Algorithms that learn to think on their feet

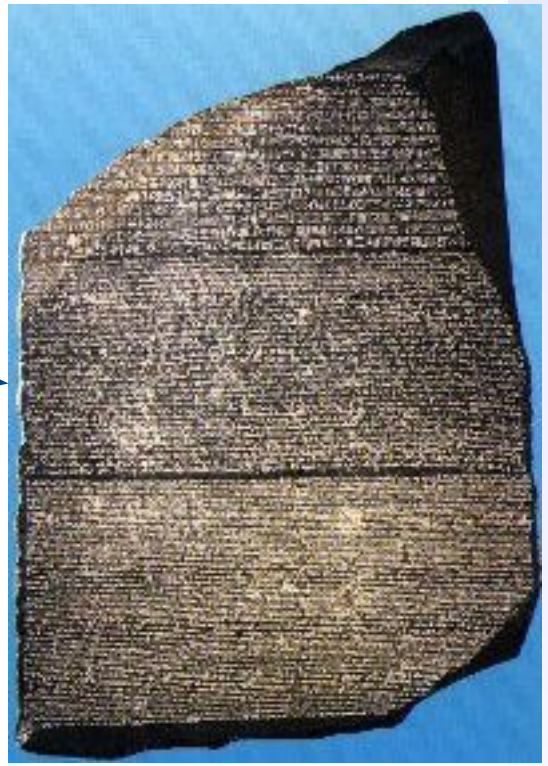
Hal Daumé III | UMD CS, UMIACS, Linguistics | me@hal3.name | @haldaume3

What is NLP?

- Fundamental goal: deep understanding of text
  - Not just string processing or keyword matching
- End systems that we want to build
  - Simple: Spelling correction, text categorization, etc.
  - Complex: Speech recognition, machine translation, information extraction, dialog interfaces, question answering
  - Unknown: human-level comprehension (more than just NLP?)

Why is language hard?

- Ambiguity abounds (some headlines)
  - Iraqi Head Seeks Arms
  - Teacher Strikes Idle Kids
  - Kids Make Nutritious Snacks
  - Stolen Painting Found by Tree
  - Local HS Dropouts Cut in Half
  - Enraged Cow Injures Farmer with Ax
  - Hospitals are Sued by 7 Foot Doctors
  - Ban on Nude Dancing on Governor's Desk
  - Scientists study whales from space
- Why are these funny?
- What does ambiguity imply about the role of learning?



Despite ambiguity, language is predictable

I like my coffee with cream and *asparagus*

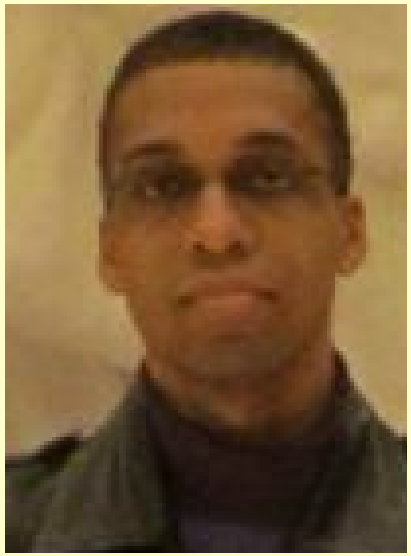
This is crummy weather for San *ta Claus*

- The brain uses this information!
- Can we use predictability to make decisions *before* all of the input is observed?

YES!!!

Outline

Quizbowl (Incremental Question Answering)



Alvin Grissom II

He He

train to Ulm

taneous

Machine

retation

Simultaneous (machine) interpretation

Nuremburg Trials

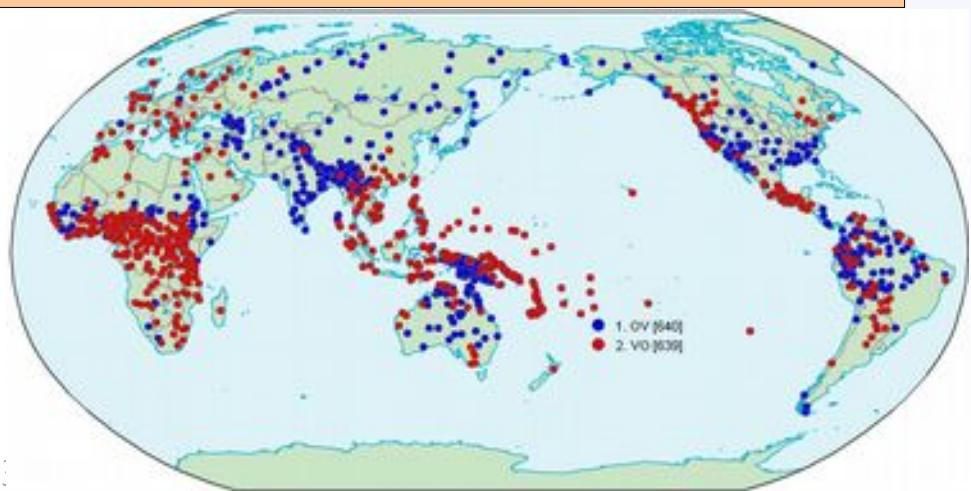
- Dozens of defendants
- Judges from four nations (three languages)
- Status quo: speak, then translate
- After Nuremberg, simultaneous translations became the norm
- Long wait → bad conversation

Why simultaneous interpretation is hard

- Human languages have vastly different word orders
  - About half are OV, the other half are VO
  - This comes with a lot more baggage than just verb-final

Running (German/English) Example:

Ich	bin	mit	dem	Zug	nach	Ulm	gefahren
I	am	with	the	train	to	Ulm	traveled
I	*(……waiting……)*						traveled by train to Ulm



Model for interpretation decisions

- We have a set of actions (predict / translate)
  - Wait
  - Predict clause-verb
  - Predict next word
  - Commit ("speak")
- In a changing environment (state)
  - The words we've seen so far
  - Our models' internal predictions
- With a well defined notion of "optimal action" at training time

Example of interpretation trajectory

Ich	bin	mit	dem	Zug	nach	Ulm	gefahren
I	am	with	the	train	to	Ulm	traveled
I	*(……waiting……)*						traveled by train to Ulm

DAgger: Dataset Aggregation

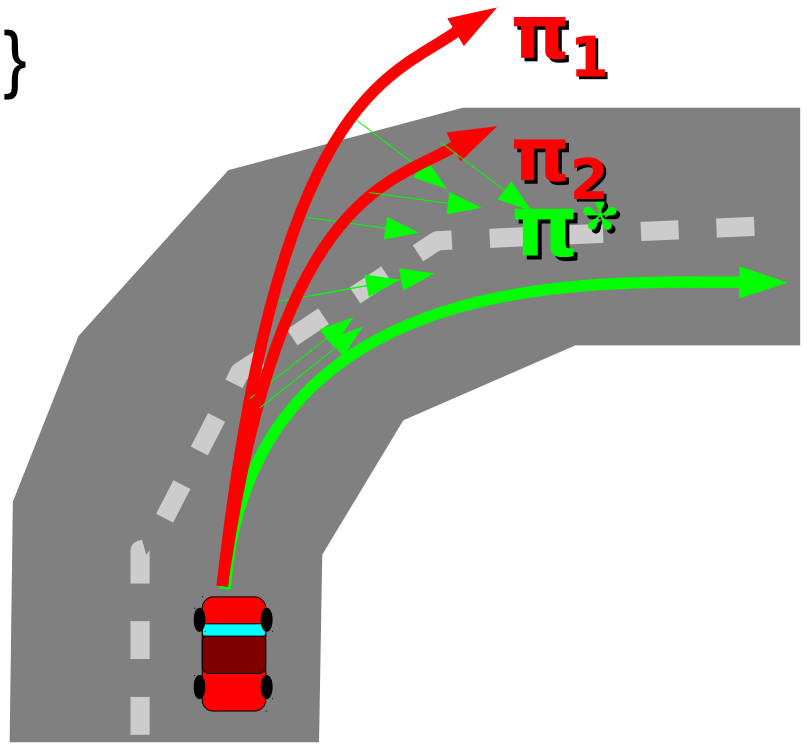
- Collect trajectories from expert π^*
- Dataset $\mathbf{D_0} = \{ (s, \pi^*(s)) \mid s \sim \pi^* \}$
- Train π_1 on $\mathbf{D_0}$
- Collect new trajectories from π_1
  - But let the *expert* steer!
- Dataset $\mathbf{D_1} = \{ (s, \pi^*(s)) \mid s \sim \pi_1 \}$
- Train π_2 on $\mathbf{D_0} \cup \mathbf{D_1}$

- In general:
  - $\mathbf{D_n} = \{ (s, \pi^*(s)) \mid s \sim \pi_n \}$
  - Train π_n on $\mathbf{U_{i<n}\, D_i}$

