Structural Decoding in Neural Text Generation

文本生成中的结构化解码

Jiacheng Xu

Beam Search

Source/Input

据悉，奥巴马在参议员唐诺利的造势场合向群众表示 […]

Seq2seq model

$P(\mathbf{y} \mid \mathbf{x})$

Target/Summary

It is reported that Obama told the crowds at a rally for Senator Donnelly, […]

Today, the Pulitzer prize for journalism went to The Post and Courier newspaper of Charleston, South Carolina […]

Seq2seq model

$P(\mathbf{y} \mid \mathbf{x})$

The smallest newspapers in the United States has won the Pulitzer Prize for journalism for the second year in a row.

Beam Search

Source/Input

据悉，奥巴马在参议员唐诺利的造势场合向群众表示 […]

Seq2seq model

$$P(\mathbf{y} \mid \mathbf{x})$$

Target/Summary

It is reported that Obama told the crowds at a rally for Senator Donnelly, […]

Today, the Pulitzer prize for journalism went to The Post and Courier newspaper of Charleston, South Carolina […]

Seq2seq model

$$P(\mathbf{y} \mid \mathbf{x})$$

The smallest newspapers in the United States has won the Pulitzer Prize for journalism for the second year in a row.

$$\textit{find} \quad \arg\max P(\mathbf{y} \mid \mathbf{x})$$

Beam search can find several high-scoring options.

Beyond Single Output

Beyond Single Output

Beyond Single Output

Find a factual summary:

What if the top summary contains errors?
(It's not the *smallest* newspaper)

Beyond Single Output

Find a factual summary:

What if the top summary contains errors? (It's not the *smallest* newspaper)

We can use the alternatives! Try something further down in the beam!

Beyond Single Output

Input Doc → **Model**

Generated Summaries (highest model score)

1. The smallest newspaper in the United States has won the Pulitzer Prize for journalism for the second year in a row.
2. The smallest newspaper in the United States has won the Pulitzer Prize for journalism.
3. The smallest newspaper in the United States has won the Pulitzer Prize for journalism for the first time.
4. The smallest newspaper in the United States was awarded the gold medal for public service.
5. ……

Find a factual summary:

What if the top summary contains errors? (It's not the *smallest* newspaper)

We can use the alternatives! Try something further down in the beam!

Beyond Single Output

Input Doc → **Model**

Generated Summaries (highest model score)

1. The smallest newspaper in the United States has won the Pulitzer Prize for journalism for the second year in a row.
2. The smallest newspaper in the United States has won the Pulitzer Prize for journalism.
3. The smallest newspaper in the United States has won the Pulitzer Prize for journalism for the first time.
4. The smallest newspaper in the United States was awarded the gold medal for public service.
5. ……

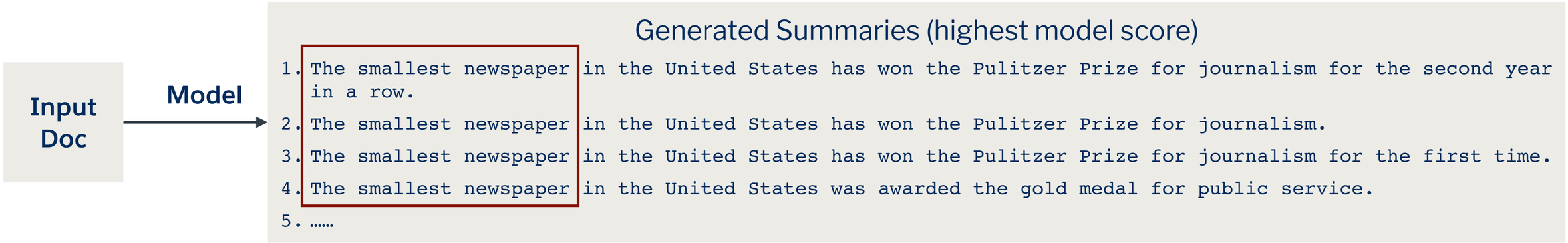
Find a factual summary:

What if the top summary contains errors? (It's not the *smallest* newspaper)

We can use the alternatives! Try something further down in the beam!

Top summaries are **similar** and **wrong**!

Beyond Single Output



Find a factual summary:

What if the top summary contains errors? (It's not the *smallest* newspaper)

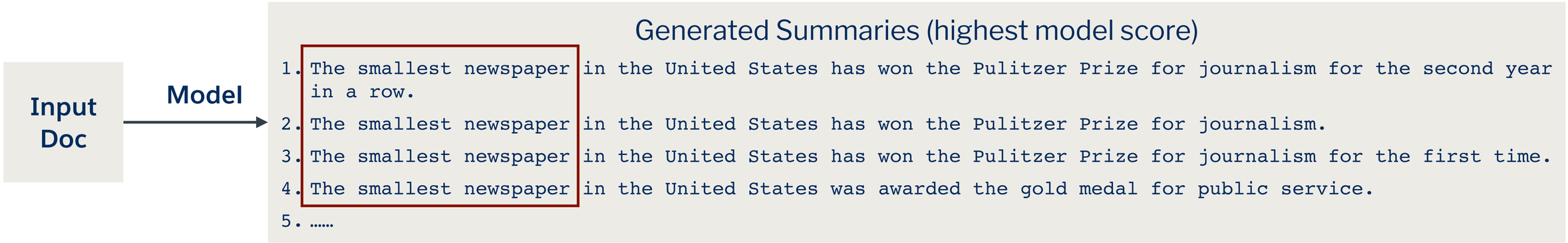
We can use the alternatives! Try something further down in the beam!

Top summaries are **similar** and **wrong**!

Customize content:

Where is the newspaper located?

Beyond Single Output



Find a factual summary:

What if the top summary contains errors? (It's not the *smallest* newspaper)

We can use the alternatives! Try something further down in the beam!

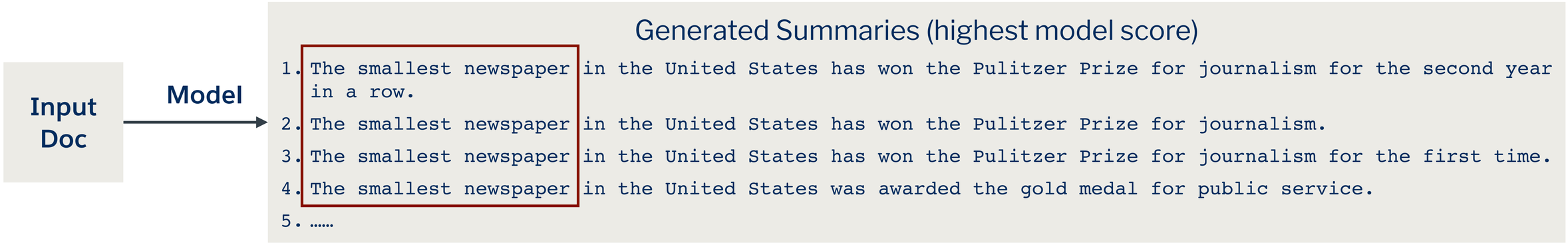
Top summaries are **similar** and **wrong**!

Customize content:

Where is the newspaper located?

None of *top* summaries says.
However, South Carolina did once show up in beam search, and got **pruned** later.

Beyond Single Output



Find a factual summary:

What if the top summary contains errors? (It's not the *smallest* newspaper)

We can use the alternatives! Try something further down in the beam!

Top summaries are **similar** and **wrong**!

Customize content:

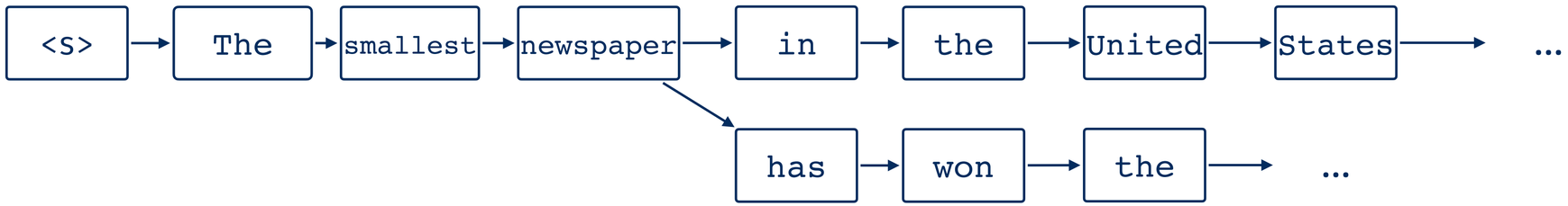
Where is the newspaper located?

None of *top* summaries says. However, South Carolina did once show up in beam search, and got **pruned** later.

These alternatives are not flexible enough for downstream generation applications.

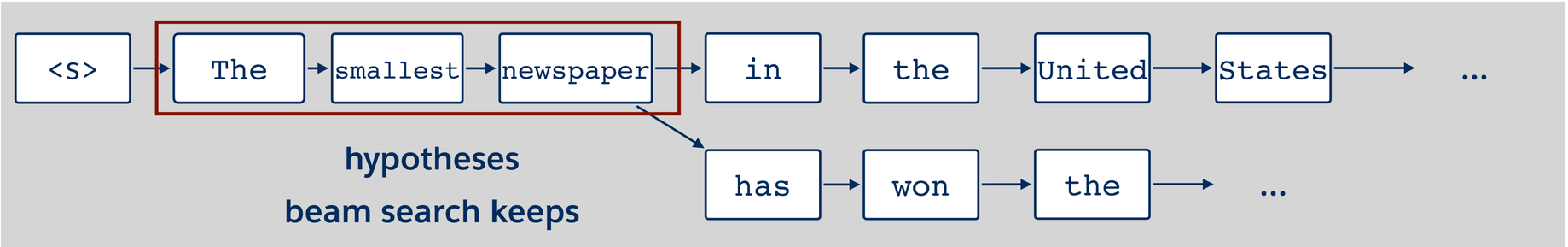
Search for Diverse Outputs

Beam Search



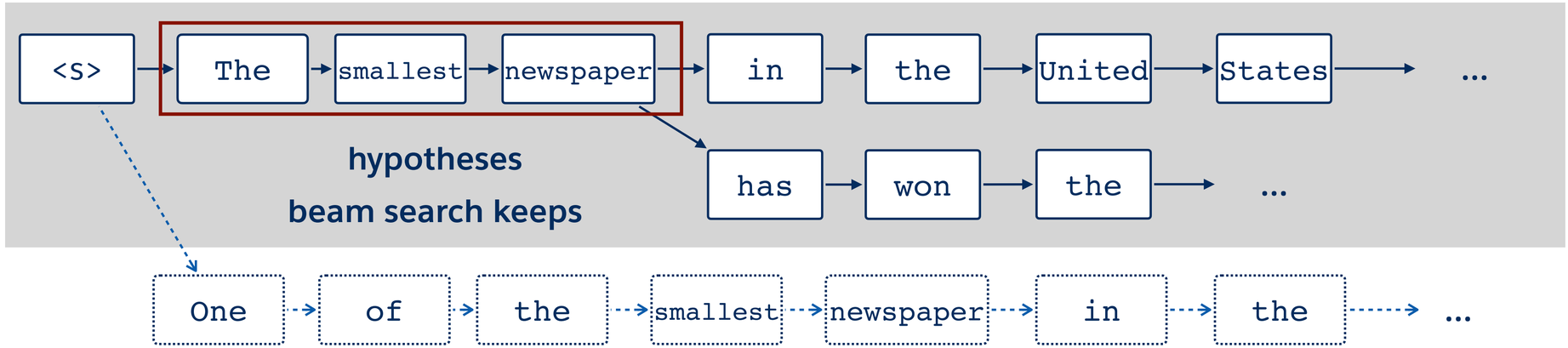
Search for Diverse Outputs

Beam Search



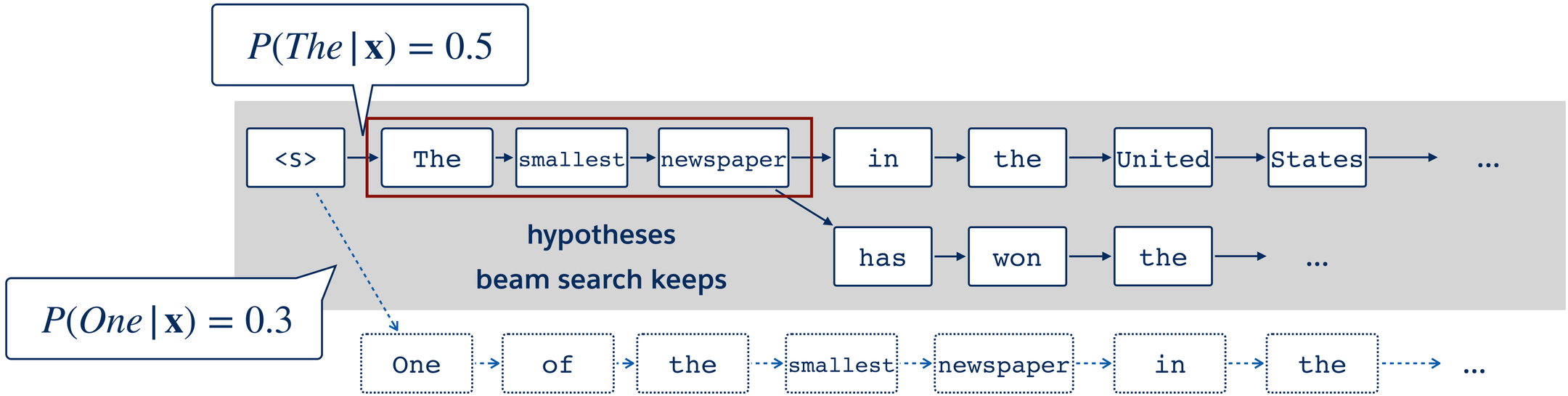
Search for Diverse Outputs

Beam Search



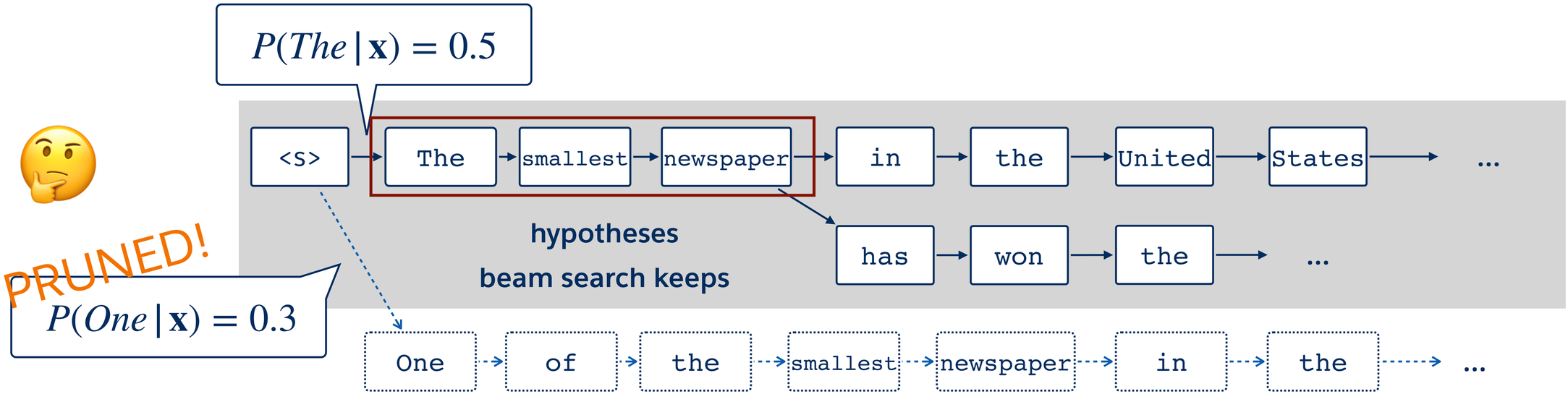
Search for Diverse Outputs

Beam Search



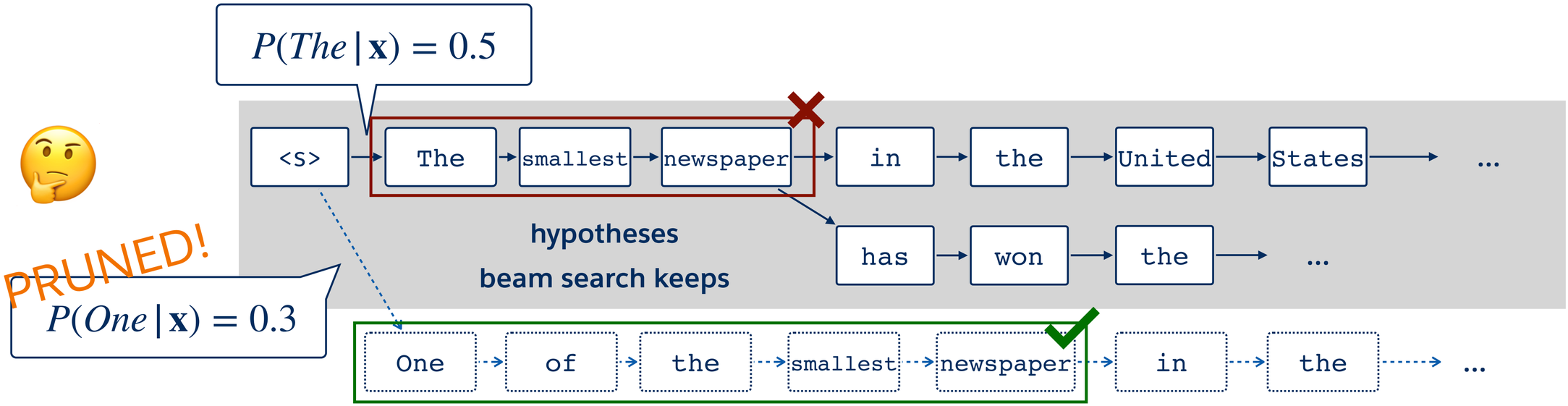
Search for Diverse Outputs

Beam Search



Search for Diverse Outputs

Beam Search



Beam search prunes many hypotheses

Some of them are great!

