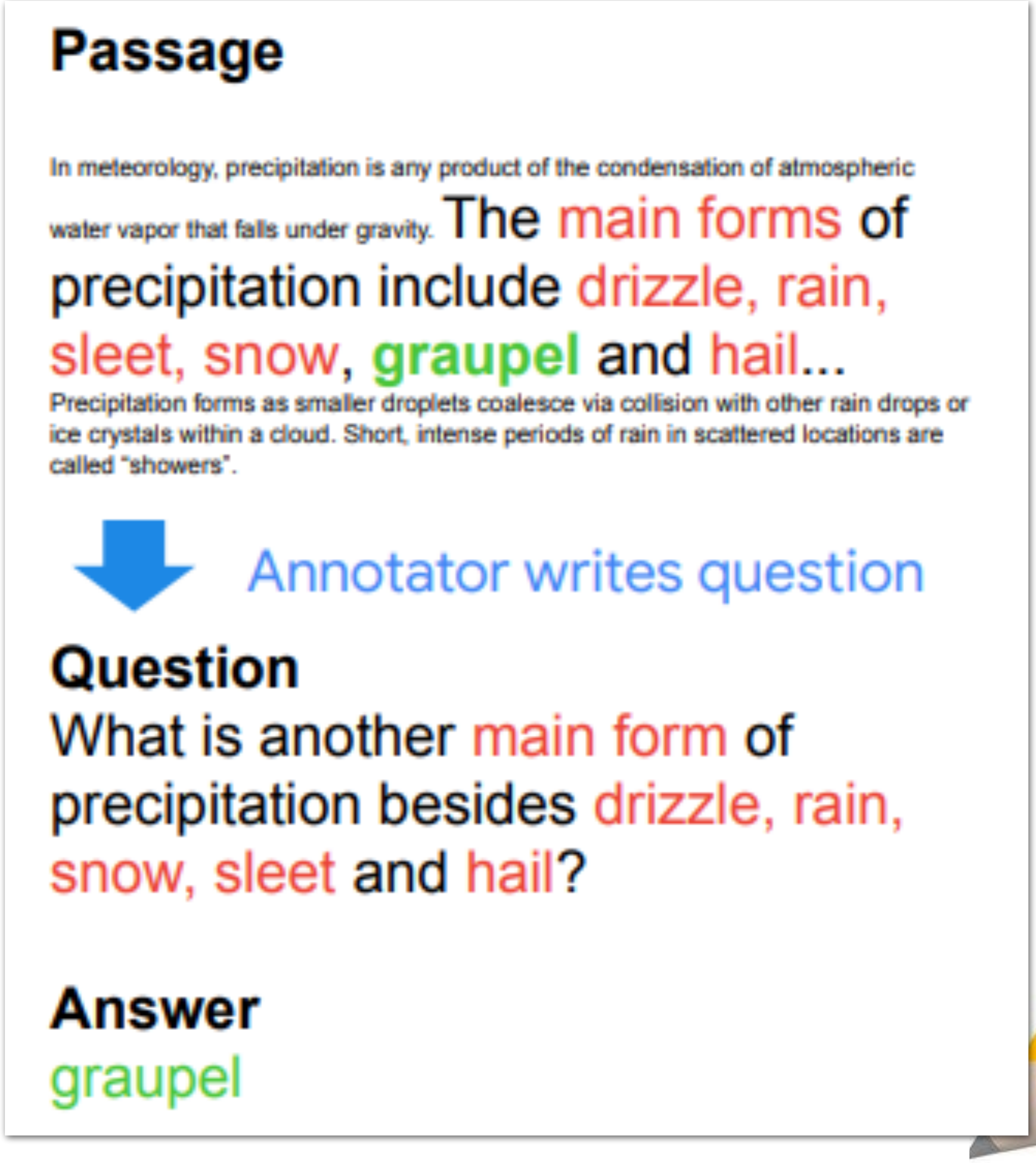
3) Where did James go after he went to the grocery store?

A) his deck
B) his freezer
[✓ C) a fast food restaurant]
D) his room

Richardson (2013)

Model-testing Queries

Questioner already knows the answer, aiming to test model's understanding



Passage

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called "showers".

Annotator writes question

Question

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?

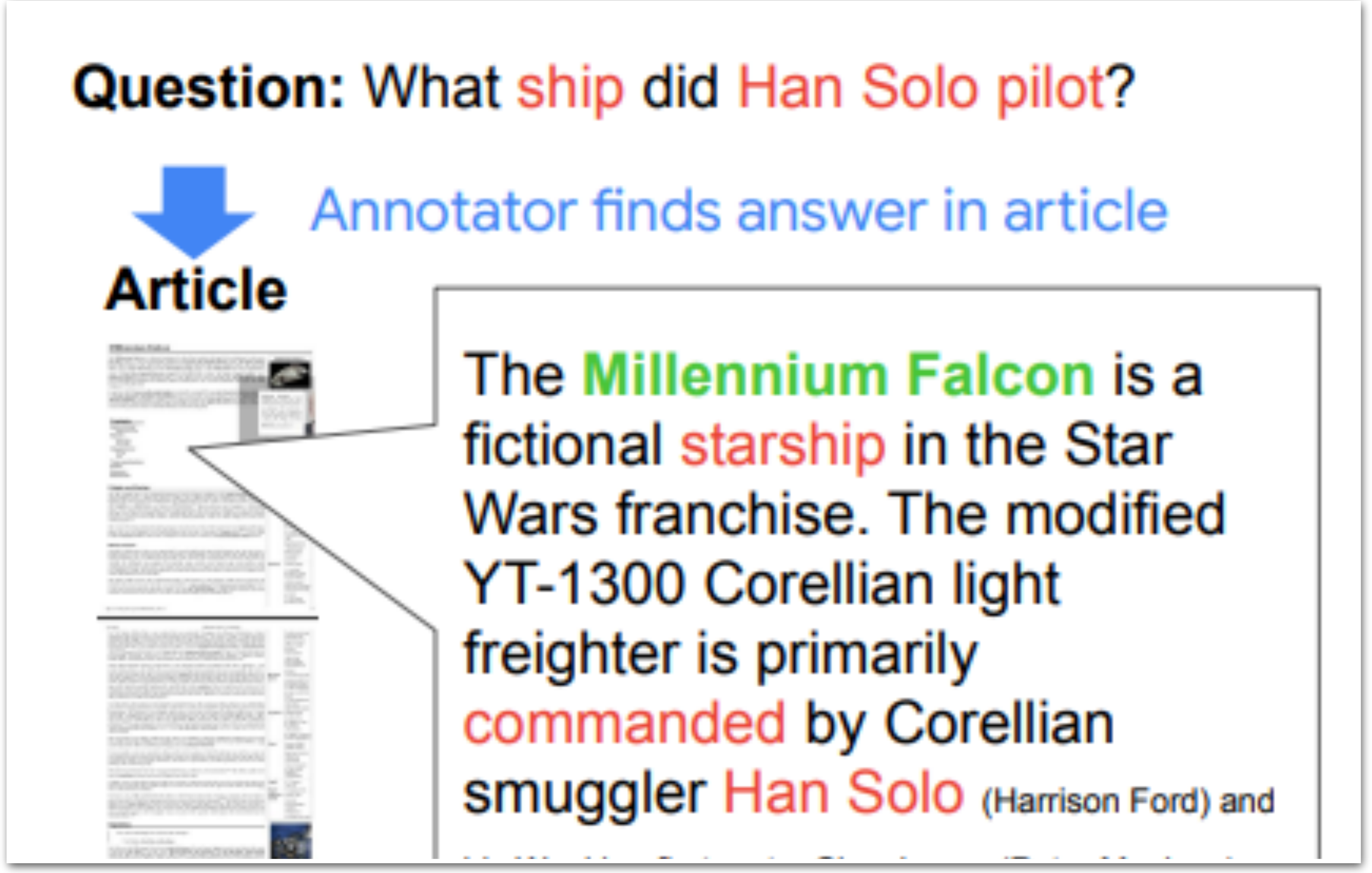
Answer

graupel

[SQuAD, MCTest, RACE, …]

