Algorithms that learn to think on their feet

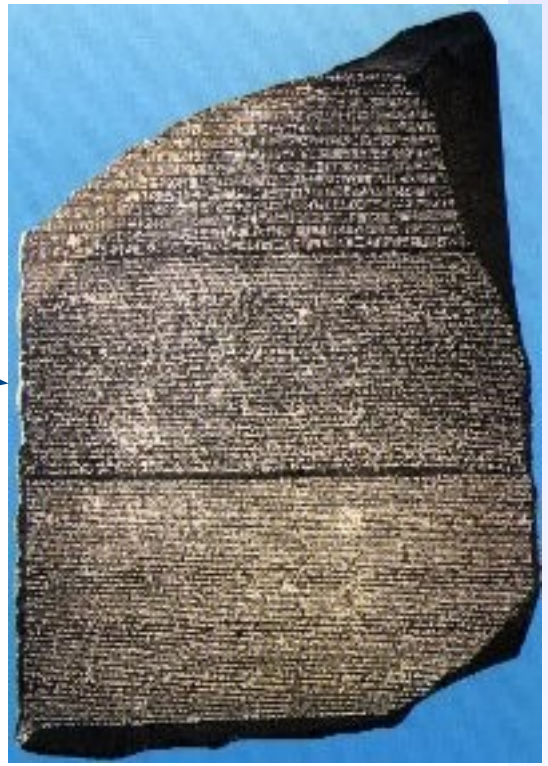
Hal Daumé III | UMD CS, UMIACS, Linguistics | me@hal3.name | @haldaume3

What is NLP?

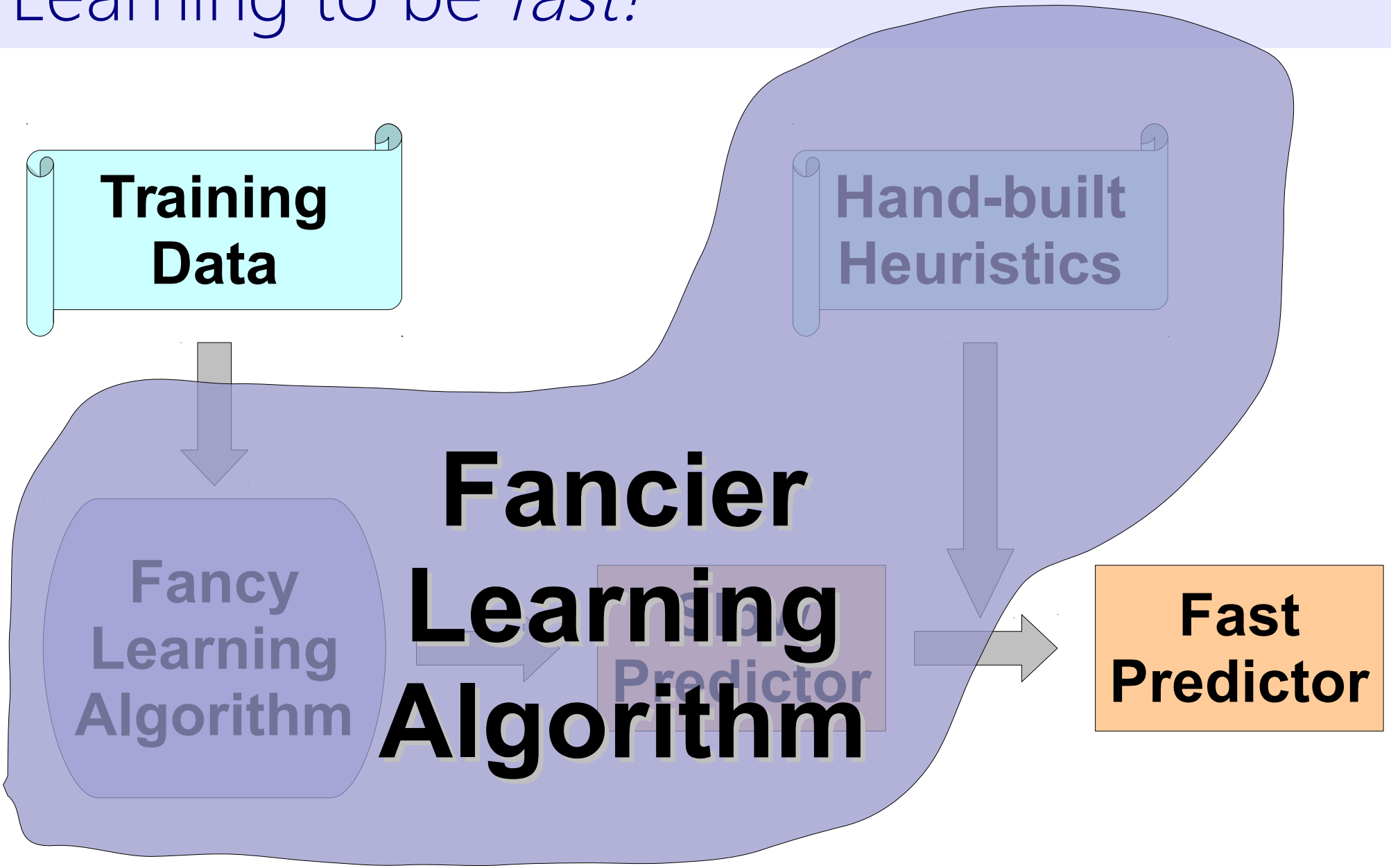
- Fundamental goal: deep understanding of text
  - Not just string processing or keyword matching
- End systems that we want to build
  - Simple: Spelling correction, text categorization, etc.
  - Complex: Speech recognition, machine translation, information extraction, dialog interfaces, question answering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

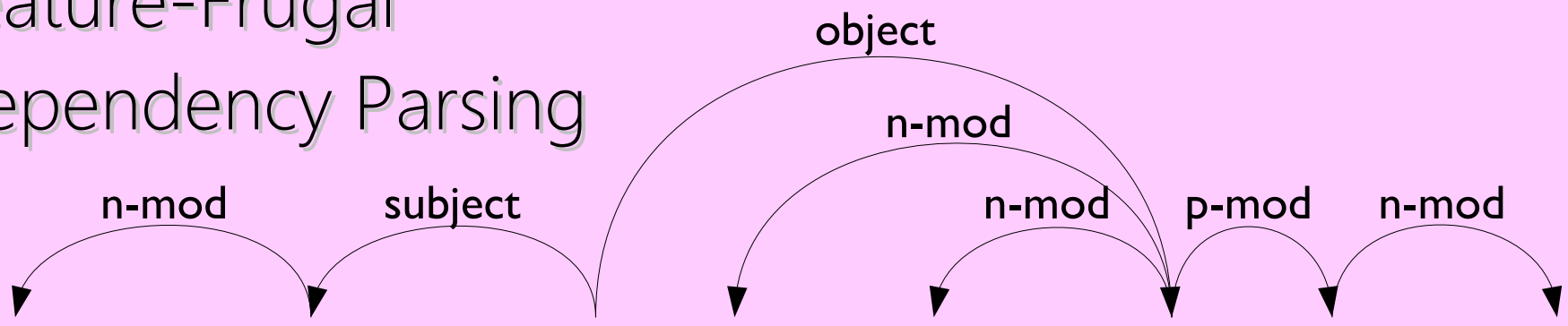


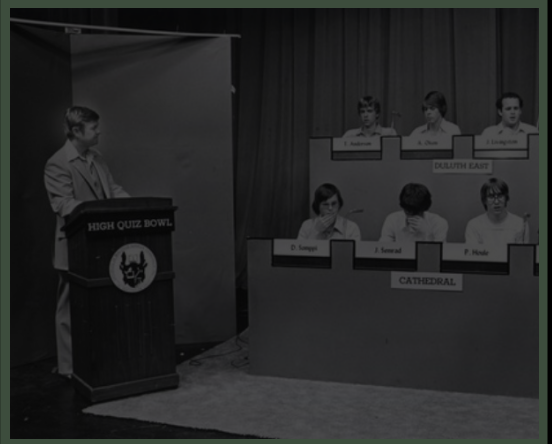
Learning to be *fast!*



Quality = tradeoff(accuracy, time)

Feature-Frugal Dependency Parsing

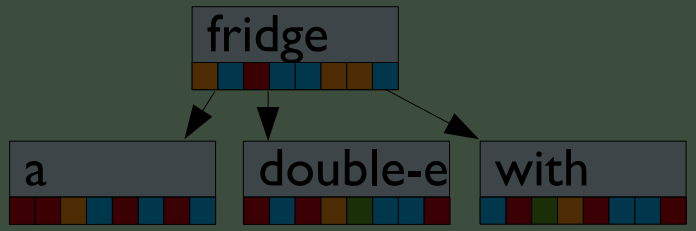




Quizbowl (Incremental Question Answering)

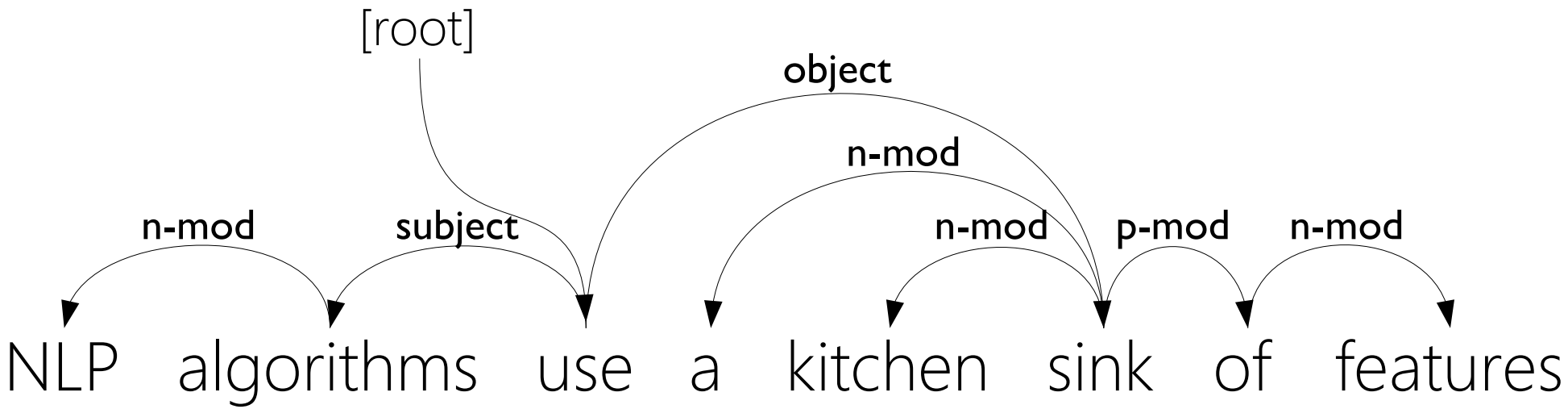
Ich bin mit dem Zug nach Ulm gefahren
I am with the train to Ulm traveled
I traveled by train to Ulm

Simultaneous Machine Interpretation

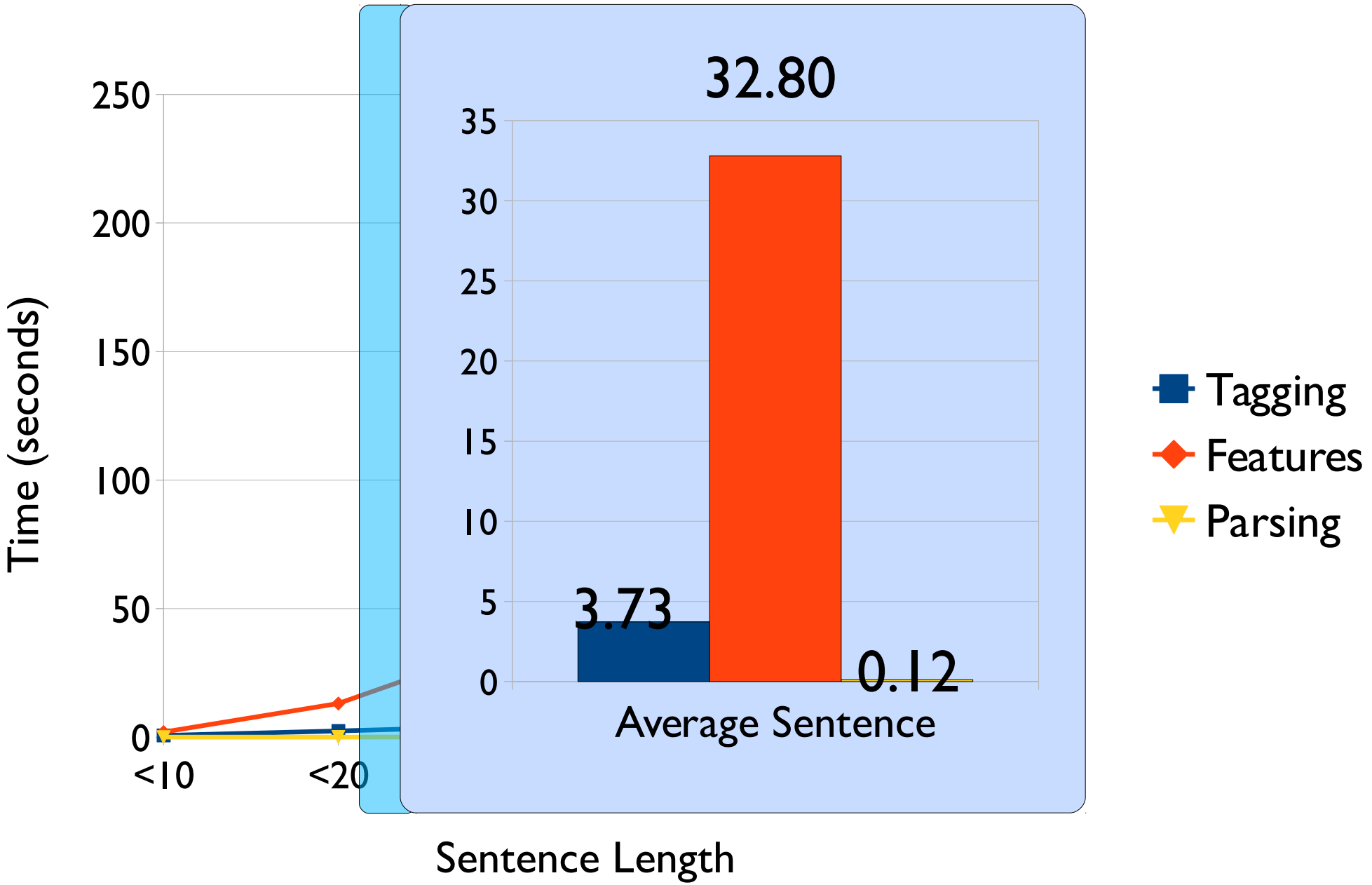


Outline

Dependency parsing

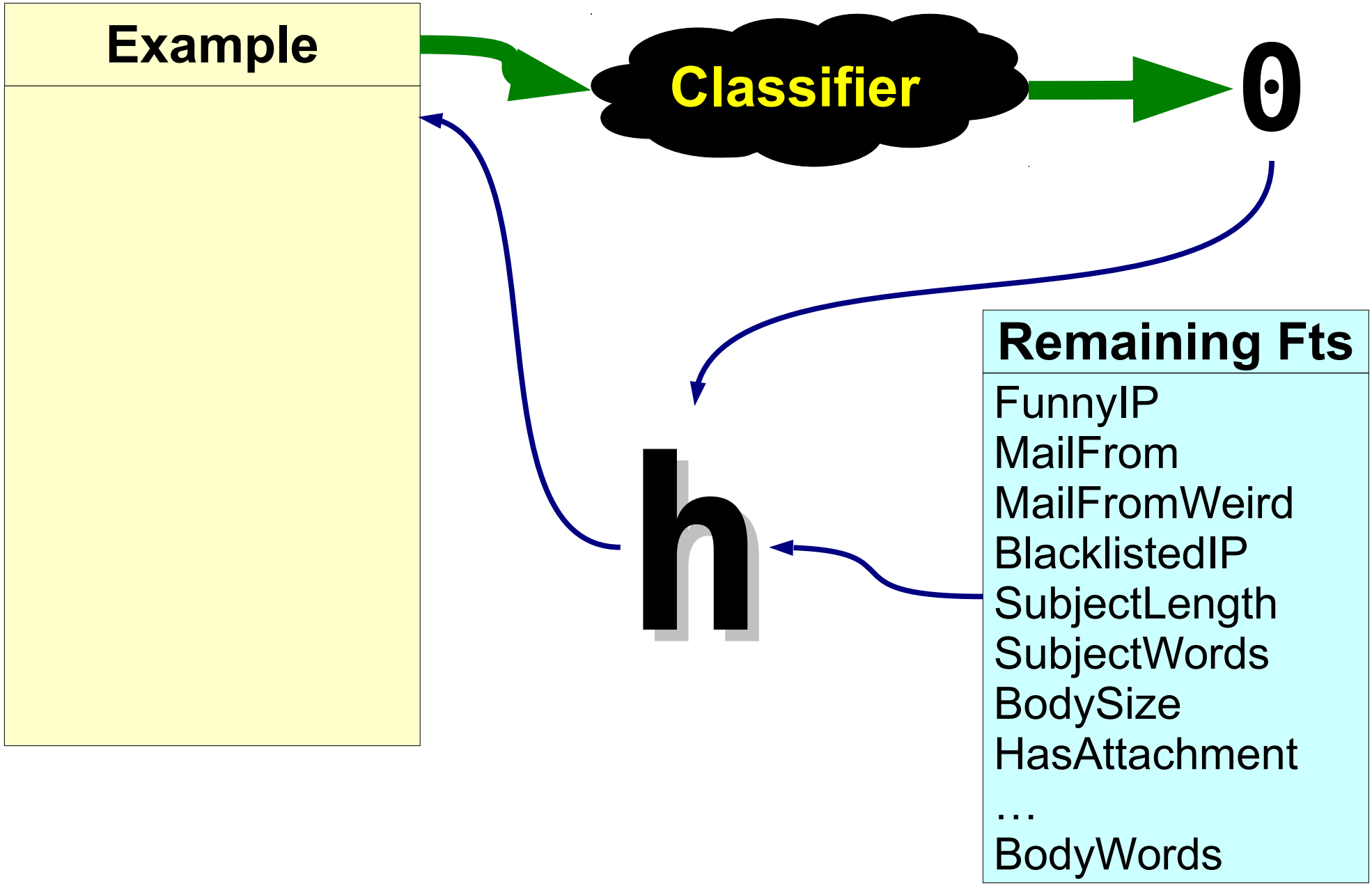


Case study: dependency parsing



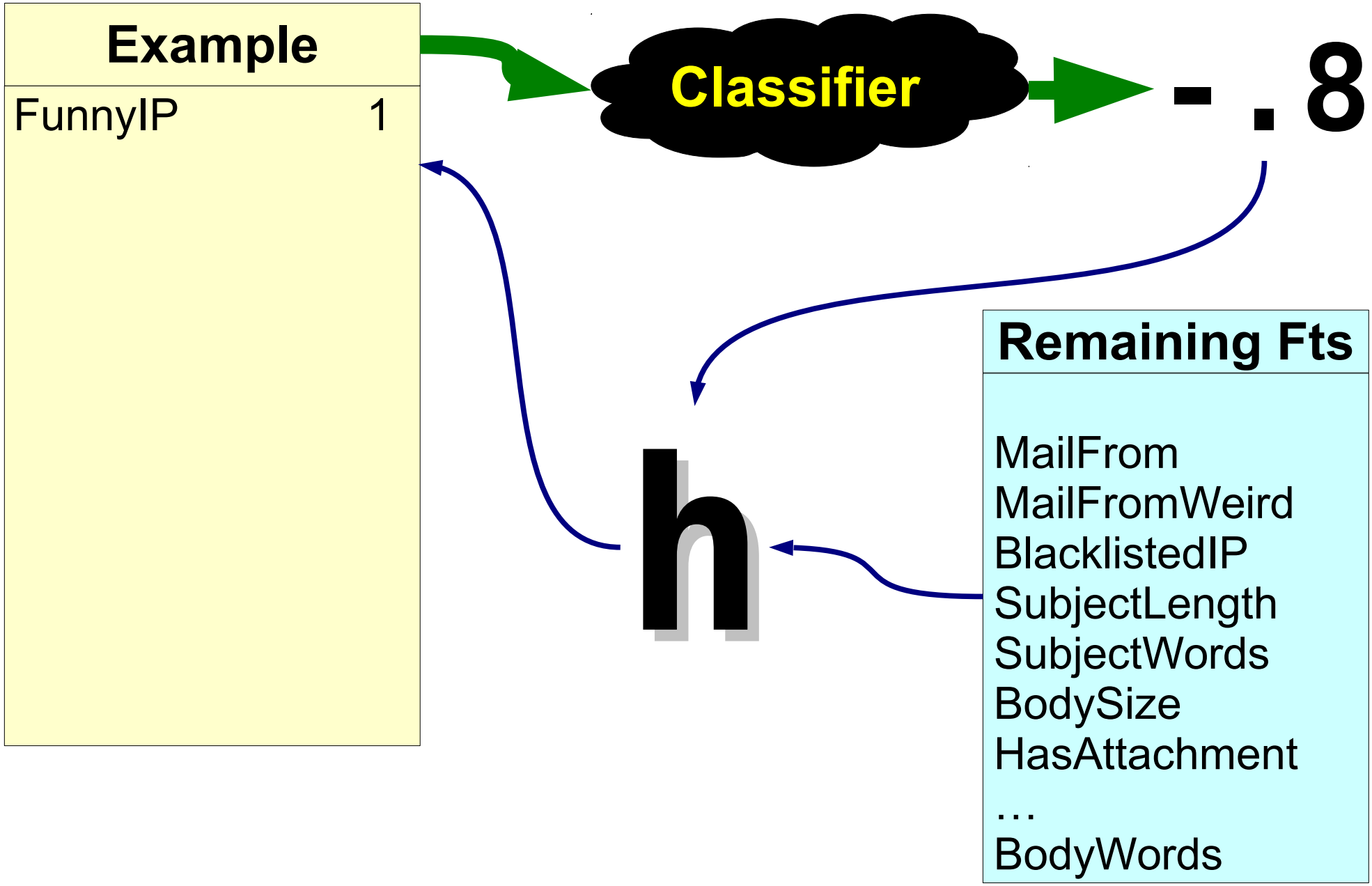
(He+Eisner+D, EMNLP 2013)