Search for Diverse Outputs

Stochastic Sampling 🎲

Task: question generation

Reference: What city is Intel located in?

Answer || context

Input: Hillsboro || Hillsboro is the fifth-largest city in the State of Oregon and is the county seat of Washington ... the city hosts many high-technology companies, such as Intel ...

Search for Diverse Outputs

Stochastic Sampling 🎲

Task: question generation

Reference: What city is Intel located in?

Answer || context

Input: Hillsboro || Hillsboro is the fifth-largest city in the State of Oregon and is the county seat of Washington ... the city hosts many high-technology companies, such as Intel ...

🎲: What is the fifth largest city in Oregon?

🎲: What is the fifth largest city in Oregon?

🎲: What is the fifth-largest city in Oregon?

🎲: What is the fifth-largest city in the State of Oregon?

Sampling sometimes causes duplication.

Search for Diverse Outputs

Stochastic Sampling 🎲

Task: question generation

Reference: What city is Intel located in?

Answer || context

Input: Hillsboro || Hillsboro is the fifth-largest city in the State of Oregon and is the county seat of Washington ... the city hosts many high-technology companies, such as Intel ...

🎲: What is the fifth largest city in Oregon?

🎲: What is the fifth largest city in Oregon?

🎲: What is the fifth-largest city in Oregon?

🎲: What is the fifth-largest city in the State of Oregon?

Sampling sometimes causes duplication.

Search for Diverse Outputs

Stochastic Sampling 🎲

Task: question generation

Reference: What city is Intel located in?

Answer || context

Input: Hillsboro || Hillsboro is the fifth-largest city in the State of Oregon and is the county seat of Washington ... the city hosts many high-technology companies, such as Intel ...

🎲: What is the fifth largest city in Oregon?

🎲: What is the fifth largest city in Oregon?

🎲: What is the fifth-largest city in Oregon?

🎲: What is the fifth-largest city in the State of Oregon?

Sampling sometimes causes duplication.

Beam search is deterministic while sampling is not.

Search for Diverse Outputs

Stochastic Sampling 🎲

Task: question generation

Reference: What city is Intel located in?

Answer || context

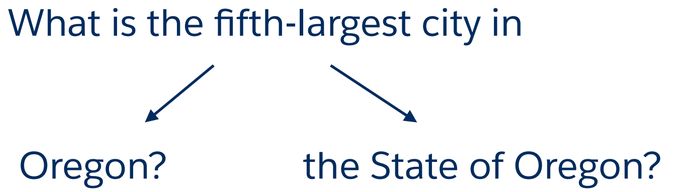
Input: Hillsboro || Hillsboro is the fifth-largest city in the State of Oregon and is the county seat of Washington ... the city hosts many high-technology companies, such as Intel ...

🎲: What is the fifth largest city in Oregon?

🎲: What is the fifth largest city in Oregon?

🎲: What is the fifth-largest city in Oregon?

🎲: What is the fifth-largest city in the State of Oregon?



Beam search is deterministic while sampling is not.

Sampling sometimes causes duplication.

Unlike beam search, sampling is hard to control.

Structural Decoding in Neural Text Generation

Roadmap

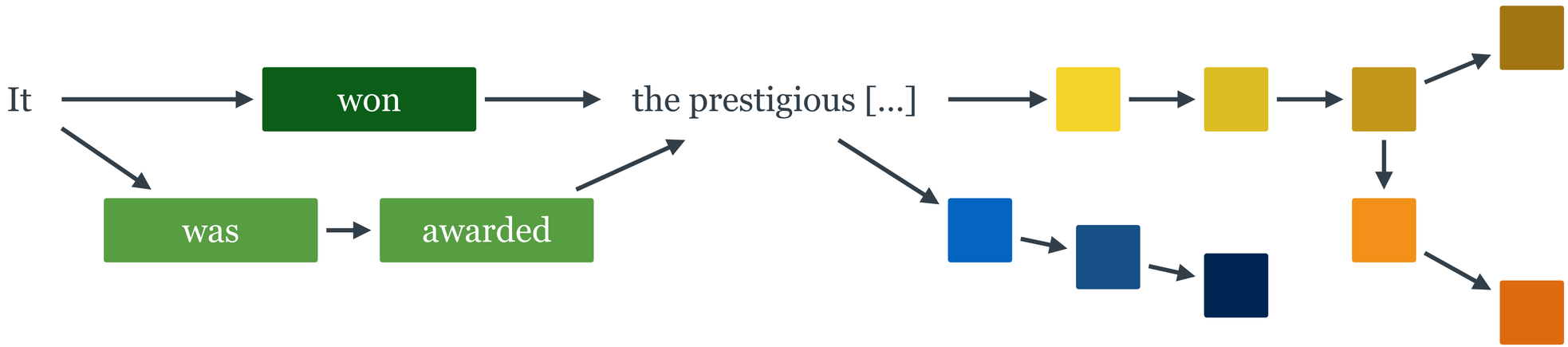
- Massive-scale Decoding for Text Generation using Lattices

  Jiacheng Xu, Siddhartha Reddy Jonnalagadda, Greg Durrett
  UT Austin, Amazon Alexa AI

- Best-k Search Algorithm for Neural Text Generation

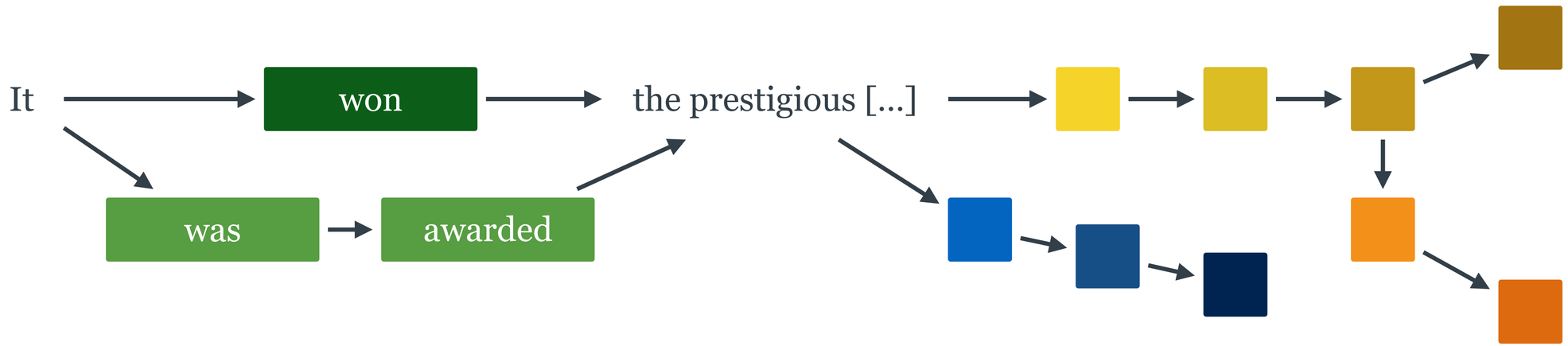
  Jiacheng Xu, Caiming Xiong, Silvio Savarese, Yingbo Zhou
  Salesforce AI Research, Palo Alto, CA

Contribution



Contribution

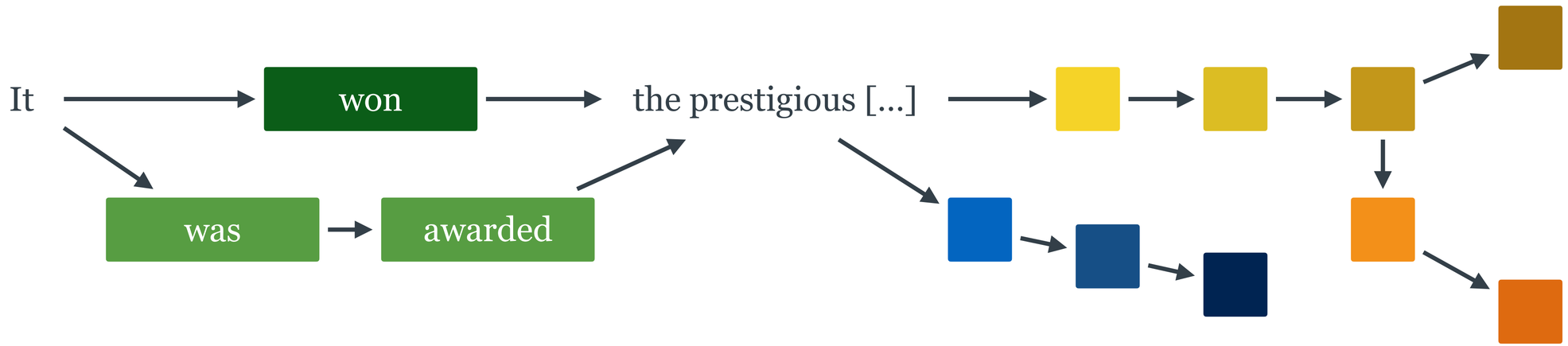
We propose a search algorithm encoding many diverse generation options



Contribution

We propose a search algorithm encoding many diverse generation options

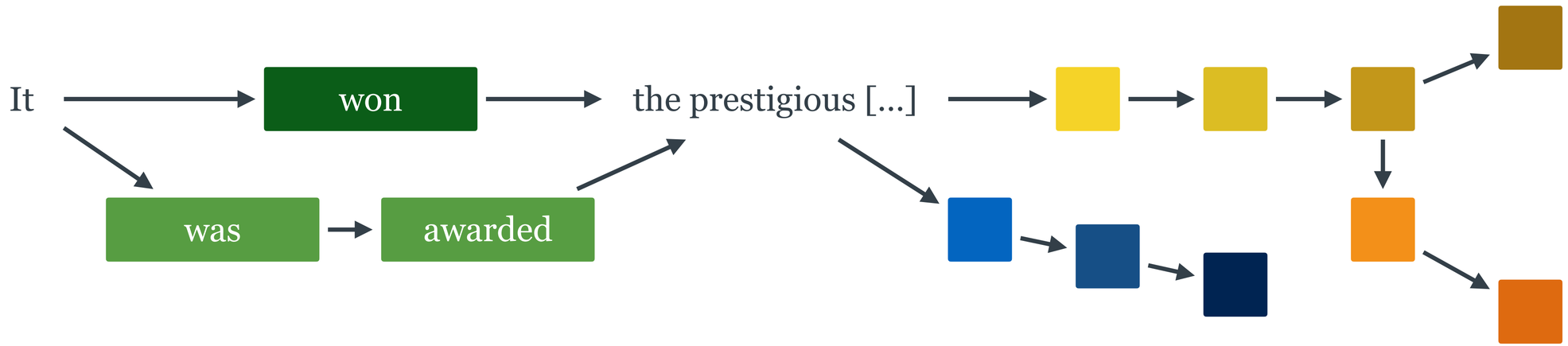
- many: 100x ~ 1000x more outputs than beam search



Contribution

We propose a search algorithm encoding many diverse generation options

- many: 100x ~ 1000x more outputs than beam search
- diverse: content, style, syntax, word choice, etc.

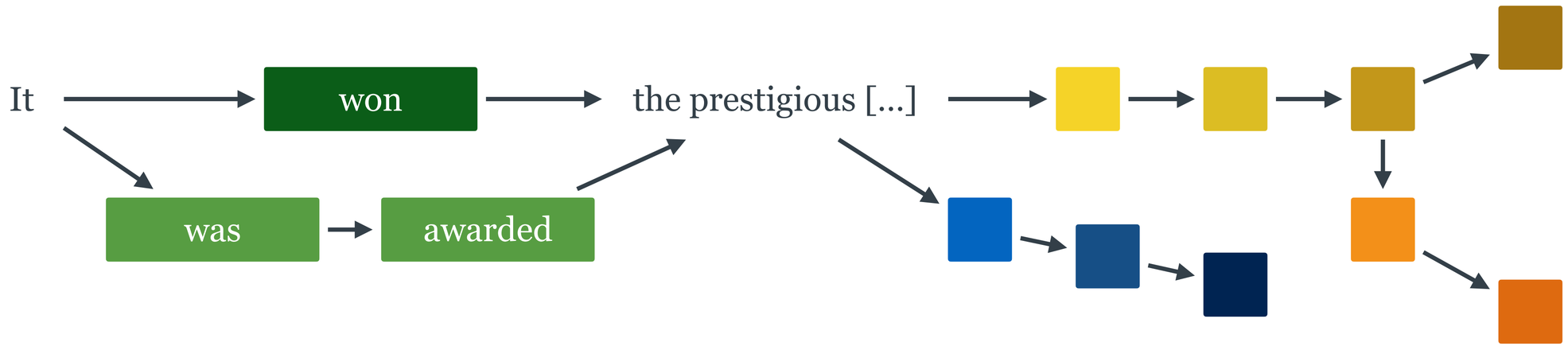


Contribution

We propose a search algorithm encoding many diverse generation options

- many: 100x ~ 1000x more outputs than beam search
- diverse: content, style, syntax, word choice, etc.

Algorithm design:



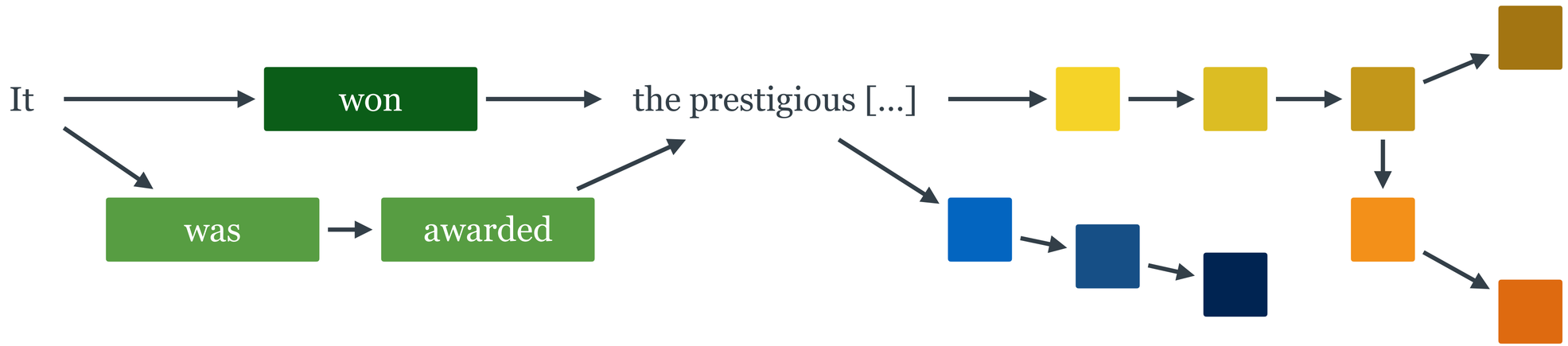
Contribution

We propose a search algorithm encoding many diverse generation options

- many: 100x ~ 1000x more outputs than beam search
- diverse: content, style, syntax, word choice, etc.