Information Seeking Queries

Questioner does not know the answer

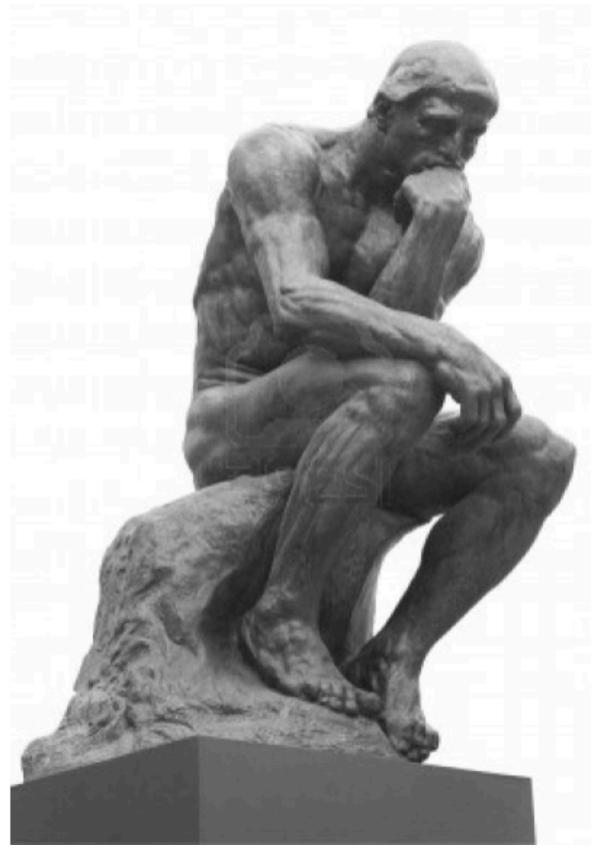


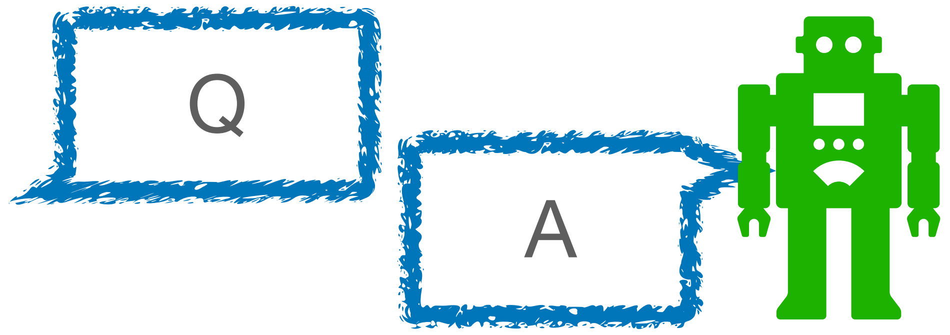
Question: What ship did Han Solo pilot?

Annotator finds answer in article

Article

The Millennium Falcon is a fictional starship in the Star Wars franchise. The modified YT-1300 Corellian light freighter is primarily commanded by Corellian smuggler Han Solo (Harrison Ford) and



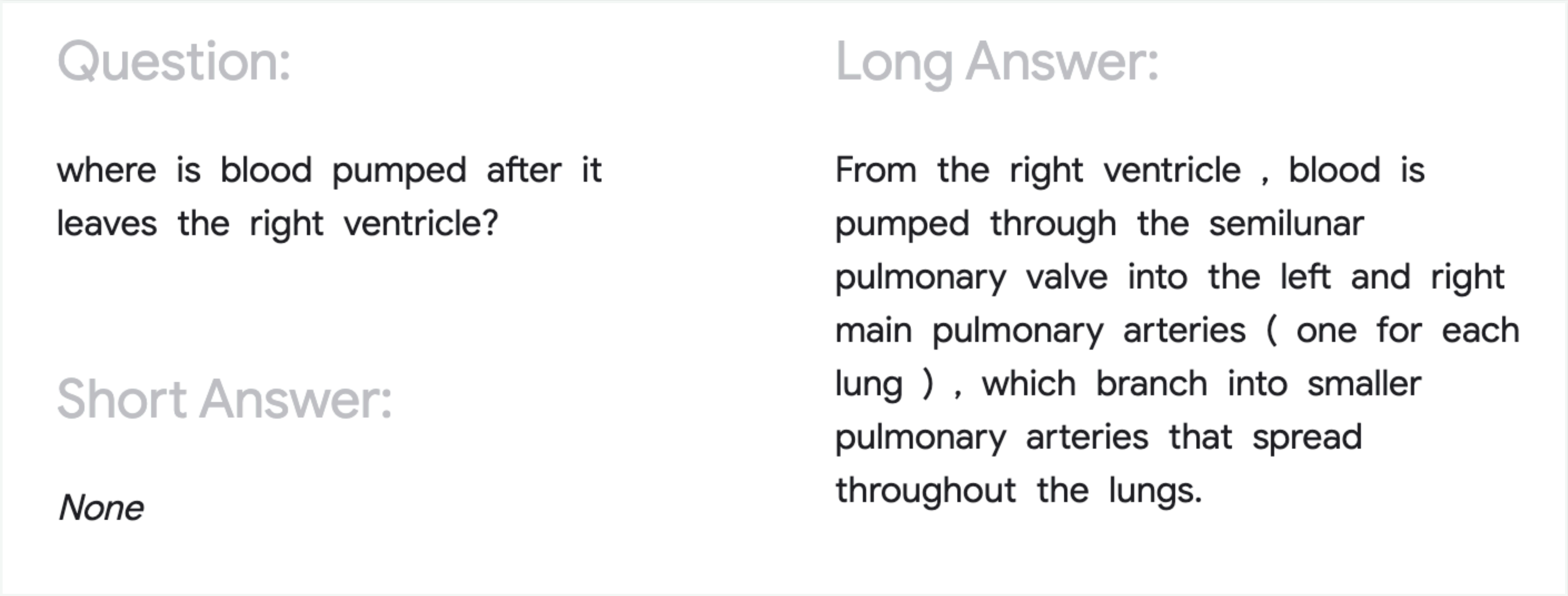


Information Seeking Queries

Trying to gain information

NaturalQuestions

- Real questions from Google, answerable with Wikipedia
- Short answers and long answers (snippets)
- Questions arose naturally



Question:

where is blood pumped after it leaves the right ventricle?