If N = T log T,

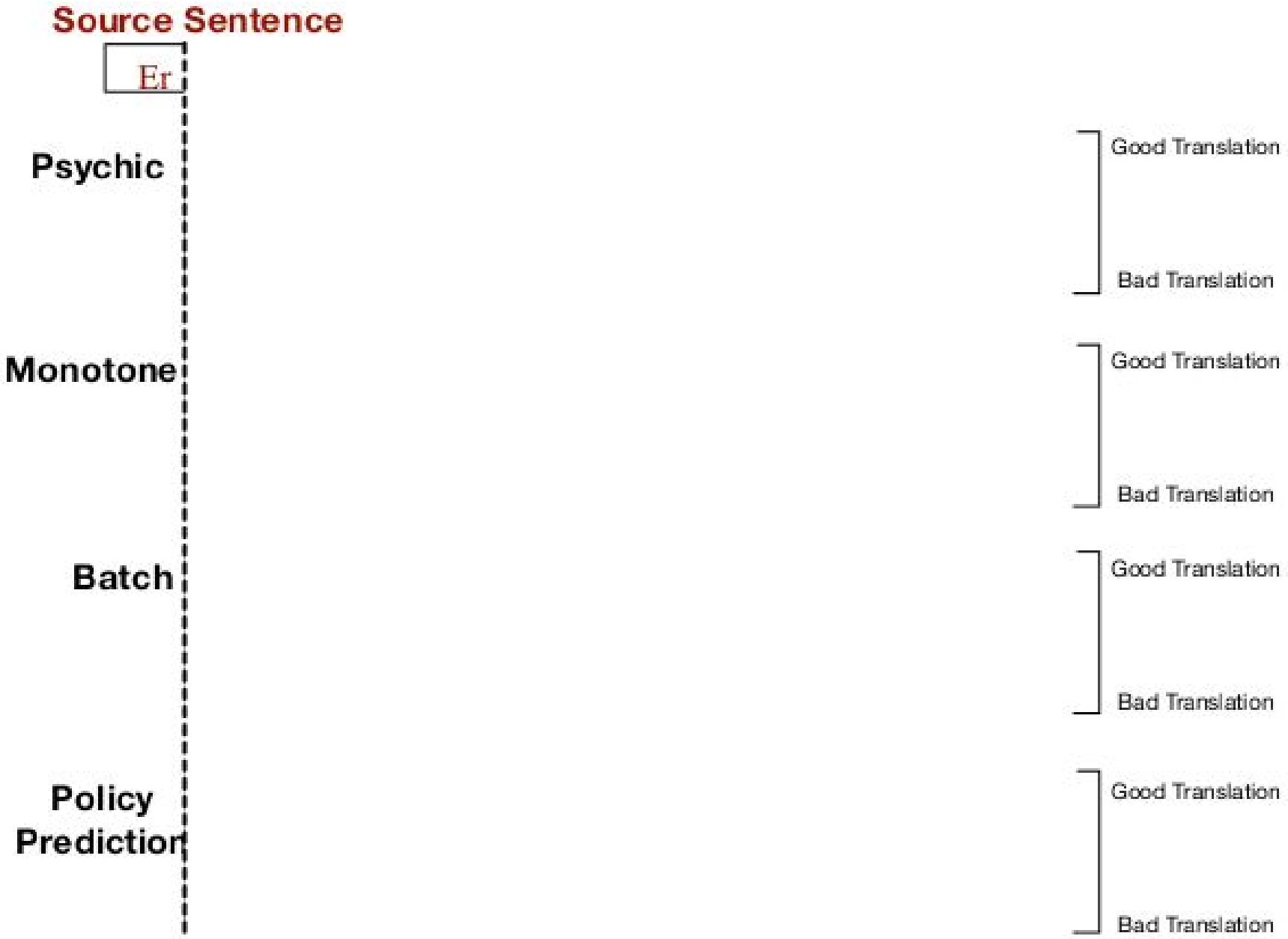
$\mathbf{L(\pi_n) < T\, \epsilon_N + O(1)}$

for some n



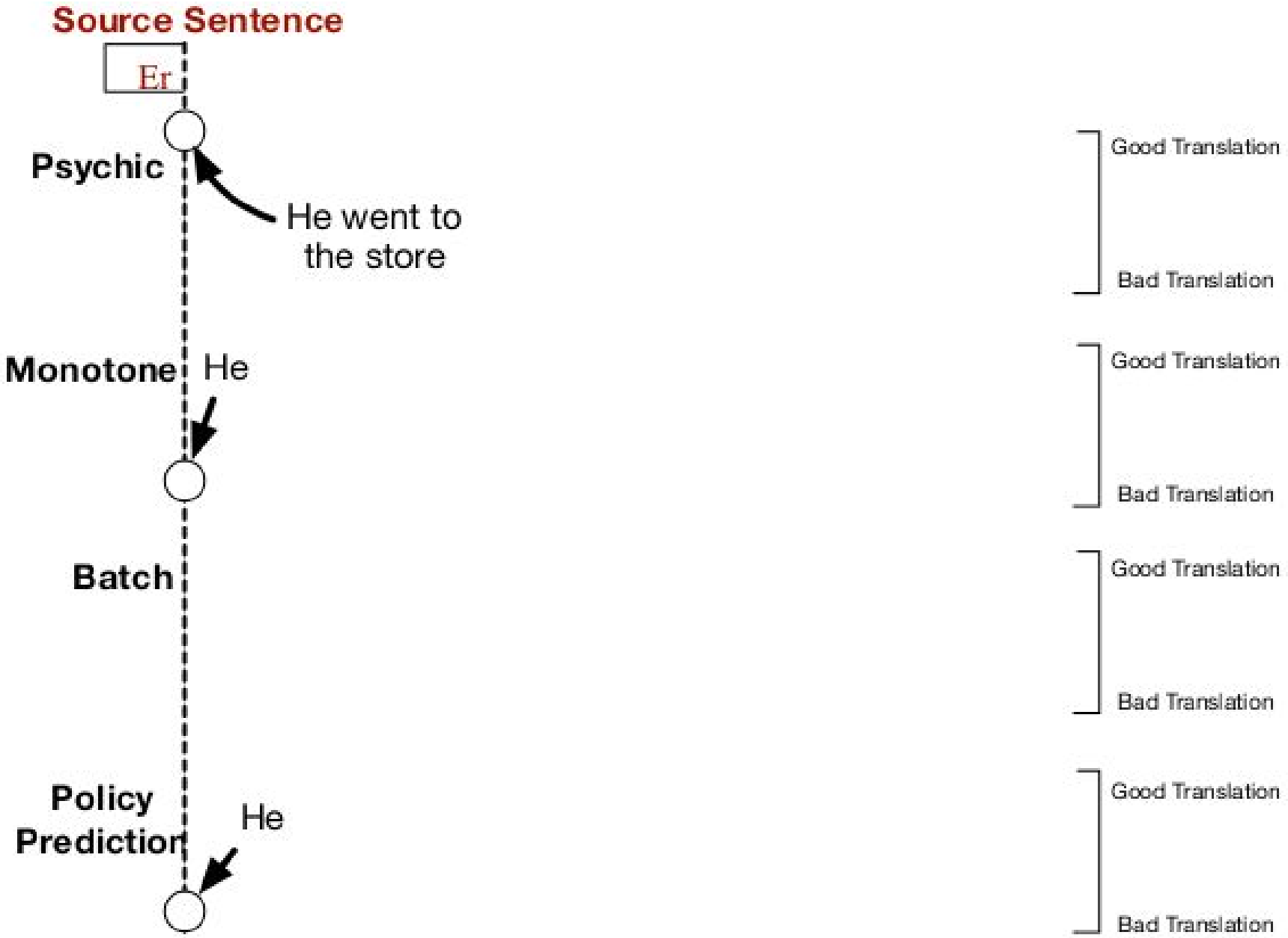
(Ross+Gordon+Bagnell, AIStats 2011)

Evaluating performance and baselines



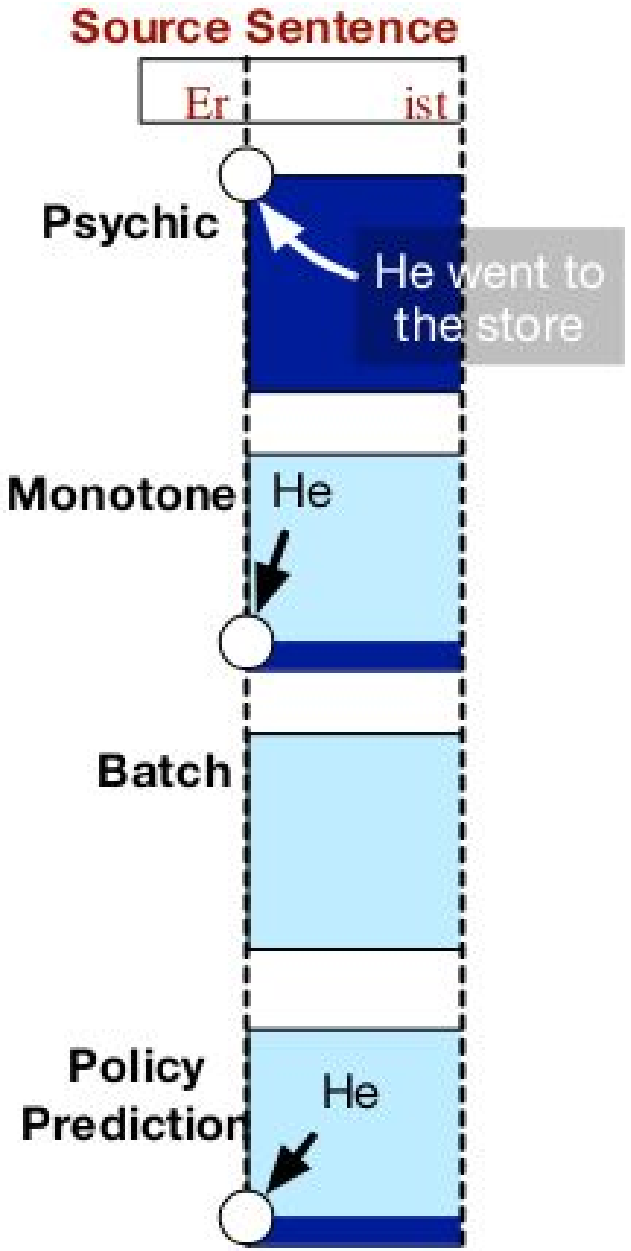
(Grissom II et al., EMNLP 2014)

Evaluating performance and baselines



(Grissom II et al., EMNLP 2014)