Algorithm design:

- Hypothesis recombination: combine similar content



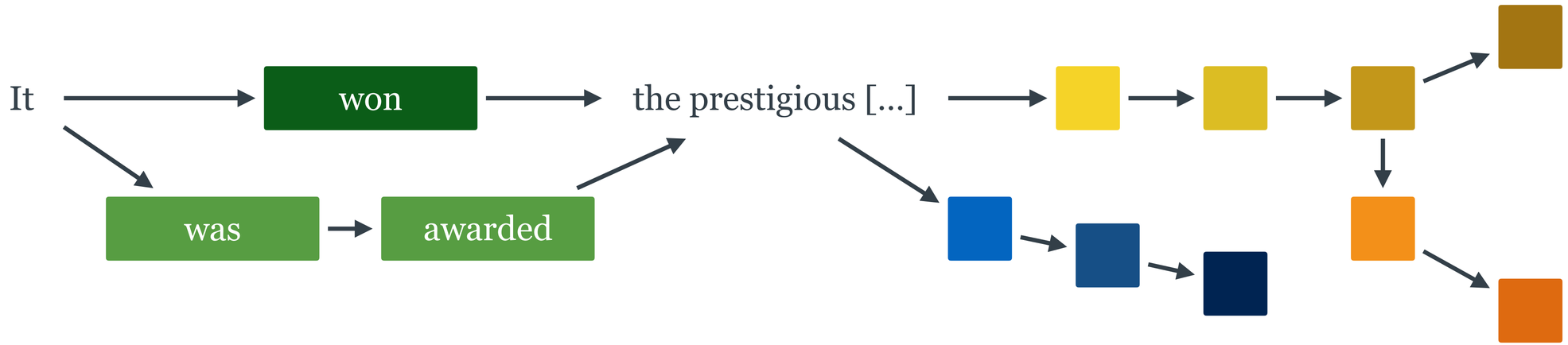
Contribution

We propose a search algorithm encoding many diverse generation options

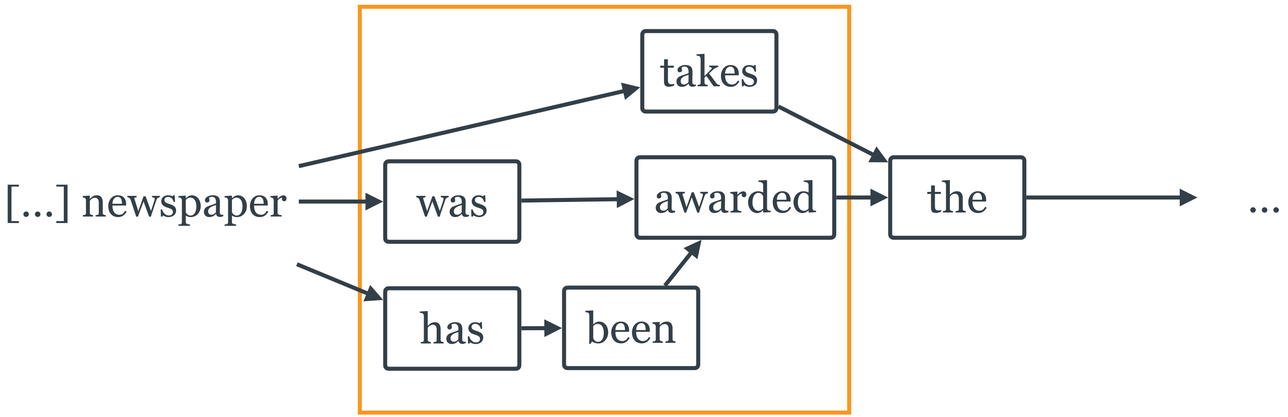
- many: 100x ~ 1000x more outputs than beam search
- diverse: content, style, syntax, word choice, etc.

Algorithm design:

- Hypothesis recombination: combine similar content
- Best-first search: flexible expansion order

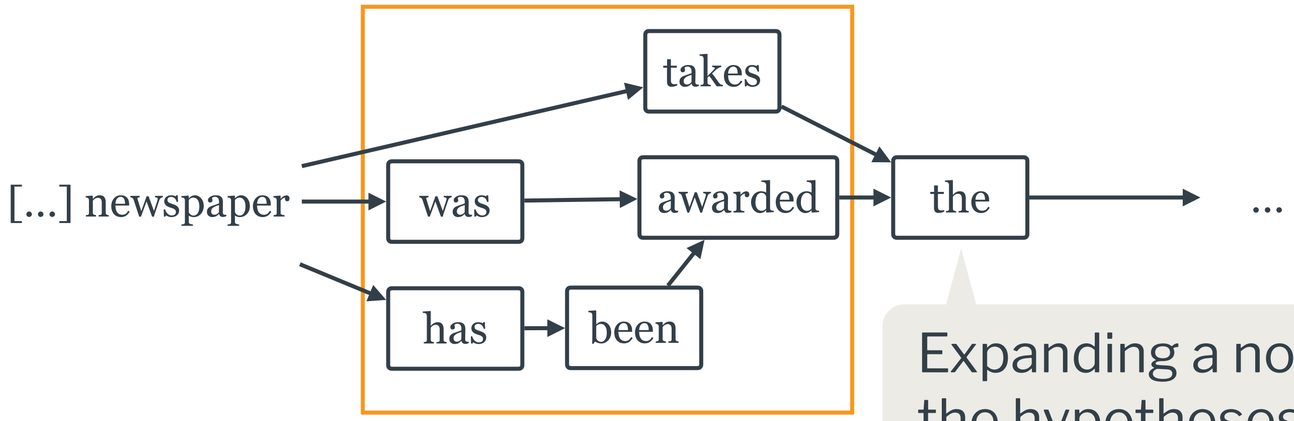


Piece 1: Hypothesis Recombination



[Och et al. 01; Zhang et al. 18]

Piece 1: Hypothesis Recombination

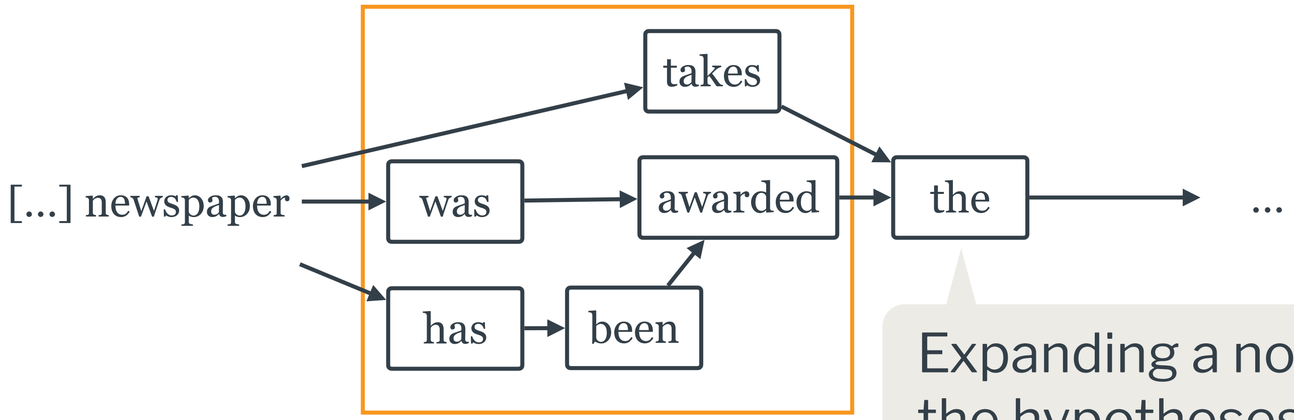


Expanding a node continues all of the hypotheses ending in that node.

Similar hypotheses can be grouped and stored in a lattice.

[Och et al. 01; Zhang et al. 18]

Piece 1: Hypothesis Recombination



Expanding a node continues all of the hypotheses ending in that node.

Similar hypotheses can be grouped and stored in a lattice.

tree lattice

[Och et al. 01; Zhang et al. 18]

Criterion for Recombination

[Zhang et al. 18]

Criterion for Recombination

Recombine partial generated hypotheses A and B if:

[Zhang et al. 18]

Criterion for Recombination

Recombine partial generated hypotheses A and B if:

- The last *n* tokens of A and B are the same (*n* = 3 or 4)

[Zhang et al. 18]

Criterion for Recombination

Recombine partial generated hypotheses A and B if:

- The last *n* tokens of A and B are the same (*n* = 3 or 4)
- A and B are roughly the same length

[Zhang et al. 18]

Criterion for Recombination

Recombine partial generated hypotheses A and B if:

- The last *n* tokens of A and B are the same (*n* = 3 or 4)
- A and B are roughly the same length

Prefix A: […] newspaper was awarded the 2015 Pulitzer

Prefix B: […] newspaper has won the 2015 Pulitzer

Assumption: if these criteria are met, the rest of the summary will be similar:

$$P(\text{prize}|A) \approx P(\text{prize}|B)$$

[Zhang et al. 18]

Criterion for Recombination

Recombine partial generated hypotheses A and B if:

- The last *n* tokens of A and B are the same (*n* = 3 or 4)
- A and B are roughly the same length

Prefix A: […] newspaper was awarded the 2015 Pulitzer

Prefix B: […] newspaper has won the 2015 Pulitzer

Assumption: if these criteria are met, the rest of the summary will be similar:

$$P(\text{prize}|A) \approx P(\text{prize}|B)$$

For summarization: **~70%** of time the greedy completion is exactly the same.

When these distributions match, merging states in the lattice is completely okay!

[Zhang et al. 18]

Executing Recombination

Recombine partial generated hypotheses A and B if:

- The last *n* tokens of A and B are the same (*n* = 3 or 4)
- A and B are roughly the same length

Running example (*n* = 3):

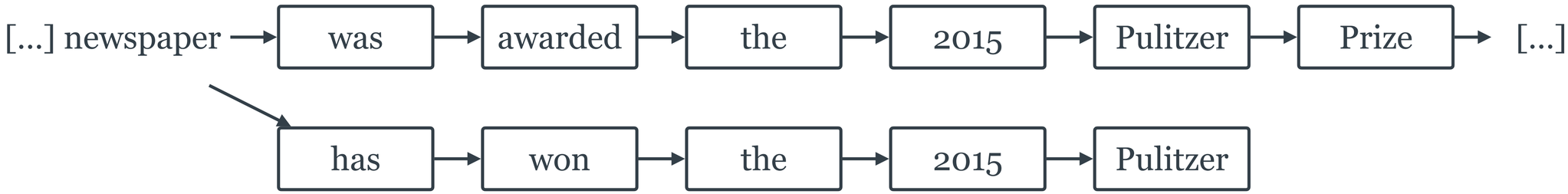
[Zhang et al. 18]

Executing Recombination

Recombine partial generated hypotheses A and B if:

- The last *n* tokens of A and B are the same (*n* = 3 or 4)
- A and B are roughly the same length

Running example (*n* = 3):



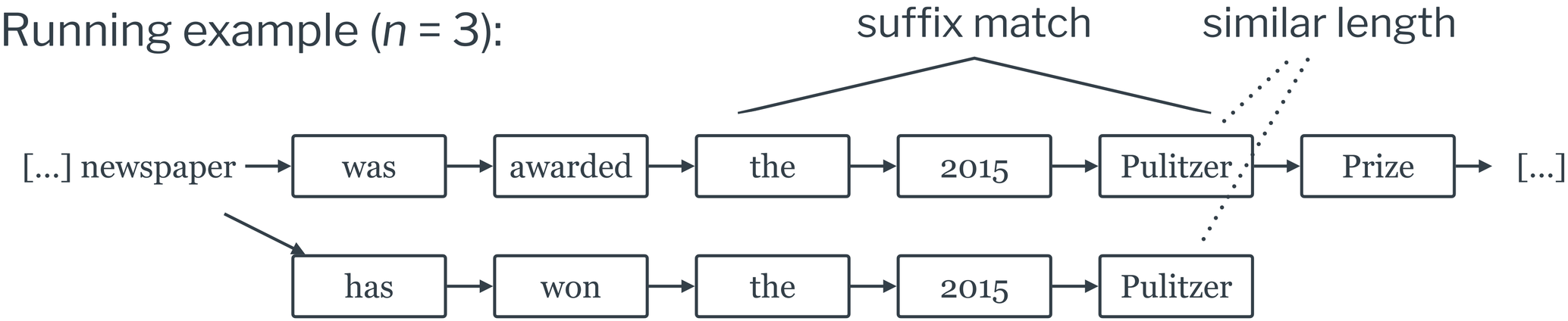
[Zhang et al. 18]

Executing Recombination

Recombine partial generated hypotheses A and B if:

- The last *n* tokens of A and B are the same (*n* = 3 or 4)
- A and B are roughly the same length

Running example (*n* = 3):



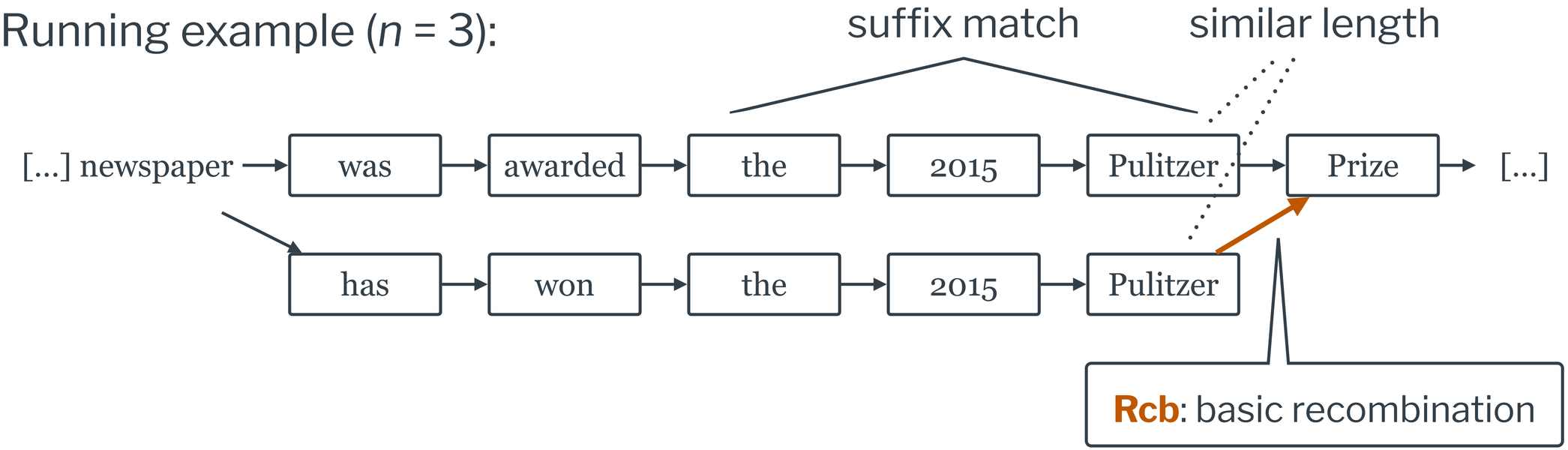
[Zhang et al. 18]

Executing Recombination

Recombine partial generated hypotheses A and B if:

- The last *n* tokens of A and B are the same (*n* = 3 or 4)
- A and B are roughly the same length