Short Answer:

None

Long Answer:

From the right ventricle , blood is pumped through the semilunar pulmonary valve into the left and right main pulmonary arteries (one for each lung) , which branch into smaller pulmonary arteries that spread throughout the lungs.

Kwiatkowski et al. (2019)

Where to get questions?

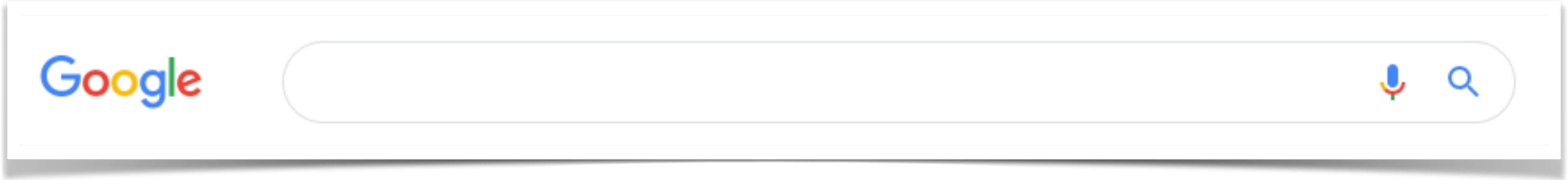
Crowdsourcing

User Queries

Given:
entity name and the first paragraph of Wikipedia page

Do:
Ask questions to learn as much as possible about this entity!

[Choi et al EMNLP 2018, Clark et al TACL 2020, Ferguson et al, EMNLP 2020]



Natural Questions [Kwiatkowski et al, TACL 2019]

[Berant et al, 2013, Yang et al, EMNLP 15,
Bajaj et al NeurIPS workshop 2018]

QA can be very broad

- Factoid QA: *what states border Mississippi?, when was Barack Obama born?*
    - Lots of this could be handled by QA from a knowledge base, if we had a big enough knowledge base
- "Question answering" as a term is so broad as to be meaningless
    - *Is P=NP?*
    - *What is 4+5?*
    - *What is the translation of [sentence] into French?* [McCann et al., 2018]

Overview

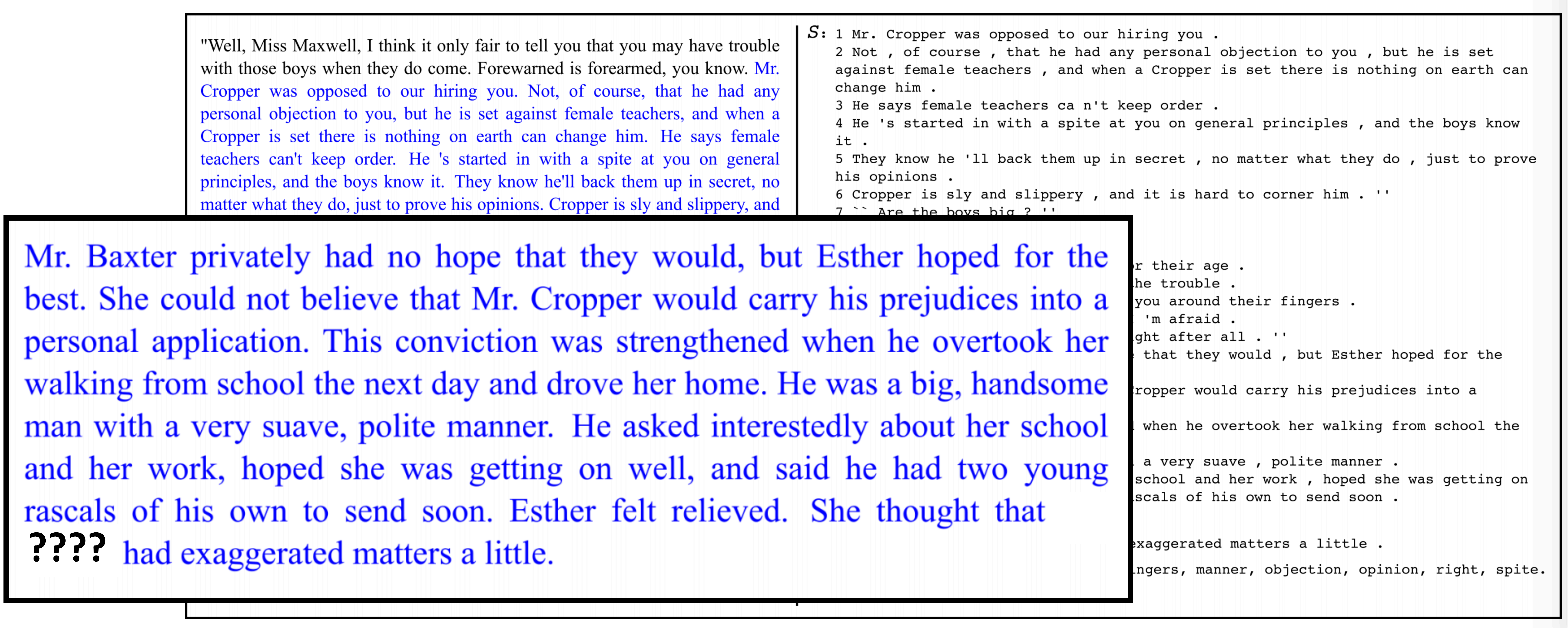
- Why QA?
- **Properties of QA datasets / formalism**
    - Output (answer) space
    - Input (evidence) space
    - Required reasoning
    - Further variations
- Presentation of answers
- Remaining challenges

Simulating QA from raw text

- Typically, question answering dataset requires human annotation
- "Cloze" task: word (often an entity) is removed from a sentence
  - Answers: multiple choice, pick from passage, or pick from vocabulary
  - Can be created automatically from things that aren't questions

[QA Dataset Explosion, Rogers et al]

Children’s Book Test



"Well, Miss Maxwell, I think it only fair to tell you that you may have trouble with those boys when they do come. Forewarned is forearmed, you know. Mr. Cropper was opposed to our hiring you. Not, of course, that he had any personal objection to you, but he is set against female teachers, and when a Cropper is set there is nothing on earth can change him. He says female teachers can't keep order. He 's started in with a spite at you on general principles, and the boys know it. They know he'll back them up in secret, no matter what they do, just to prove his opinions. Cropper is sly and slippery, and

S: 1 Mr. Cropper was opposed to our hiring you .
2 Not , of course , that he had any personal objection to you , but he is set against female teachers , and when a Cropper is set there is nothing on earth can change him .
3 He says female teachers ca n't keep order .
4 He 's started in with a spite at you on general principles , and the boys know it .
5 They know he 'll back them up in secret , no matter what they do , just to prove his opinions .
6 Cropper is sly and slippery , and it is hard to corner him . ''
7 `` Are the boys big ? ''