Dynamic feature selection

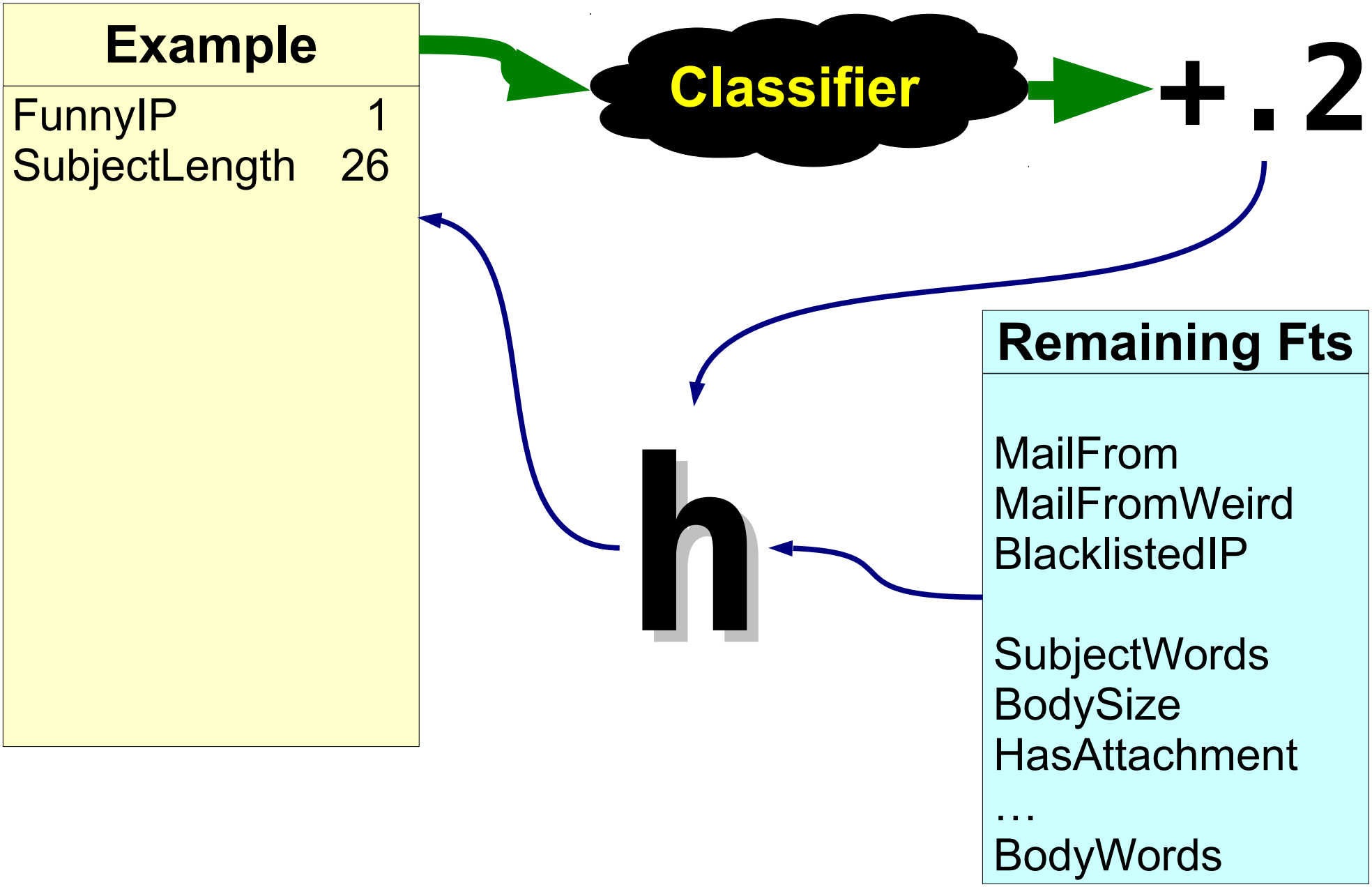


(He+Eisner+D, NIPS 2012)

Dynamic feature selection

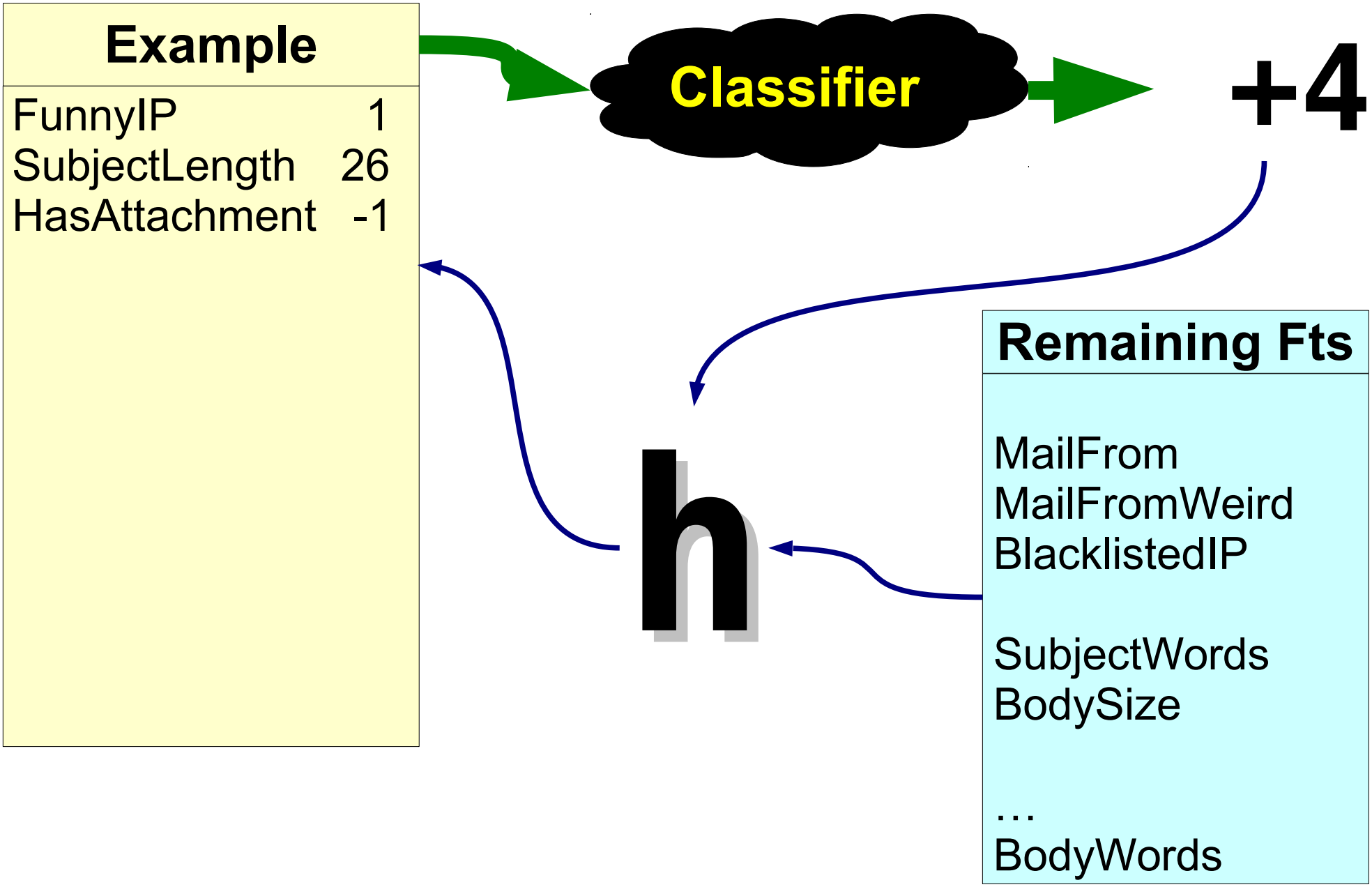


Dynamic feature selection



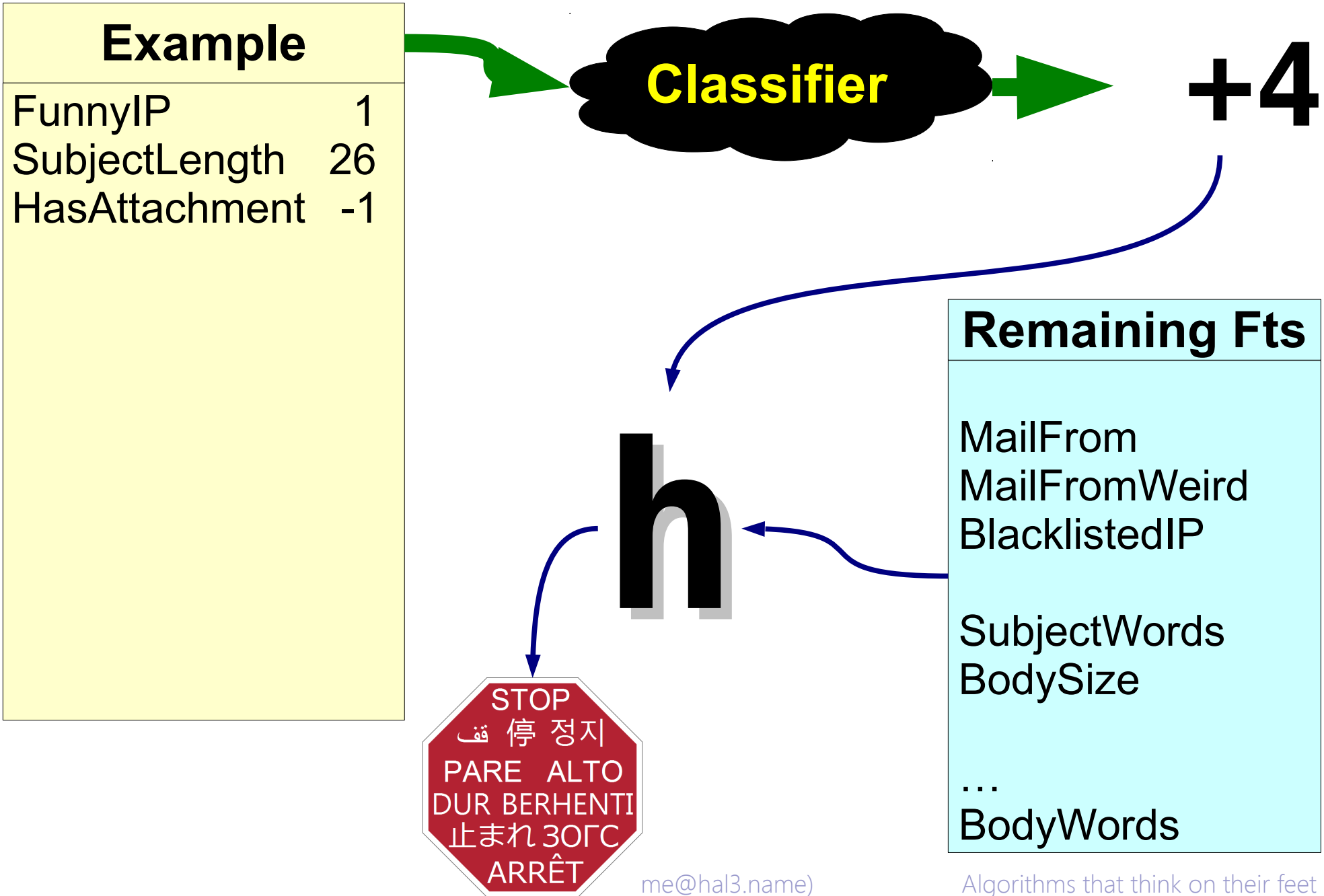
(He+Eisner+D, NIPS 2012)

Dynamic feature selection



(He+Eisner+D, NIPS 2012)

Dynamic feature selection



(He+Eisner+D, NIPS 2012)

DAgger: Dataset Aggregation

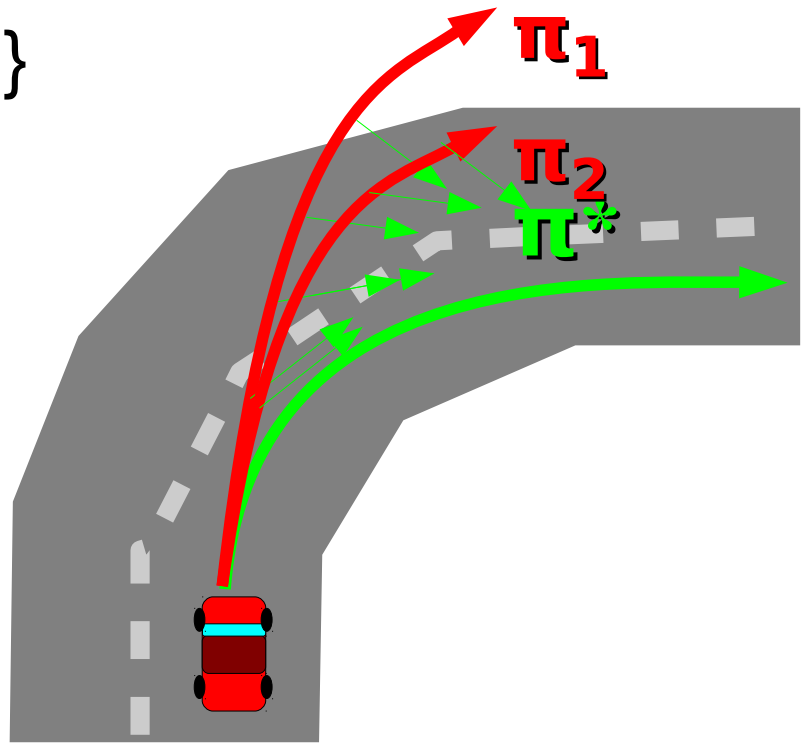
- Collect trajectories from expert π^*
- Dataset $\mathbf{D_0} = \{ (s, \pi^*(s)) \mid s \sim \pi^* \}$
- Train π_1 on $\mathbf{D_0}$
- Collect new trajectories from π_1
  - But let the *expert* steer!
- Dataset $\mathbf{D_1} = \{ (s, \pi^*(s)) \mid s \sim \pi_1 \}$
- Train π_2 on $\mathbf{D_0} \cup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

- 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)

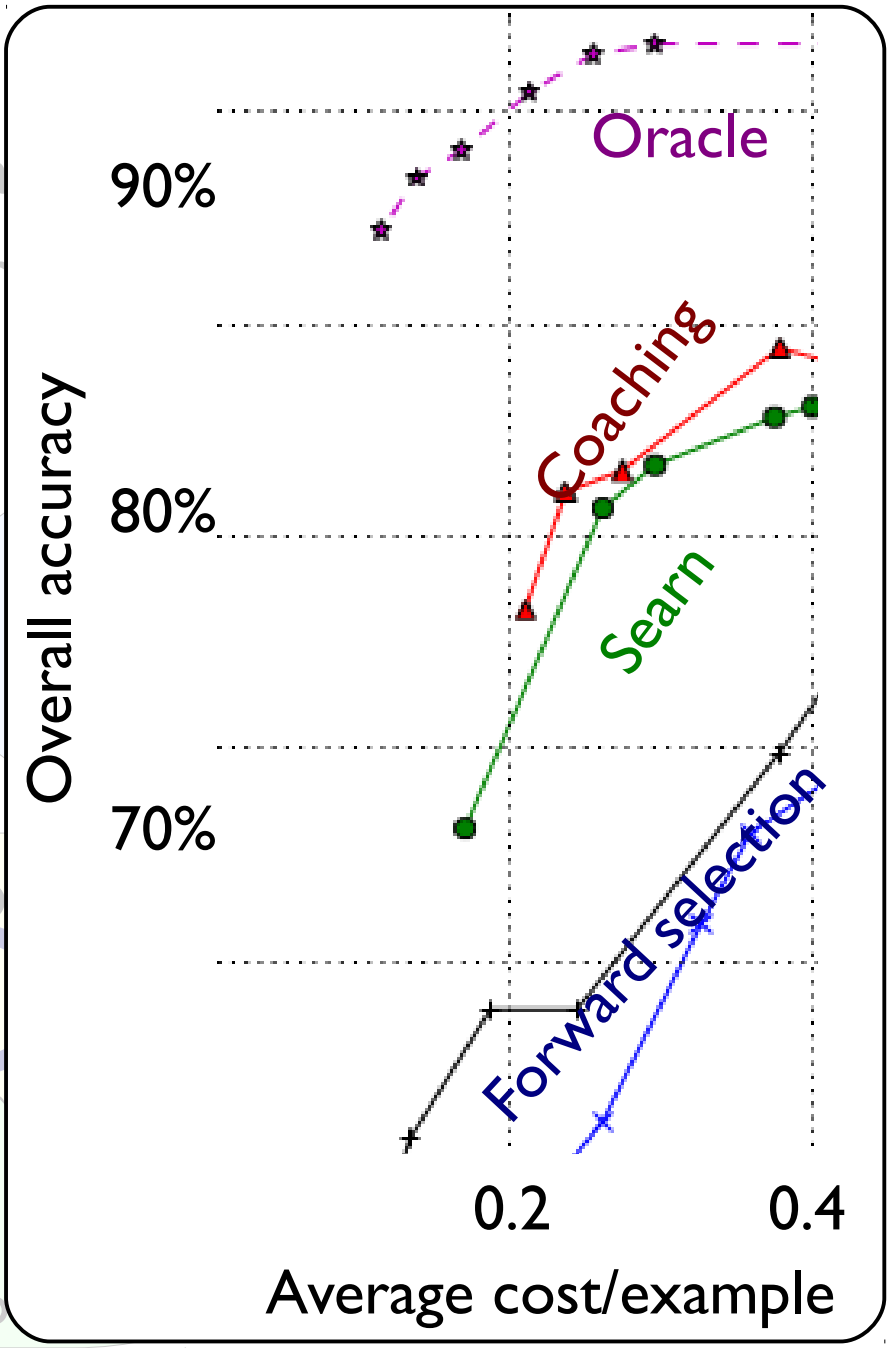
The oracle too good!