Running example (*n* = 3):



[Zhang et al. 18]

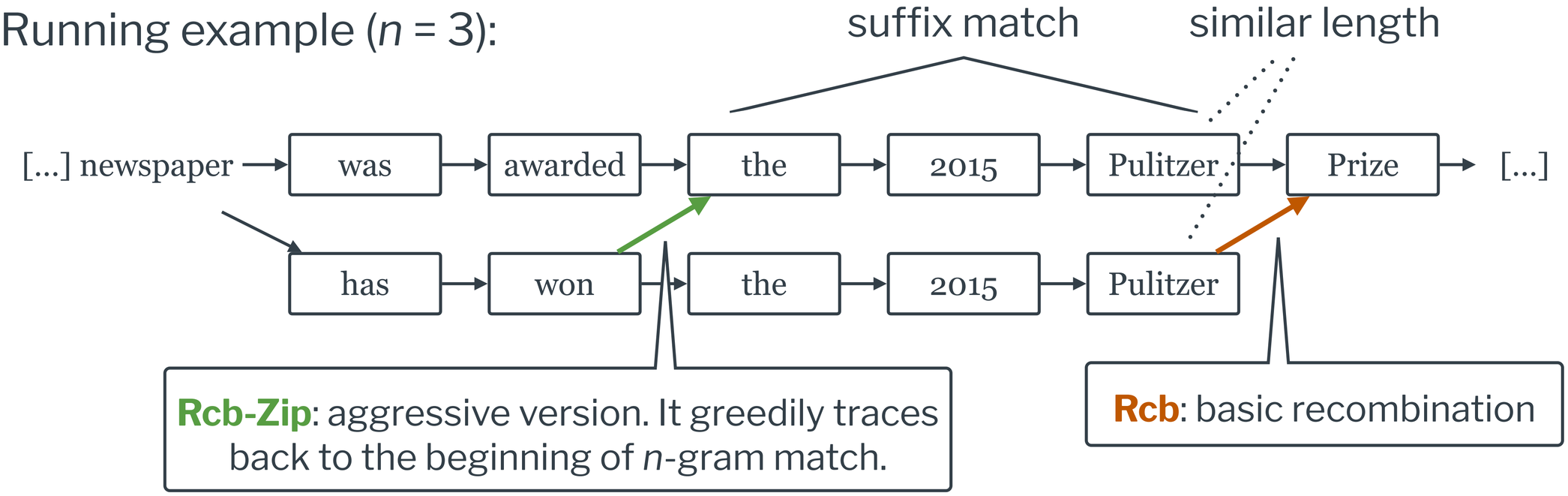
Executing Recombination

Recombine partial generated hypotheses A and B if:

- The last *n* tokens of A and B are the same (*n* = 3 or 4)
- A and B are roughly the same length

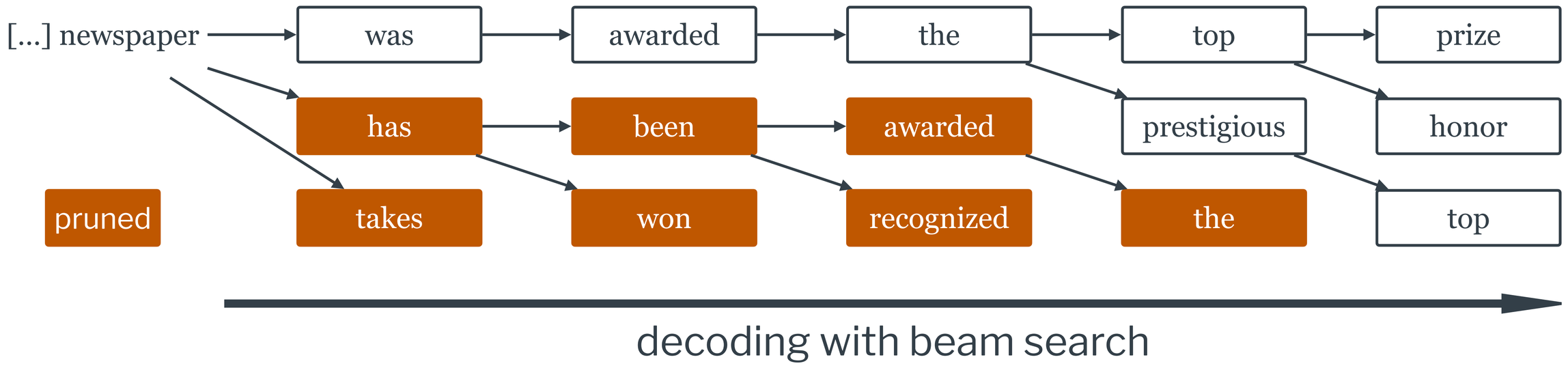
Check the paper for more formal and algorithmic description!

Running example (*n* = 3):



[Zhang et al. 18]

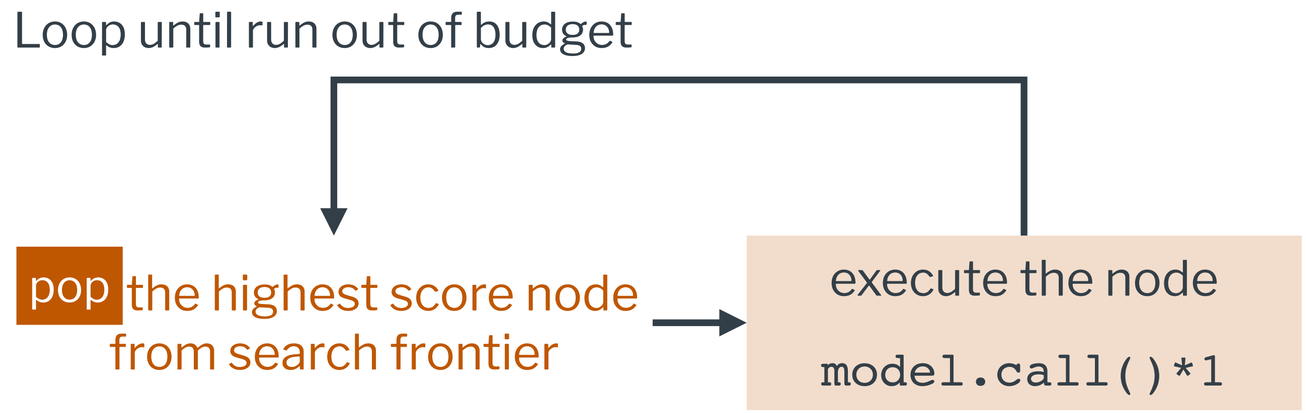
Heavy Pruning in Beam Search



Every **orange step** ultimately led to a prediction that got pruned.

Piece 2: Best-first Search

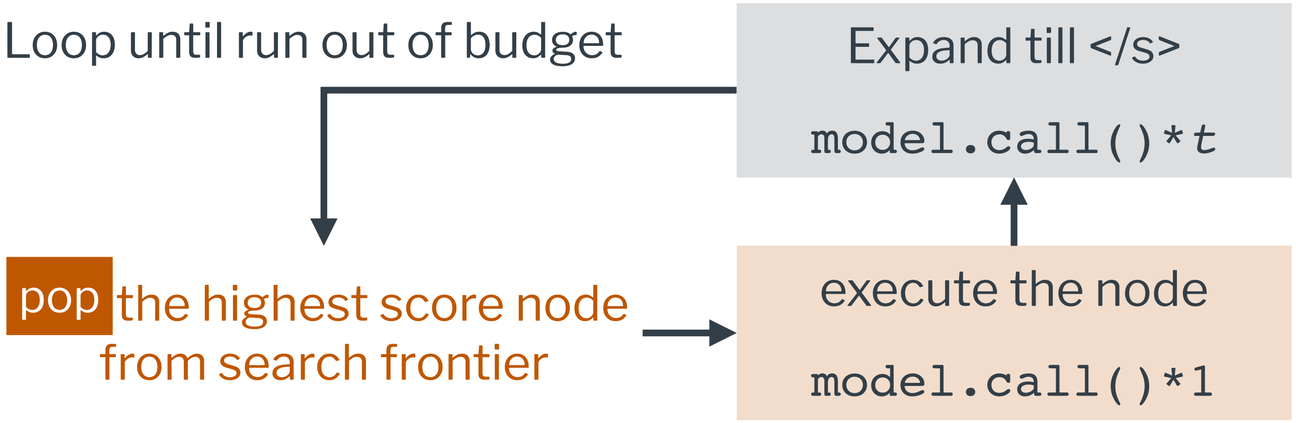
best-first search



Piece 2: Best-first Search

- Desired: **every explored state** is on some finished path
- Solution: greedily expansion till </s>

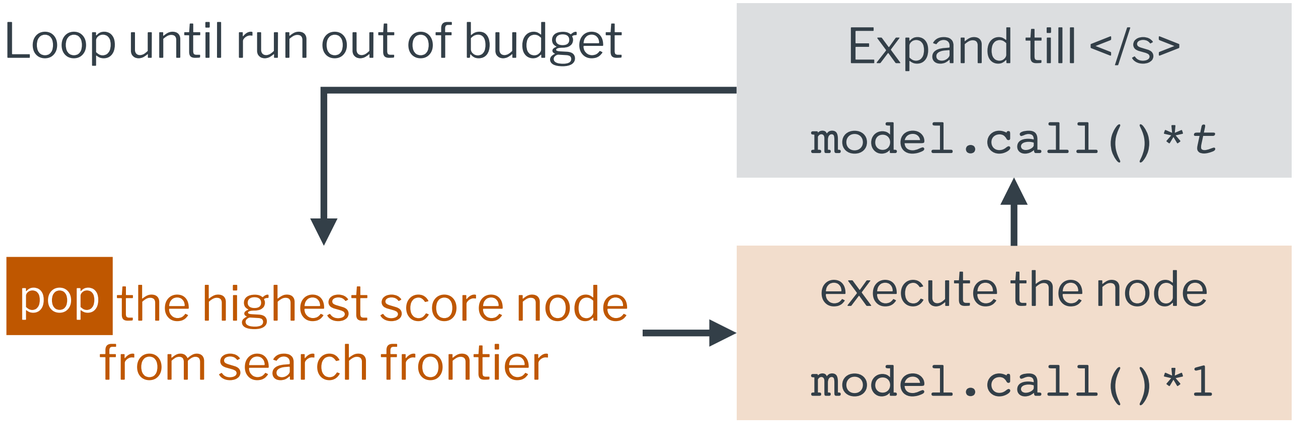
Modified best-first search
with **depth-first completion**



Piece 2: Best-first Search

- Desired: **every explored state** is on some finished path
- Solution: greedily expansion till </s>

Modified best-first search
with **depth-first completion**

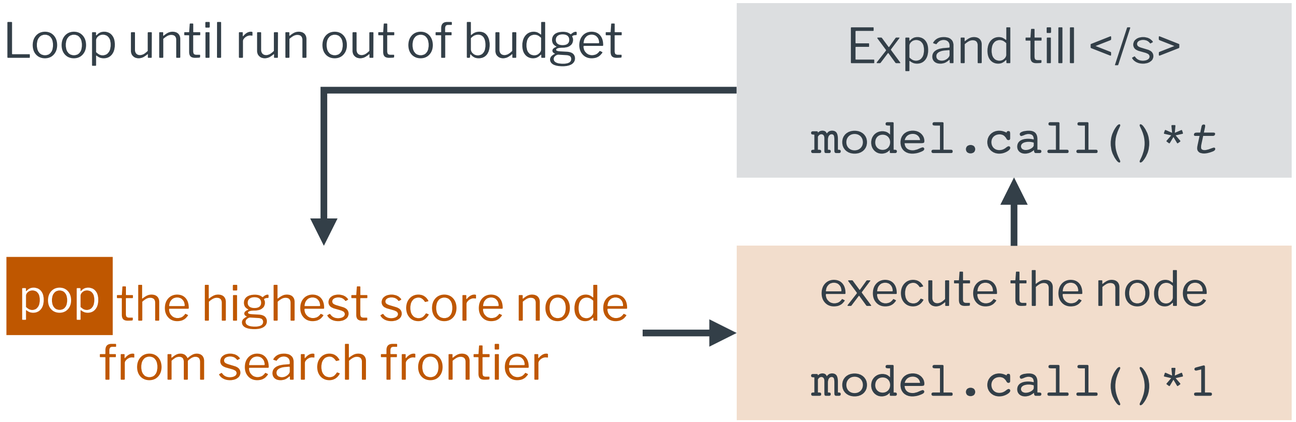


(0.0, <s>)

Piece 2: Best-first Search

- Desired: **every explored state** is on some finished path
- Solution: greedily expansion till </s>

Modified best-first search
with **depth-first completion**



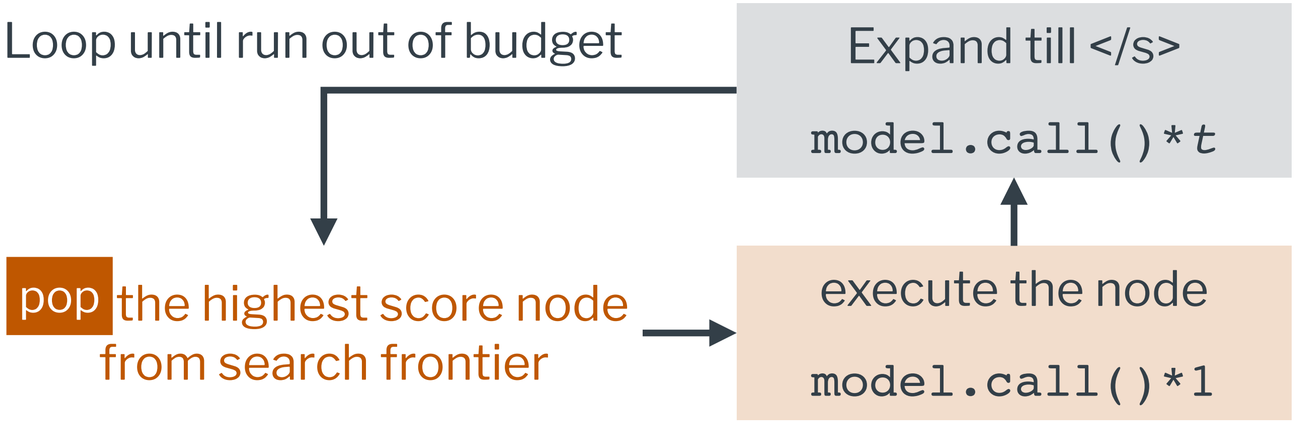
pop

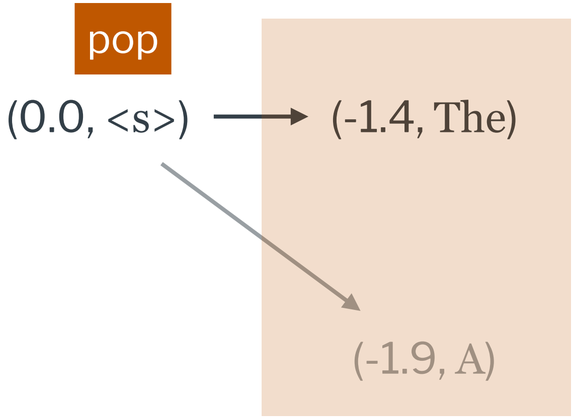
(0.0, <s>)

Piece 2: Best-first Search

- Desired: **every explored state** is on some finished path
- Solution: greedily expansion till </s>

Modified best-first search
with **depth-first completion**





Piece 2: Best-first Search

- Desired: **every explored state** is on some finished path
- Solution: greedily expansion till </s>