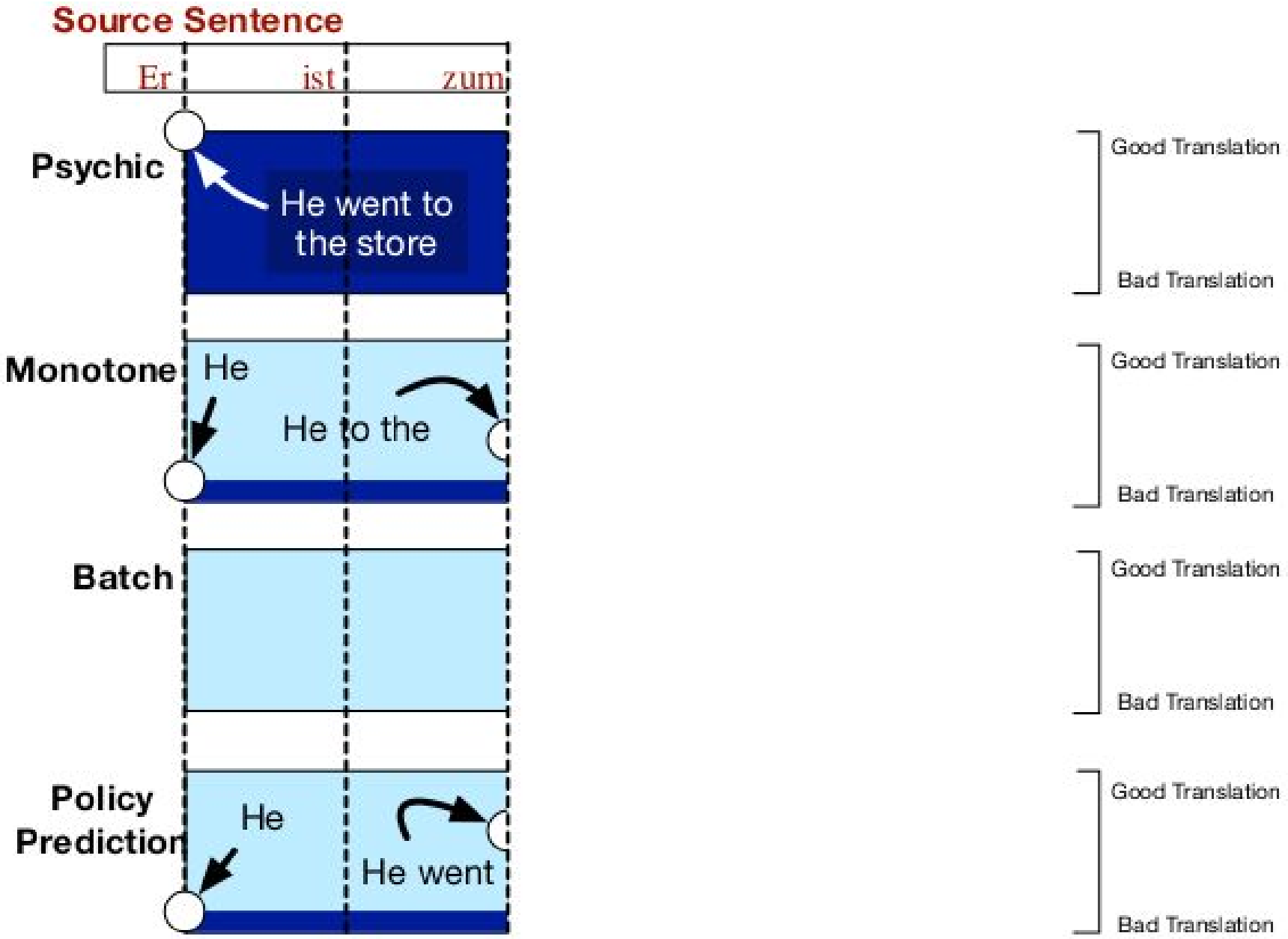
Evaluating performance and baselines



Good Translation
Bad Translation
Good Translation
Bad Translation
Good Translation
Bad Translation
Good Translation
Bad Translation

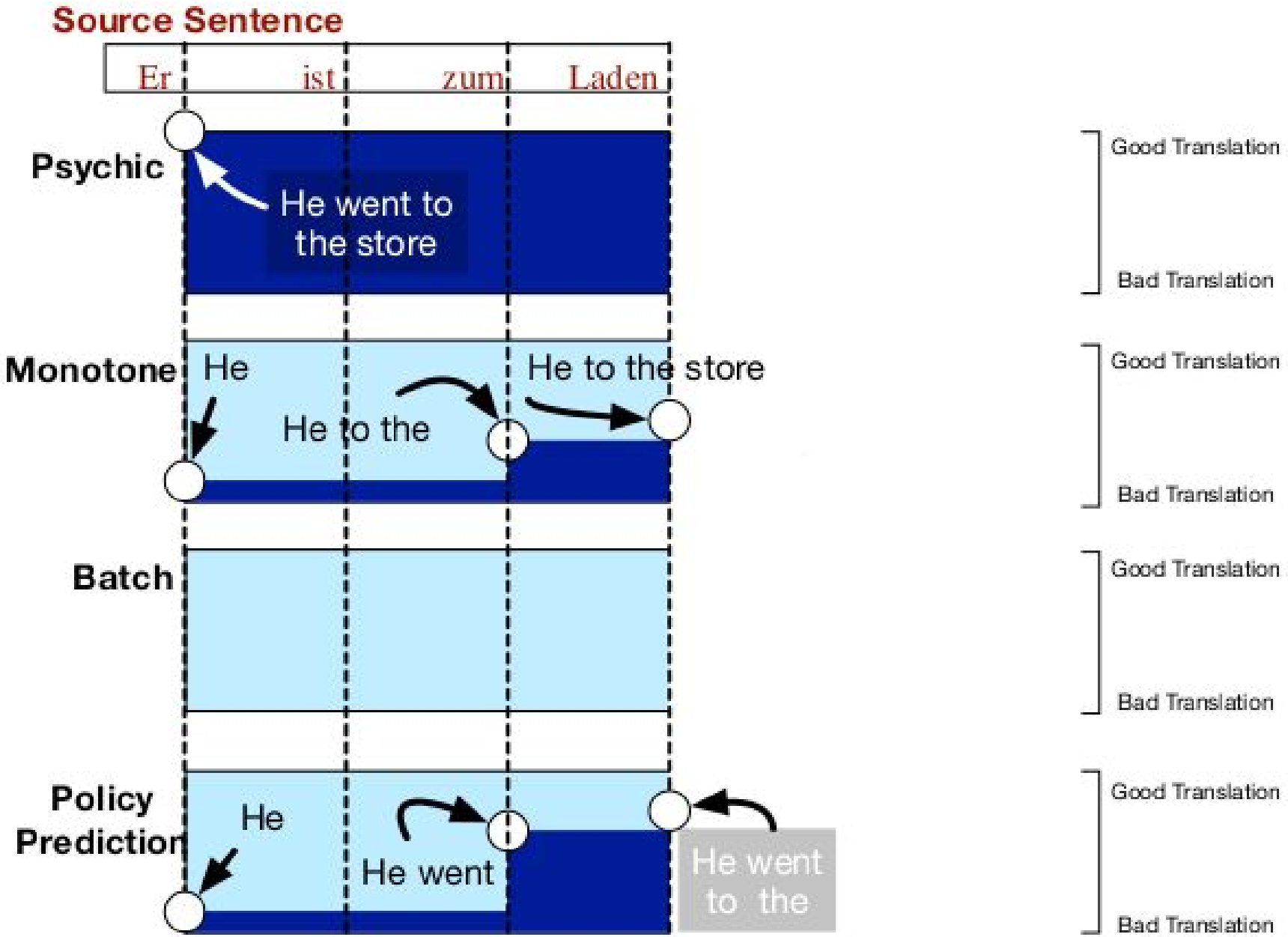
(Grissom II et al., EMNLP 2014)

Evaluating performance and baselines



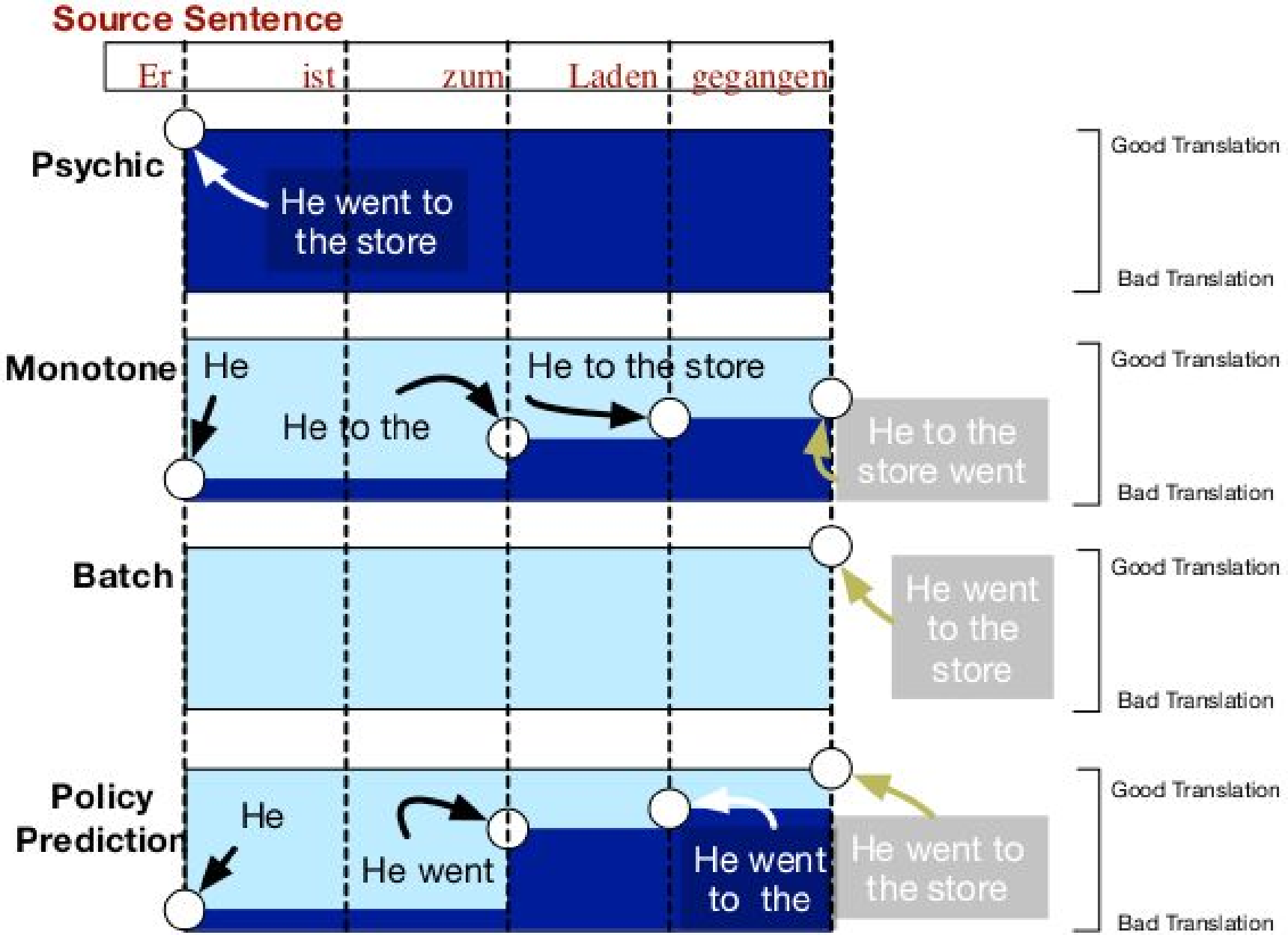
(Grissom II et al., EMNLP 2014)