- The oracle *knows the label*
  - Picks feature with highest y*va
  - Ends after selecting one featur
- Coach says how to improve,

Psst! You should choose
argmin **E**[l(a)]
a

If N = T lo
$L(\pi_n) < T \epsilon_N$
for son

Provably smaller
oracle's epsilo



(He+Eisner+D, NIPS 2012)

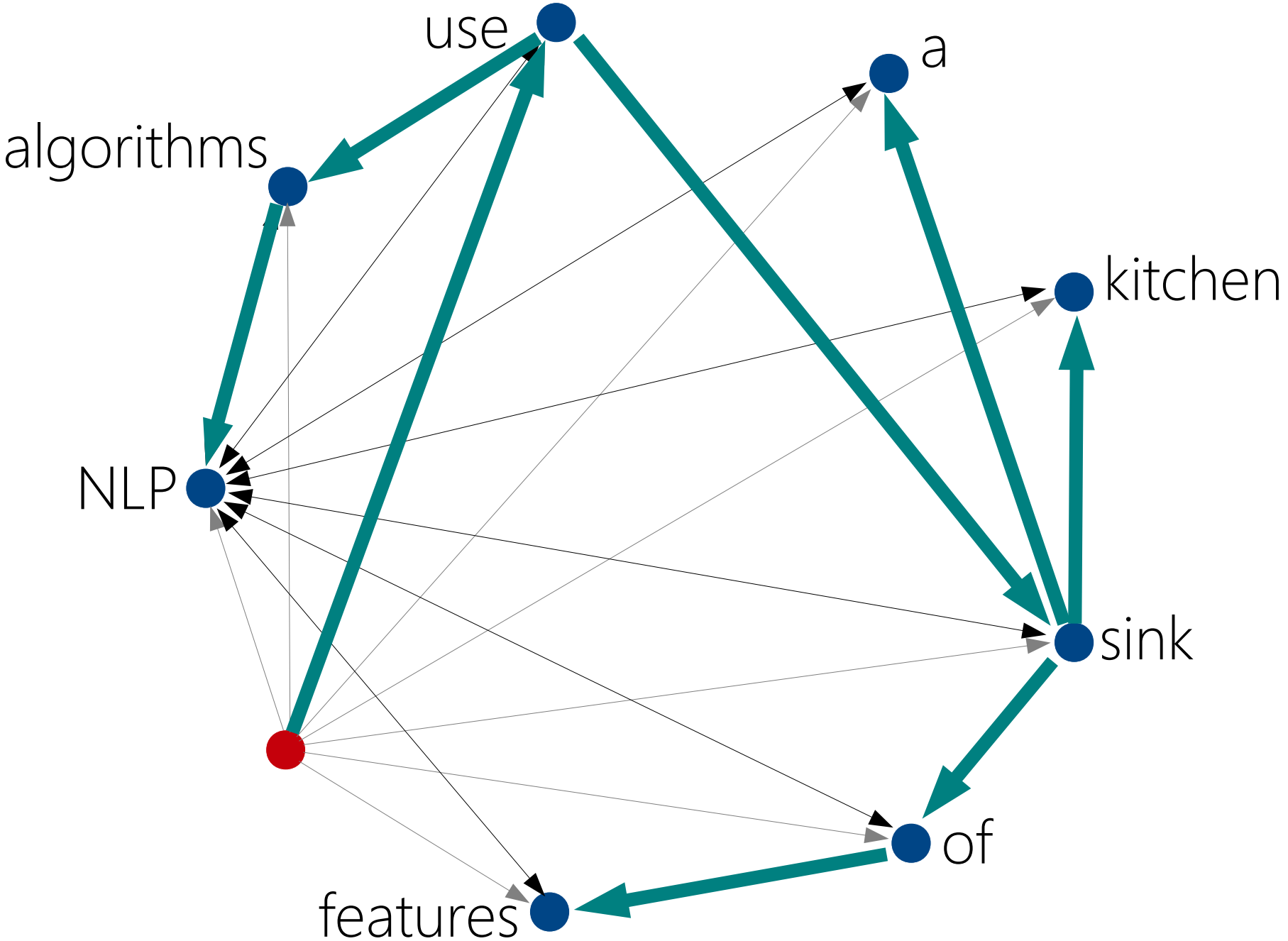
Dependency parsing

[root]

NLP algorithms use a kitchen sink of features

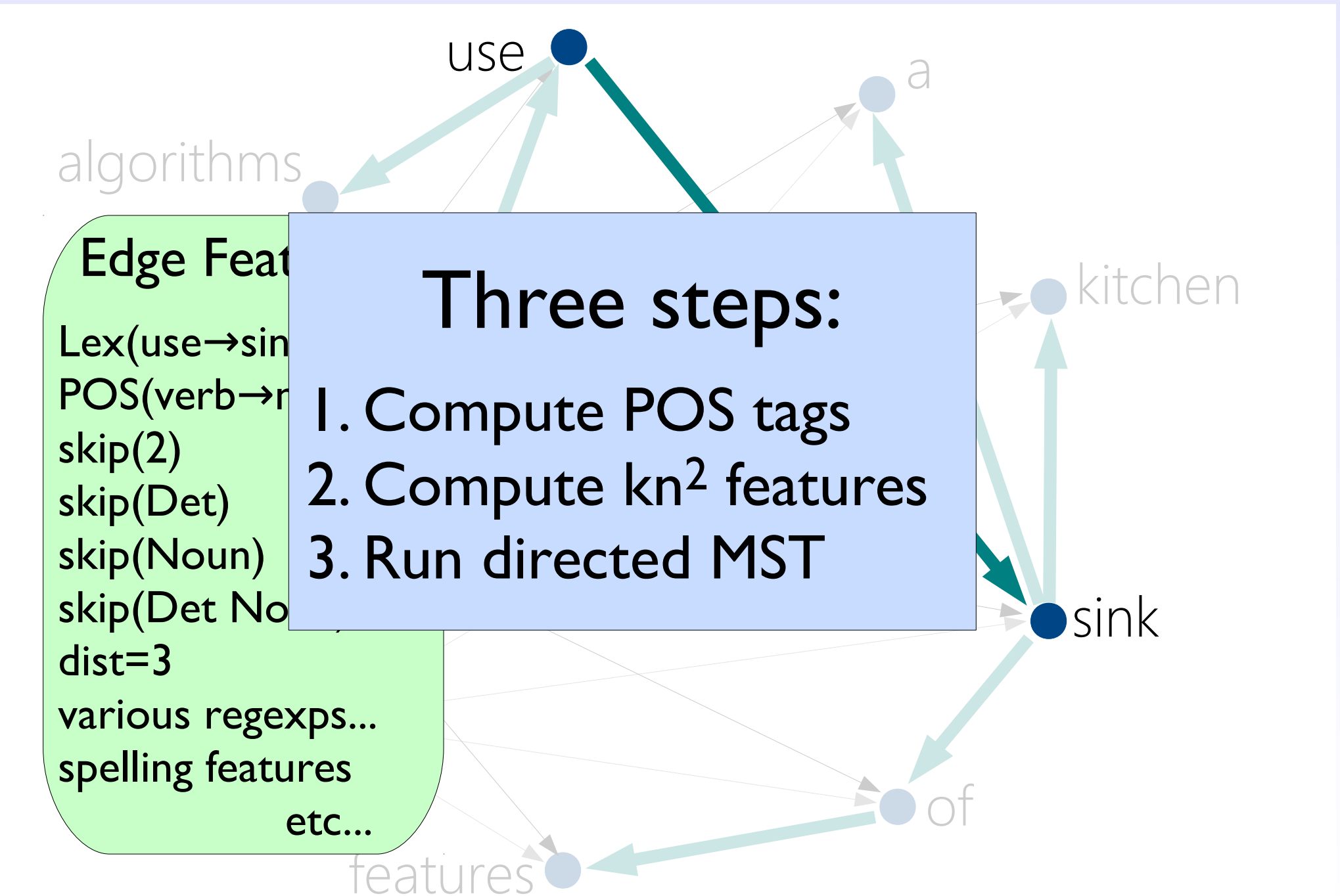
(He+Eisner+D, EMNLP 2013)

Dependency parsing



(He+Eisner+D, EMNLP 2013)

Dependency parsing

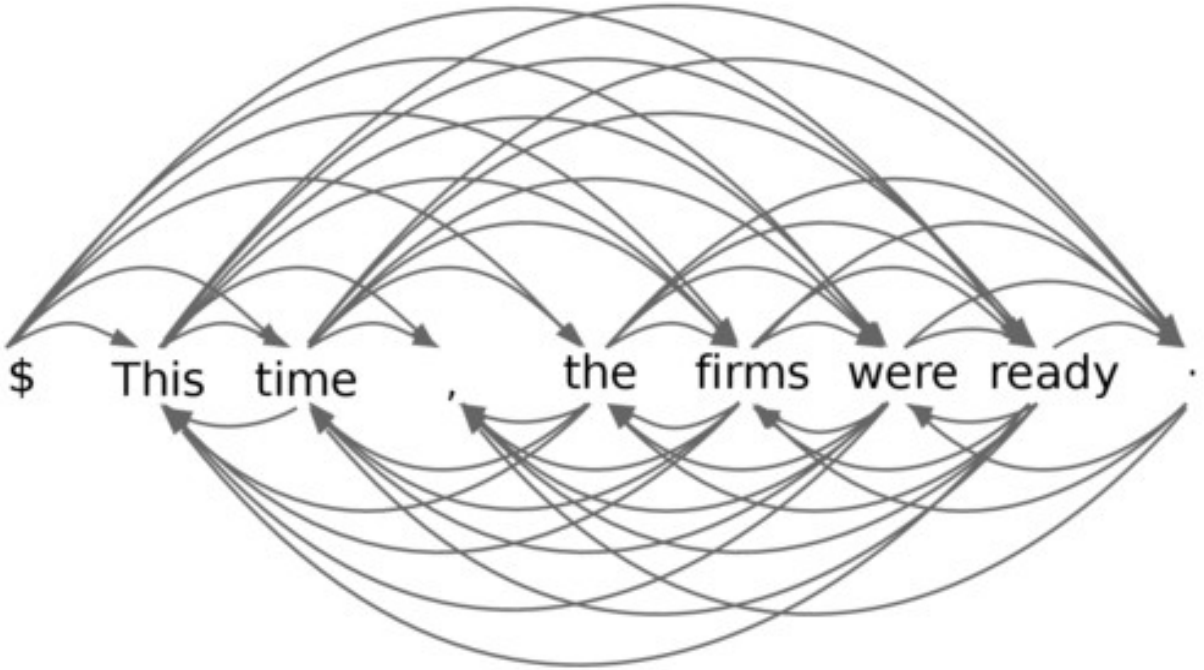


(He+Eisner+D, EMNLP 2013)

The system we learn to control

+ first feature group 5 features per gray edge

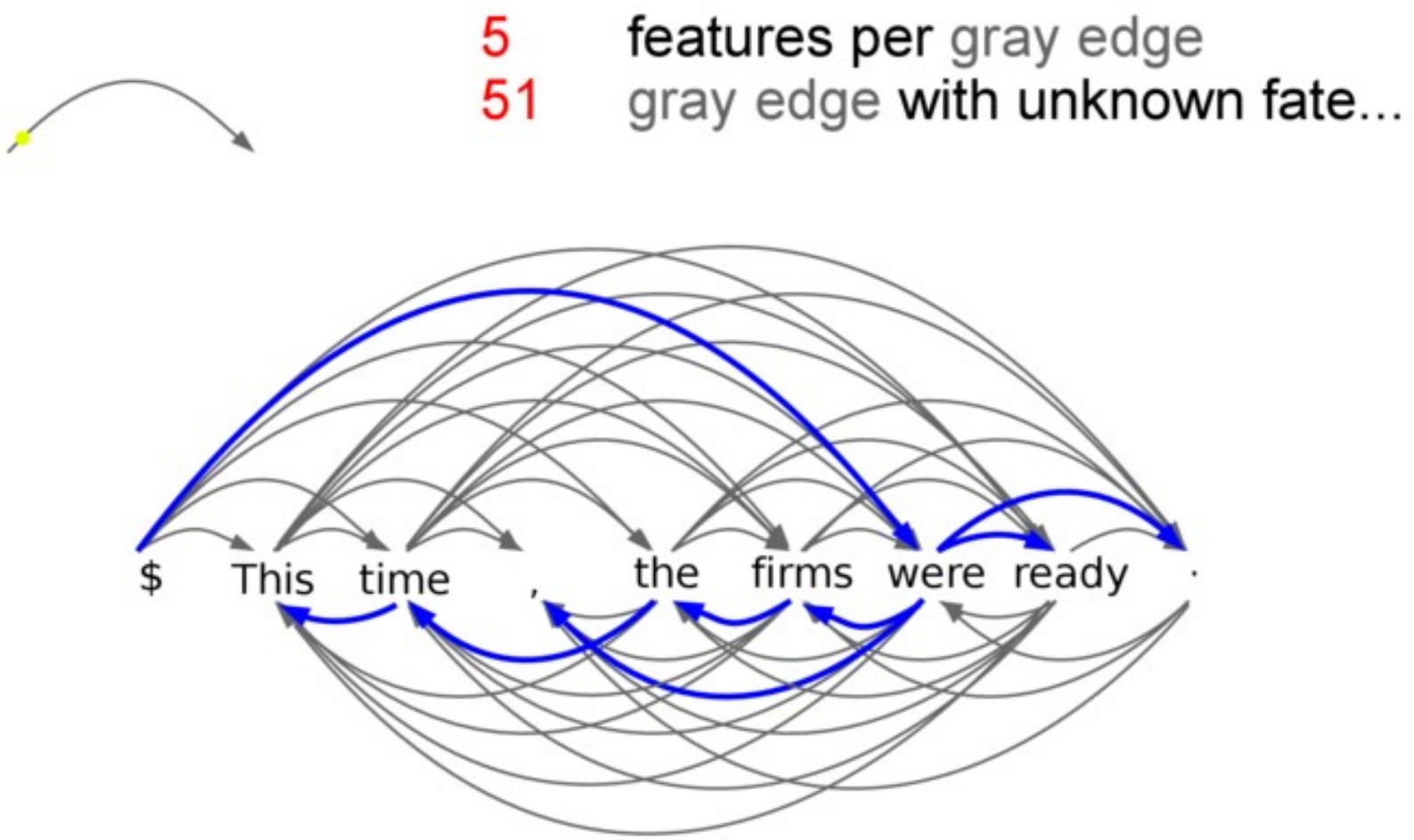
51 gray edge with unknown fate...



Undetermined edge
Current 1-best tree
Winner edge
Loser edge

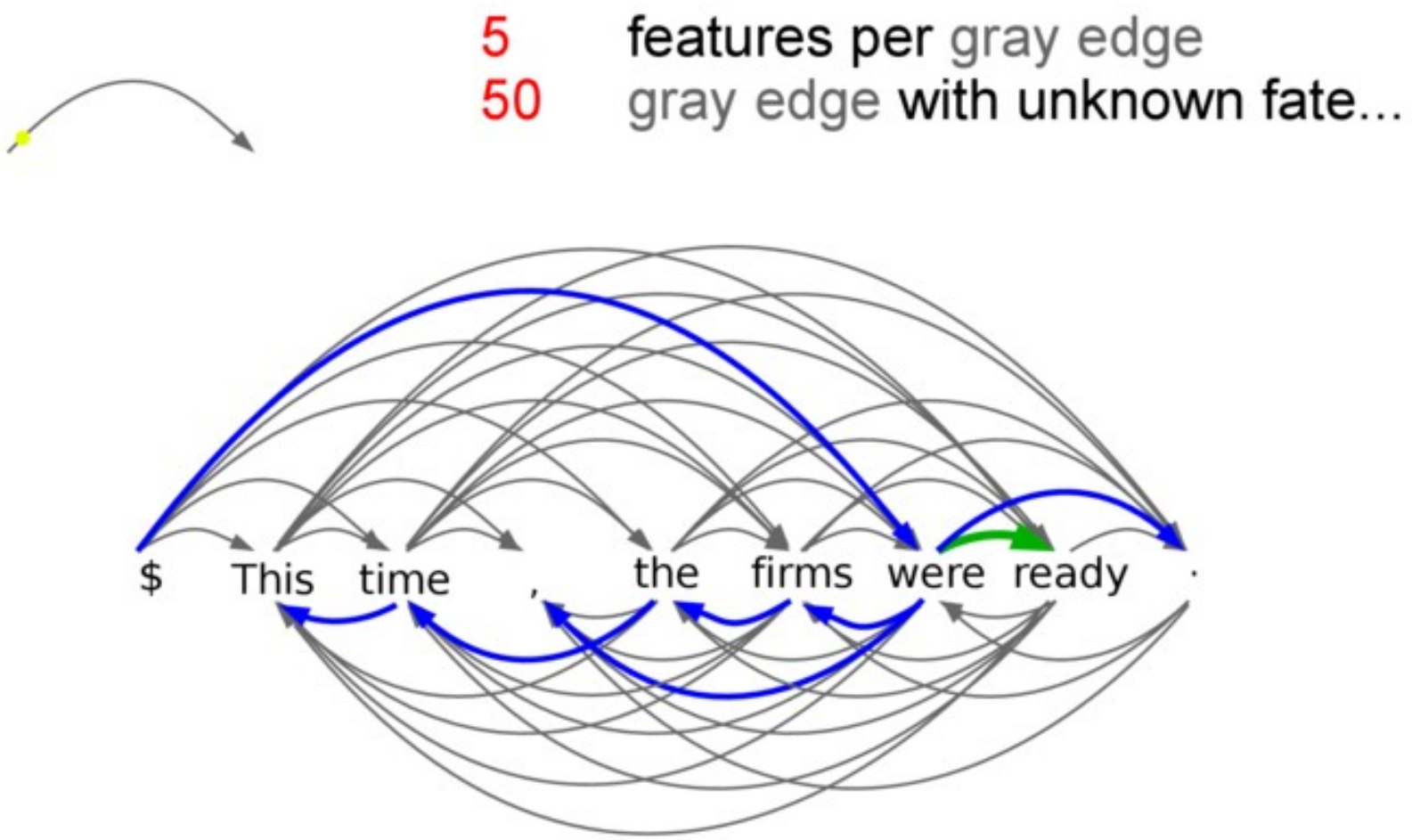
(He+Eisner+D, EMNLP 2013)

The system we learn to control



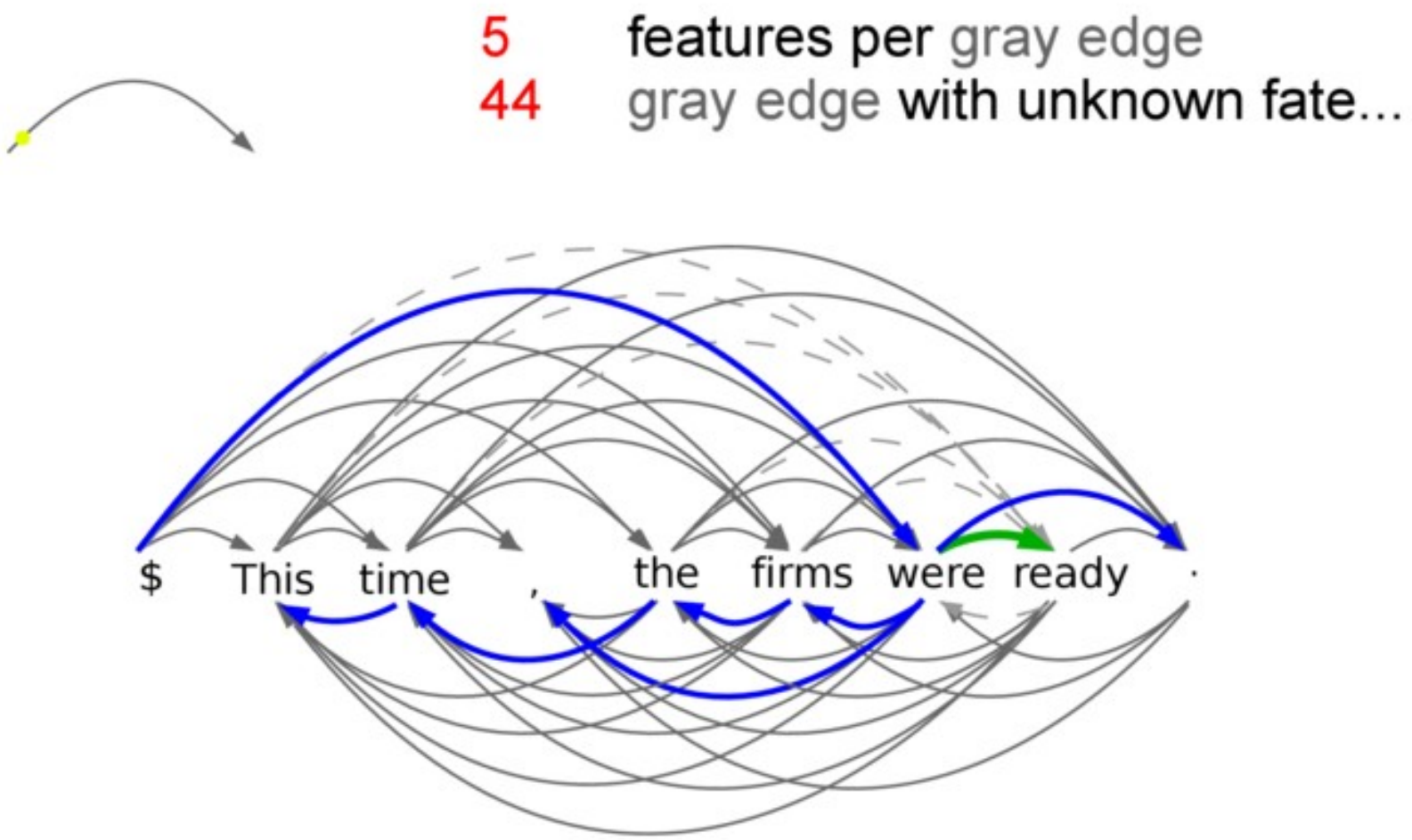
(He+Eisner+D, EMNLP 2013)