Loop until run out of budget

Expand till </s>

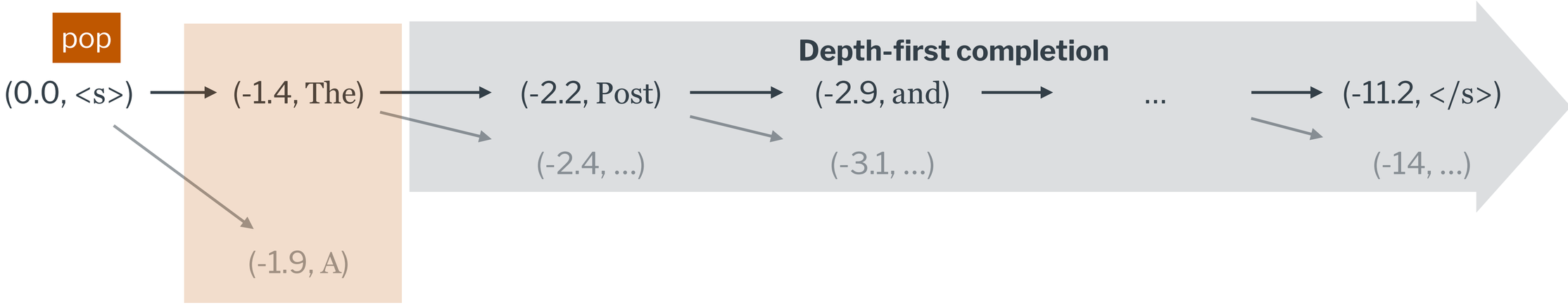
`model.call()*t`

pop the highest score node from search frontier

execute the node

`model.call()*1`

Modified best-first search with **depth-first completion**



Piece 2: Best-first Search

- Desired: **every explored state** is on some finished path
- Solution: greedily expansion till </s>

Loop until run out of budget

Expand till </s>

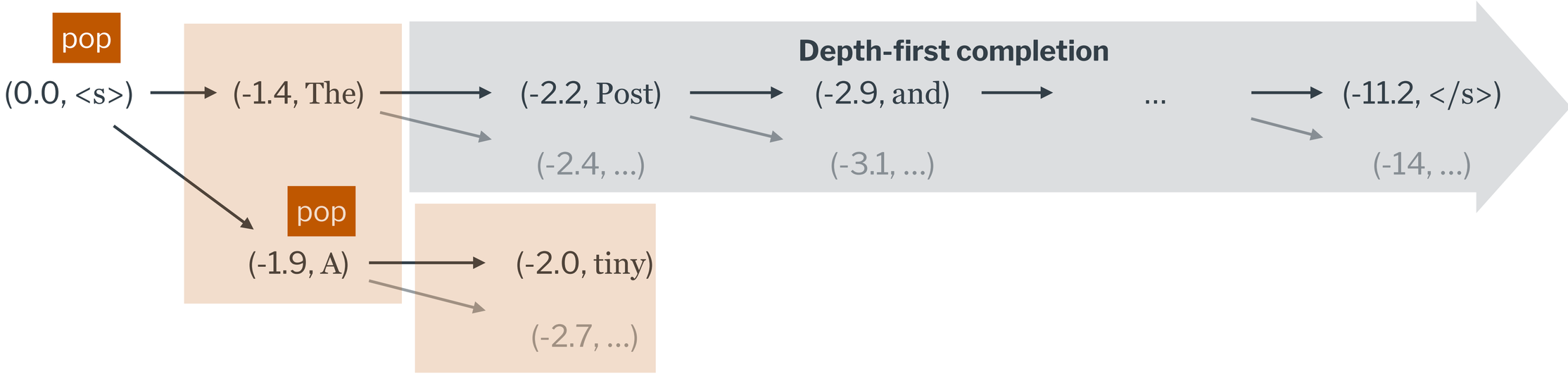
`model.call()*t`

Modified best-first search
with **depth-first completion**

pop the highest score node from search frontier

execute the node

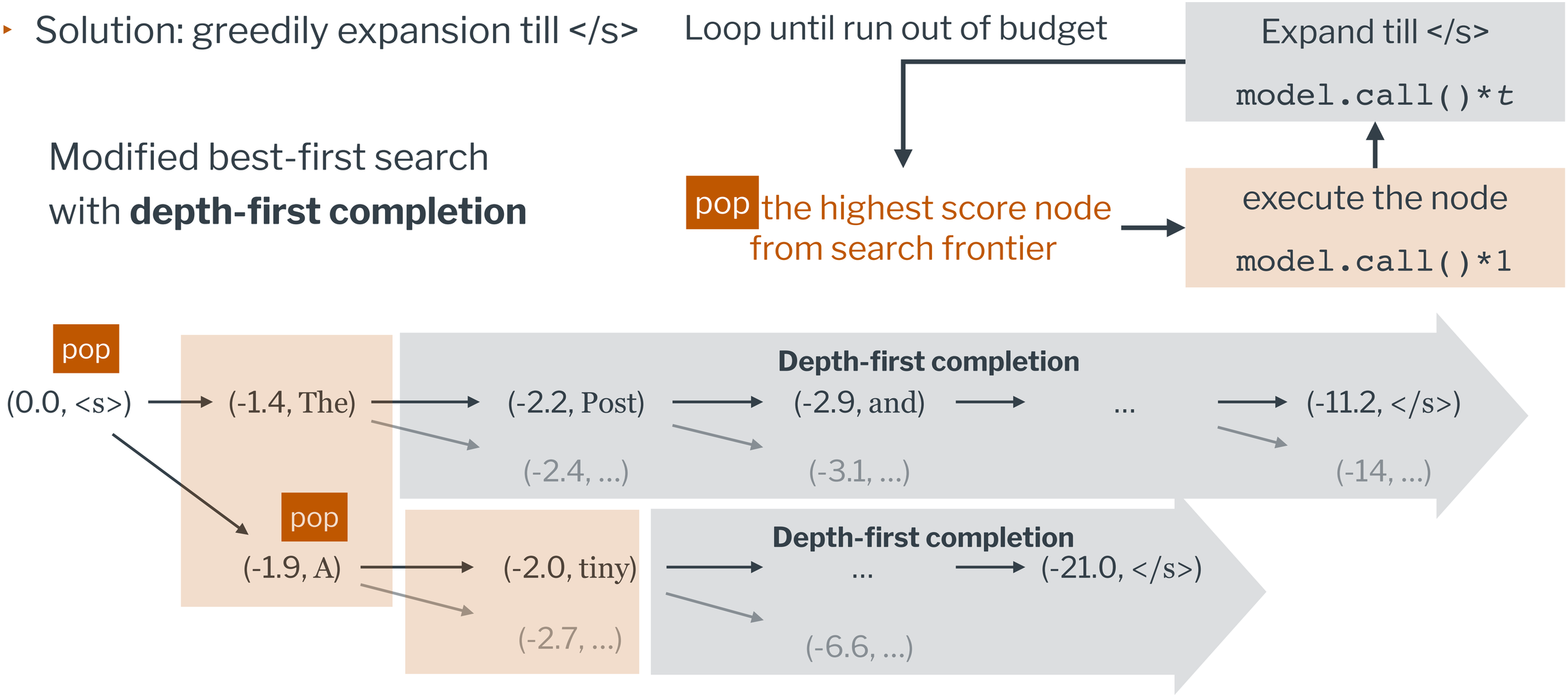
`model.call()*1`



Piece 2: Best-first Search

- Desired: **every explored state** is on some finished path
- Solution: greedily expansion till </s>

Modified best-first search
with **depth-first completion**



Putting it all together

- Full algorithm: path recombination + best-first search
- Running example (adjacent nodes are combined for visualization)

Putting it all together

- Full algorithm: path recombination + best-first search
- Running example (adjacent nodes are combined for visualization)

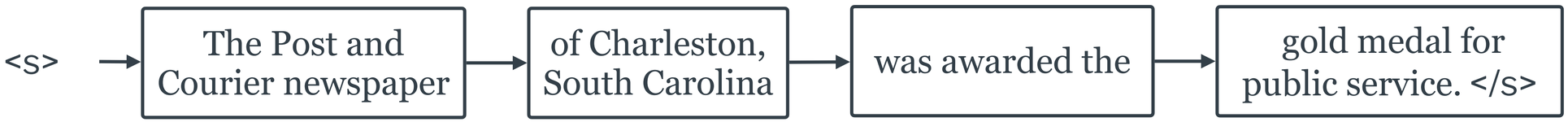
<s>

Putting it all together

- Full algorithm: path recombination + best-first search
- Running example (adjacent nodes are combined for visualization)

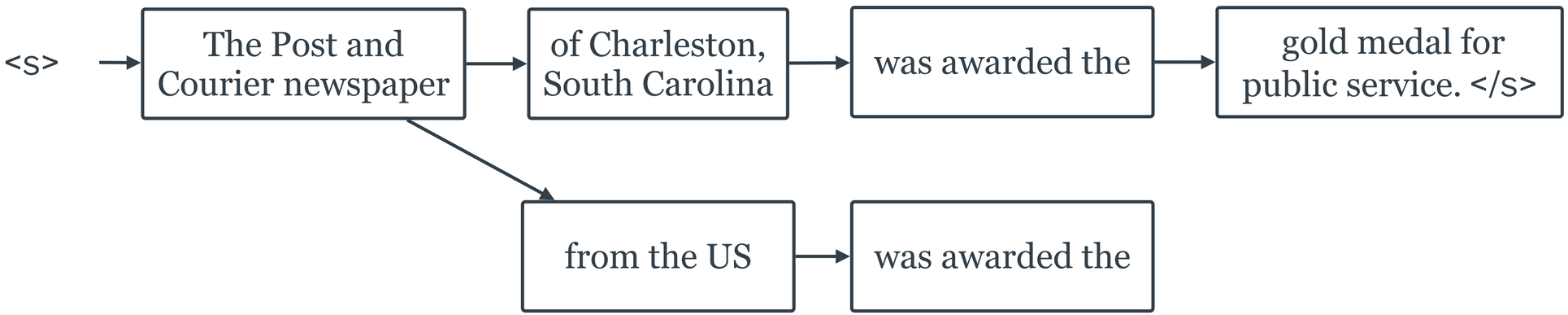
Putting it all together

- Full algorithm: path recombination + best-first search
- Running example (adjacent nodes are combined for visualization)



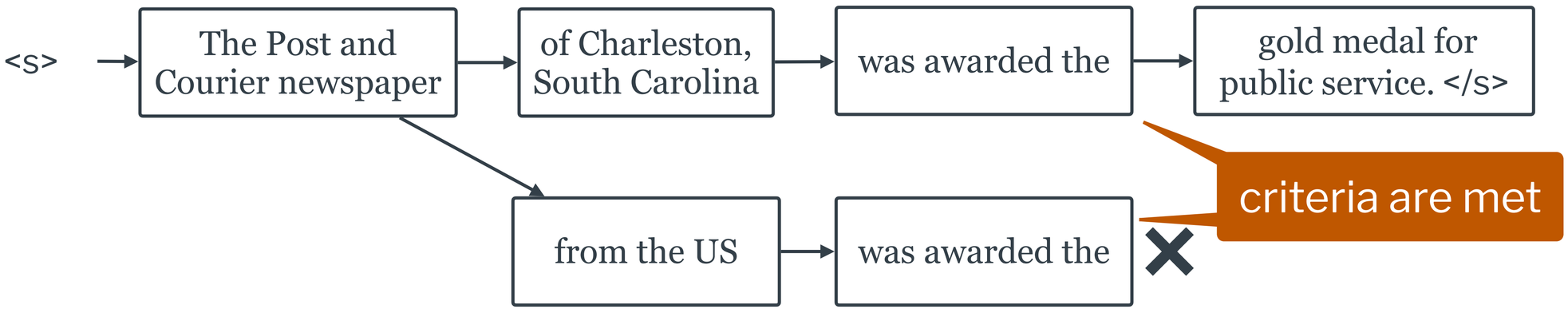
Putting it all together

- Full algorithm: path recombination + best-first search
- Running example (adjacent nodes are combined for visualization)



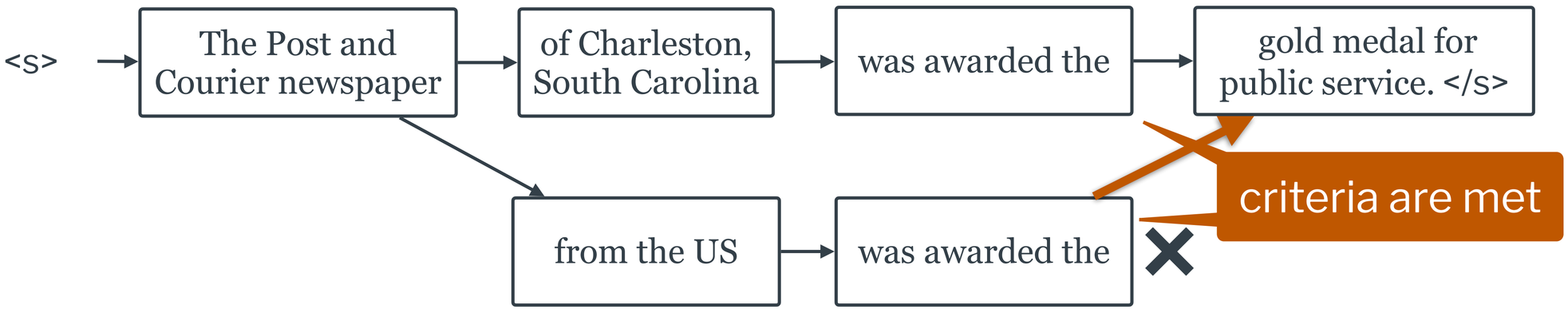
Putting it all together

- Full algorithm: path recombination + best-first search
- Running example (adjacent nodes are combined for visualization)



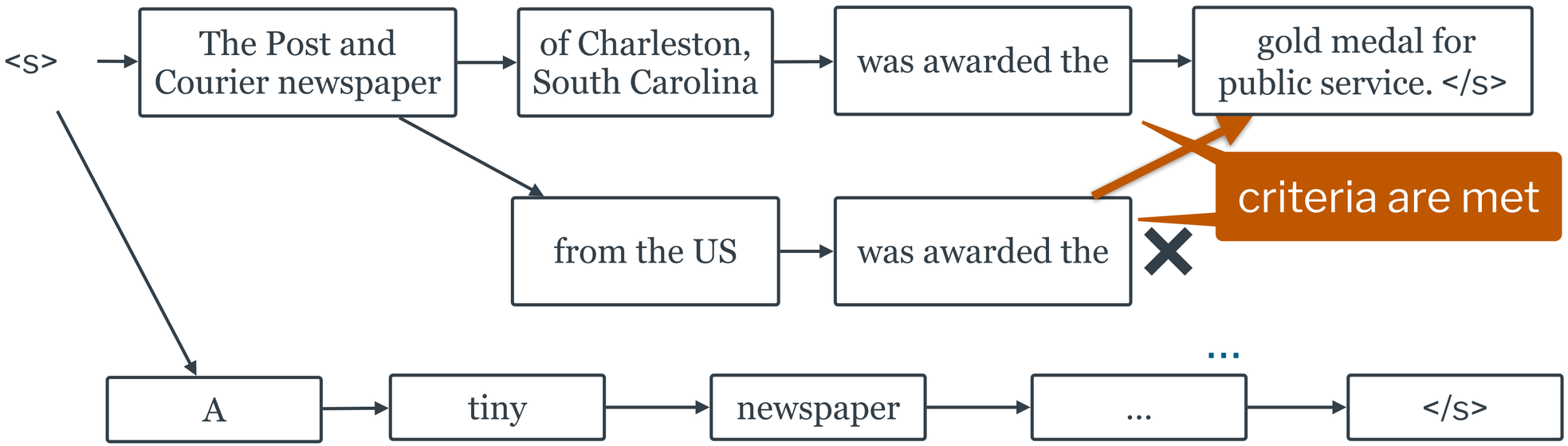
Putting it all together

- Full algorithm: path recombination + best-first search
- Running example (adjacent nodes are combined for visualization)



Putting it all together

- Full algorithm: path recombination + best-first search
- Running example (adjacent nodes are combined for visualization)



Evaluation

Goal: single best output

Evaluation

Goal: ~~single~~ best output

↓

many

Diversity

Def: many = large number, unique

Metrics:

1. # of unique paths found ↑
2. Self-BLEU
3. # of novel *n*-gram
4. …

Evaluation

Goal: ~~single~~ ~~best~~ output

many good

Diversity

Def: many = large number, unique

Metrics:

1. # of unique paths found ↑
2. Self-BLEU
3. # of novel *n*-gram
4. …

Quality

Def: good = relevant, grammatical, high oracle

Metrics:

1. Oracle ROUGE/BLEU
2. Average sample ROUGE/BLEU
3. Grammatical errors (%)

Evaluation

Goal: ~~single~~ ~~best~~ output

many good

Diversity

Def: many = large number, unique

Metrics:

1. # of unique paths found ↑
2. Self-BLEU
3. # of novel *n*-gram
4. …

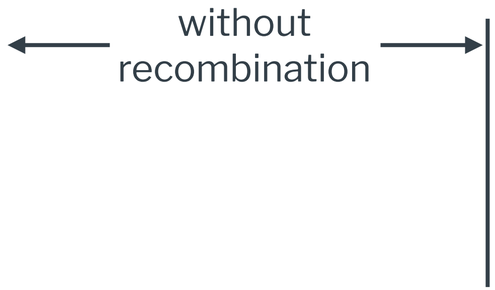
Quality

Def: good = relevant, grammatical, high oracle

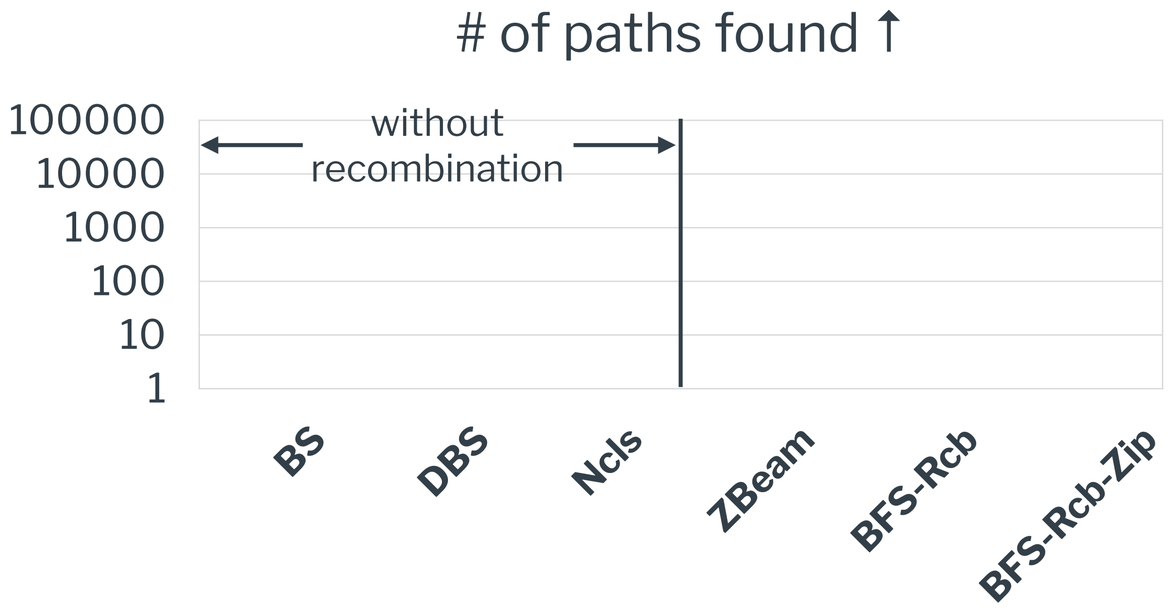
Metrics:

1. Oracle ROUGE/BLEU
2. Average sample ROUGE/BLEU
3. Grammatical errors (%)

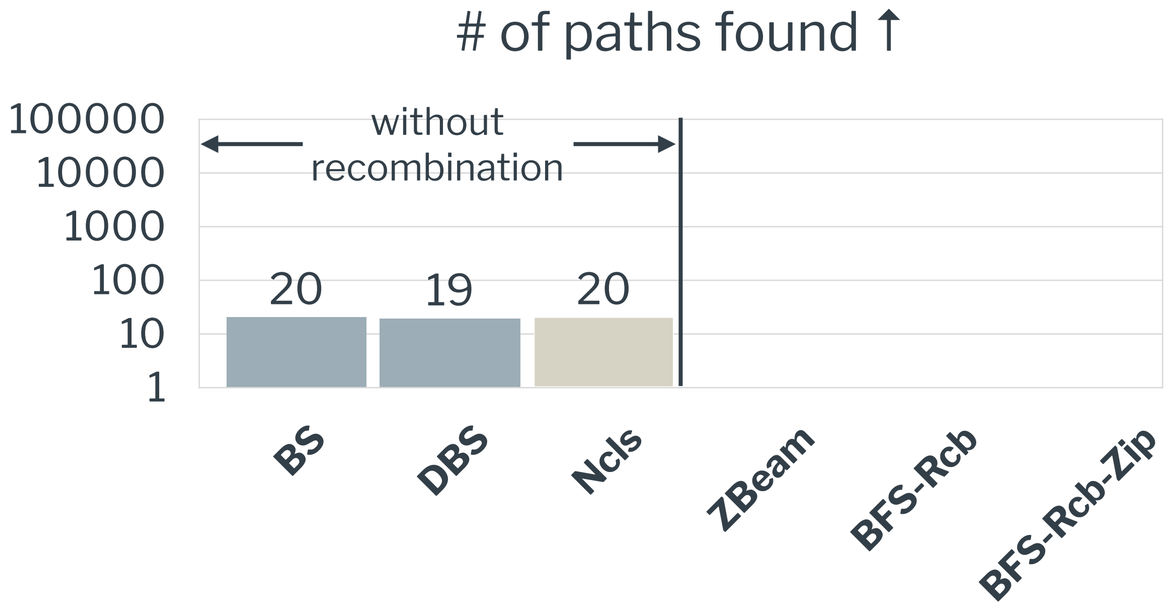
Results on Text Summarization



Results on Text Summarization



Results on Text Summarization

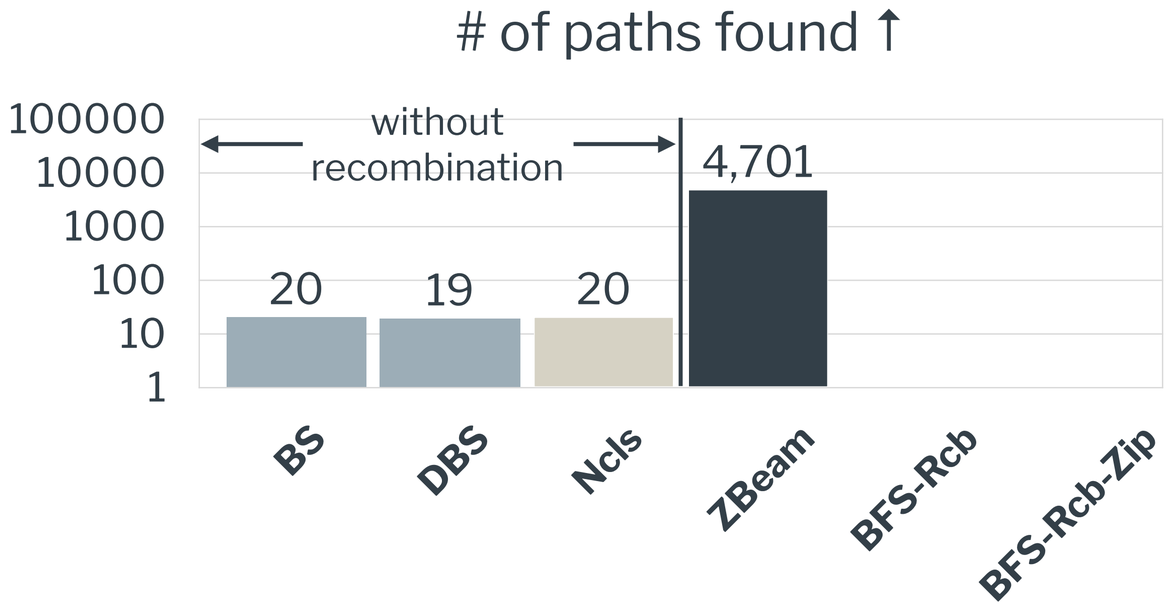


BS: beam search

DBS: diverse beam search [Vijayakumar et al. 18]

Ncls: nucleus sampling [Holtzman et al. 19]

Results on Text Summarization



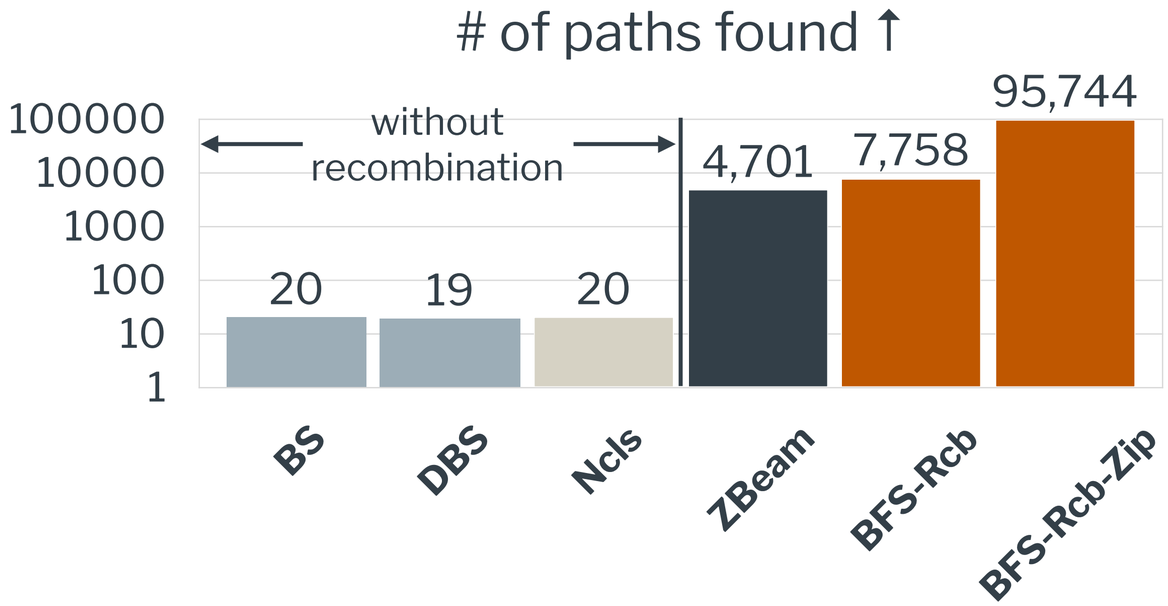
BS: beam search

DBS: diverse beam search [Vijayakumar et al. 18]

Ncls: nucleus sampling [Holtzman et al. 19]

ZBeam: beam search + path recombination [Zhang et al. 18]

Results on Text Summarization



BS: beam search

DBS: diverse beam search [Vijayakumar et al. 18]

Ncls: nucleus sampling [Holtzman et al. 19]

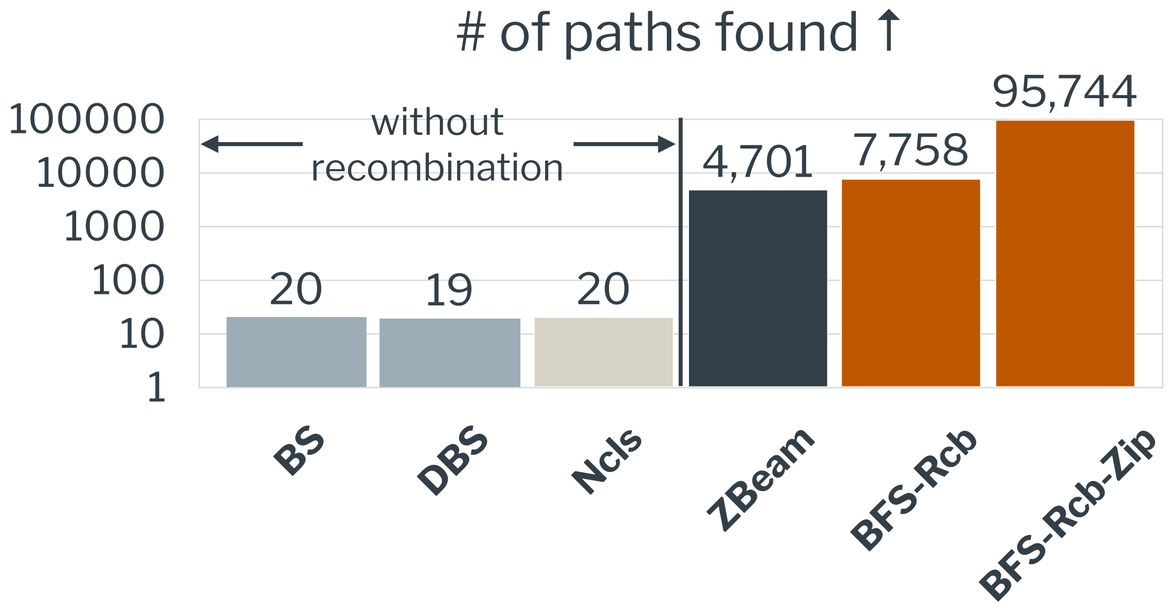
ZBeam: beam search + path recombination [Zhang et al. 18]

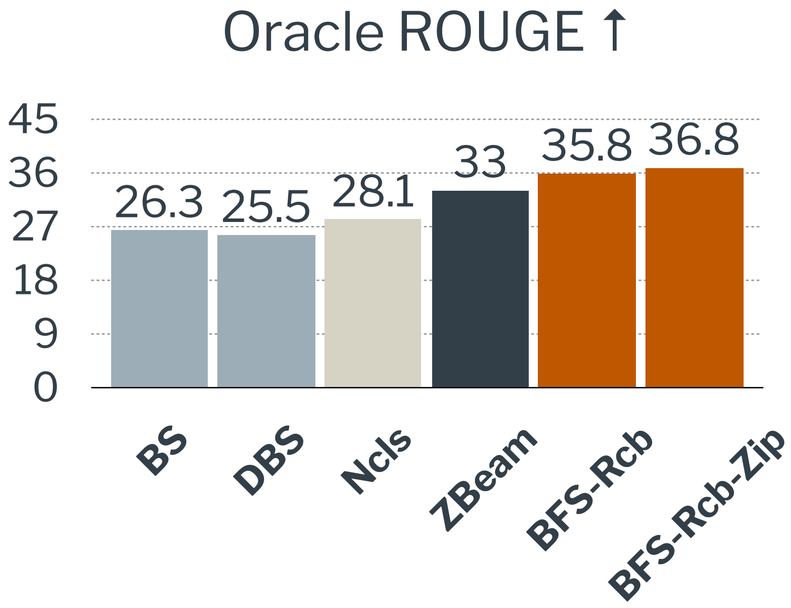
BFS-Rcb: BFS with path recombination

BFS-Rcb-Zip: aggressive version of BFS-Rcb

- A lot more unique paths found!