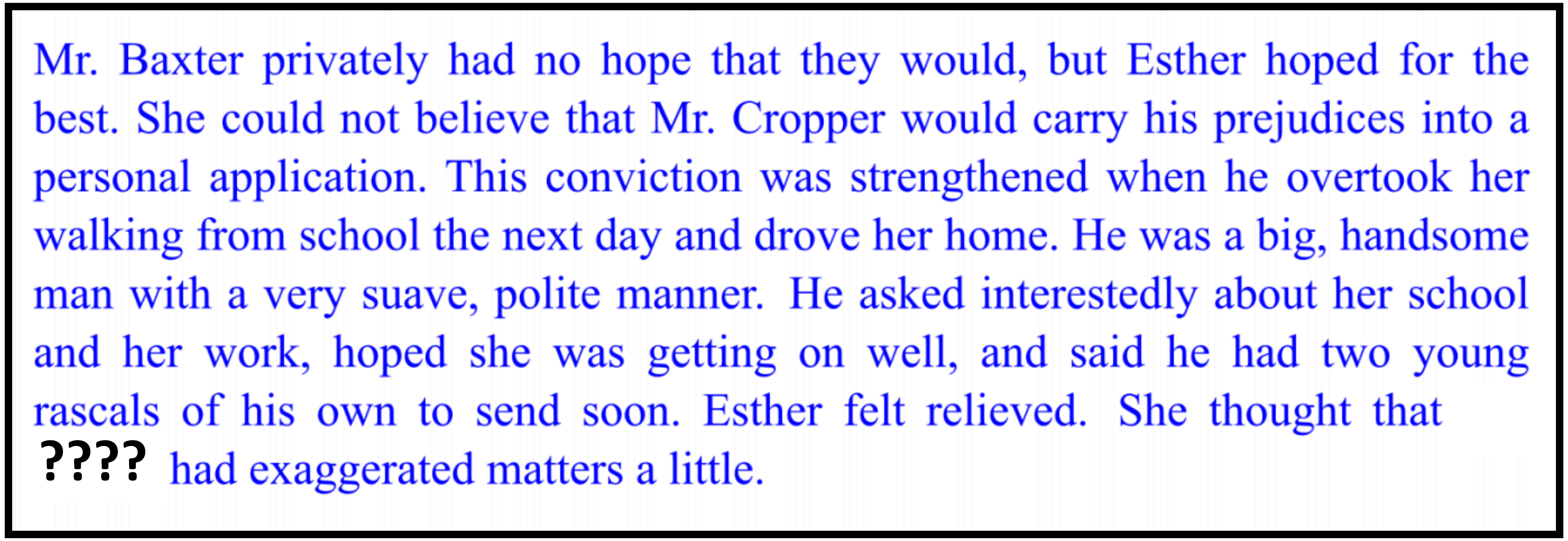
r their age .
he trouble .
you around their fingers .
'm afraid .
ght after all . ''
that they would , but Esther hoped for the
ropper would carry his prejudices into a
when he overtook her walking from school the
a very suave , polite manner .
school and her work , hoped she was getting on
scals of his own to send soon .
xaggerated matters a little .
ngers, manner, objection, opinion, right, spite.

Mr. Baxter privately had no hope that they would, but Esther hoped for the best. She could not believe that Mr. Cropper would carry his prejudices into a personal application. This conviction was strengthened when he overtook her walking from school the next day and drove her home. He was a big, handsome man with a very suave, polite manner. He asked interestedly about her school and her work, hoped she was getting on well, and said he had two young rascals of his own to send soon. Esther felt relieved. She thought that **????** had exaggerated matters a little.

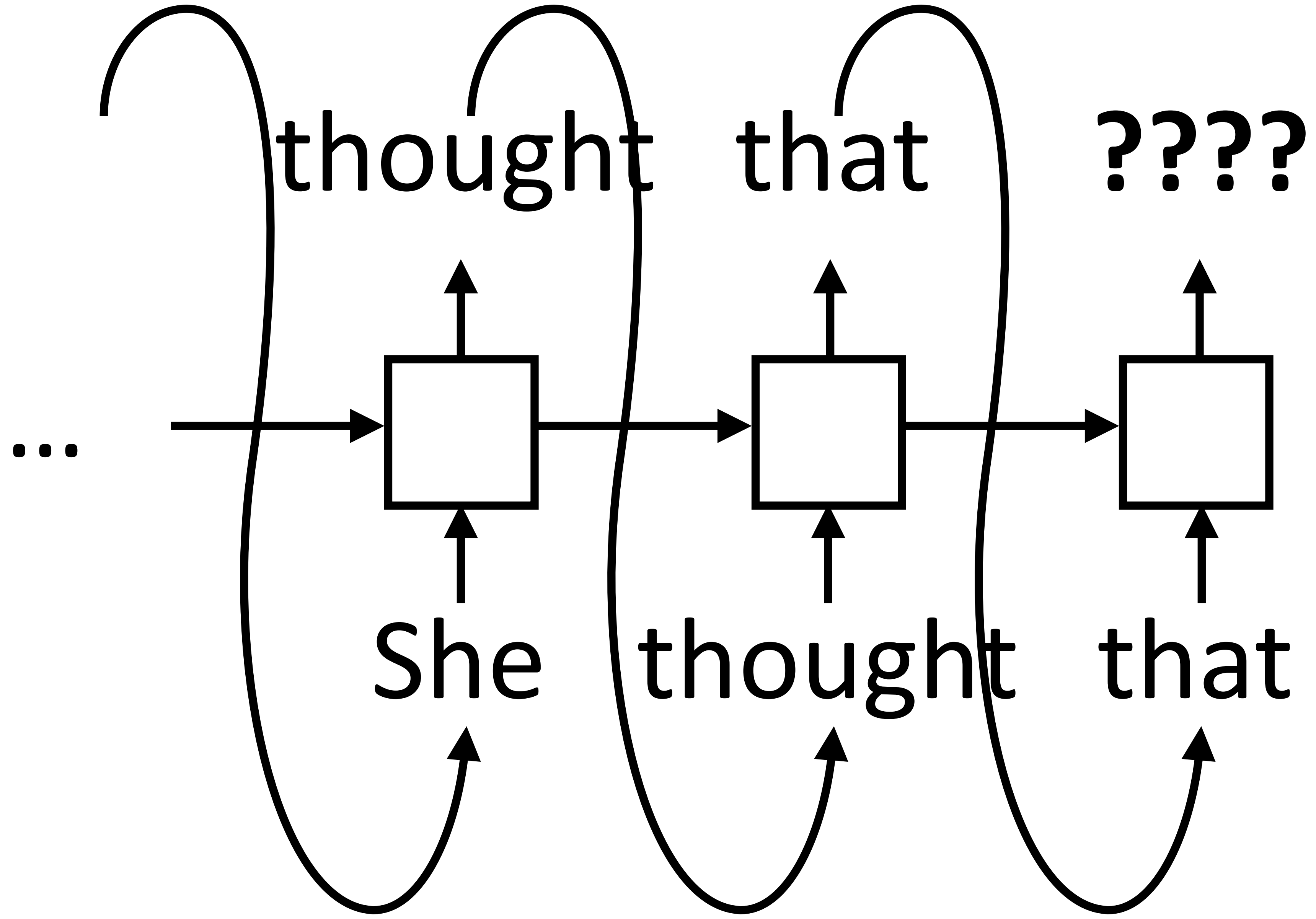
- Children’s Book Test: take a section of a children’s story, block out an entity and predict it (one-doc multi-sentence cloze task)

Hill et al. (2015)

LSTM Language Models



Mr. Baxter privately had no hope that they would, but Esther hoped for the best. She could not believe that Mr. Cropper would carry his prejudices into a personal application. This conviction was strengthened when he overtook her walking from school the next day and drove her home. He was a big, handsome man with a very suave, polite manner. He asked interestedly about her school and her work, hoped she was getting on well, and said he had two young rascals of his own to send soon. Esther felt relieved. She thought that **????** had exaggerated matters a little.