The system we learn to control



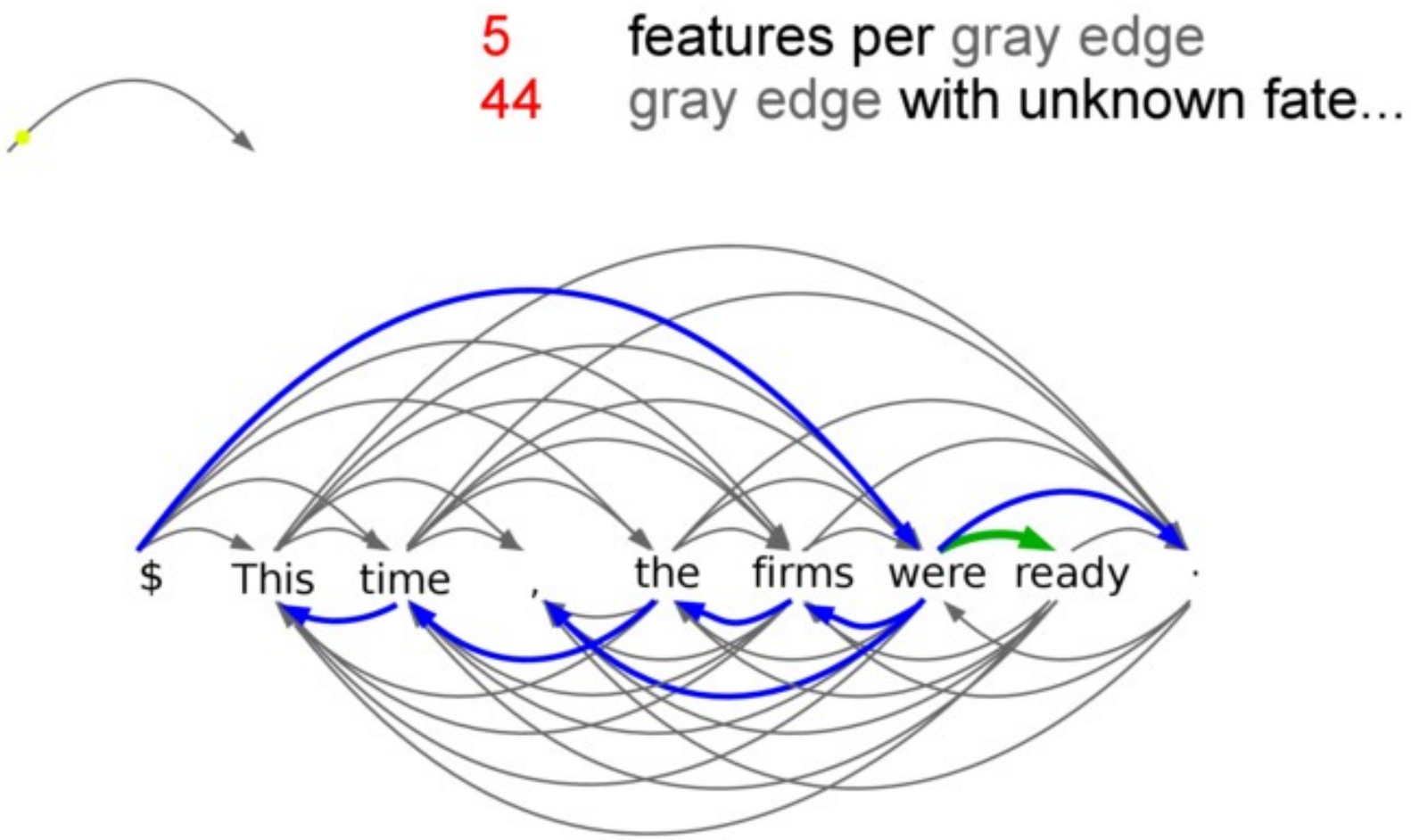
(He+Eisner+D, EMNLP 2013)

The system we learn to control



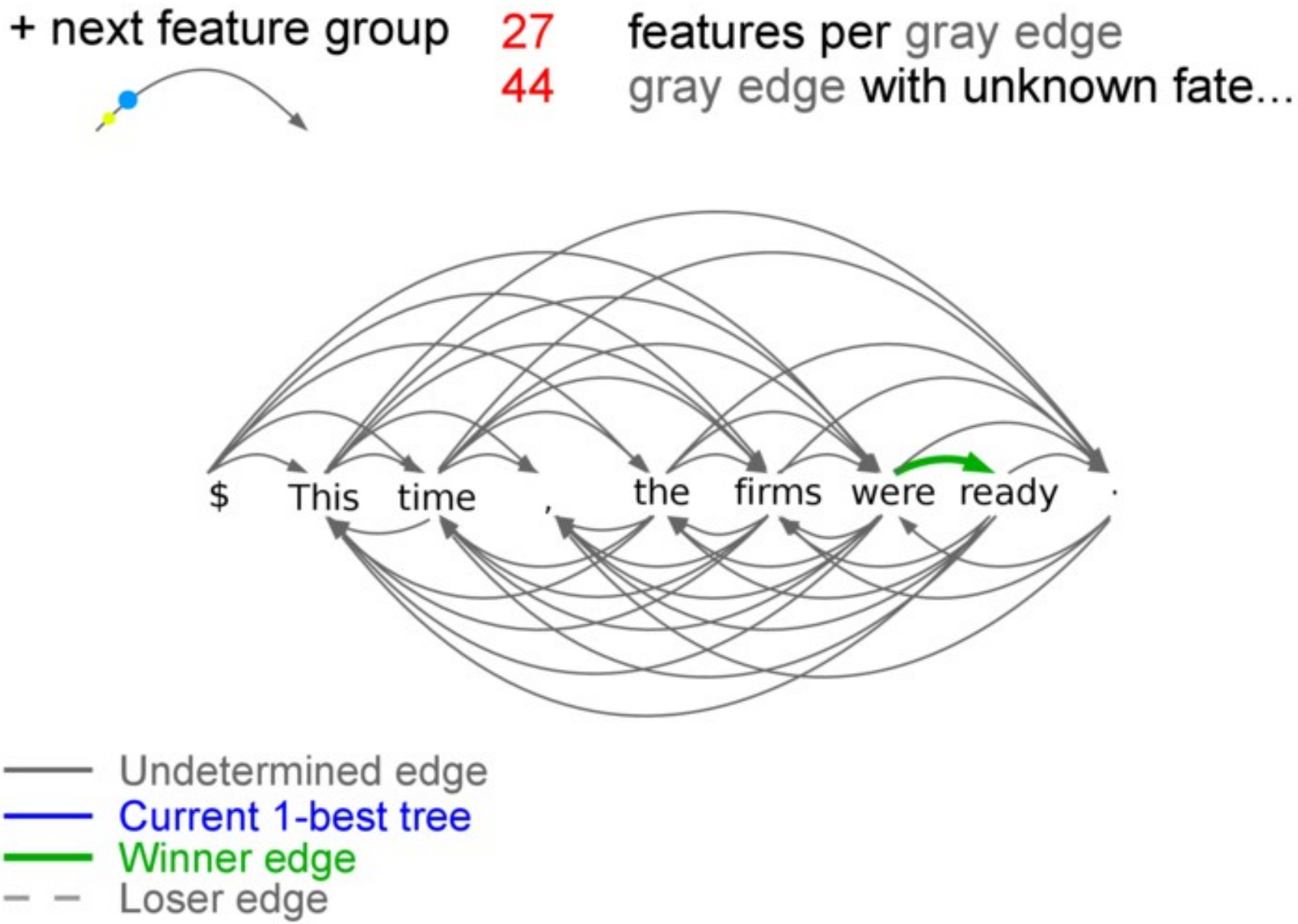
(He+Eisner+D, EMNLP 2013)

The system we learn to control



(He+Eisner+D, EMNLP 2013)

The system we learn to control



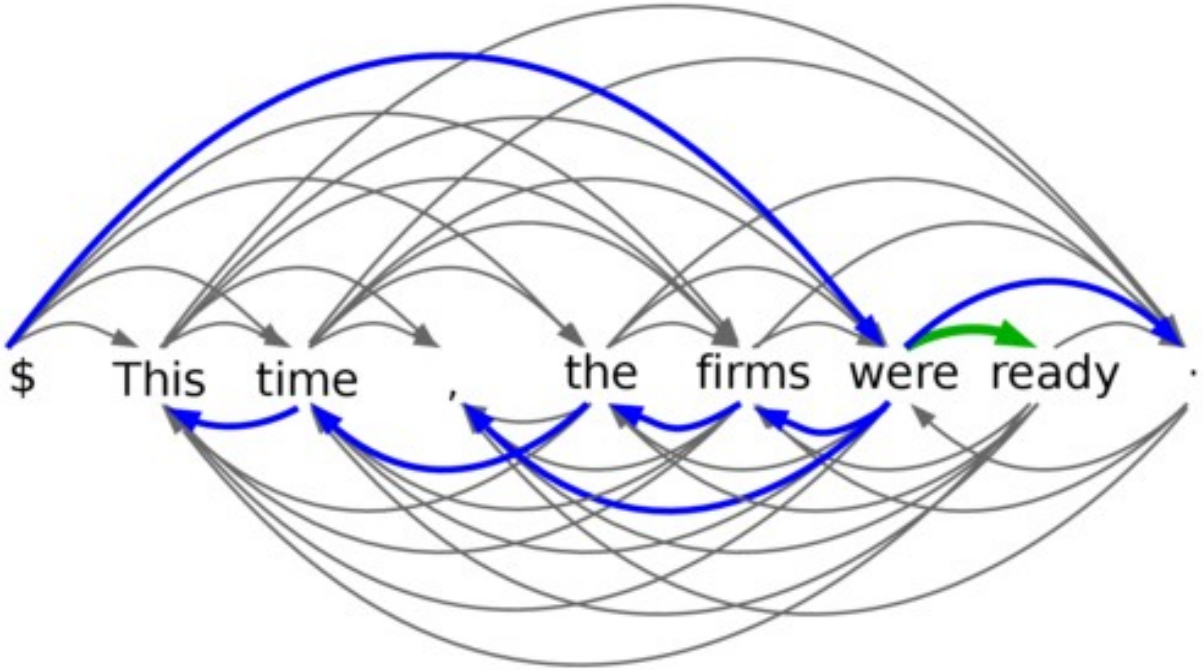
(He+Eisner+D, EMNLP 2013)

The system we learn to control

+ next feature group

27 features per gray edge

44 gray edge with unknown fate...

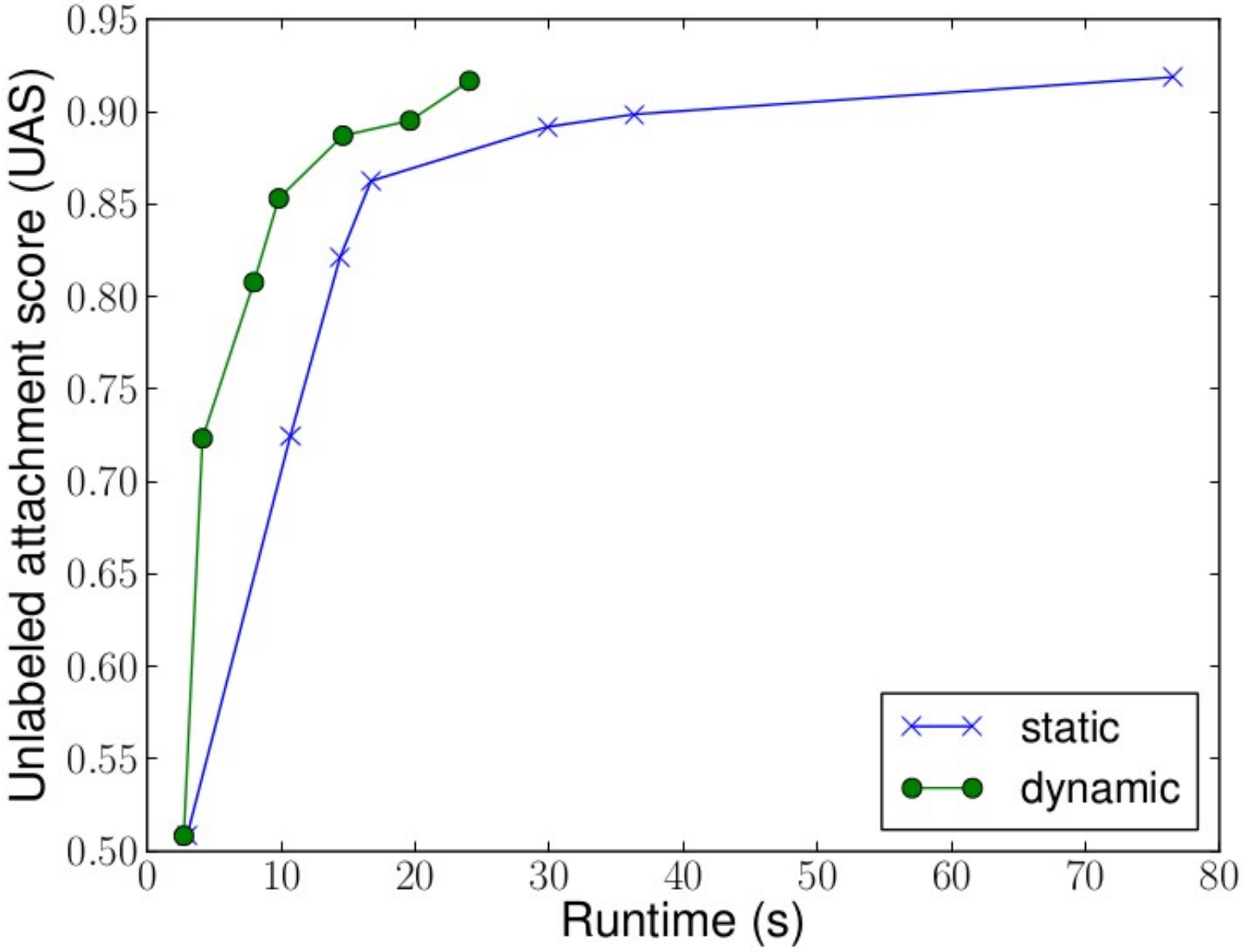


Undetermined edge
Current 1-best tree
Winner edge
Loser edge

Non-projective decoding

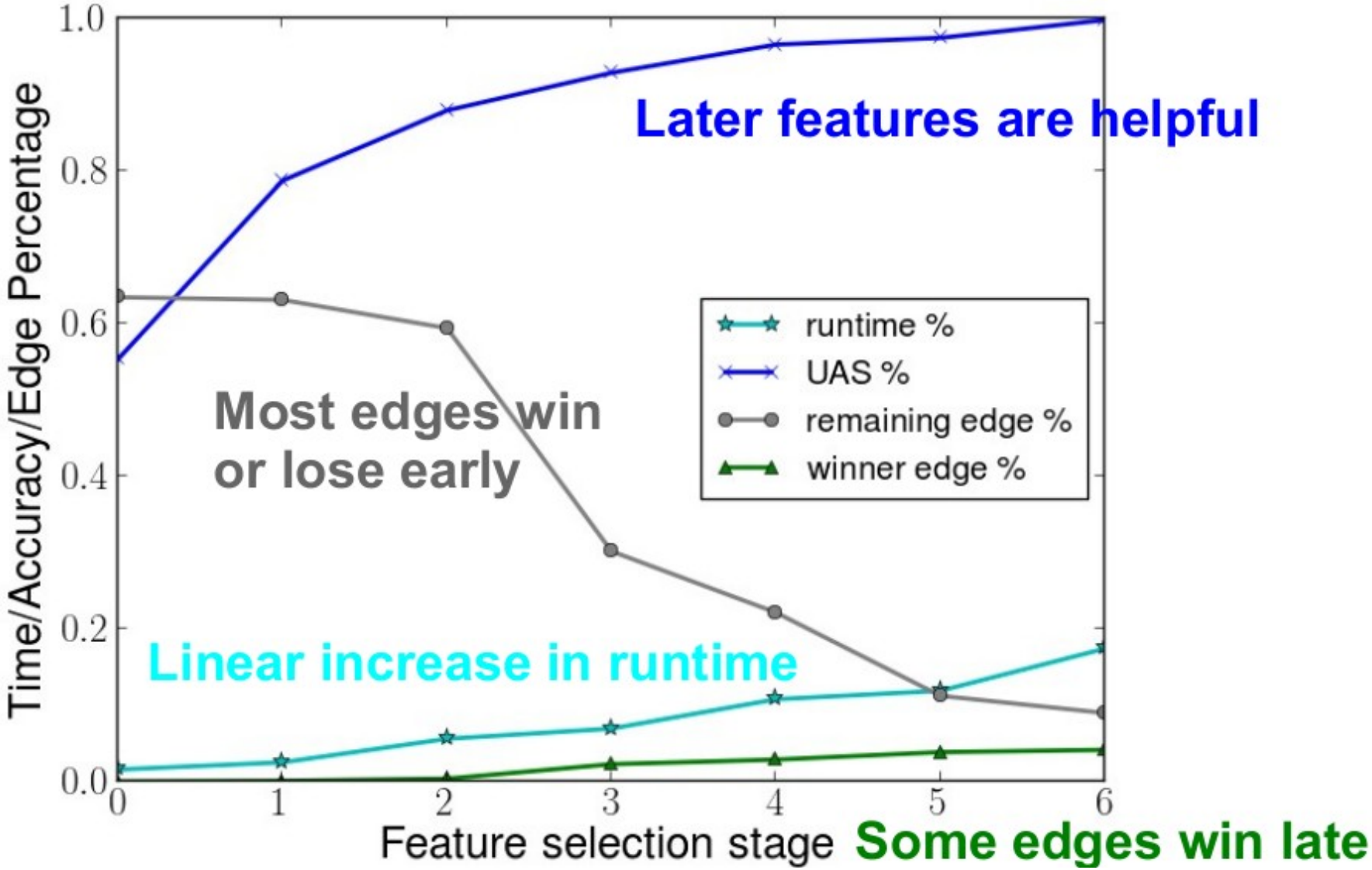
(He+Eisner+D, EMNLP 2013)

Static versus dynamic feature selection



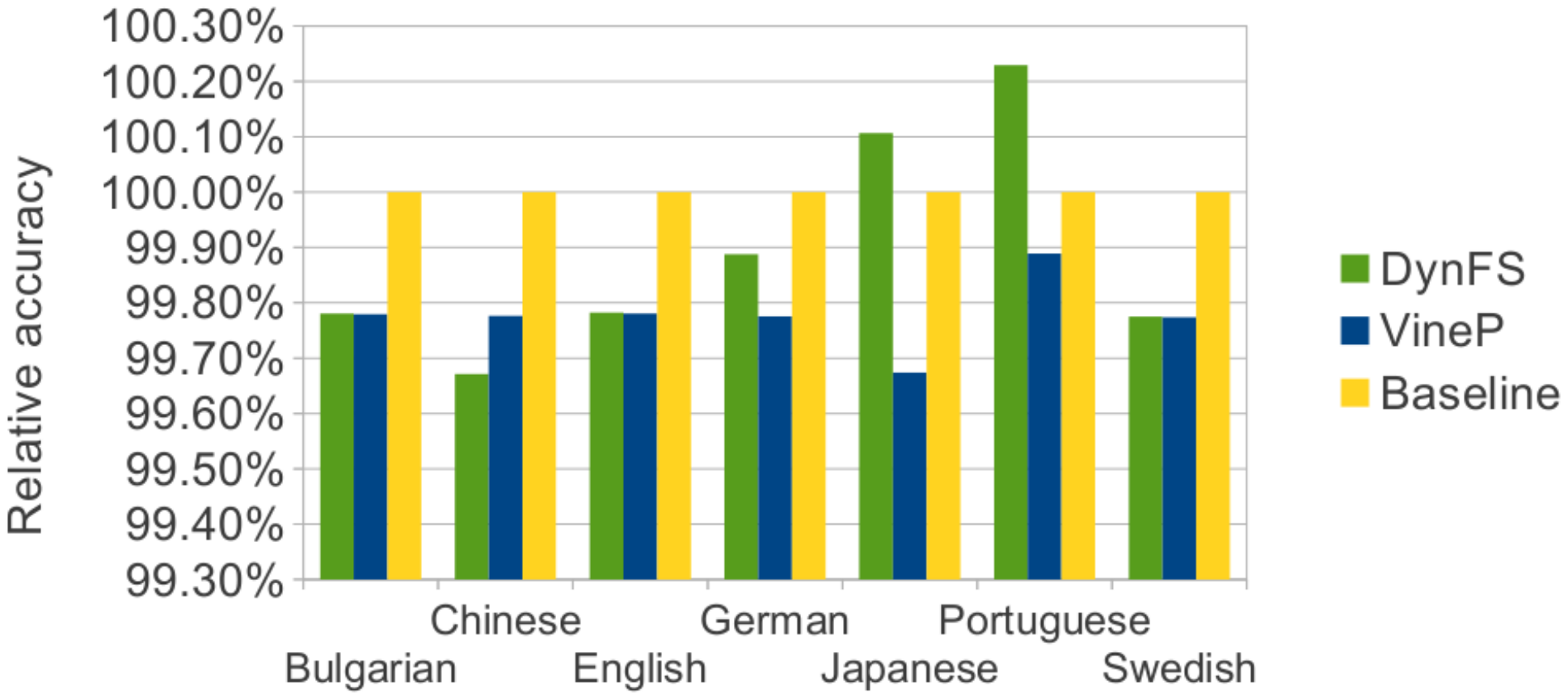
(He+Eisner+D, EMNLP 2013)