Results on Text Summarization





BS: beam search

DBS: diverse beam search [Vijayakumar et al. 18]

Ncls: nucleus sampling [Holtzman et al. 19]

ZBeam: beam search + path recombination [Zhang et al. 18]

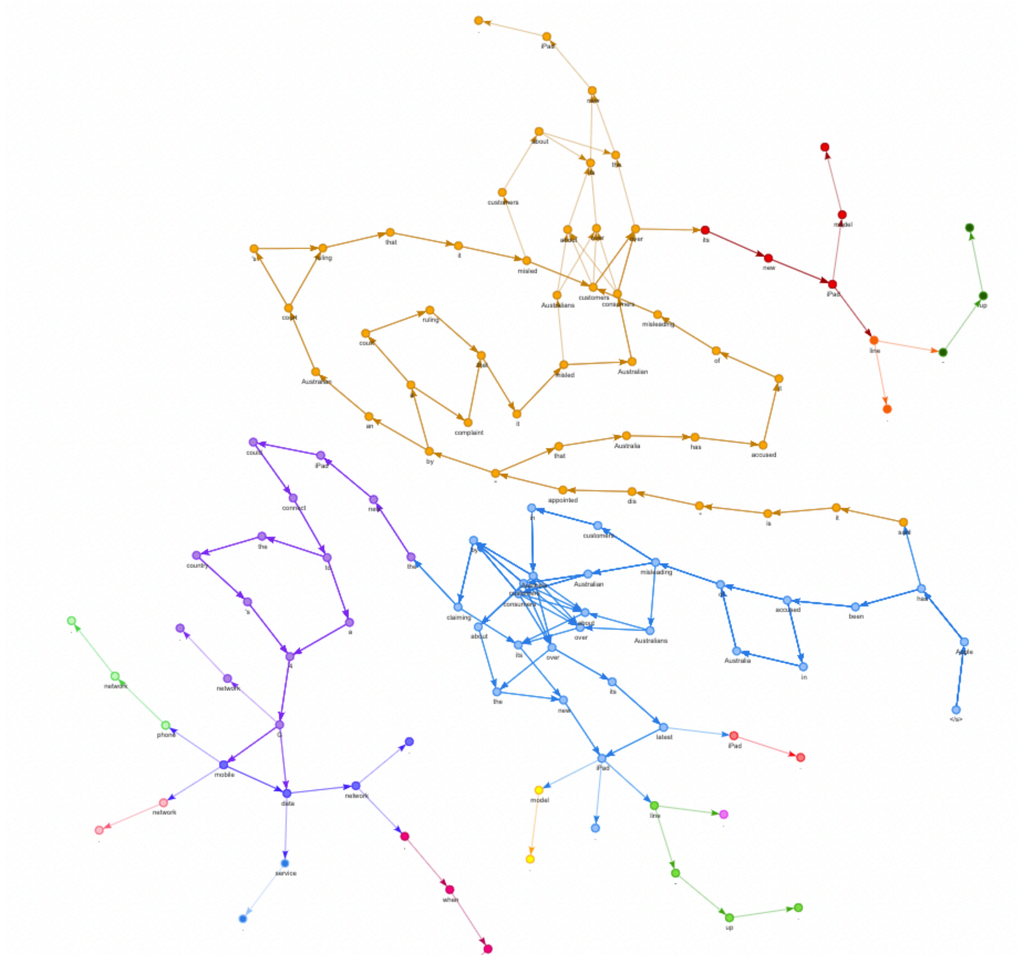
BFS-Rcb: BFS with path recombination

BFS-Rcb-Zip: aggressive version of BFS-Rcb

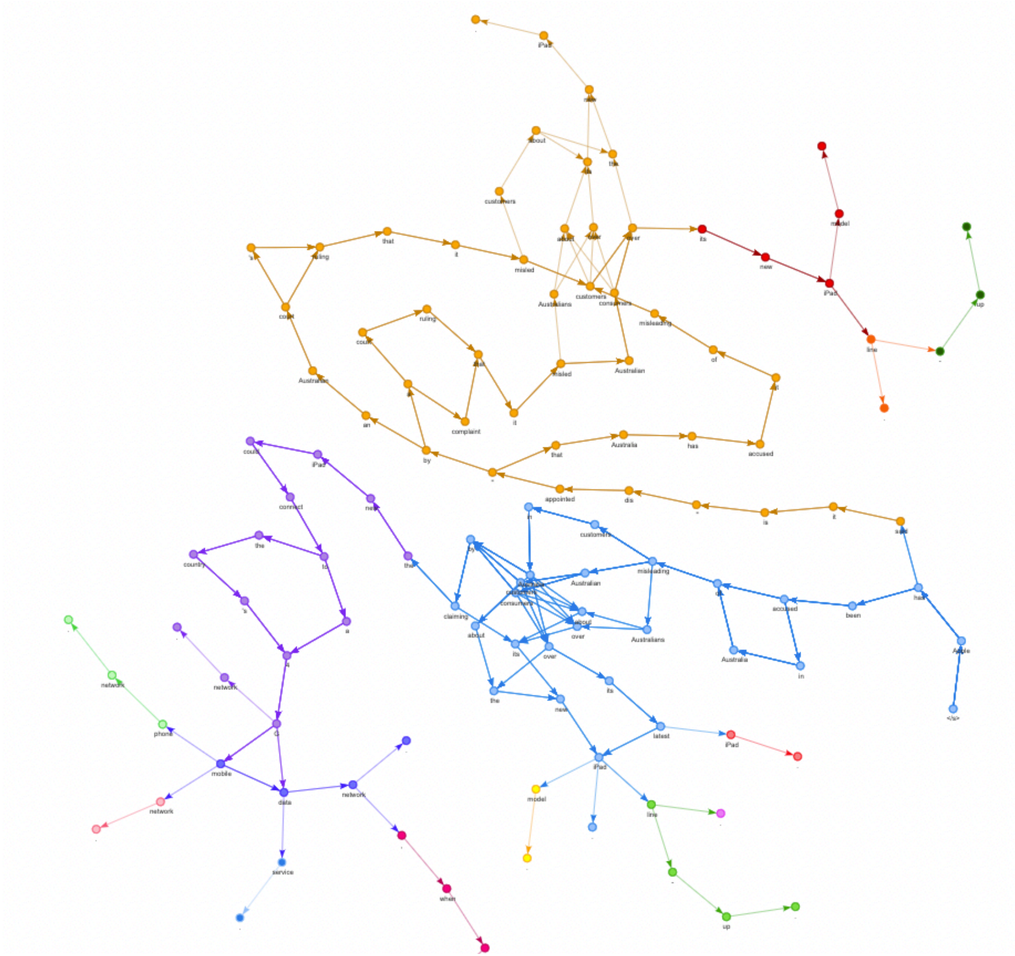
- A lot more unique paths found!
- High-quality oracle summaries encoded

BFS with recombination are overall the best in diversity

Visualization

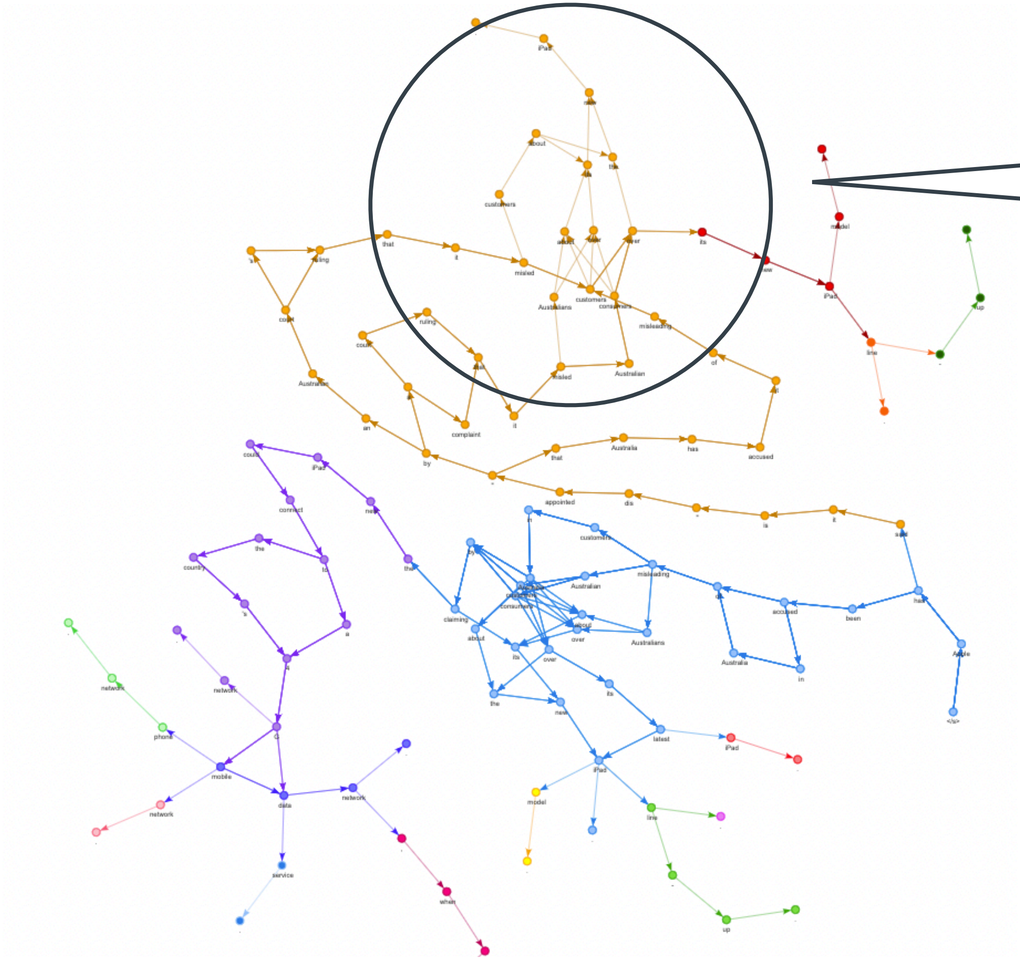


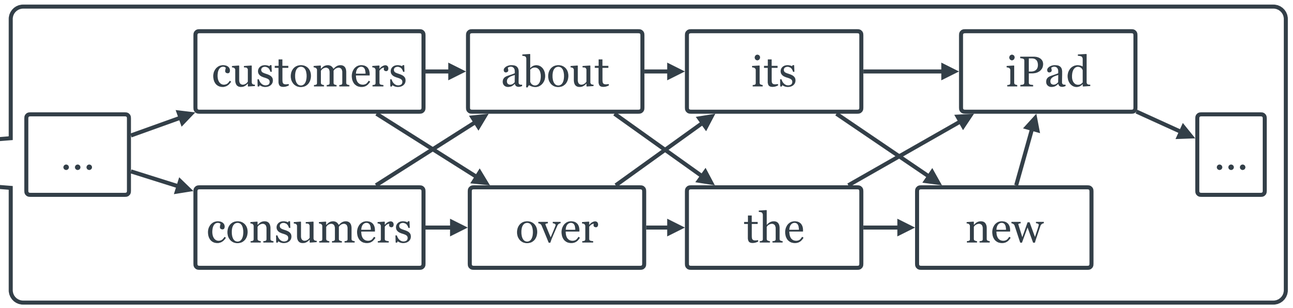
Visualization



Encode exponentially many summaries in a compact space

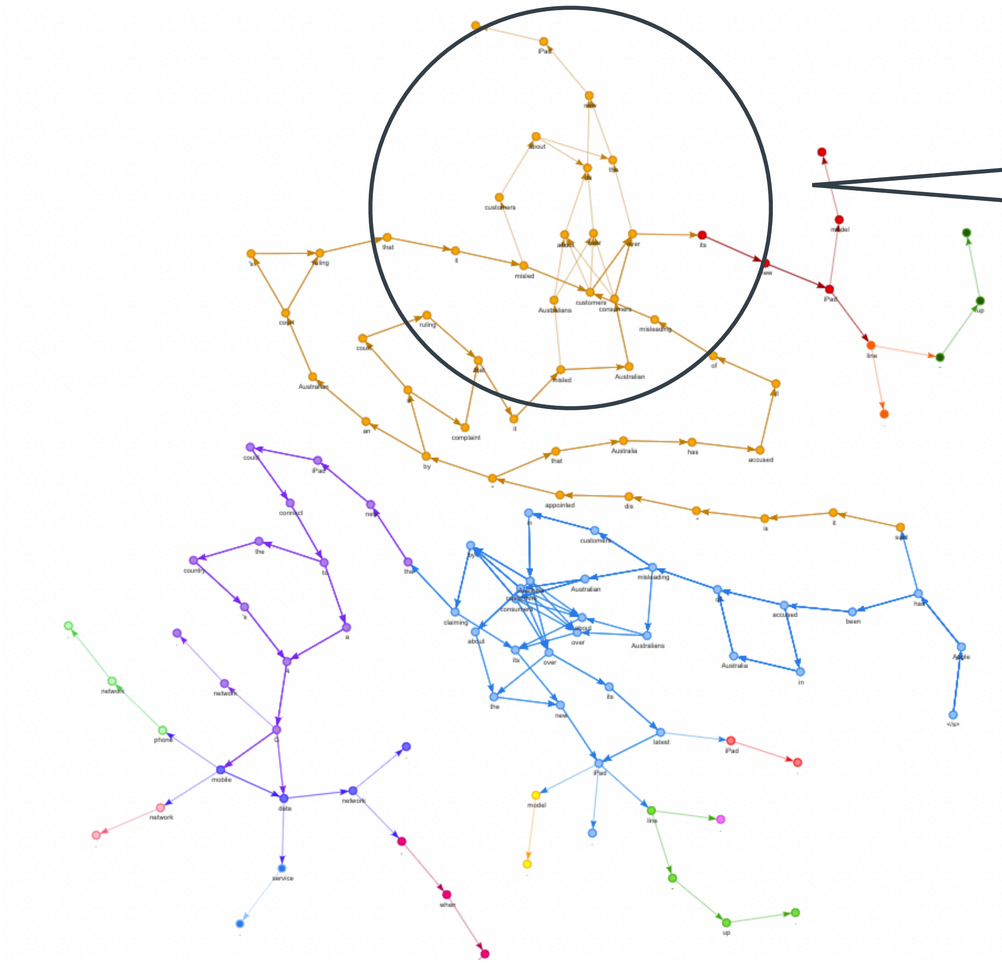
Visualization

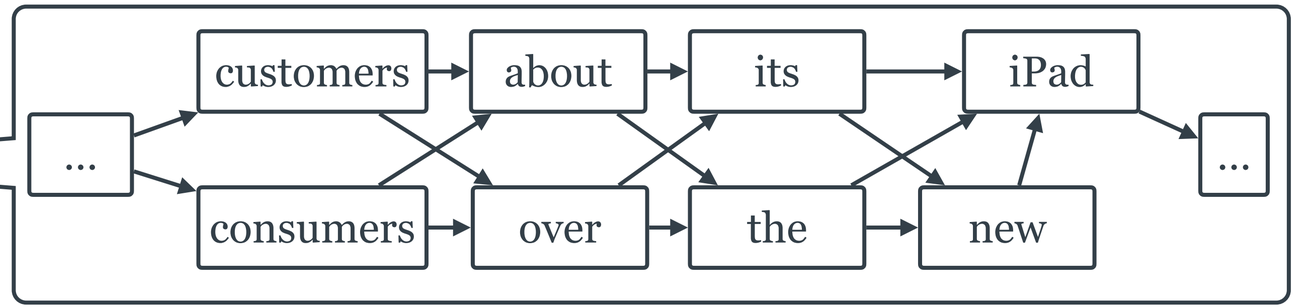




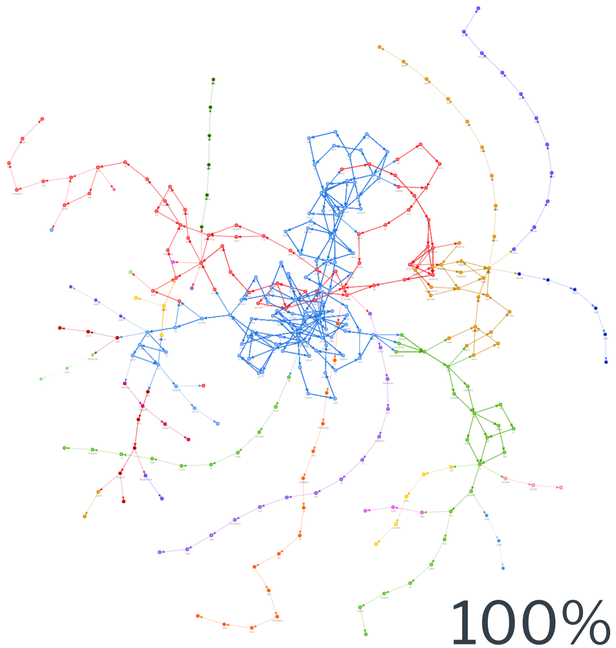
Encode exponentially many summaries in a compact space