Evaluating performance and baselines



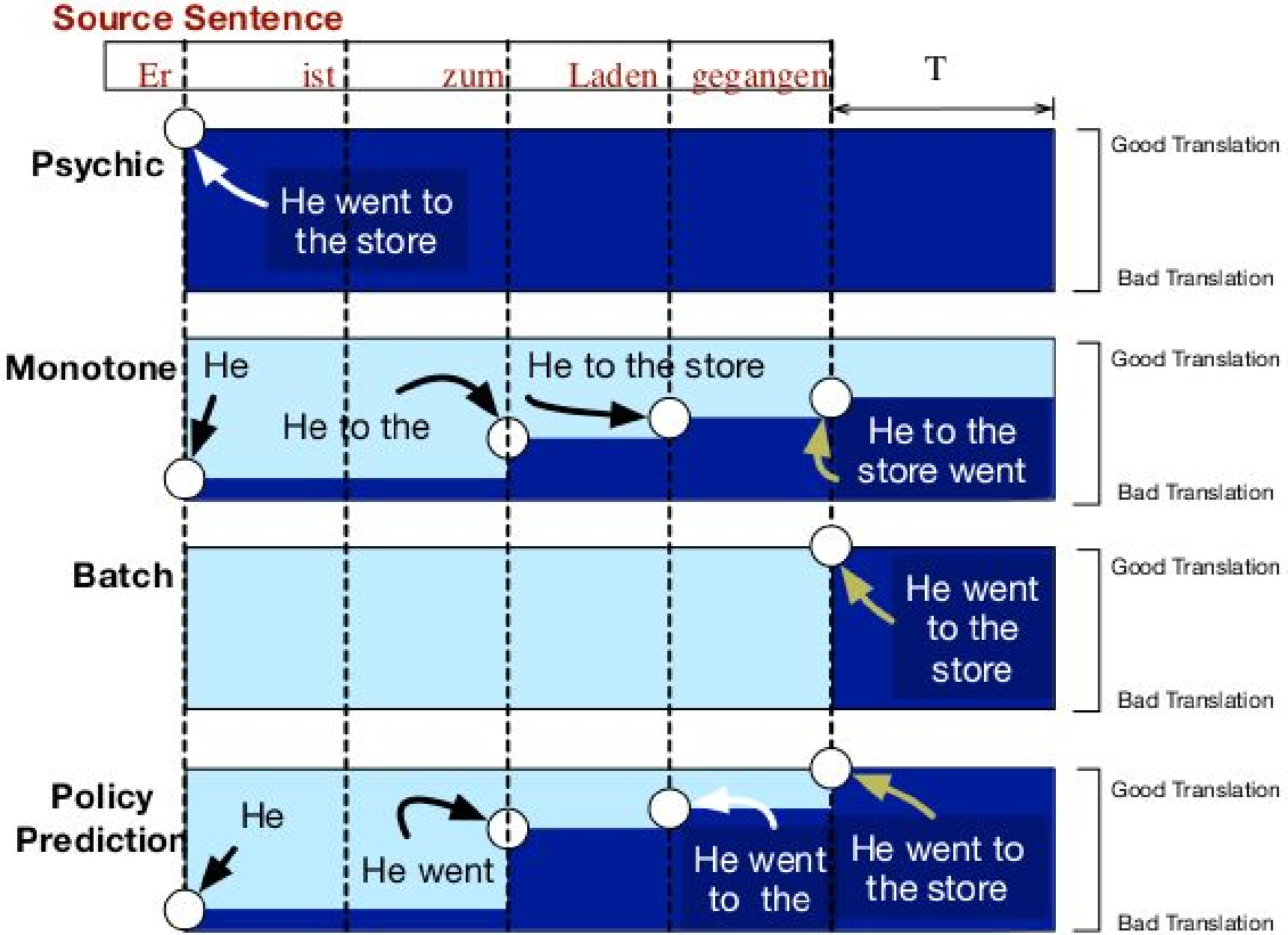
(Grissom II et al., EMNLP 2014)

Evaluating performance and baselines



(Grissom II et al., EMNLP 2014)

Evaluating performance and baselines



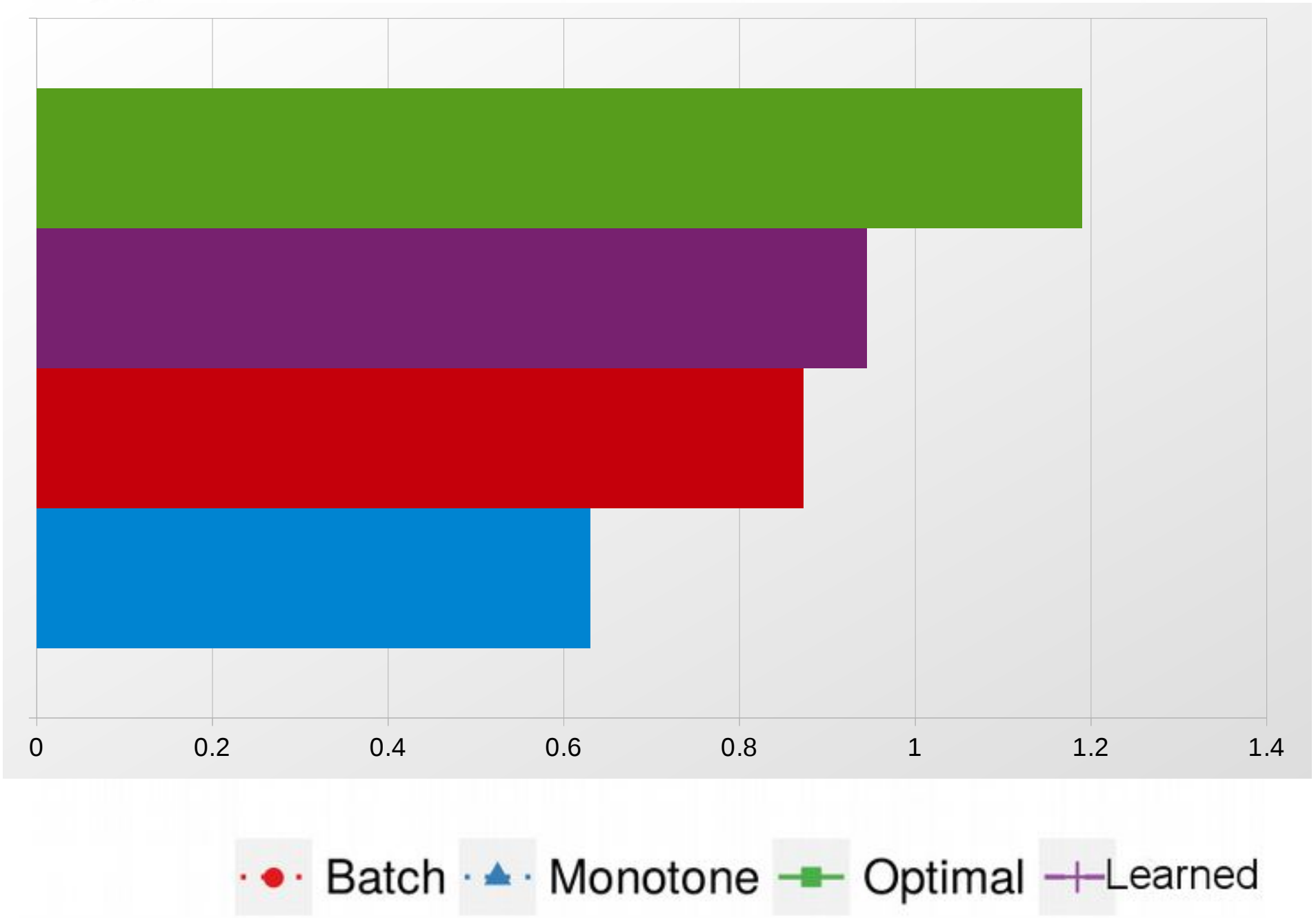
(Grissom II et al., EMNLP 2014)

Training the policy

- Actions:
  - Commit `translate(revealed words)`
  - Predict (verb/next) `translate(revealed + predicted)`
  - Wait `get_next_words()`
- Delayed feedback: latency BLEU
- Features:
  - Output & confidence of predictors
  - Internal translation / language model scores
  - Previous decisions made by policy

(Grissom II et al., EMNLP 2014)

Evaluating performance



(Grissom II et al., EMNLP 2014)

Outline

Quizbowl (Incremental Question Answering)

Mohit Iyyer

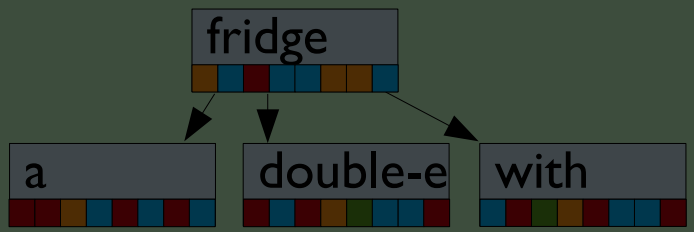
He He

train to Ulm

taneous

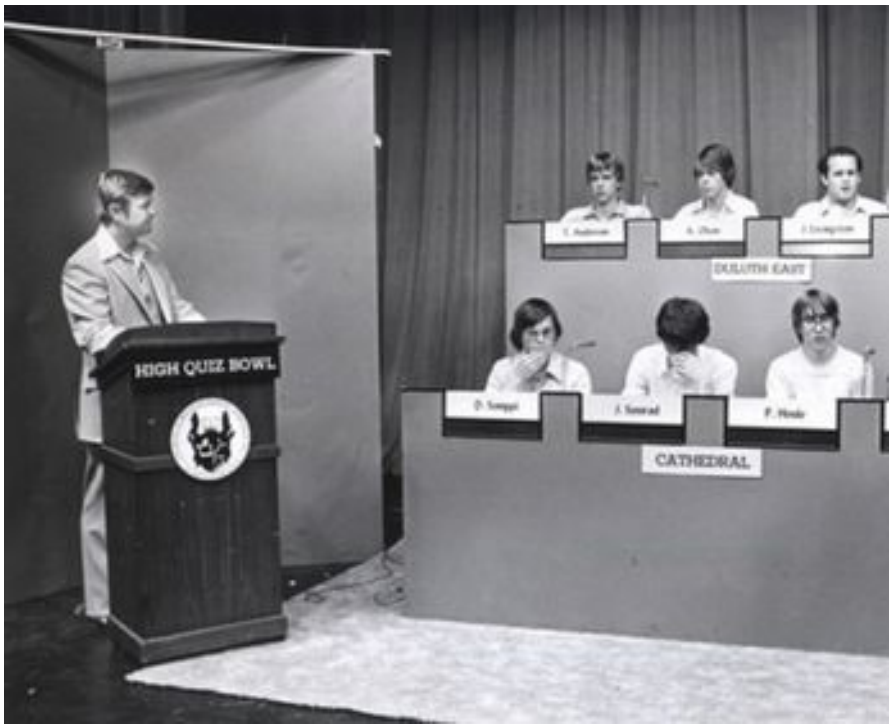
Machine

retation



Humans doing incremental prediction

- Game called "quiz bowl"
- Two teams play each other
  - Moderator reads a question
  - When a team knows the answer, they buzz in
  - **If right**, they get points; **otherwise**, rest of the question is read to the other team
- Hundreds of teams in the US alone
- Example …