Looking inside the box…



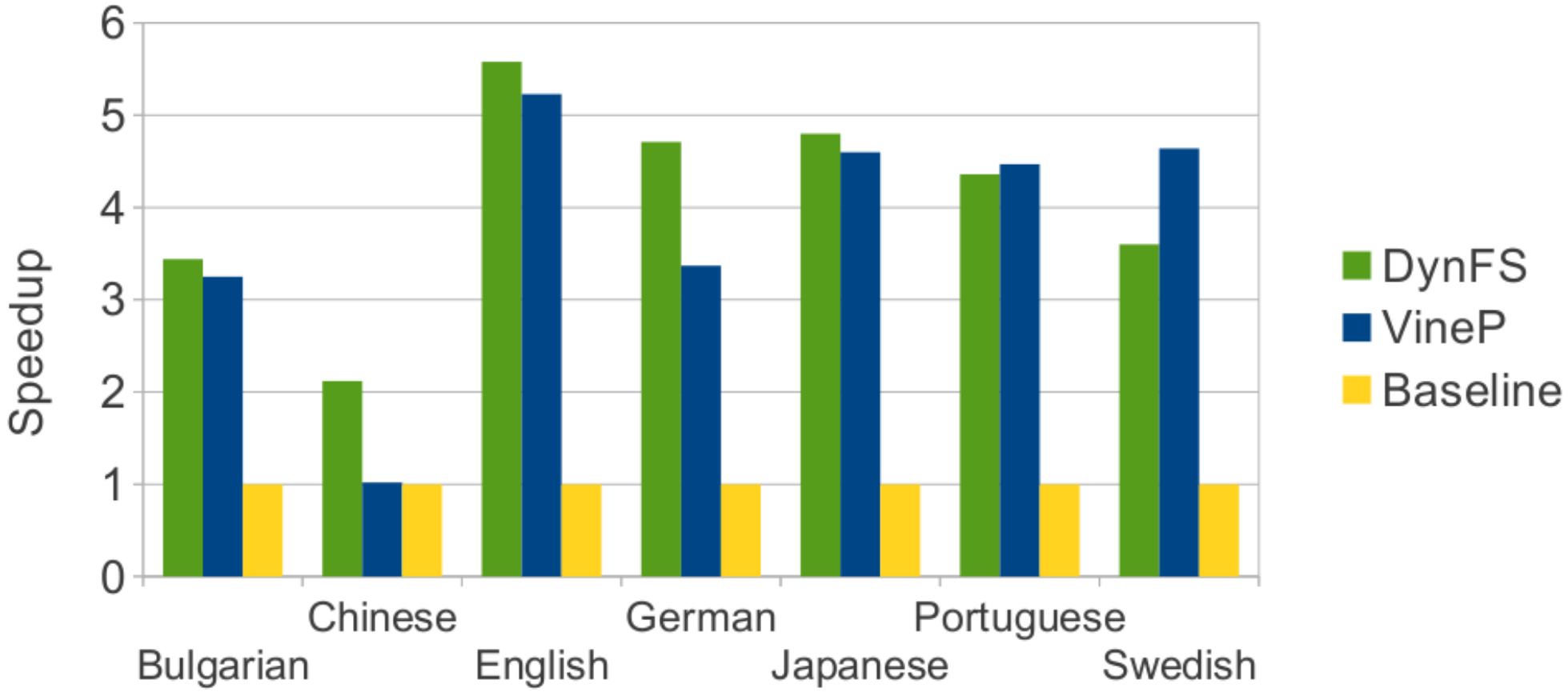
(He+Eisner+D, EMNLP 2013)

Accuracy is (essentially) uneffected



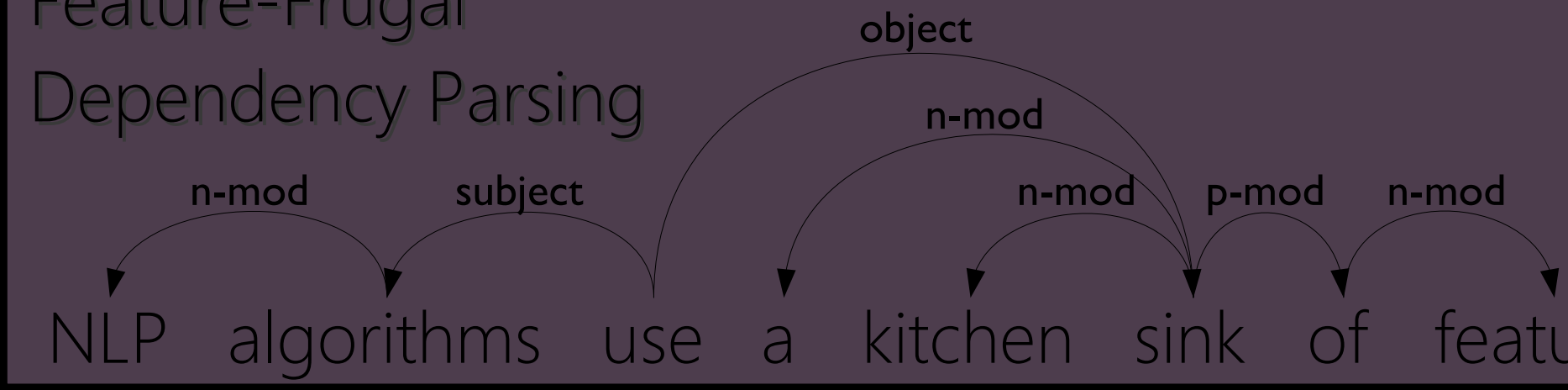
(He+Eisner+D, EMNLP 2013)

...but we get a lot faster



(He+Eisner+D, EMNLP 2013)

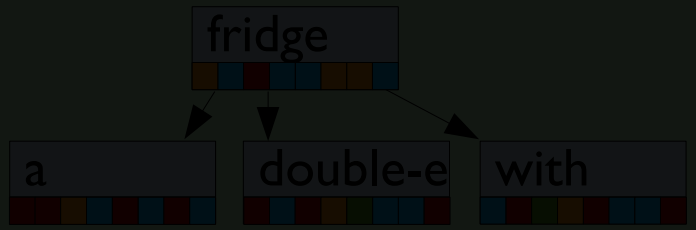
Feature-Frugal Dependency Parsing



Quizbowl (Incremental Question Answering)

Ich bin mit dem Zug nach Ulm gefahren
I am with the train to Ulm traveled
I traveled by train to Ulm

Simultaneous Machine Interpretation



Outline

Simultaneous (machine) interpretation

Nuremburg Trials

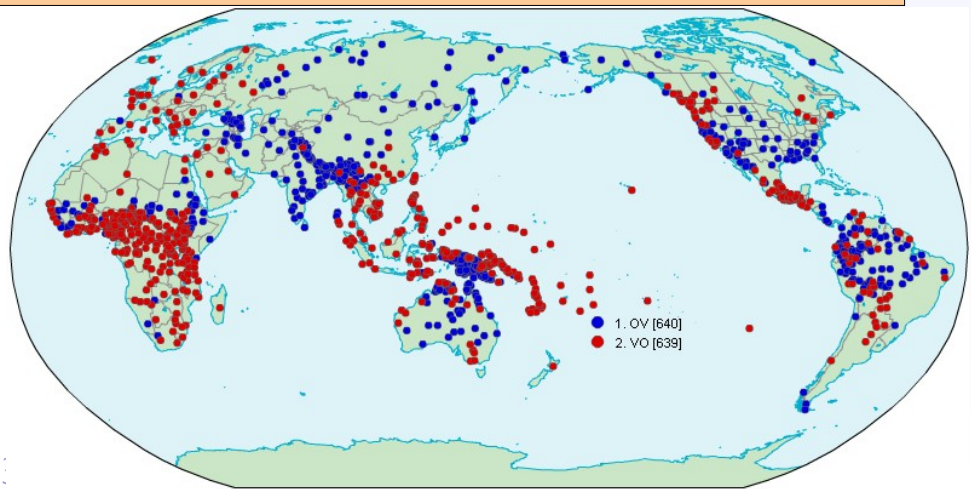
- Dozens of defendants
- Judges from four nations (three languages)
- Status quo: speak, then translate
- After Nuremberg, simultaneous translations became the norm
- Long wait → bad conversation

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:

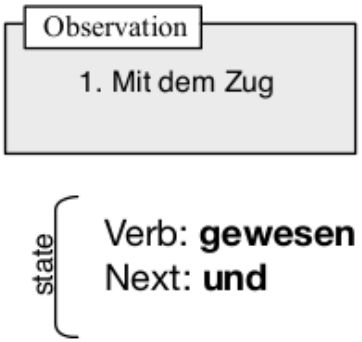
Ich	bin	mit	dem	Zug	nach	Ulm	gefahren
I	am	with	the	train	to	Ulm	traveled
I	*(……waiting……)*						traveled by train to Ulm



Model for interpretation decisions

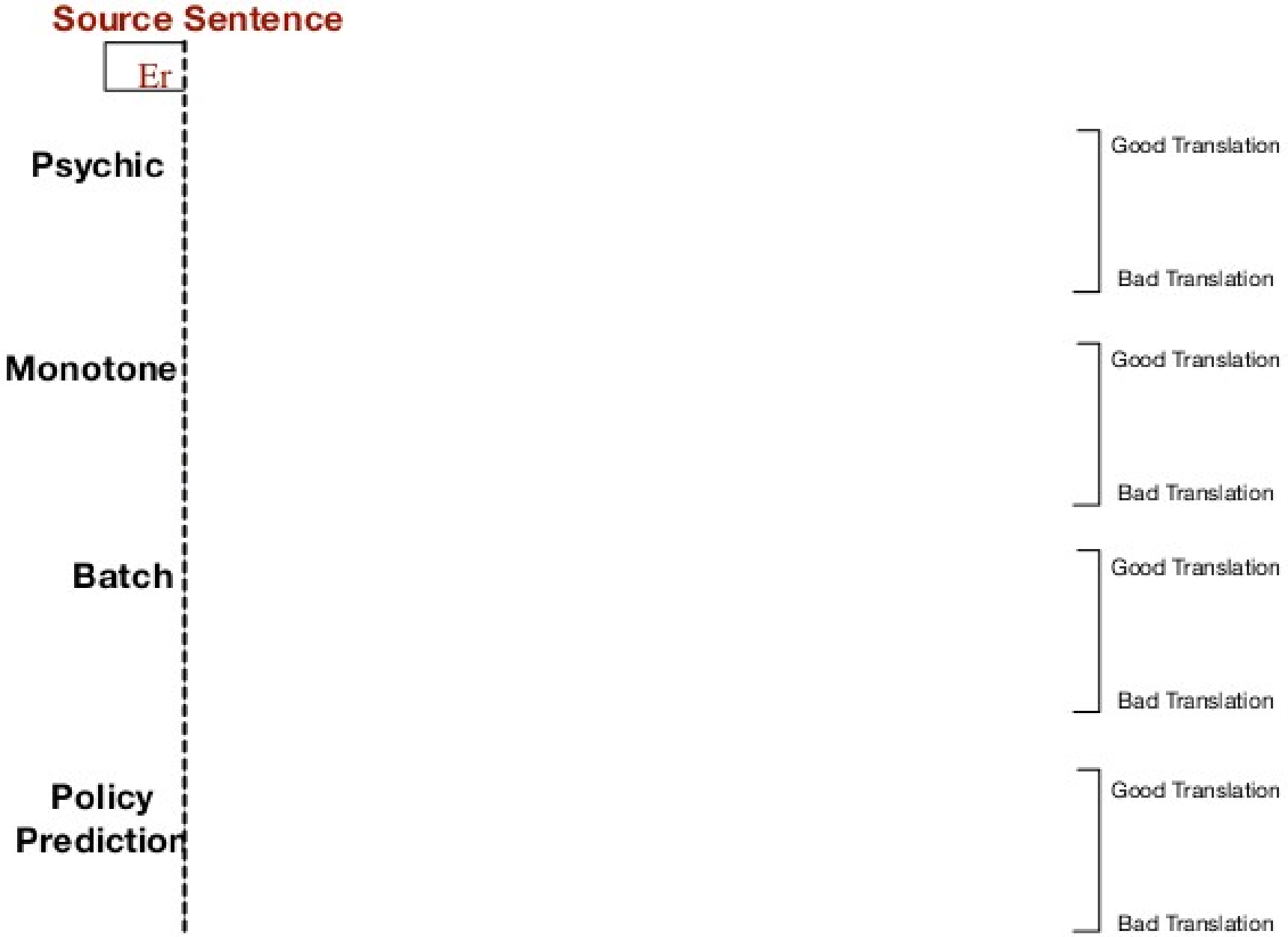
- We have a set of actions (predict / translate)
  - Wait
  - Predict clause-verb
  - Predict next word
  - Commit ("speak")
- In a changing environment (state)
  - The words we've seen so far
  - Our models' internal predictions
- With a well defined oracle

Example of interpretation trajectory

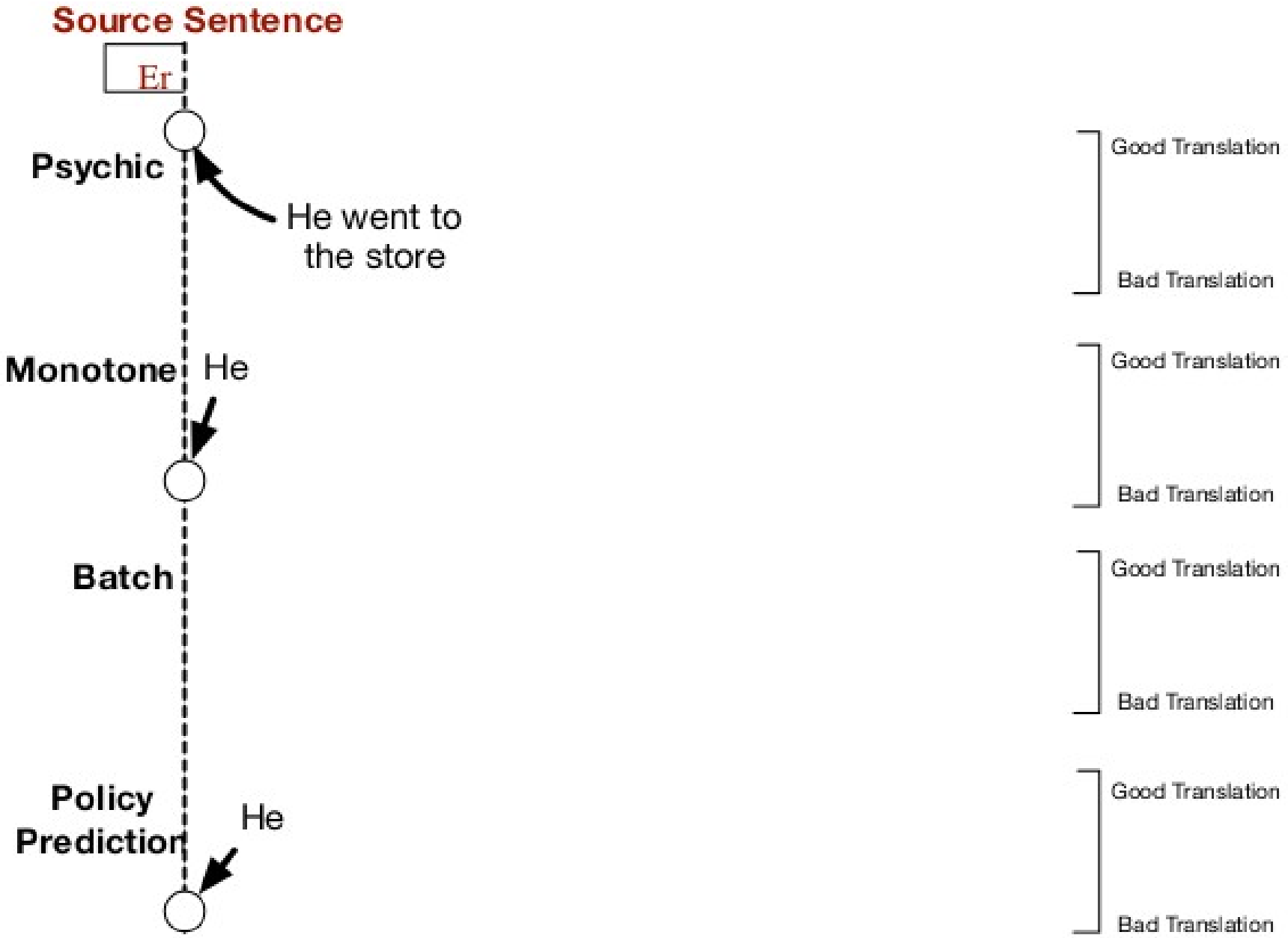


Ich	bin	mit	dem	Zug	nach	Ulm	gefahren
I	am	with	the	train	to	Ulm	traveled
I	*(……waiting……)*						traveled by train to Ulm