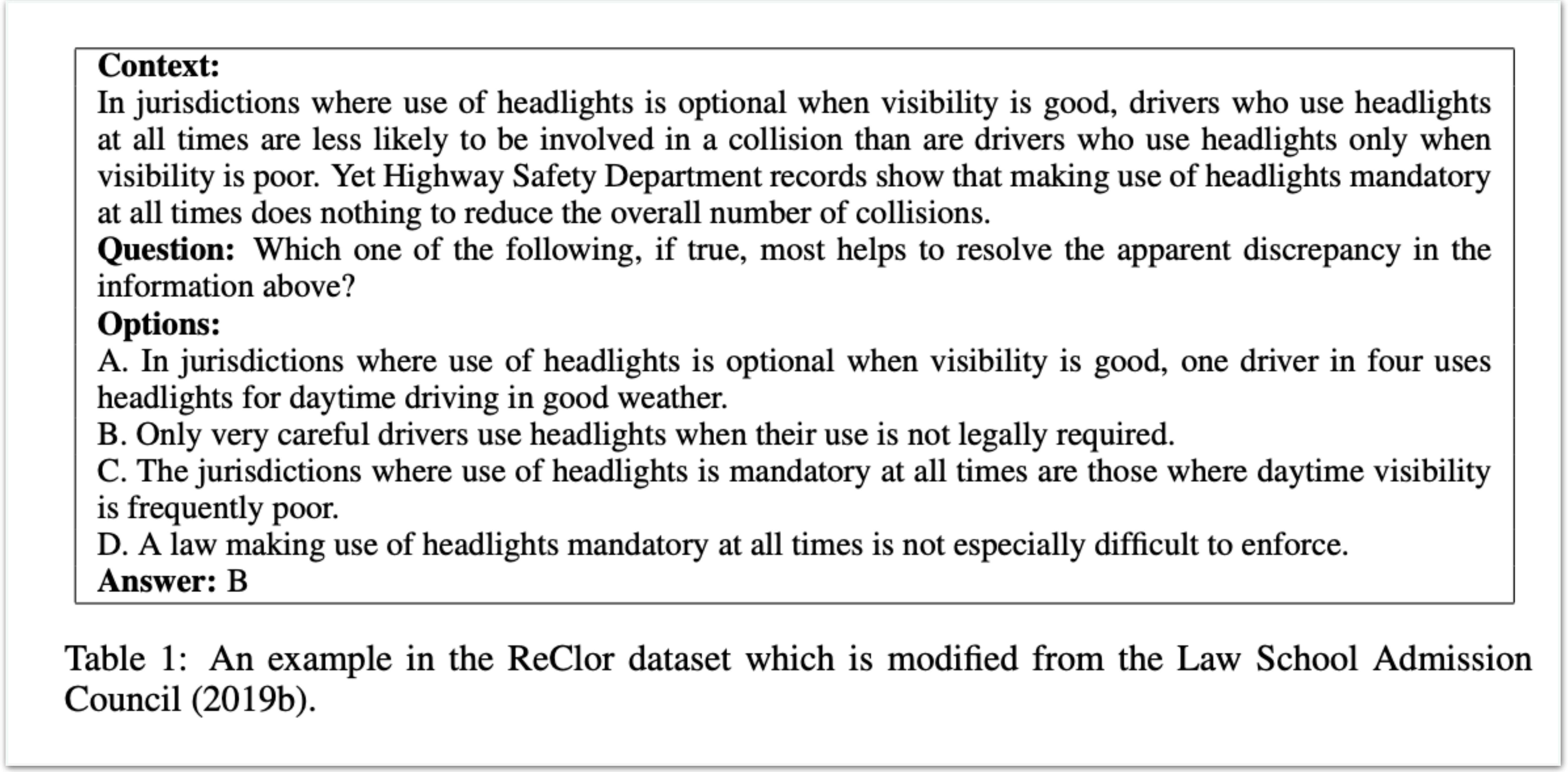
- Predict next word with LSTM LM
- Context: either just the current sentence (query) or the whole document up to this point (query+context)

Hill et al. (2015)

Dataset Properties

- Axis 1: what's the output space?
    - cloze task (fill in blank)

Multiple-Choice datasets



Context:
In jurisdictions where use of headlights is optional when visibility is good, drivers who use headlights at all times are less likely to be involved in a collision than are drivers who use headlights only when visibility is poor. Yet Highway Safety Department records show that making use of headlights mandatory at all times does nothing to reduce the overall number of collisions.
Question: Which one of the following, if true, most helps to resolve the apparent discrepancy in the information above?
Options:
A. In jurisdictions where use of headlights is optional when visibility is good, one driver in four uses headlights for daytime driving in good weather.
B. Only very careful drivers use headlights when their use is not legally required.
C. The jurisdictions where use of headlights is mandatory at all times are those where daytime visibility is frequently poor.
D. A law making use of headlights mandatory at all times is not especially difficult to enforce.
Answer: B

Table 1: An example in the ReClor dataset which is modified from the Law School Admission Council (2019b).

- Can capture complex semantics
- Evaluation is straightforward
- But is it realistic?

ReCLOR dataset (ICLR 2021) https://openreview.net/pdf?id=HJgJtT4tvB

Span-based prediction

Question : What shift happened in animal regulation in 1963 in U.S?