(Iyyer et al., ACL 2014)

Quizbowl example

With Leo Szilard, he invented a doubly-eponymous

(Iyyer et al., ACL 2014)

Quizbowl example

With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory

(Iyyer et al., ACL 2014)

Quizbowl example

With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory of heat capacity, so

(Iyyer et al., ACL 2014)

Quizbowl example

With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory of heat capacity, so Debye adjusted the theory for low temperatures. His summation convention automatically sums repeated indices in tensor products. His name is attached to the A and B coefficients

(Iyyer et al., ACL 2014)

Quizbowl example

With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory of heat capacity, so Debye adjusted the theory for low temperatures. His summation convention automatically sums repeated indices in tensor products. His name is attached to the A and B coefficients for spontaneous and stimulated emission, the subject of one of his multiple groundbreaking 1905 papers. He further developed the model of statistics sent to him by

(Iyyer et al., ACL 2014)

Quizbowl example

With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory of heat capacity, so Debye adjusted the theory for low temperatures. His summation convention automatically sums repeated indices in tensor products. His name is attached to the A and B coefficients for spontaneous and stimulated emission, the subject of one of his multiple groundbreaking 1905 papers. He further developed the model of statistics sent to him by Bose to describe particles with integer spin. For 10 points, who is this German physicist best known for formulating

(Iyyer et al., ACL 2014)

Quizbowl example

With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory of heat capacity, so Debye adjusted the theory for low temperatures. His summation convention automatically sums repeated indices in tensor products. His name is attached to the A and B coefficients for spontaneous and stimulated emission, the subject of one of his multiple groundbreaking 1905 papers. He further developed the model of statistics sent to him by Bose to describe particles with integer spin. For 10 points, who is this German physicist best known for formulating the special and general theories of relativity?

(Iyyer et al., ACL 2014)

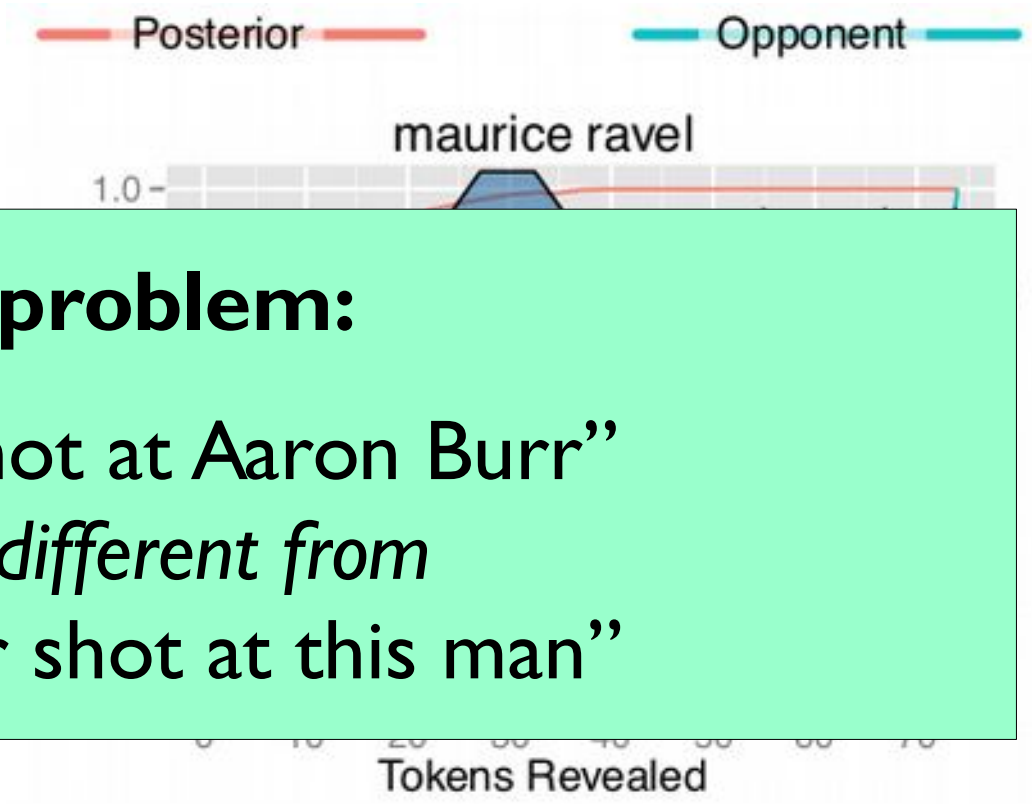
Solving incrementally

- Action: buzz now or wait
  - Content Model is constantly generating guesses
  - Oracle provides examples where it is correct
  - The Policy generalizes to test data
  - Features represent our state

(Iyyer et al., ACL 2014)

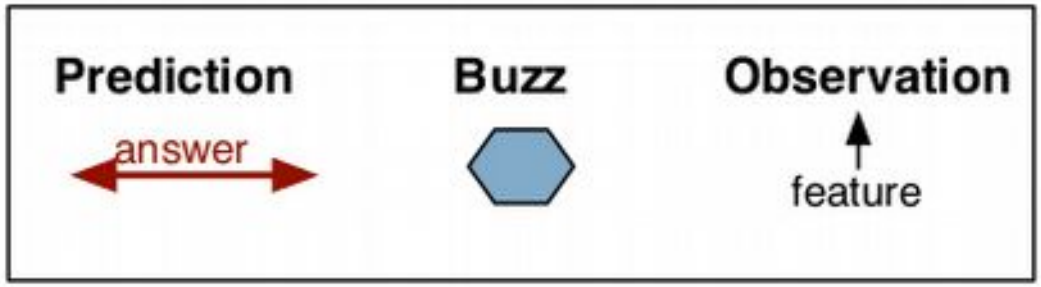
Evaluation methodology

- Mechanical Turk to collect human data
- 7000 questions were
- Total of 461 unique users
- Leaderboard to encourage users



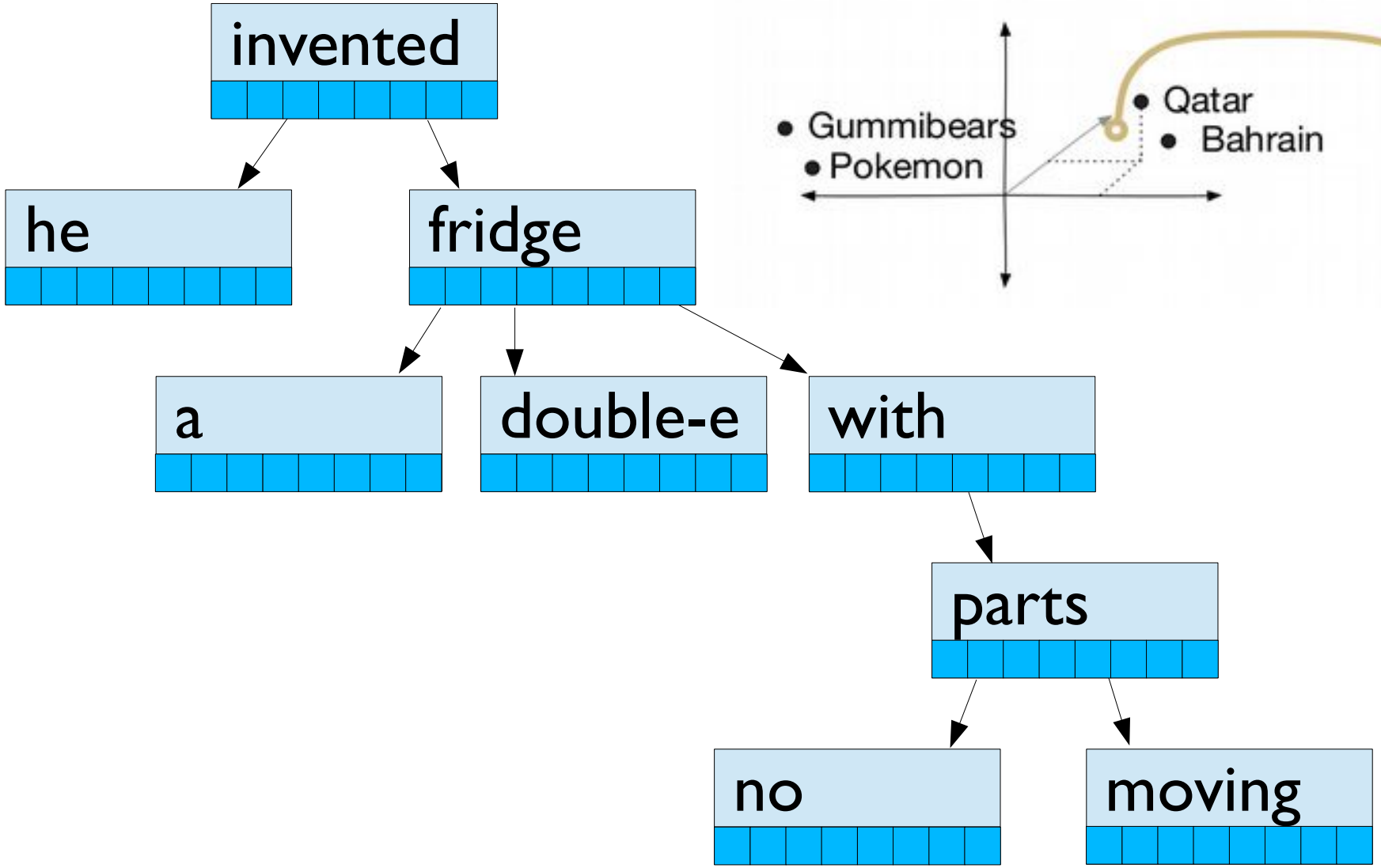
Big problem:

"this man shot at Aaron Burr"
is very different from
"Aaron Burr shot at this man"



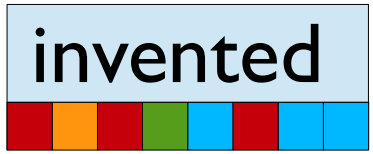
(Iyyer et al., ACL 2014)

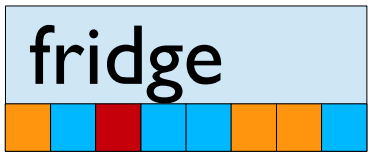
Challenge: modeling compositionality

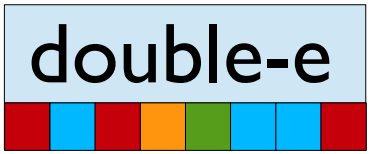


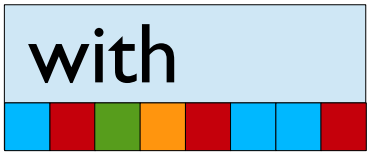
(Iyyer et al., ACL 2014)

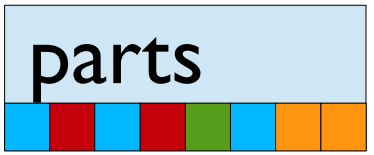
Challenge: modeling compositionality

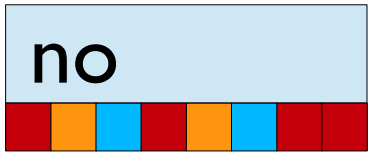


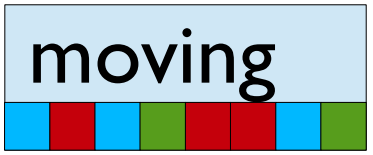


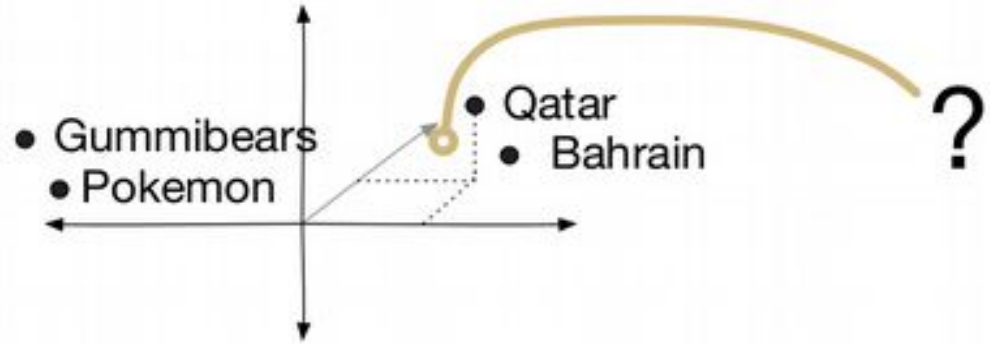






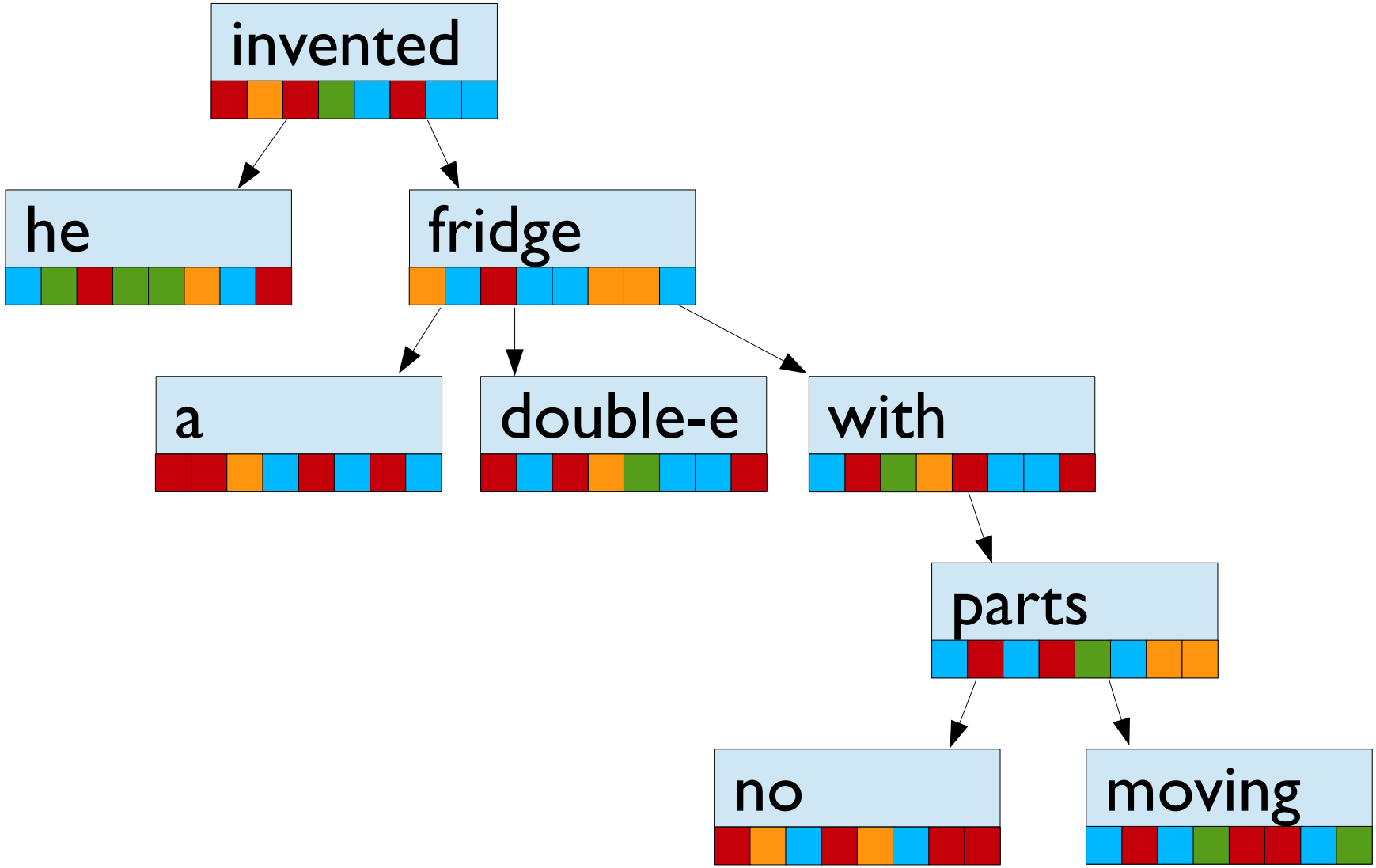






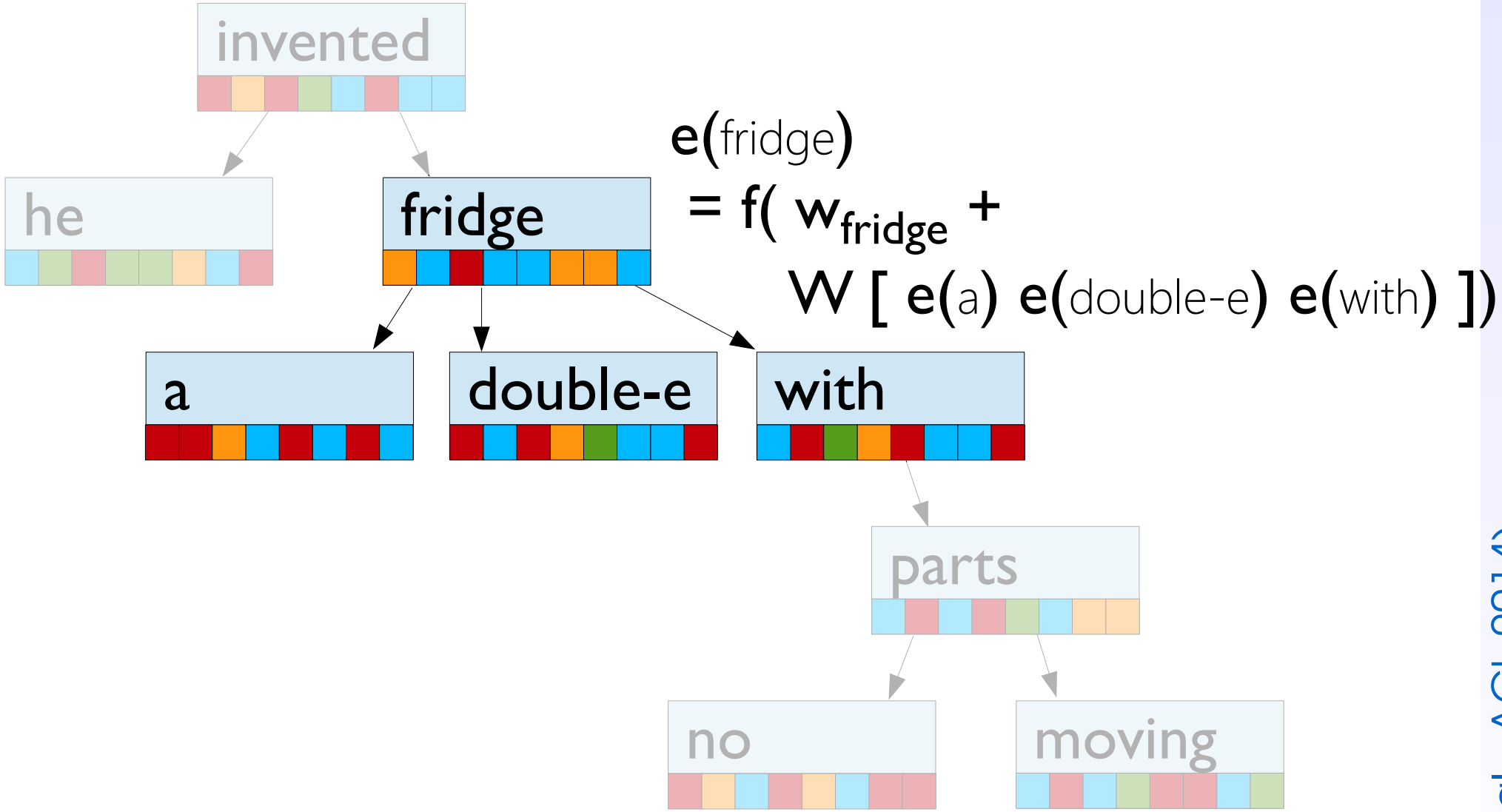
(Iyyer et al., ACL 2014)