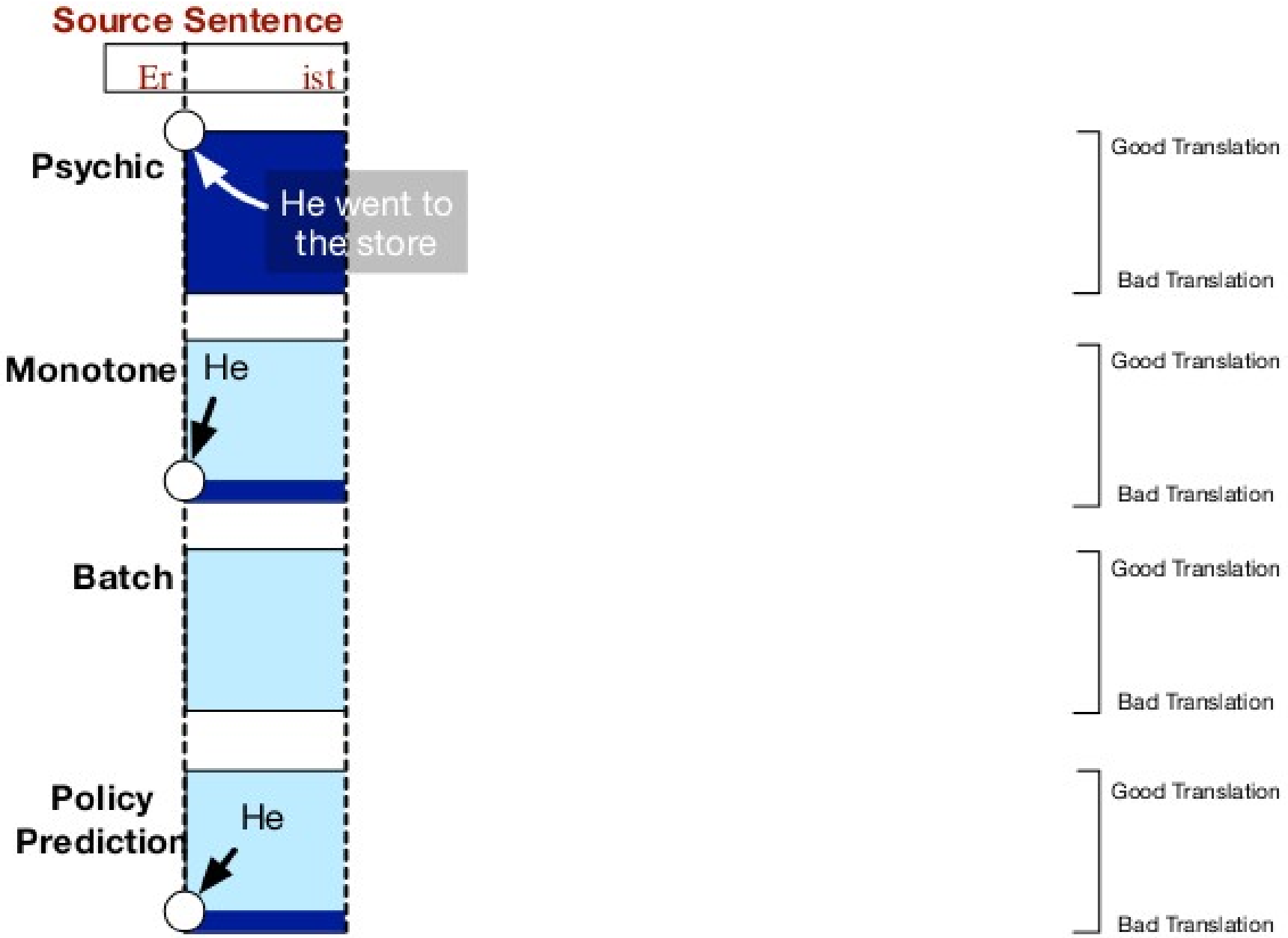
Evaluating performance and baselines



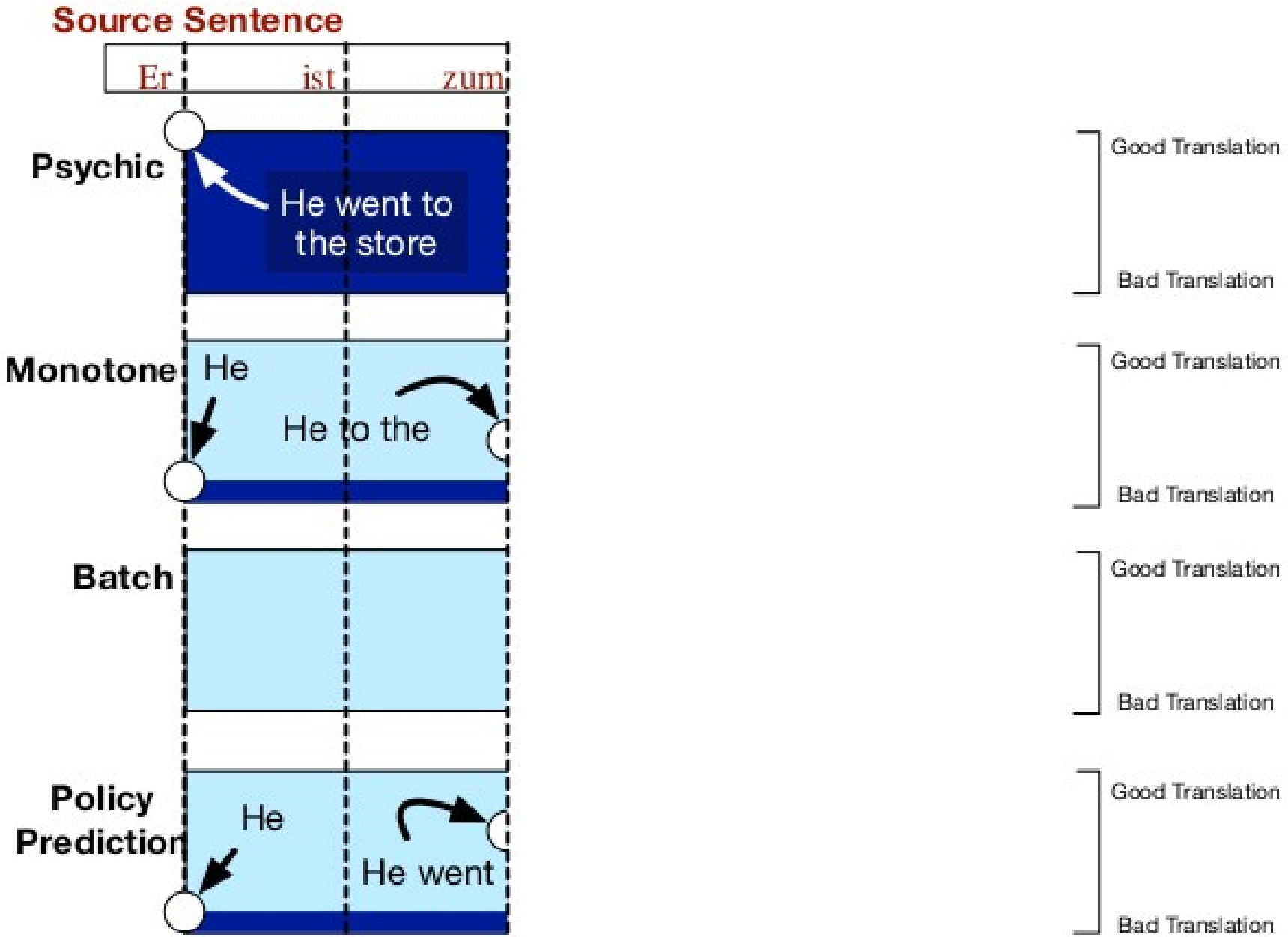
Evaluating performance and baselines



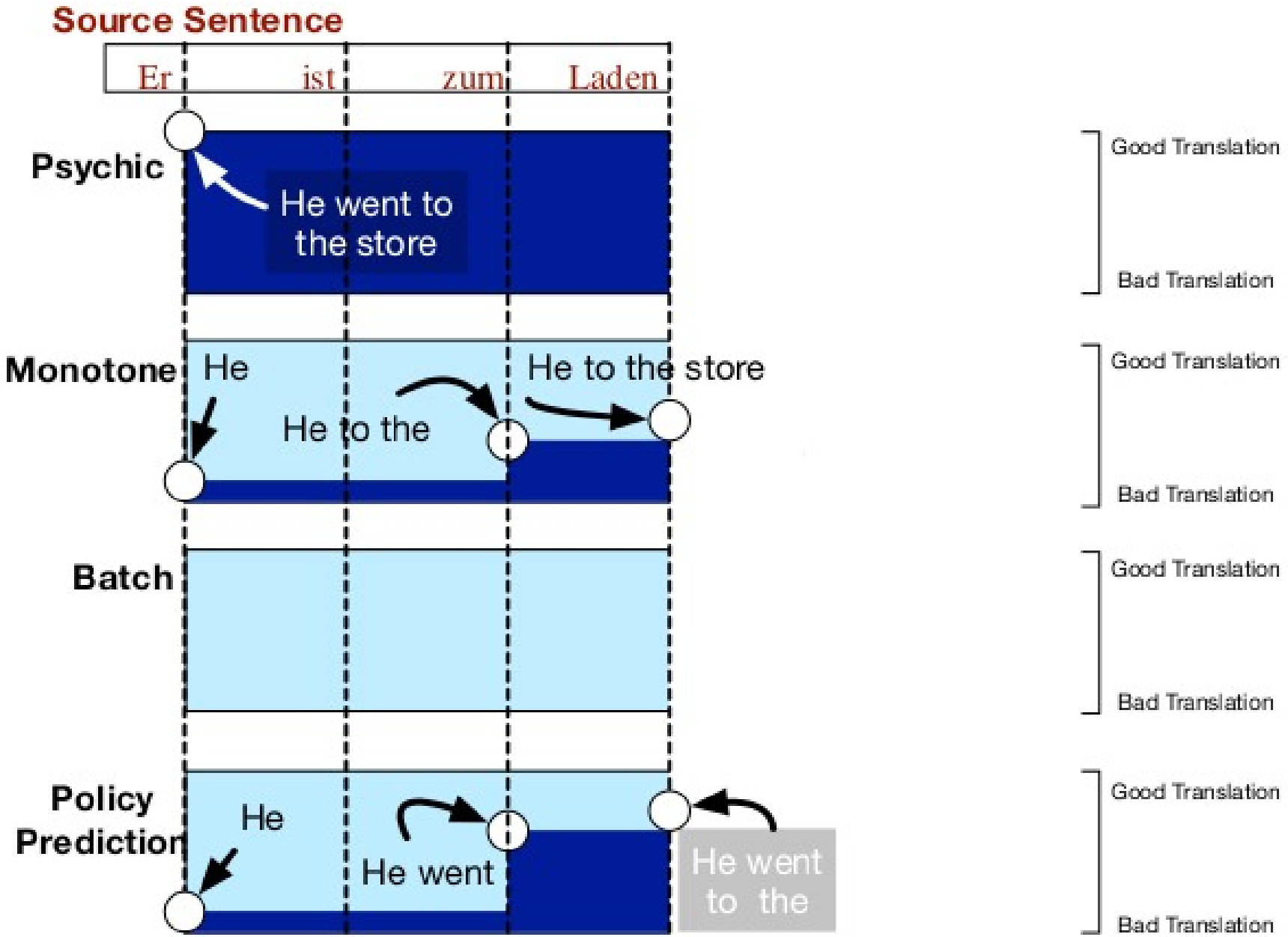
Evaluating performance and baselines



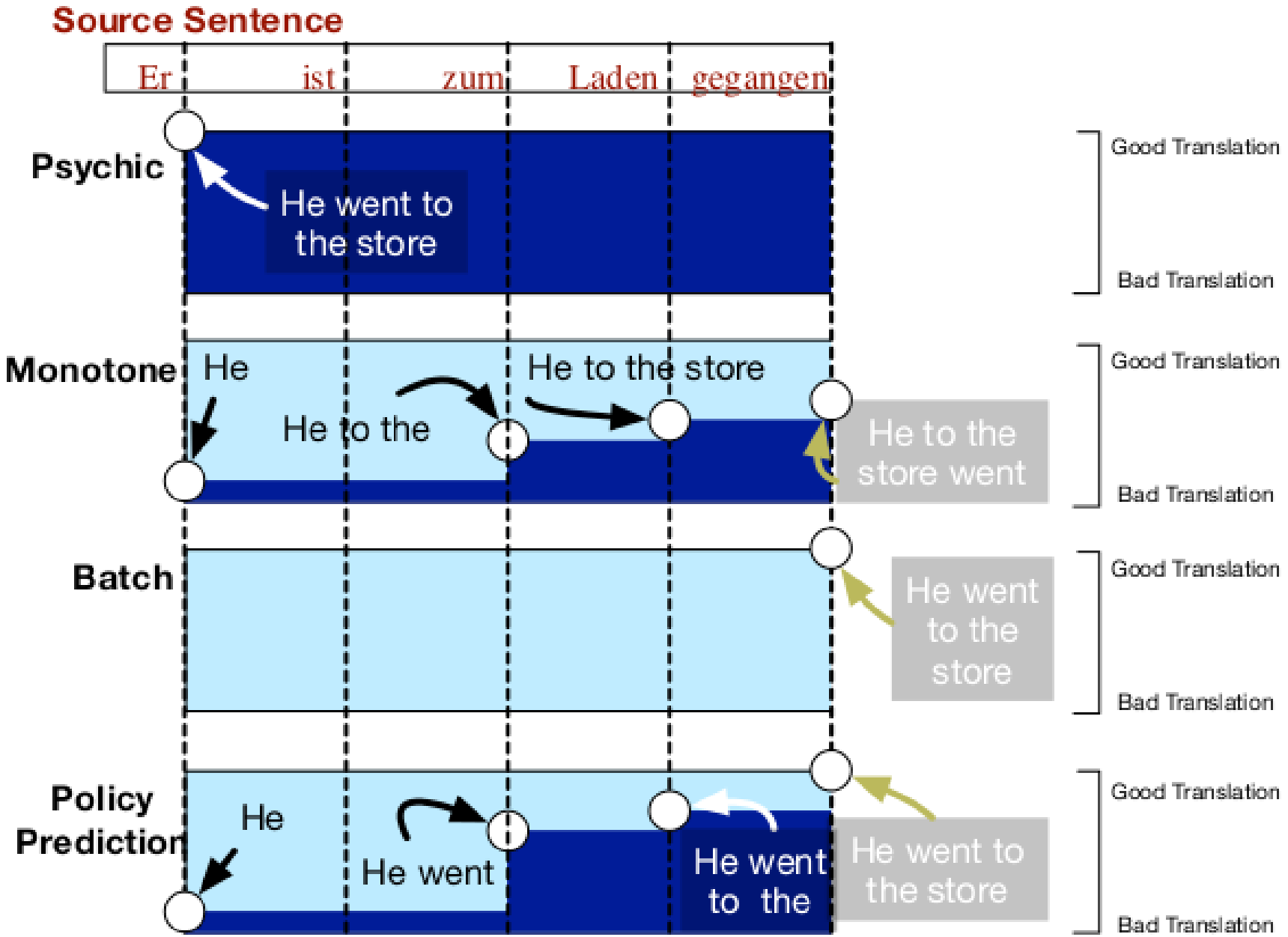
Evaluating performance and baselines



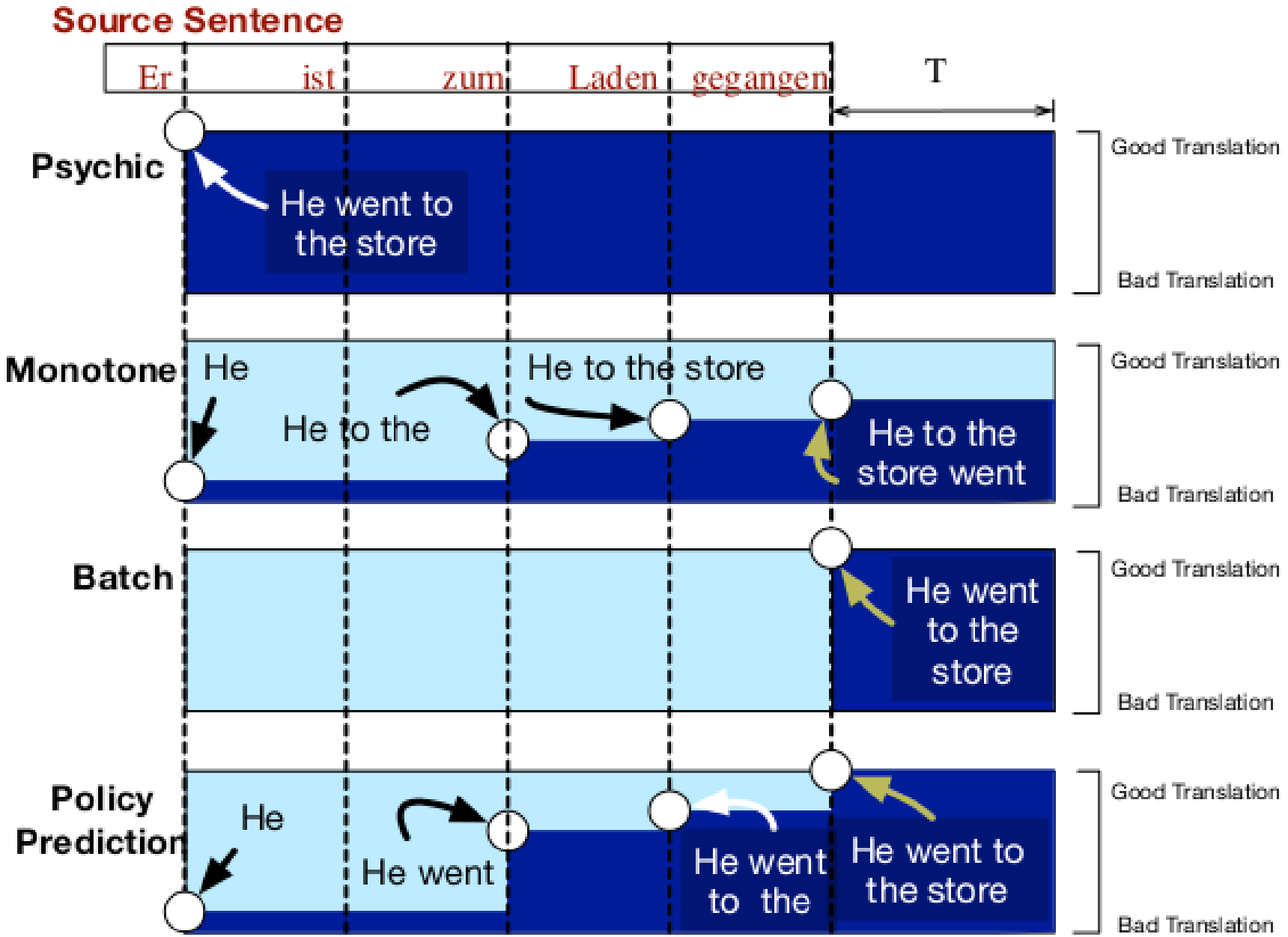
Evaluating performance and baselines



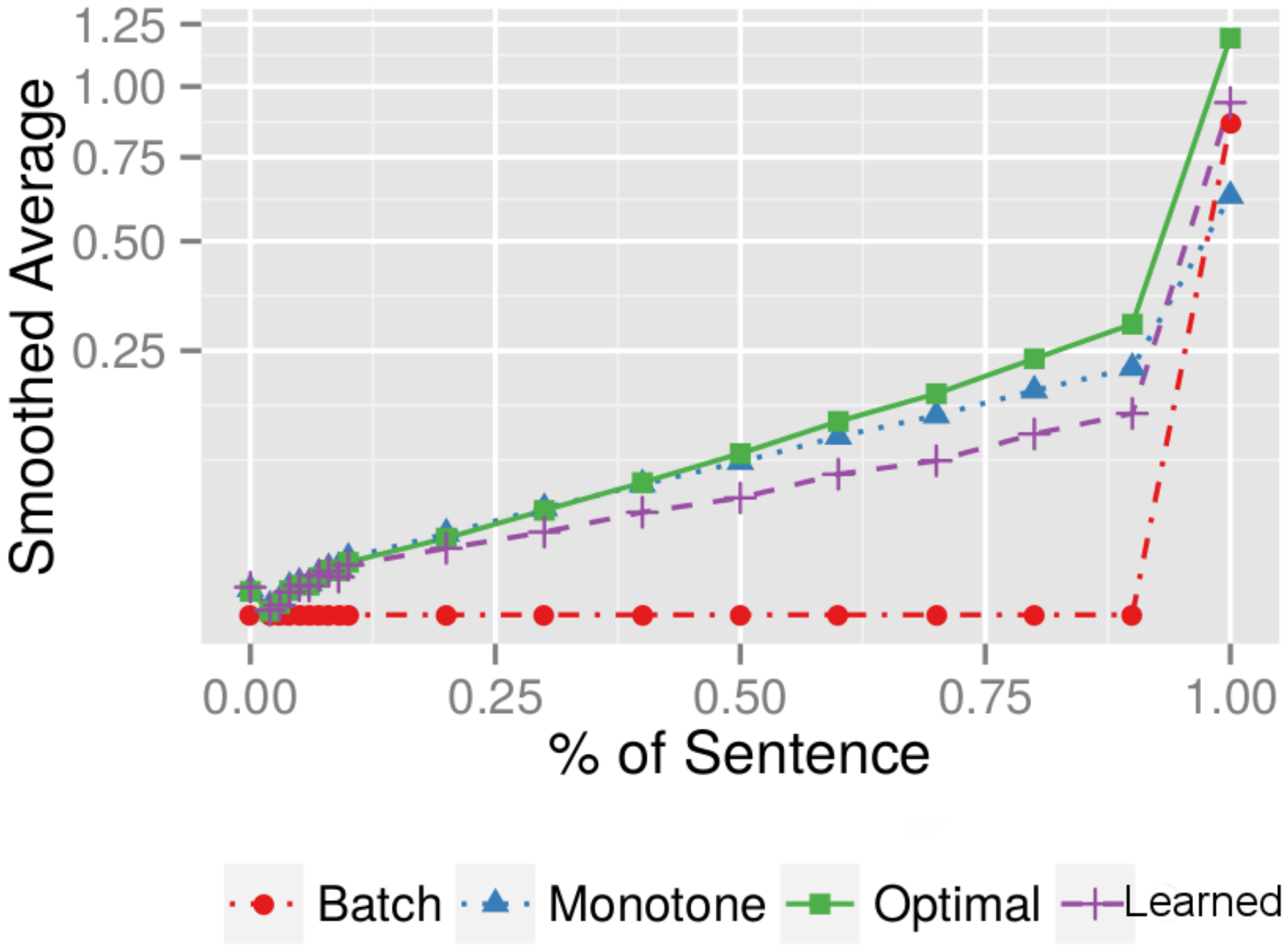
Evaluating performance and baselines



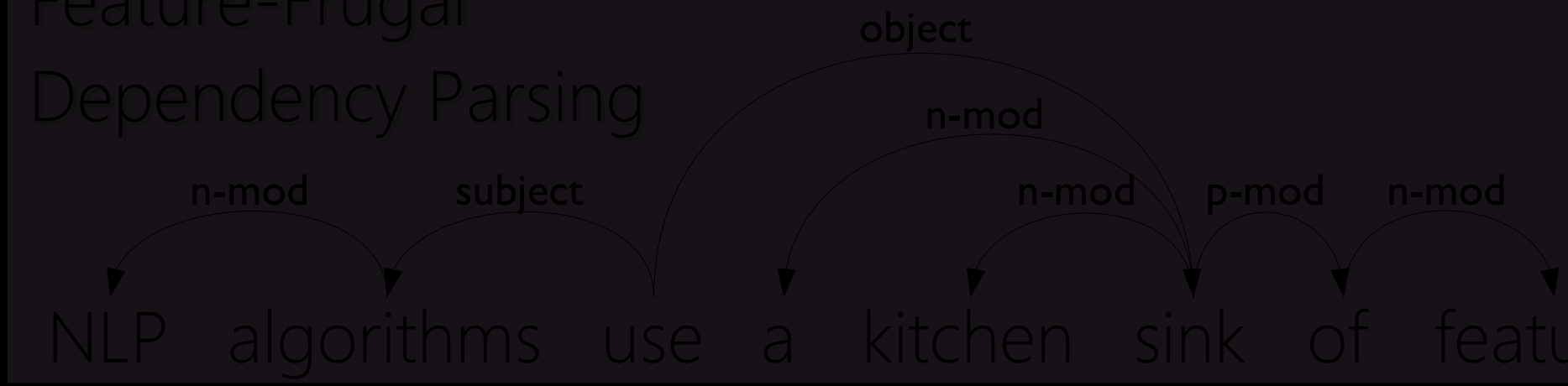
Evaluating performance and baselines

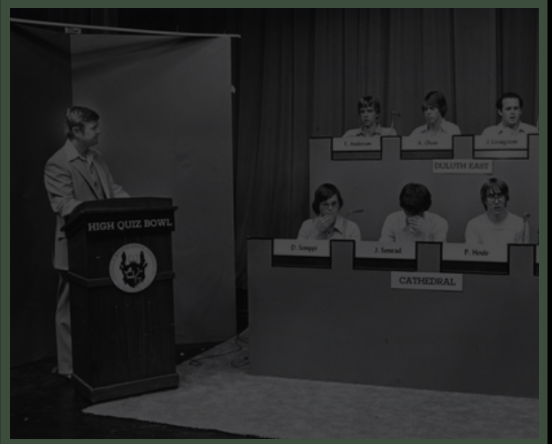


Evaluating performance and baselines



Feature-Frugal Dependency Parsing

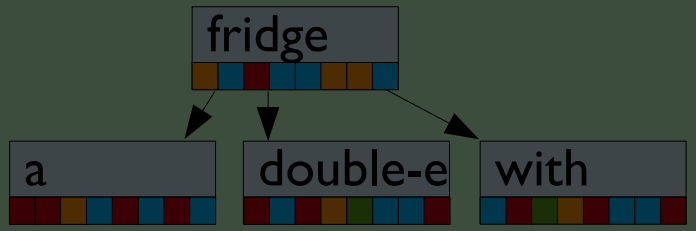




Quizbowl (Incremental Question Answering)

Ich bin mit dem Zug nach Ulm gefahren
I am with the train to Ulm traveled
I traveled by train to Ulm

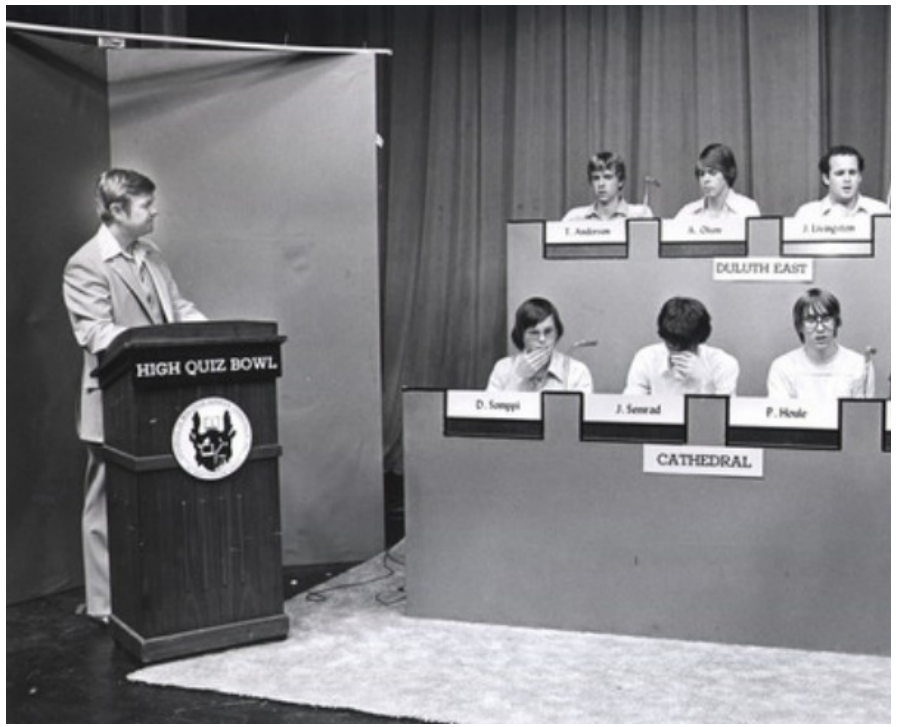
Simultaneous Machine Interpretation



Outline

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 . . .