Document Context :

The Lacey Act of 1900 was the first federal law that regulated commercial animal markets. It prohibited interstate commerce of animals killed in violation of state game laws, and covered all wildlife. Whereas the Lacey Act dealt with game animal management and market commerce species, a major shift in focus occurred by 1963 to habitat preservation instead of take regulations. A provision was added by Congress in the Land and Water Conservation Fund Act of…

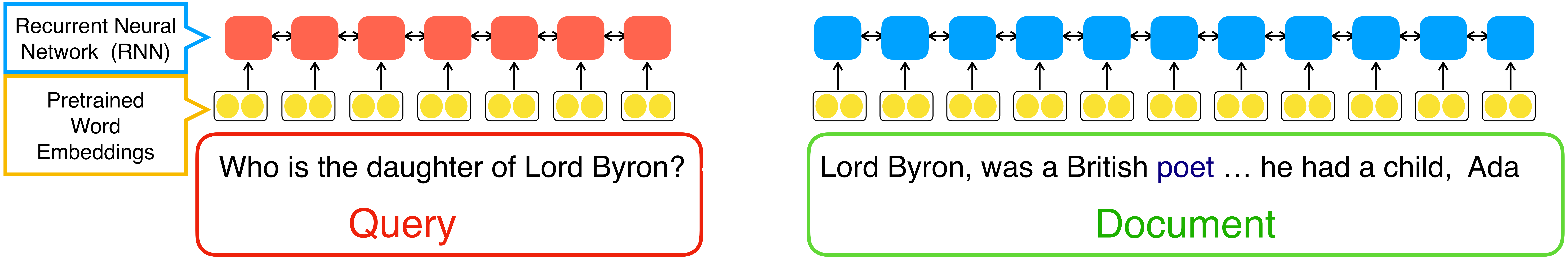
Answer is span in the original document

- Can capture various semantics
- Evaluation is somewhat straightforward
- More realistic than multiple choice

[Rajpurkar et al 2016]

Model: BiDAF (Bi-directional Attention Flow)

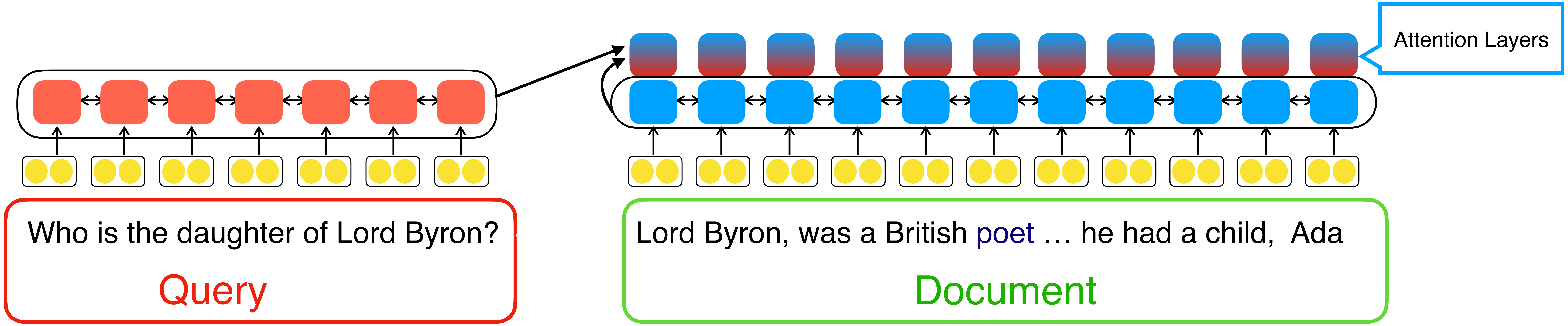
- Encode text and question with recurrent neural network


Recurrent Neural Network (RNN)
Pretrained Word Embeddings
Who is the daughter of Lord Byron?
Query
Lord Byron, was a British poet … he had a child, Ada
Document

[Seo et al, ICLR 17]

Model: BiDAF (Bi-directional Attention Flow)

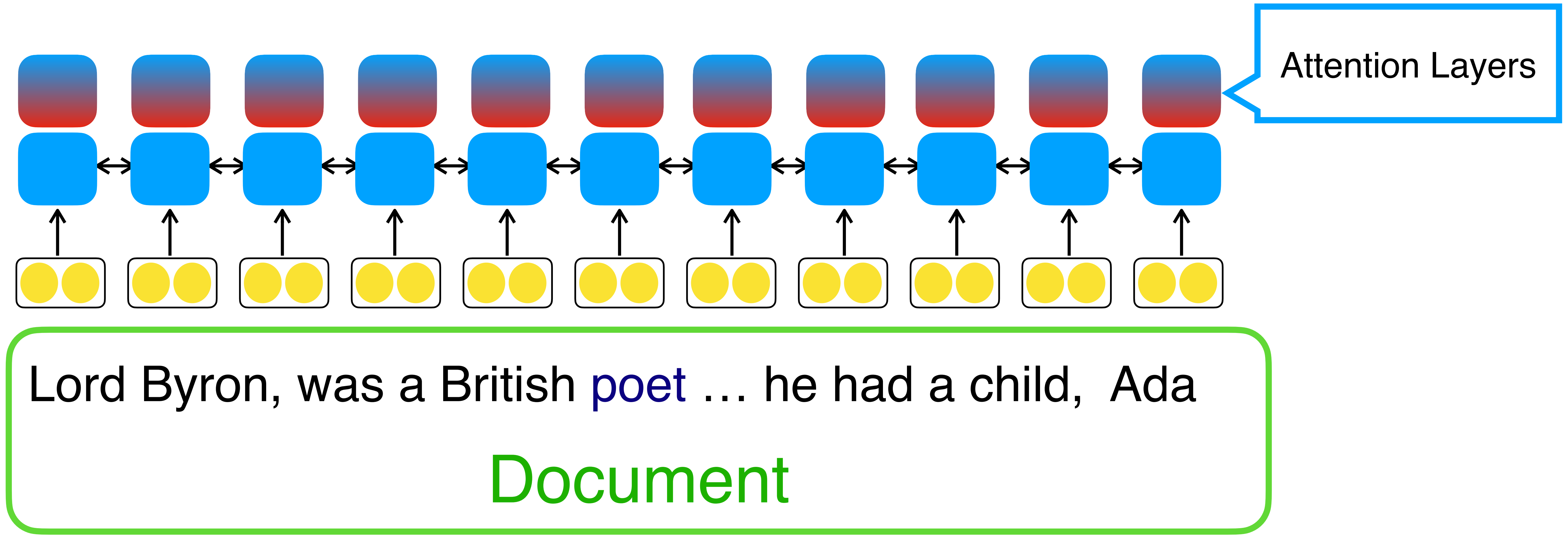
- Encode text and question with recurrent neural network
- Compute inter-sentence alignment with attention


Attention Layers
Who is the daughter of Lord Byron?
Query
Lord Byron, was a British poet … he had a child, Ada
Document

[Seo et al, ICLR 17]

Model: BiDAF (Bi-directional Attention Flow)