Challenge: modeling compositionality



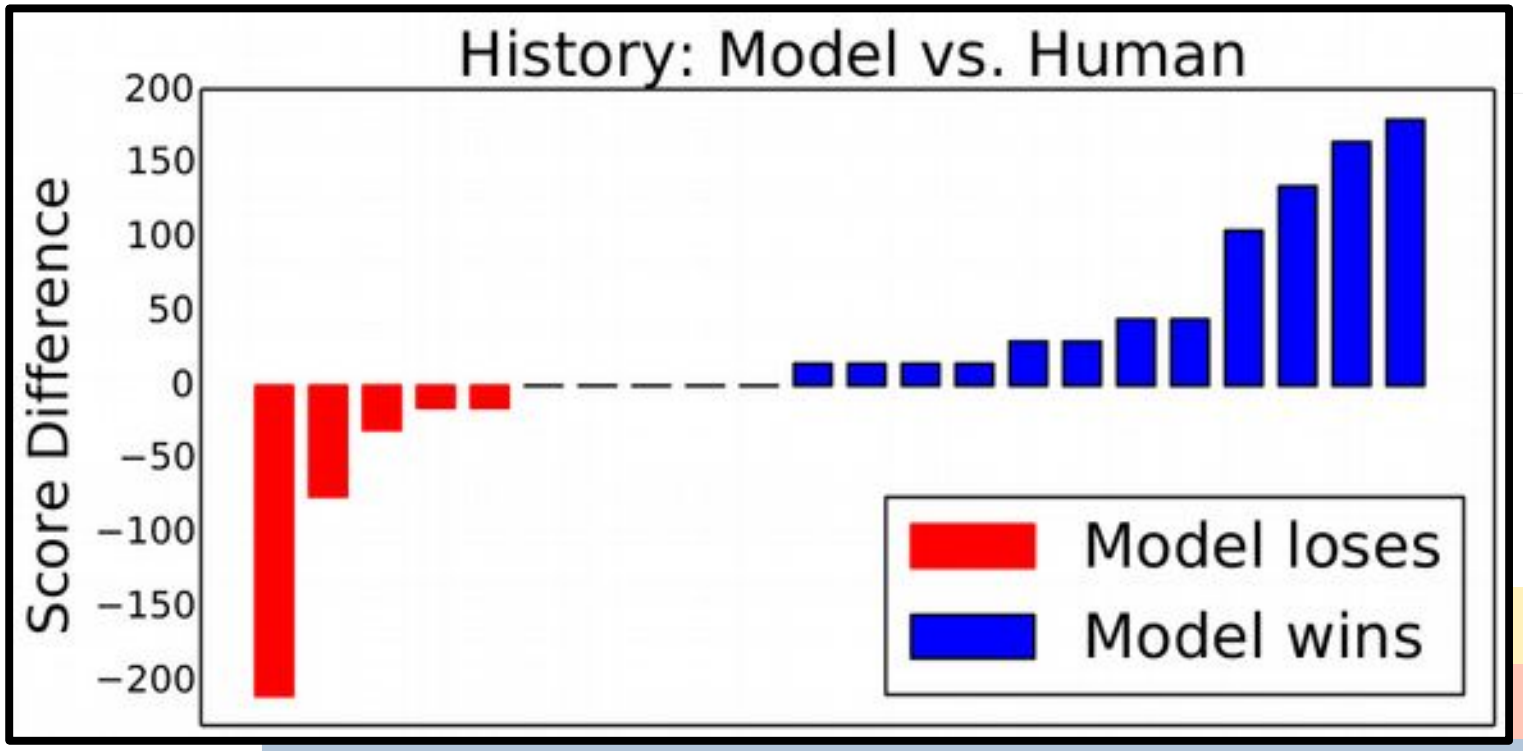
(Iyyer et al., ACL 2014)

Challenge: modeling compositionality



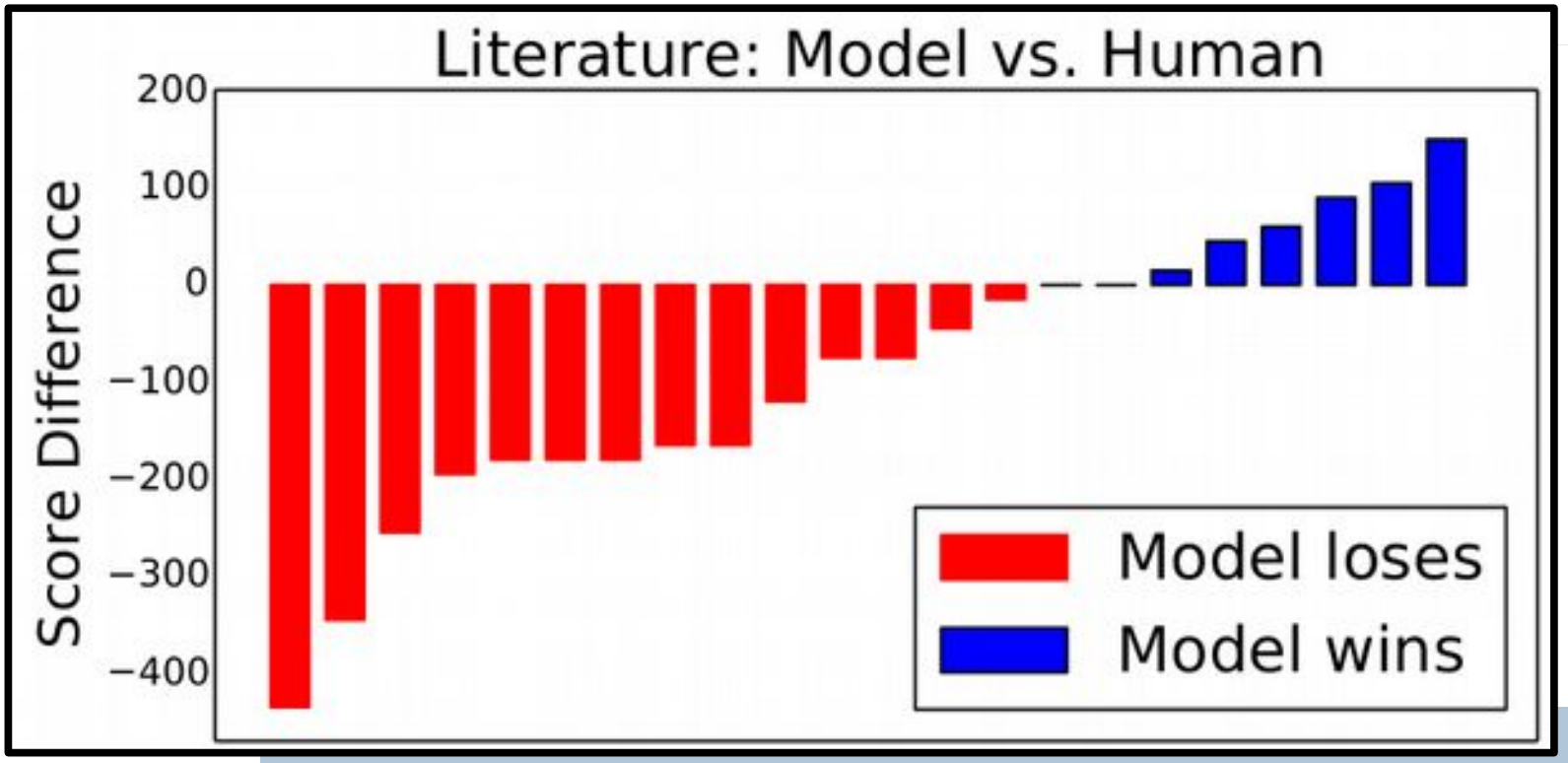
(Iyyer et al., ACL 2014)

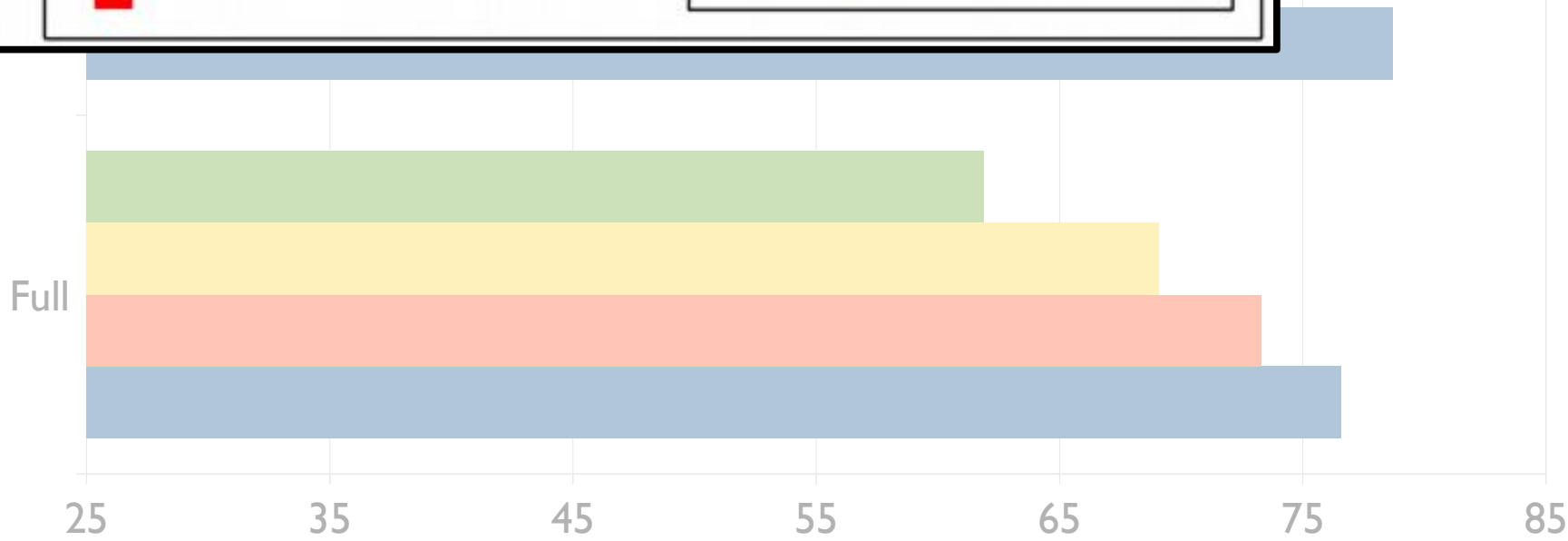
Results on question-answering task



(Iyyer et al., ACL 2014)