Quizbowl example

With Leo Szilard, he invented a doubly-eponymous

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

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

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

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

Quizbowl example

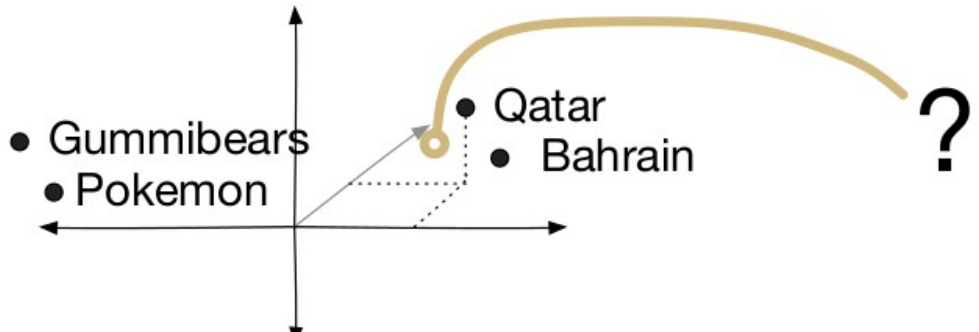
With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory of heat capacity, so Debye adjusted the theory for low temperatures. His summation convention automatically sums repeated indices in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

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?

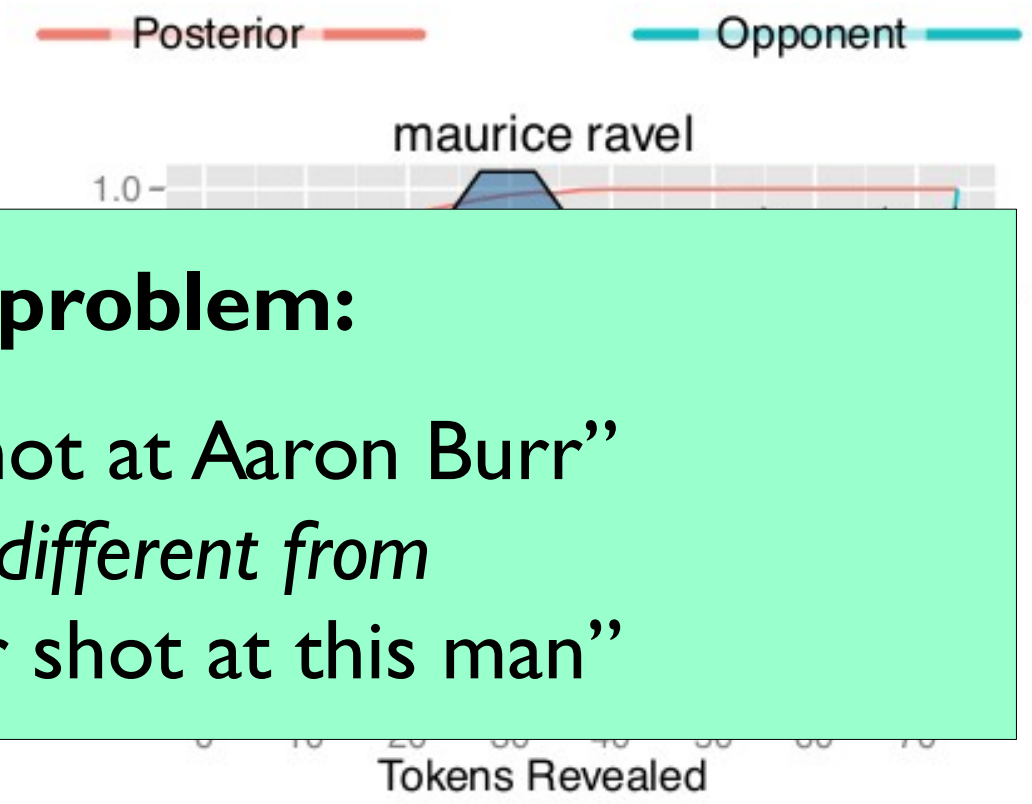
Solving incrementally

- 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



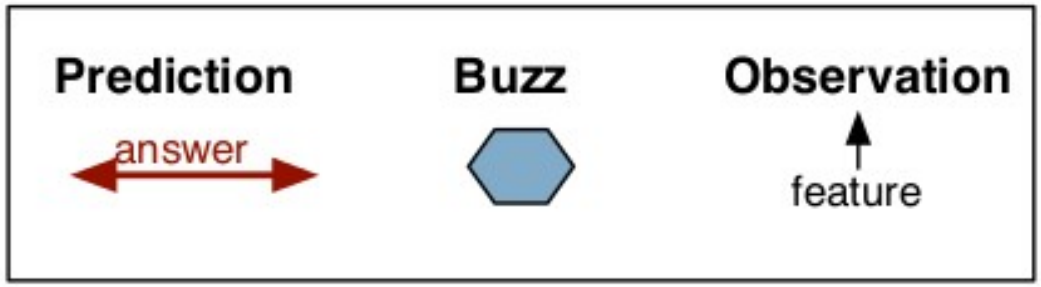
Evaluation methodology

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

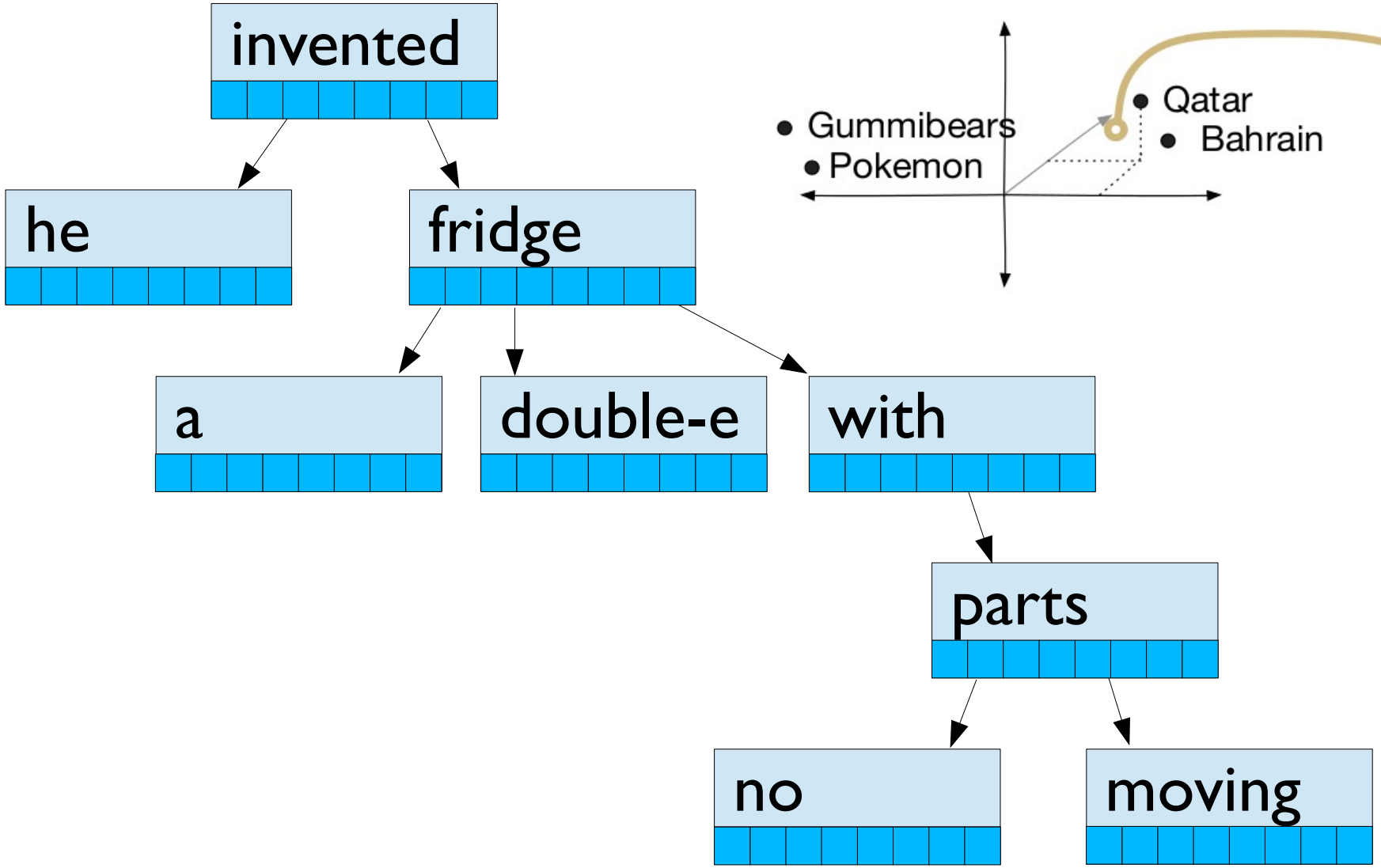


Big problem:

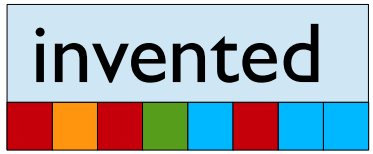
"this man shot at Aaron Burr"
is very different from
"Aaron Burr shot at this man"