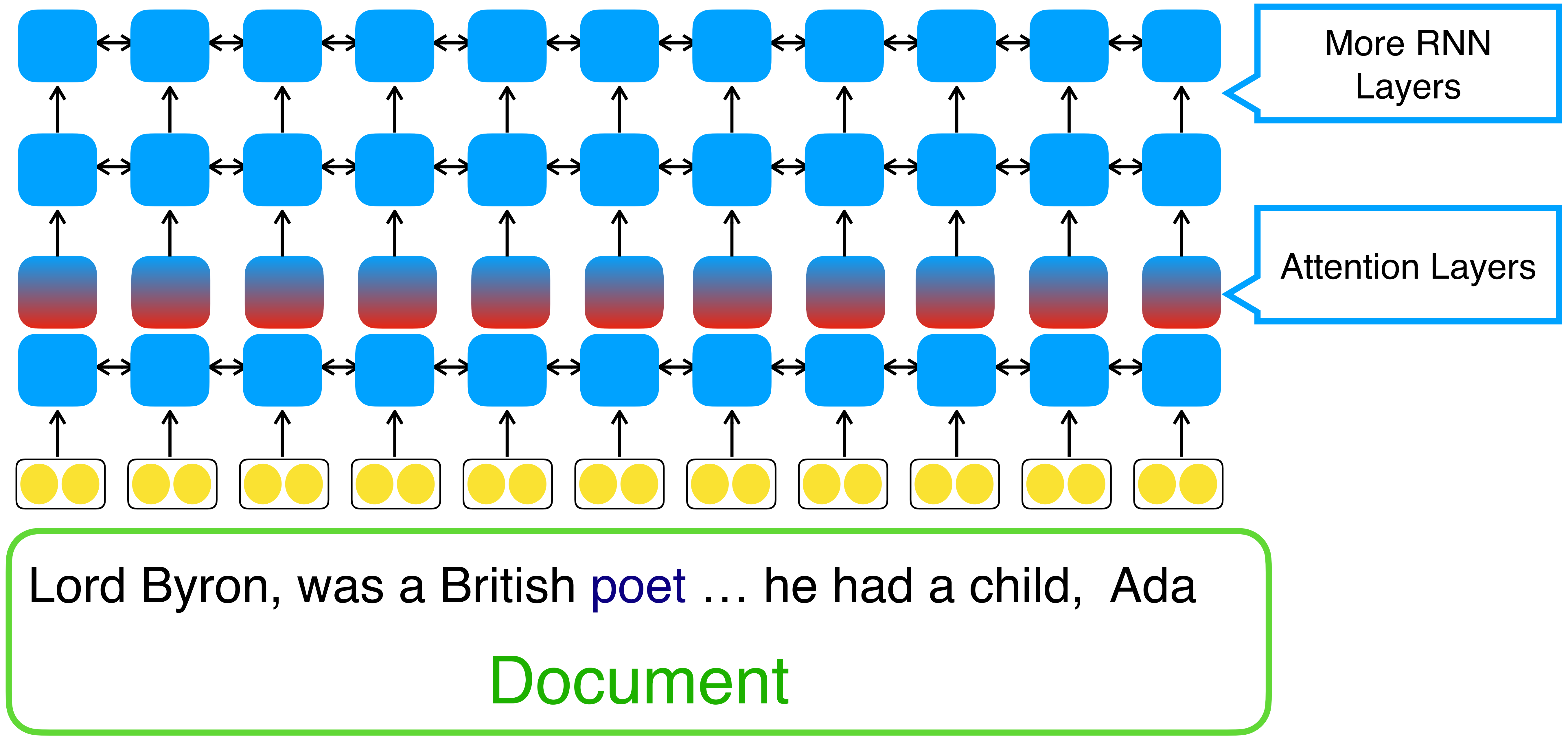
- Encode text and question with recurrent neural network
- Compute inter-sentence alignment with attention



[Seo et al, ICLR 17]

Model: BiDAF (Bi-directional Attention Flow)

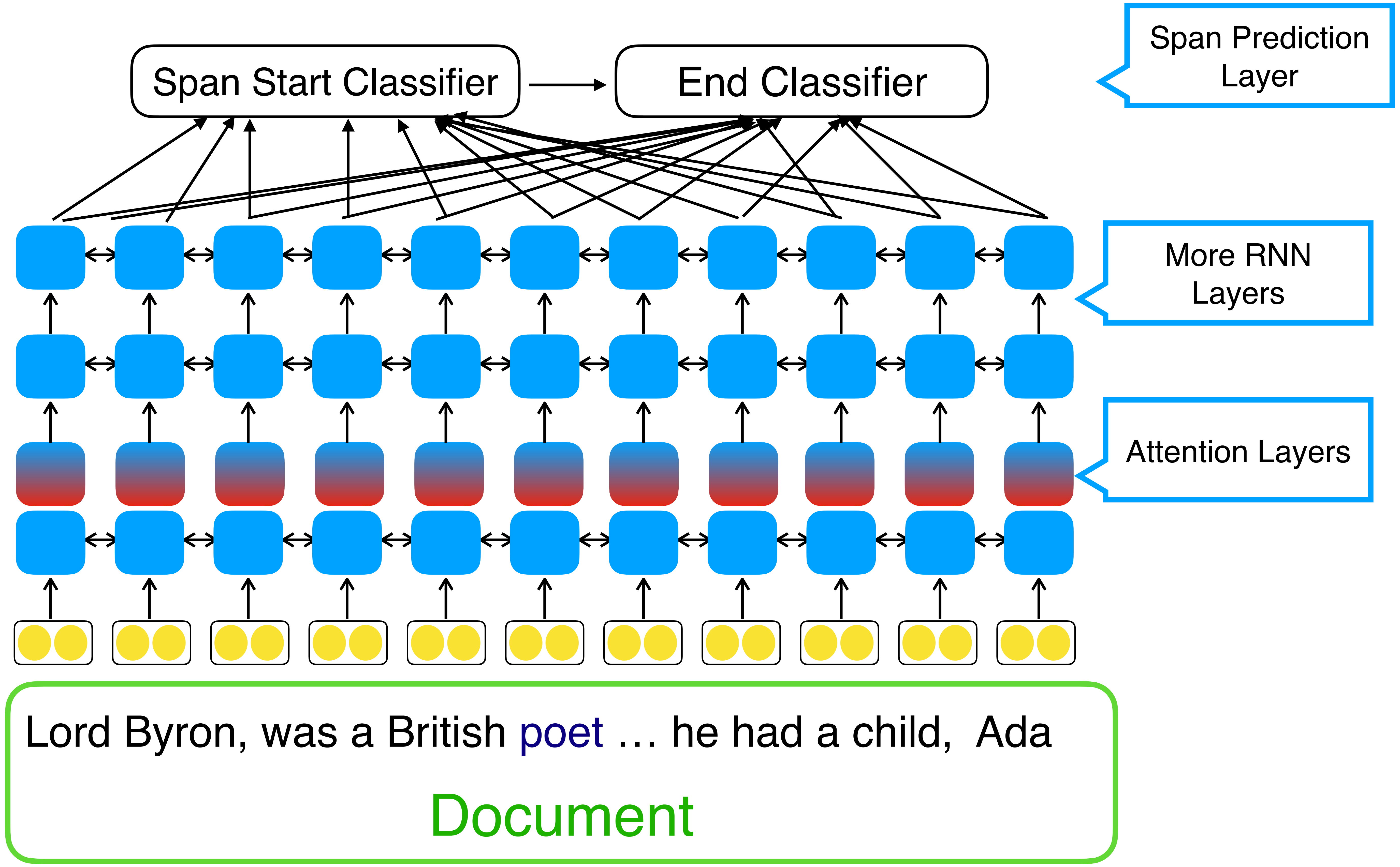
- Encode text and question with recurrent neural network
- Compute inter-sentence alignment with attention