Results on question-answering task



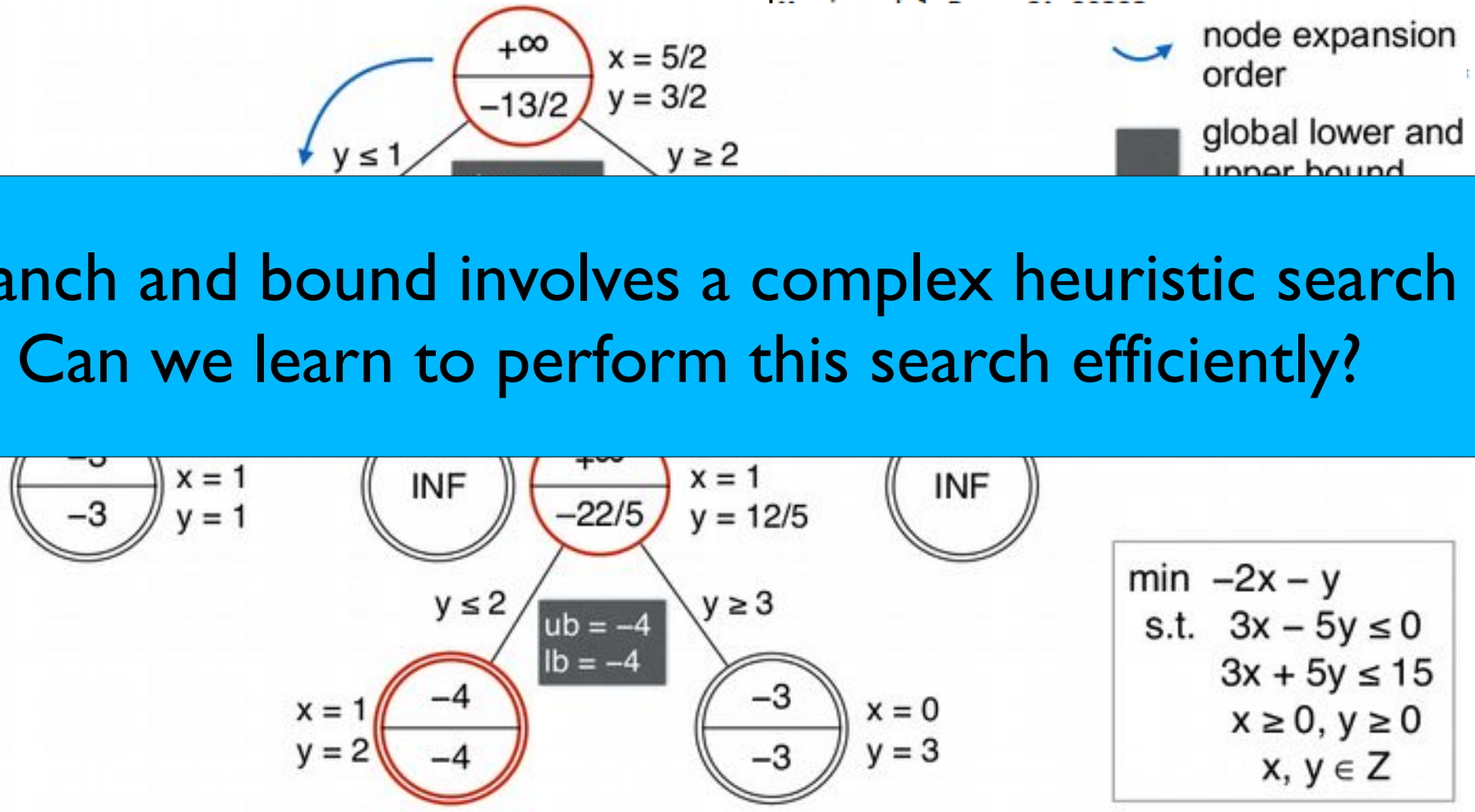


(Iyyer et al., ACL 2014)

Moving to more general frameworks

- Lots of NLP (+al) problems can be cast *at test time* as integer linear programs
- ILPs are usually solved using

```
Thursday, March 6, 2003
10:30am - 12:00pm
11th Floor Large Conference Room
USC/Information Sciences Institute
4676 Admiralty Way, Suite 1001
```

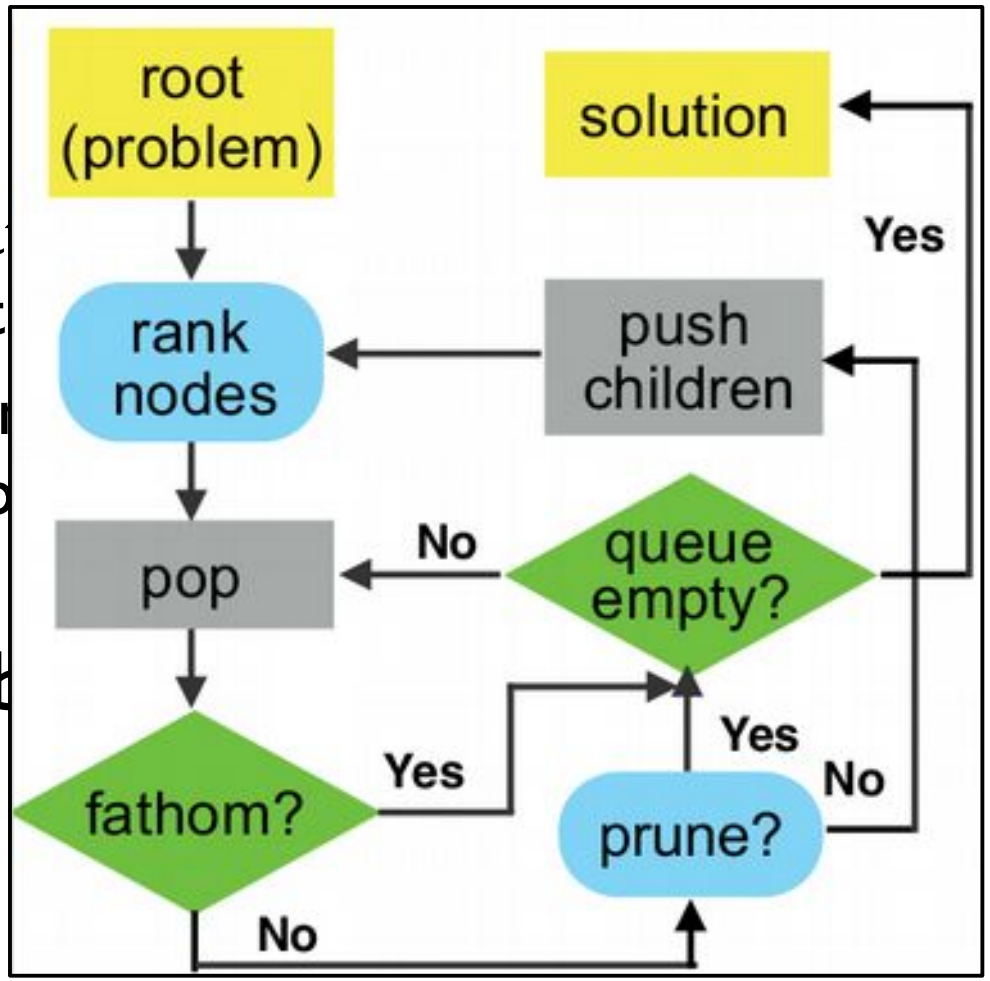


Branch and bound involves a complex heuristic search
Can we learn to perform this search efficiently?

(He+Eisner+D, NIPS 2014)

Some intuition

- A good search strategy should:
  - find a good incumbent solution early
  - identify non-promising nodes before expansion
- "Good" varies depending
  - DFS should only be used a
    good feasible solution t
  - Best-bound-first search ca
    nodes, but should not b
- We will learn a heuristic b
  capture this intuition



(He+Eisner+D, NIPS 2014)

Training and experiments

- Same algo
- Four (stan
- Comparis
  - DFS (baseline)
  - Gurobi (thousands of person-hours of effort)
- Measures:
  - Optimality Gap, Integrality Gap, and improvement from initial heuristic solution

We achieve less than 1.2 optimality gap while exploring 0.05%, 1.5%, 5.1% and 47% of the nodes explored by Gurobi!

Dataset	Ours(DAgger training)			DFS			Gurobi		
	OGap	IGap	Impr	OGap	IGap	Impr	OGap	IGap	Impr
MIK	**0.23**	16.63	**4.39**	6.74	35.48	0.00	**0.17**	**15.24**	0.36
Regions1	**0.54**	**4.53**	**10.57**	3.07	8.48	8.61	2.24	7.20	0.60
Regions2	**1.22**	**6.76**	**19.36**	4.75	11.38	15.12	1.65	7.48	2.15
Hybrid	**0.87**	**20.28**	**24.46**	1.69	23.08	23.53	**1.37**	23.49	1.58

(He+Eisner+D, NIPS 2014)


Jordan B-G


Jason Eisner


Alvin Grissom II


He He


Mohit Iyyer


John Langford

- Reasoning with incomplete information is useful for *speed* and *modeling*
- *Imitation learning* can help us build such systems
  - Plug: even when you can't construct a perfect oracle (see *LOLS*, ICML 2015)
- Wide range of new, interesting problems to work on!
  - How to learn from human interpreters?
  - How to learn to compete?
  - How to *not need* BOW in deepNN models?

Thanks! Questions?