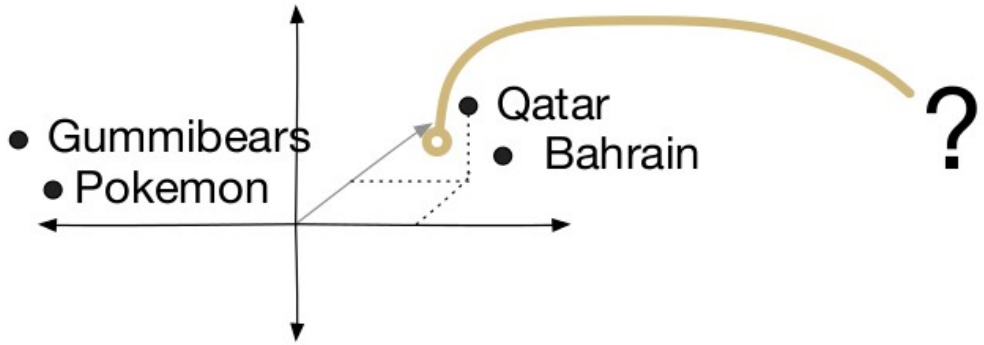


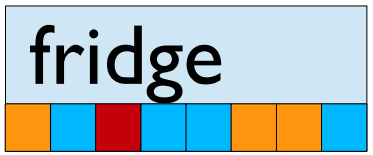
Challenge: modeling compositionality

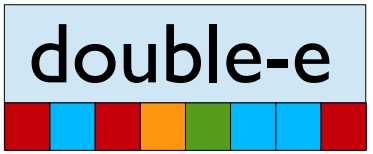


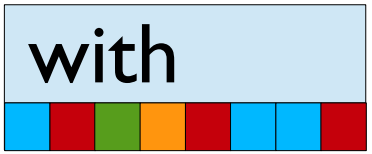
Challenge: modeling compositionality

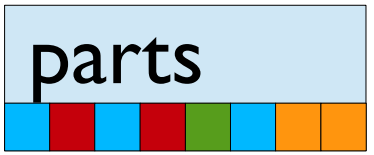


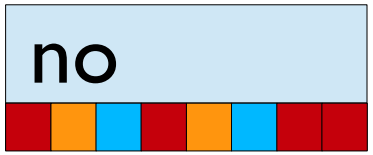


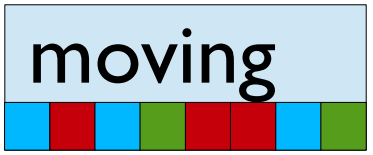




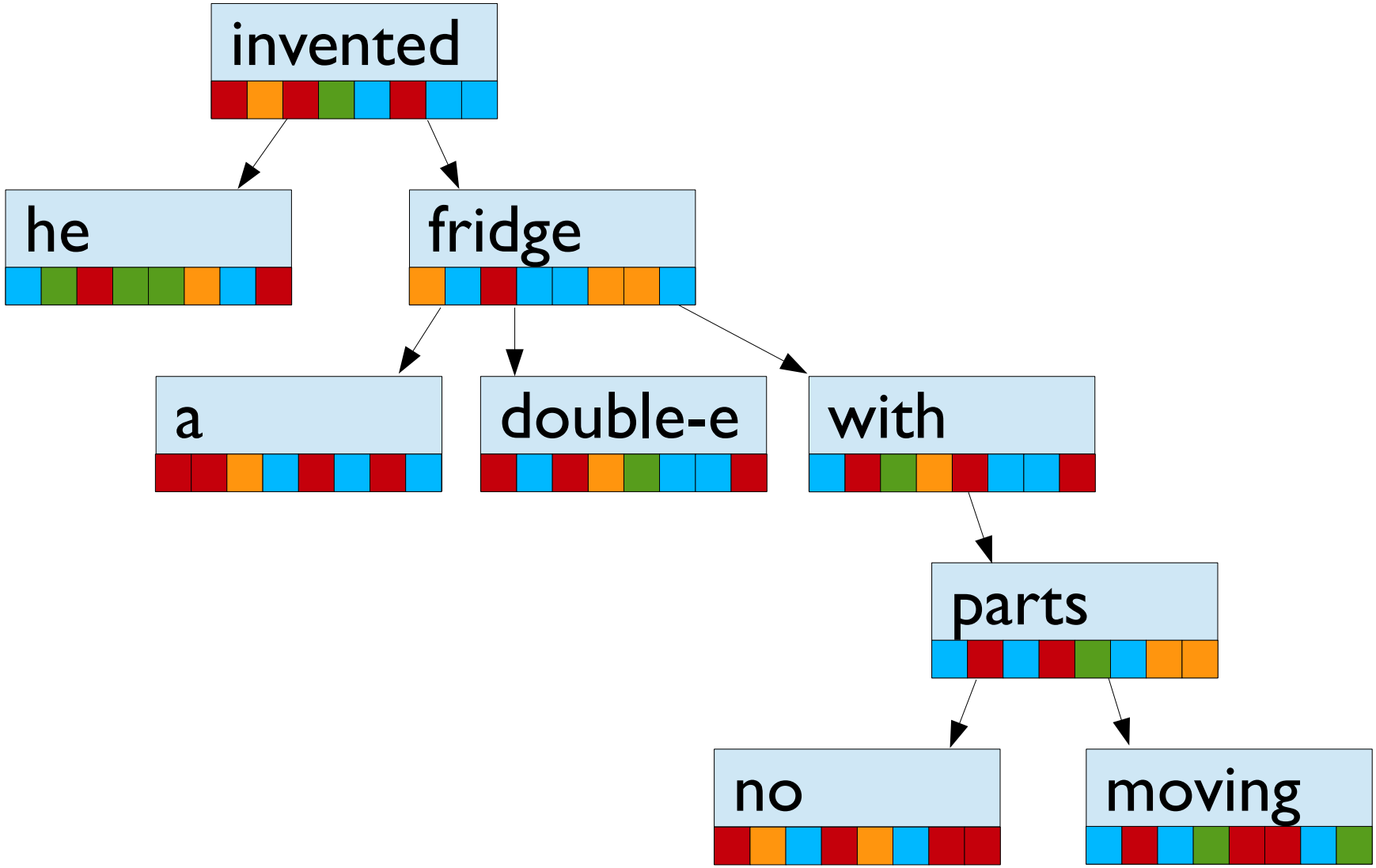




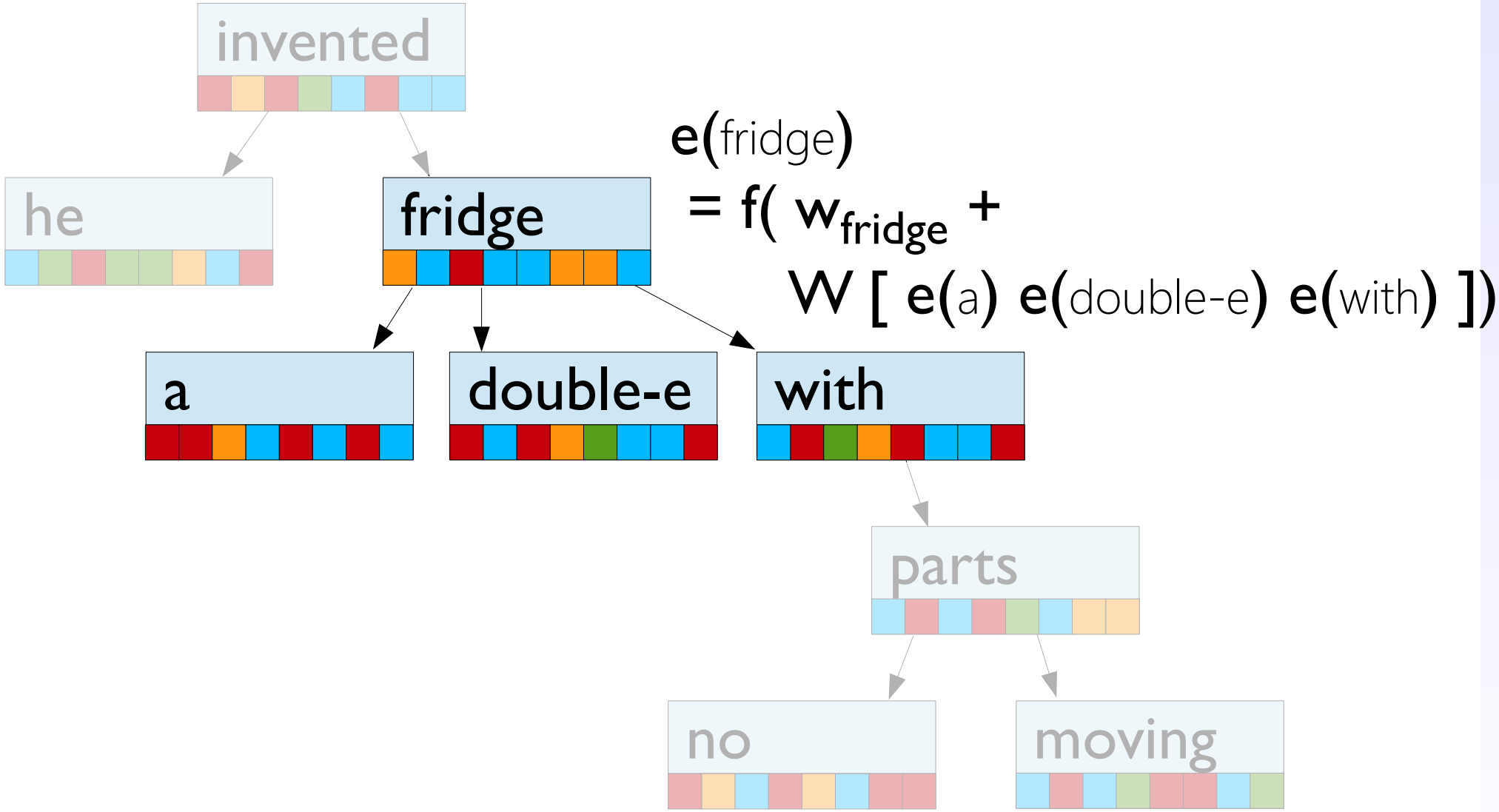




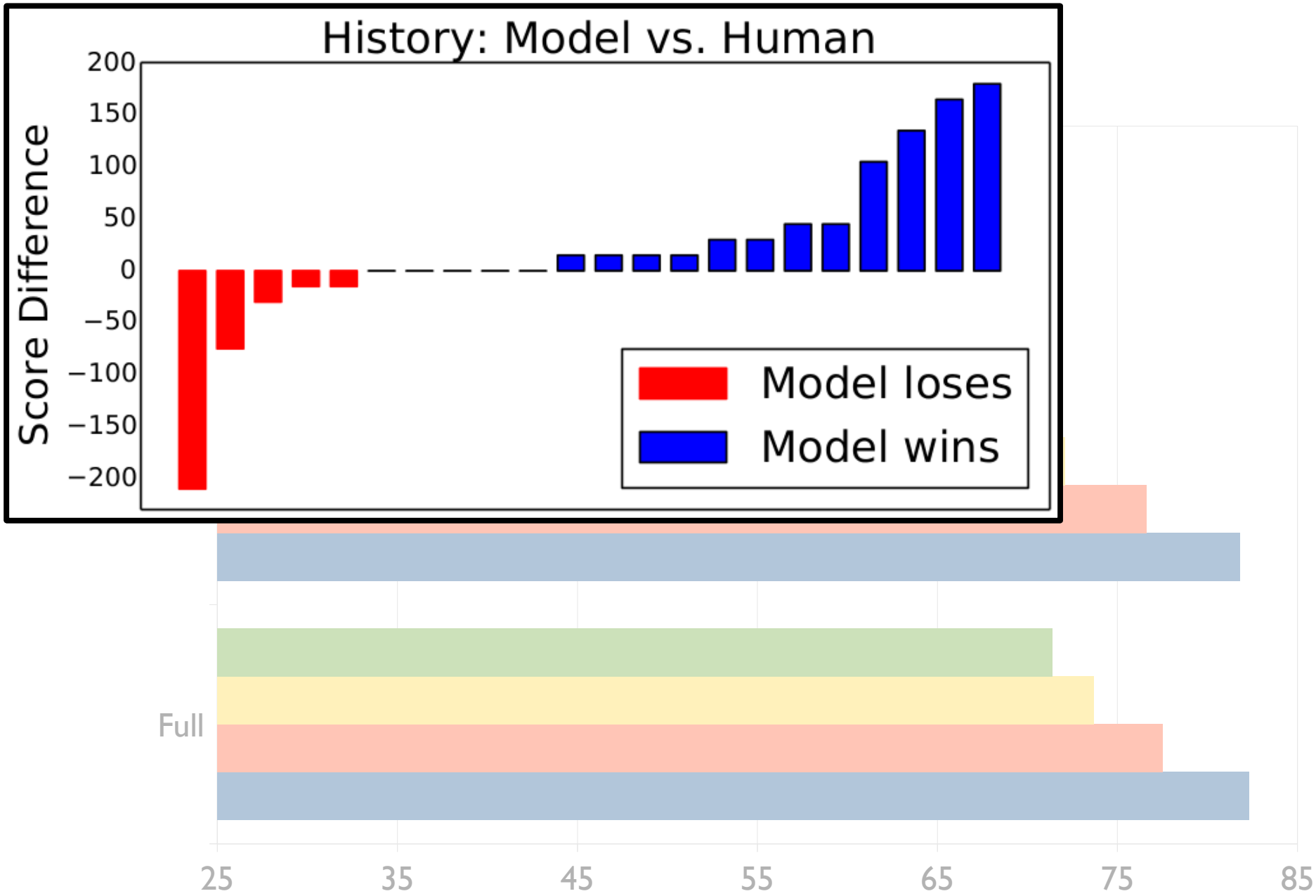
Challenge: modeling compositionality



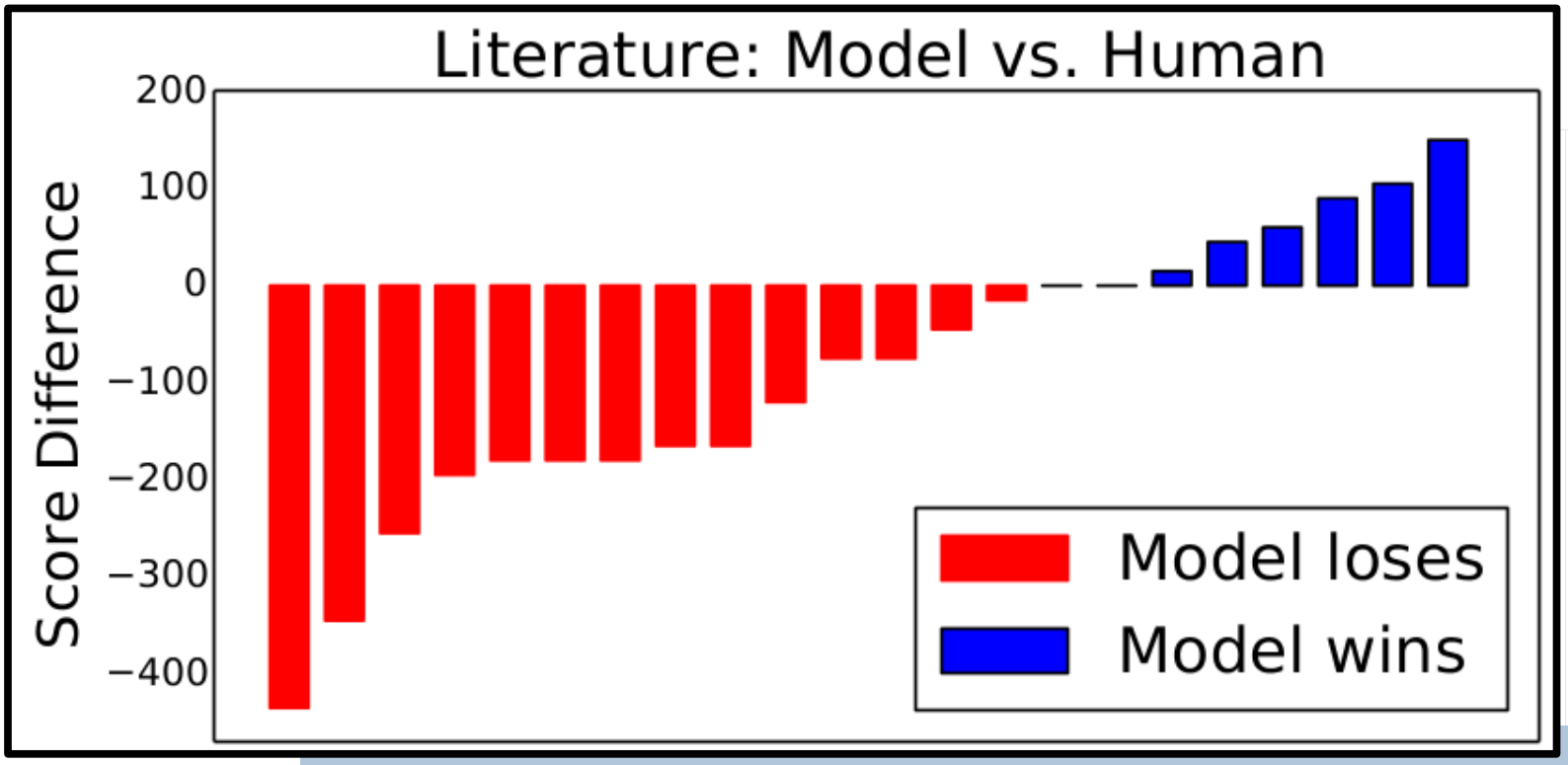
Challenge: modeling compositionality

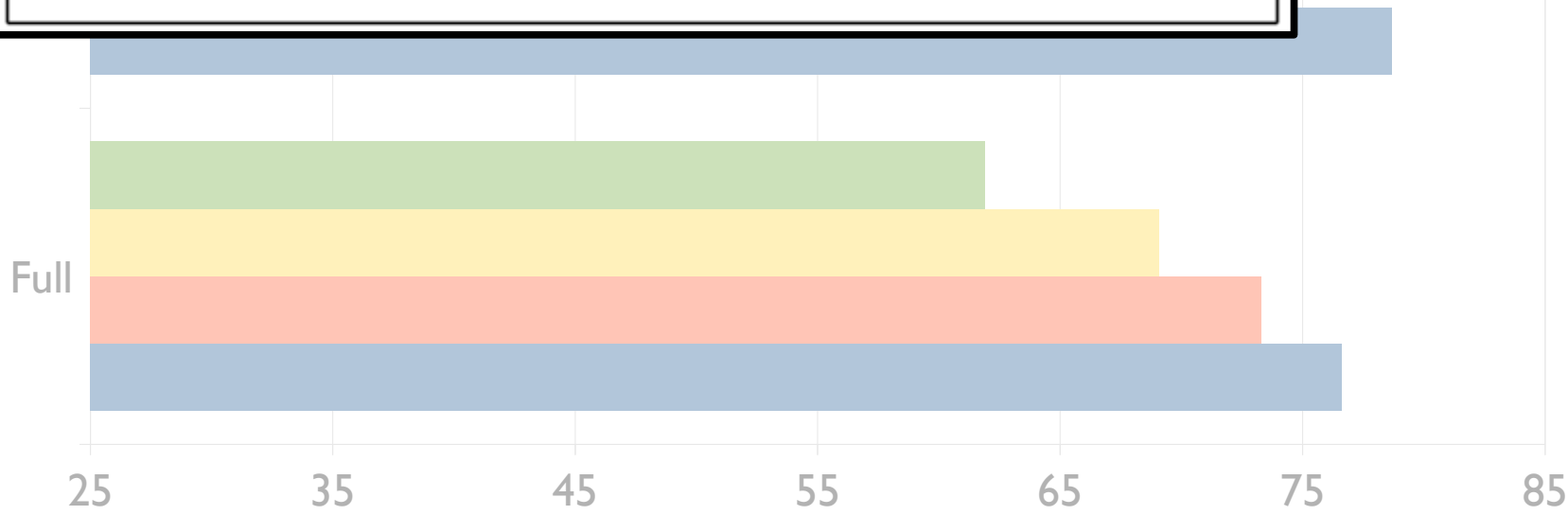


Results on question-answering task



Results on question-answering task

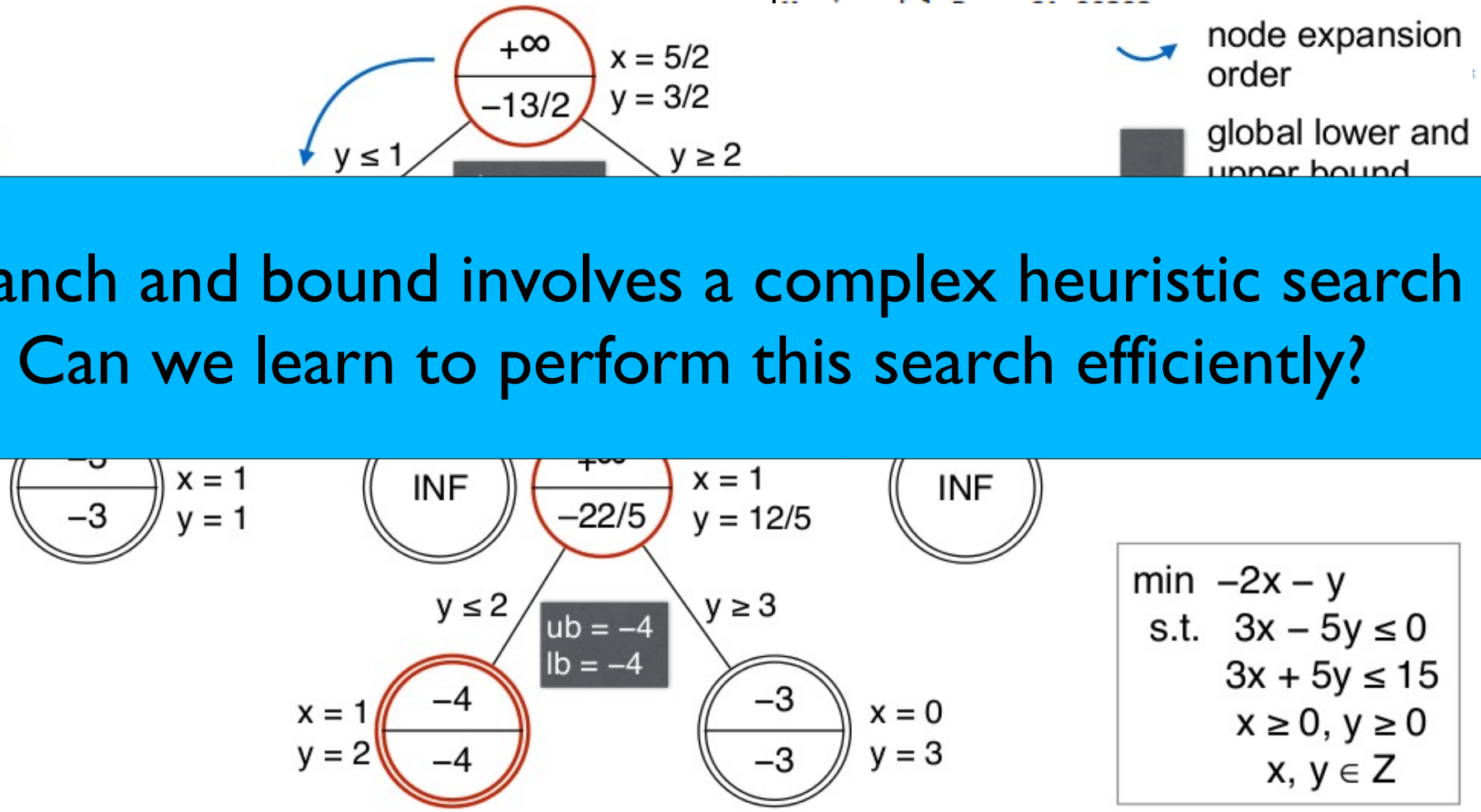




Moving to more general frameworks

- 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
```

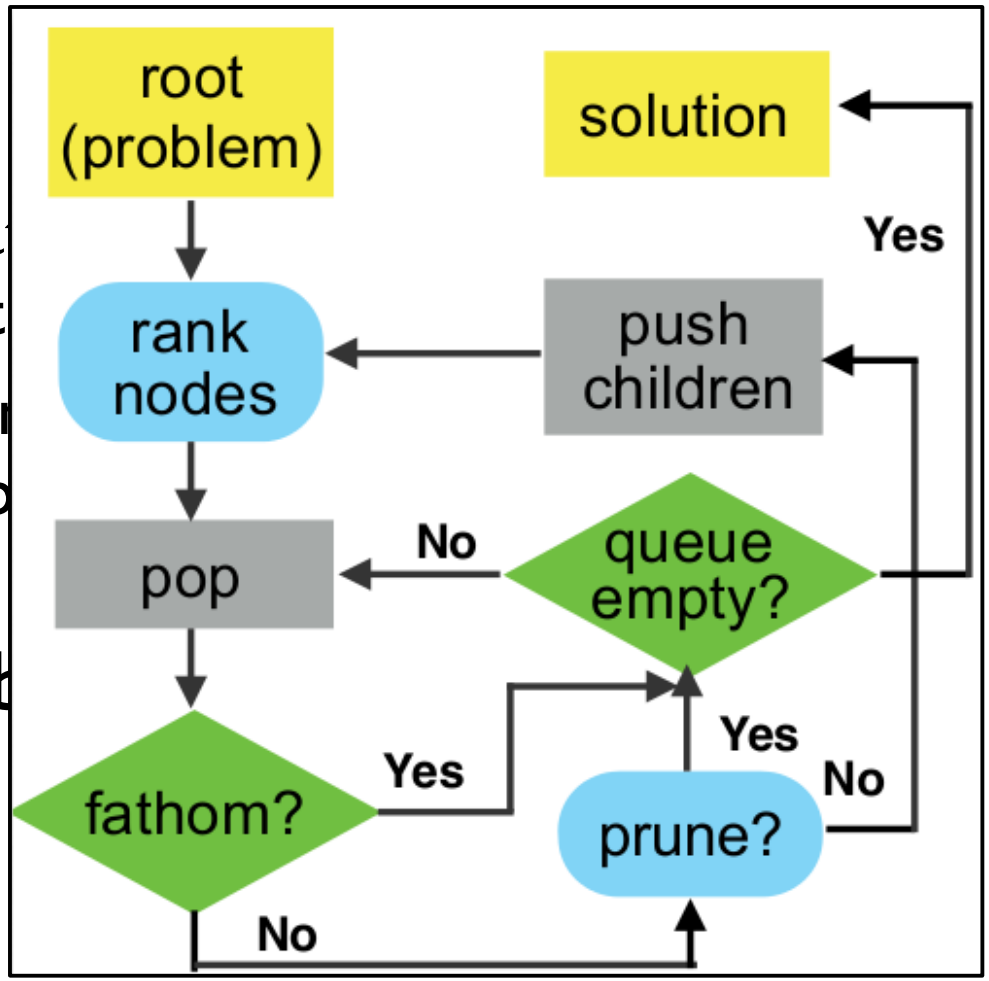


Branch and bound involves a complex heuristic search
Can we learn to perform this search efficiently?

(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)

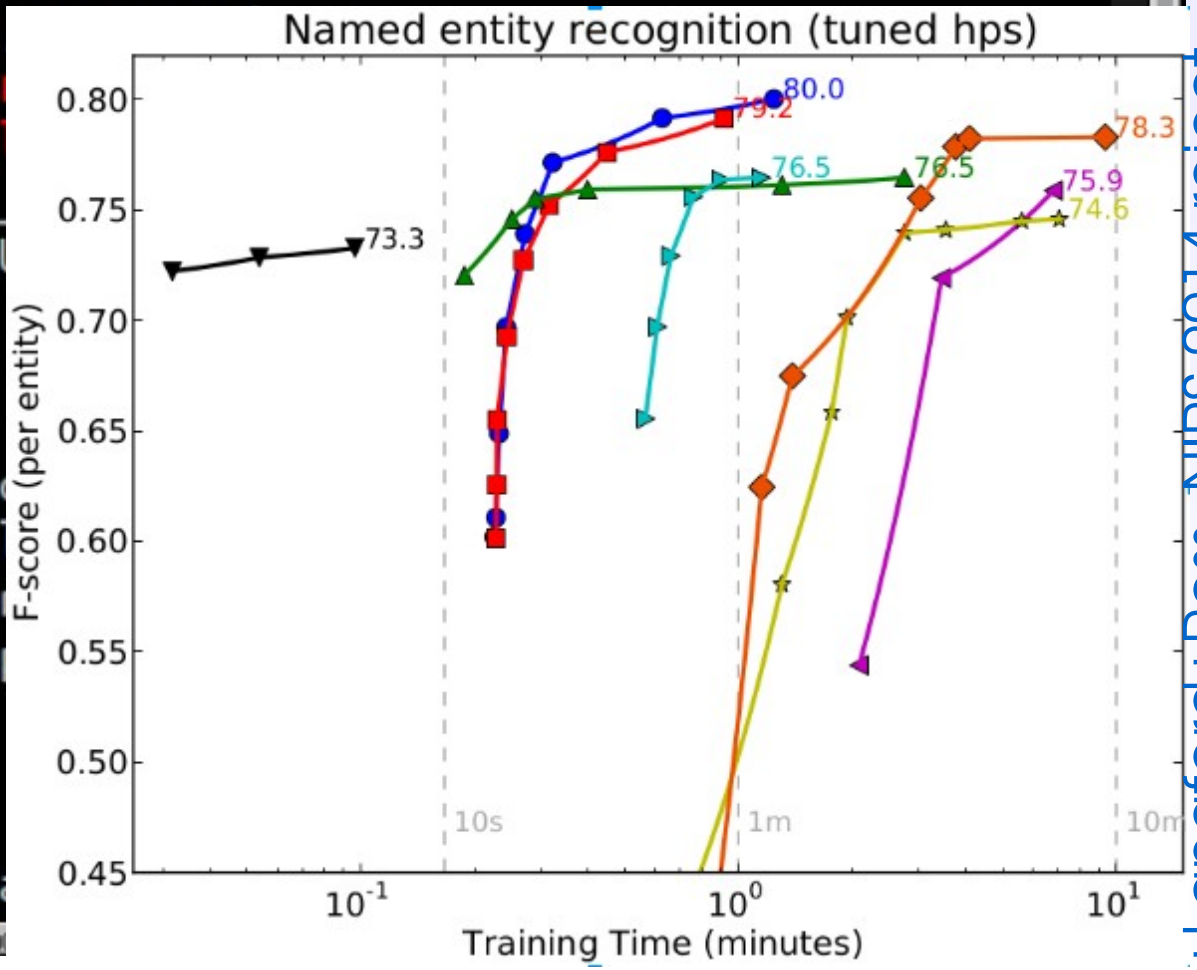
Some other fun stuff…

```
test_search.py
File  Edit  Options  Buffers  Tools  Python  YASnippet  Development  Help
import sys
import pyvw

class SequenceLabeler(pyvw.SearchTask):
    def __init__(self, vw, srn
        # you must must must i
        # this will automatica
        pyvw.SearchTask.__init_
        srn.set_options( srn.A

    def _run(self, sentence):
        output = []
        for tag,word in senten
            with self.vw.examp
                pred = self.sr
                output.append(
        return output

for curPass in range(10):
    sequenceLabeler.learn(my_d
U:**-  test_search.py   Top (1,0)      (Python BufFa
```



(D+Langford+Ross, NIPS 2014 reject – on arxiv)


Jordan B-G


Jason Eisner


Alvin Grissom


He He


Mohit Iyyer


John Langford

- Reasoning with imcomplete information is useful for *speed* and *modeling*
- *Imitation learning* can help us build such systems
- Wide range of new, interesting problems to work on!

Thanks! Questions?