[Seo et al, ICLR 17]

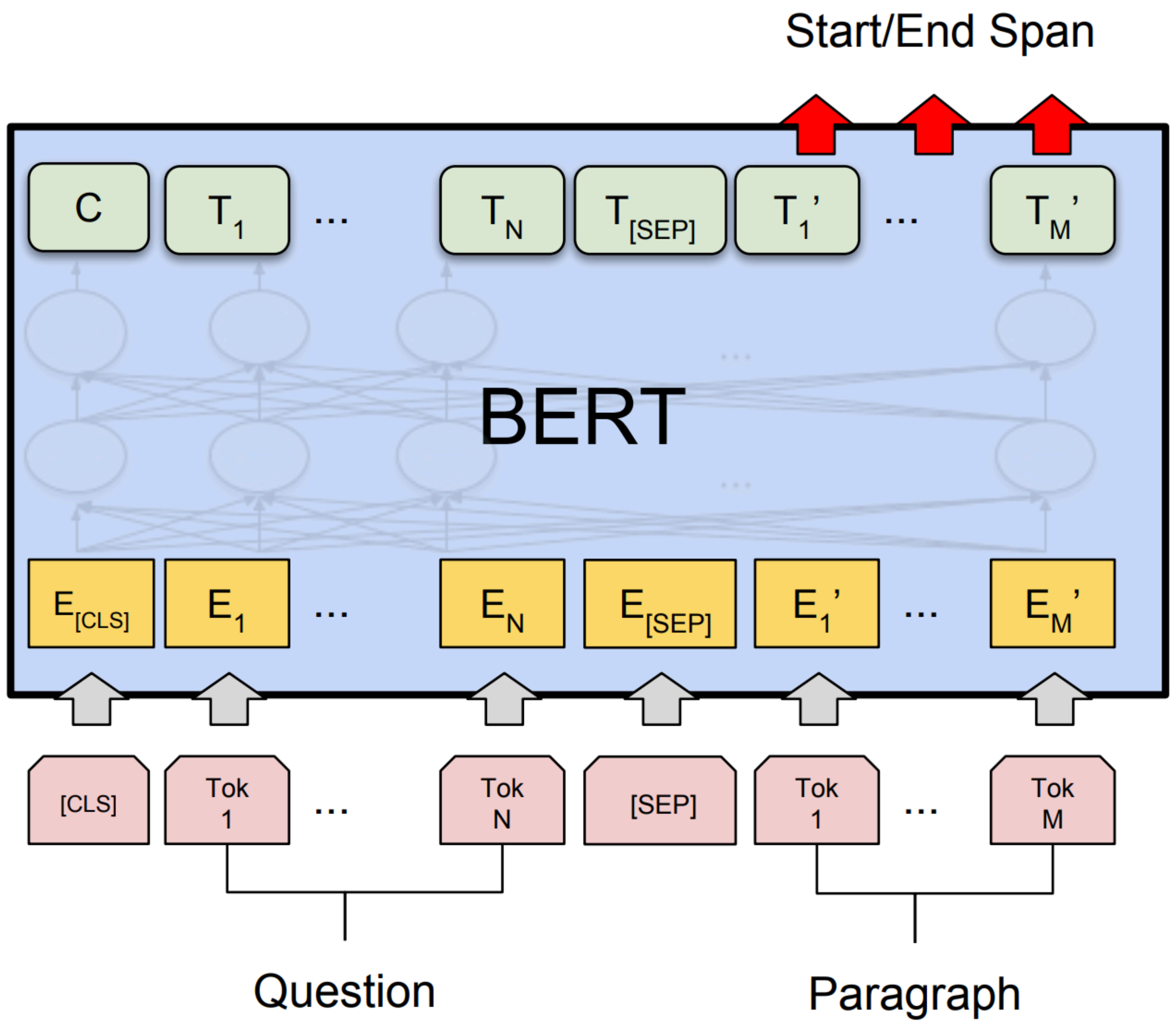
Model: BiDAF (Bi-directional Attention Flow)

- Encode text and question with recurrent neural network
- Compute inter-sentence alignment with attention
- Optimize for the log likelihood of finding the correct start and end positions



[Seo et al, ICLR 17]

Span-based QA with BERT



What was Marie Curie the first female recipient of ? [SEP] One of the most famous people born in Warsaw was Marie …

- Simplified architecture — just concatenate question and context!

Devlin et al. (2019)

Free-form answer generation

Question: Why does salt bring out the flavor in most foods?

Answer: Salt does a couple of things that add to the flavor of foods. First off, it makes things salty. That sounds simplistic, but salt is one of the 5 basic tastes, so it tends to taste good simply all by itself. Secondly, salt will lessen sensitivity to other tastes, especially bitter. This is somewhat true of all the tastes, but adding salt will balance out a dish and not let one taste overpower the others. Thirdly, it's been shown to increase that aromatic effects of many types of food. A good deal of your "taste" of a food actually comes from the smell of that food (which is why things tend to taste so bland when you nose is congested, like when you have the flu).

- Can capture complex semantics
- Evaluation??

Explain Like I'm Five (ELI5), from Fan et al. 2019

Difficulty of evaluating LFQA

Lexical matching based automatic metrics (e.g. ROUGE) are used, but not meaningful.

Q: Can you protect electronics from EMPs/solar flares? If so, how?

Random answer (ROUGE-L: 19.4)

The fast lane/slow lane is a bit of a misnomer. It gives the impression that new, faster lanes are being built. In reality, normal speed will be the new "fast lane", which will cost extra, and everything else will be in the throttled "slow lane".

Gold answer (ROUGE-L: 18.6)

I'll start with the grounding question, because that's the easiest to answer: Doesn't help a bit. All that matters is that the metal container is conductive and doesn't have gaps...completely seal your Faraday cage. Consider soldering the lid on to that paint can... look at little baggie it comes in. Sealed mylar. That protected that chip from air travel at 35,000 feet, land travel through rural, urban, and suburban areas, and all the electromagnetic radiation that the trip entails... No lead shielding. No safes…

Krishna et al. 2021

Understanding LFQA

Question: Can the capacity of our brains be roughly measured in bytes?

💡 Summary
✏️ Answer
💬 Example

📖 Auxiliary Info
🧩 Org sentence

Human written answer: This is a hard question to answer.Of course, since we occupy finite space, our properties must be finite.But the exceedingly complex structure of the brain and other systems […] makes it hard to calculate the amount of "data" that we can store.We don't have a way to measure resolution of life or the quality of everyday noises.[…]Most of our memories are vague recollections,[…].However, we can estimate average capacity, although it has no clear meaning in our terms of cognition. That number would be 2.5 petabytes […]Even so, this number varies as well from person to person.[…]

- Not all sentences in long form answer convey answer information
- About 40% of sentences serve other roles

[How do we answer complex questions: Discourse structure of long form answers Xu et al, ACL 2022]

Dataset Properties

- Axis 1: what's the output space?
  - cloze task (fill in blank)
  - multiple choice
  - span-based prediction
  - freeform generation
- Complex output space allows answering more complex queries, but evaluation becomes very tricky…