Visualization

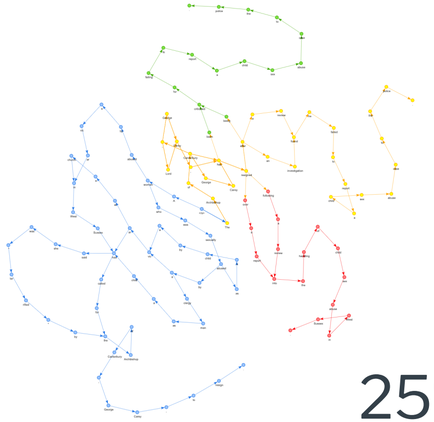




Encode exponentially many summaries in a compact space



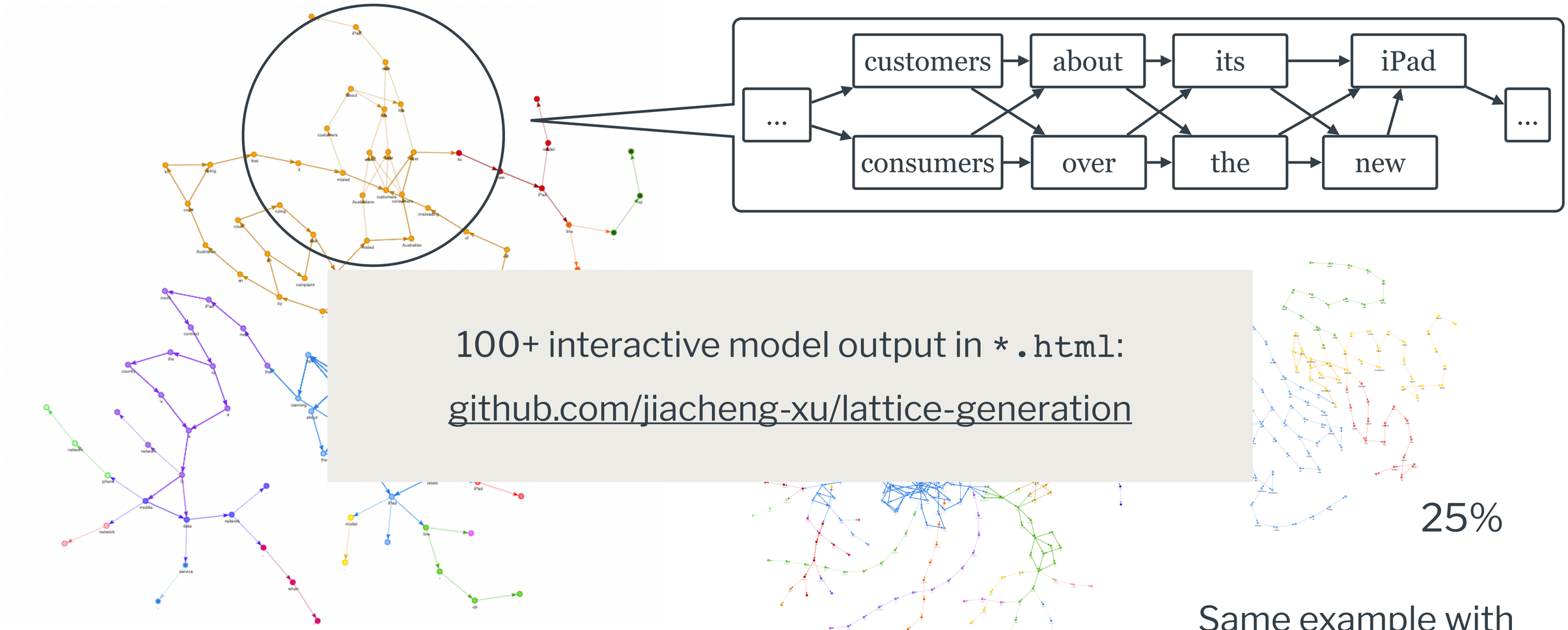
100%



25%

Same example with different budget

Visualization



100+ interactive model output in `*.html`:

github.com/jiacheng-xu/lattice-generation

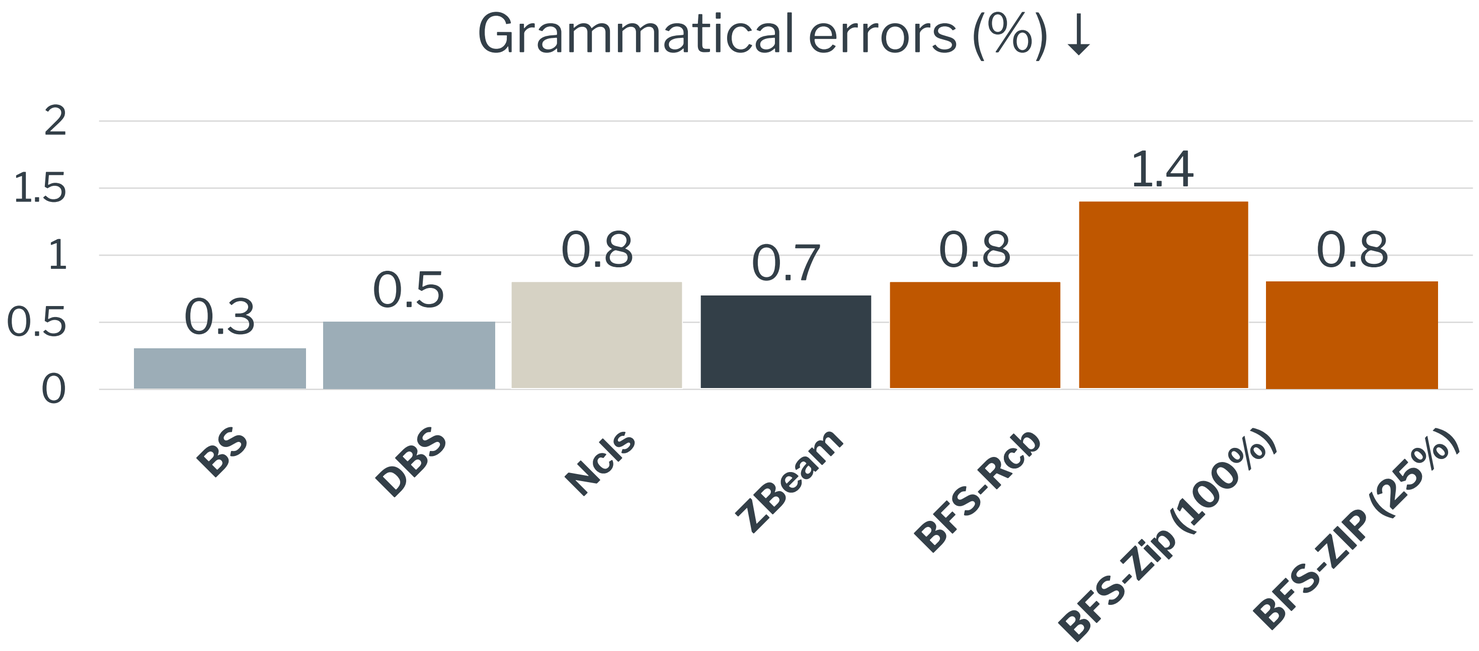
Encode exponentially many summaries in a compact space

100%

25%

Same example with different budget

Tradeoff between quality & diversity



Our most aggressive merging does introduce some grammatical errors. We can …

- choose *right* budget
- use better merging heuristics

Goals for Lattices

Our generation systems can already encode lots of good options.

Goals for Lattices

Our generation systems can already encode lots of good options.

- Rerank our generated summaries/translations and pick out the desired ones?
- Users control/correct the system on-the-fly, with the system learning those?

Goals for Lattices

Our generation systems can already encode lots of good options.

- Rerank our generated summaries/translations and pick out the desired ones?
- Users control/correct the system on-the-fly, with the system learning those?

Applications: factuality, controllable dialogue, diverse paraphrasing, and more!

Code, visualization, more:
github.com/jiacheng-xu/lattice-generation

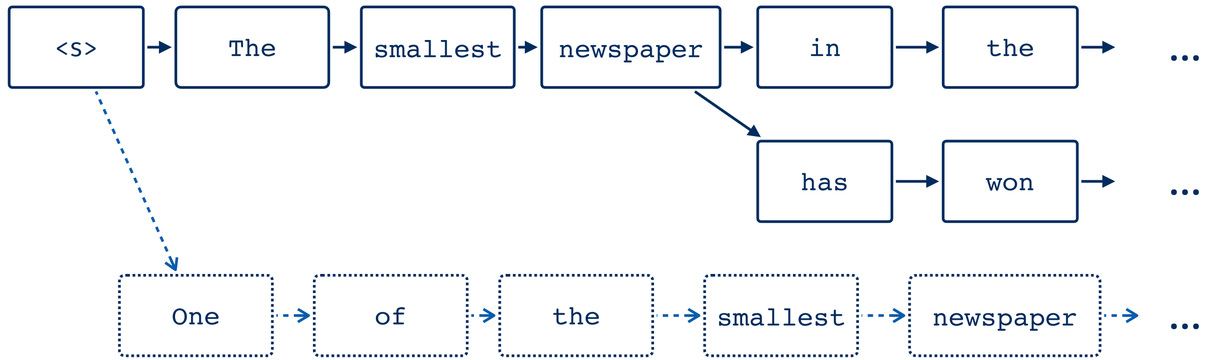
Best-*k* Search Algorithm for Neural Text Generation

Jiacheng Xu, *Caiming Xiong, Silvio Savarese, Yingbo Zhou*

Salesforce AI Research

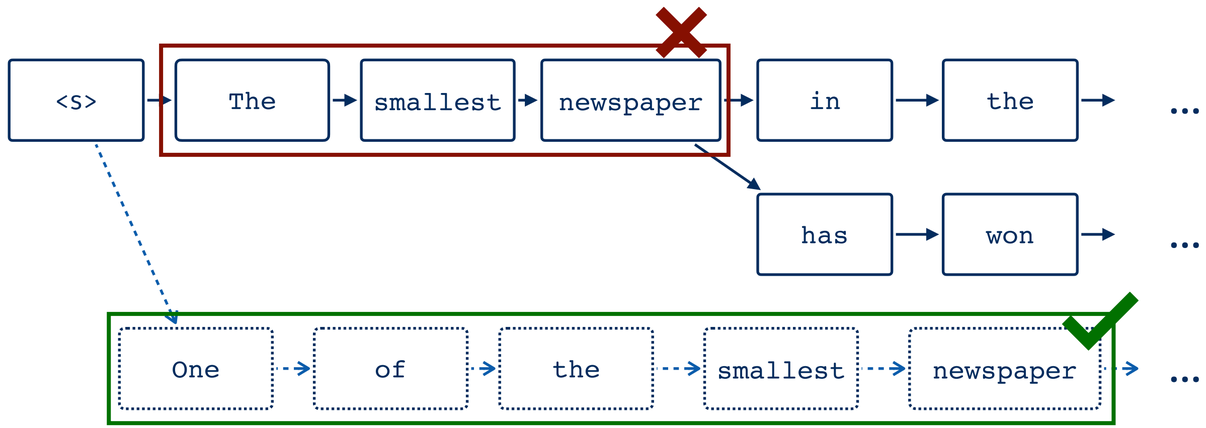
Search for Diverse Outputs

Existing approaches



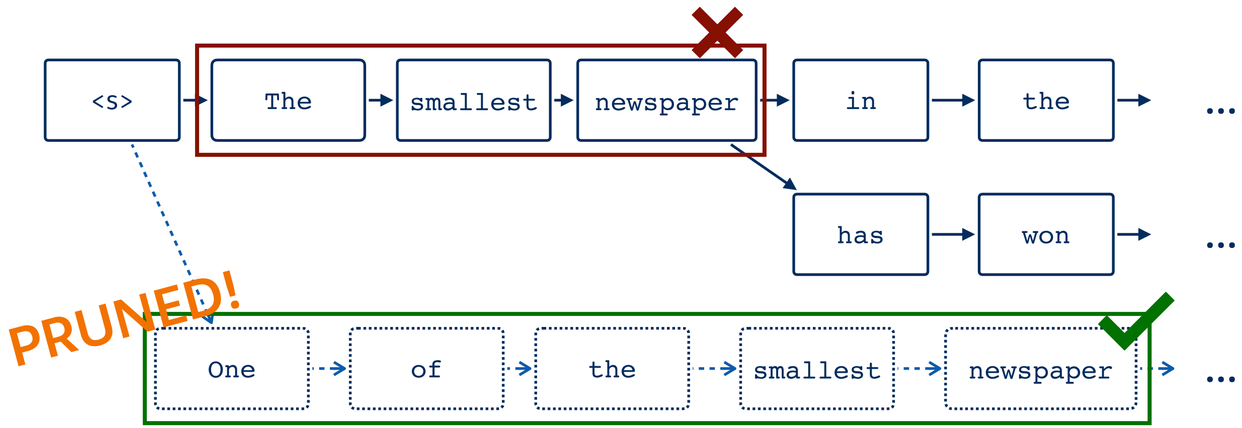
Search for Diverse Outputs

Existing approaches



Search for Diverse Outputs

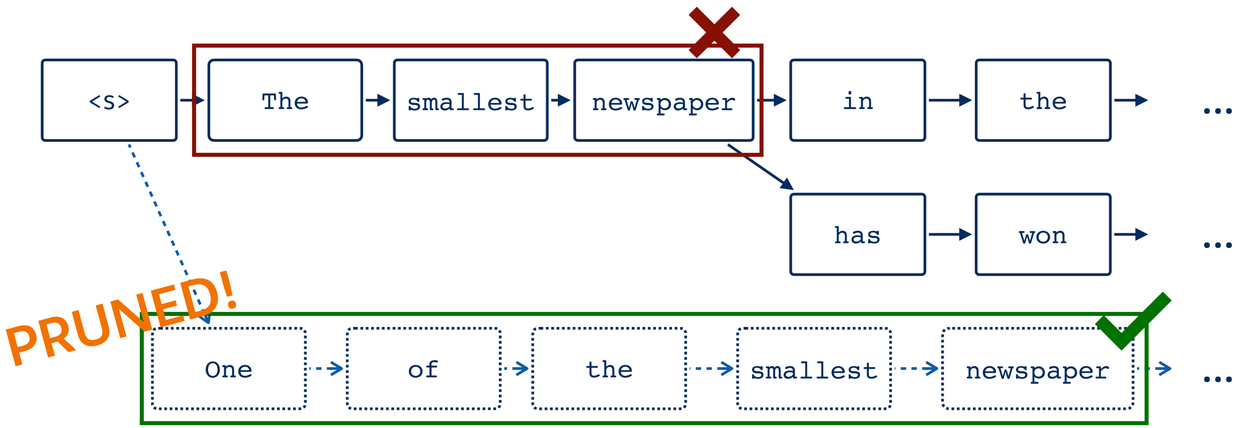
Existing approaches



Beam search prunes valuable and diverse hypotheses.

Search for Diverse Outputs

Existing approaches



Beam search prunes valuable and diverse hypotheses.

🎲: What is the fifth largest city in Oregon?

🎲: What is the fifth largest city in Oregon?

🎲: What is the fifth-largest city in Oregon?

🎲: What is the fifth-largest city in the State of Oregon?

Sampling is hard to control and sometimes causes duplication.

Search for Diverse Outputs

What are we looking for?

	Beam Search	Sampling
Text Quality outputs are high-quality and natural	✓	✓
Flexibility low pruning, switching to other nodes	✗	✗
Controllability deterministic, expanding a state to multiple states	✓	✗
Diversity a pool of outputs with great diversity and low duplication	✗	✓

Search for Diverse Outputs

What are we looking for?

	Beam Search	Sampling	?
Text Quality outputs are high-quality and natural	✓	✓	✓
Flexibility low pruning, switching to other nodes	✗	✗	✓
Controllability deterministic, expanding a state to multiple states	✓	✗	✓
Diversity a pool of outputs with great diversity and low duplication	✗	✓	✓

Search for Diverse Outputs

What are we looking for?

	Beam Search	Sampling	Best-First Search
Text Quality outputs are high-quality and natural	✓	✓	✓
Flexibility low pruning, switching to other nodes	✗	✗	✓
Controllability deterministic, expanding a state to multiple states	✓	✗	✓
Diversity a pool of outputs with great diversity and low duplication	✗	✓	✓

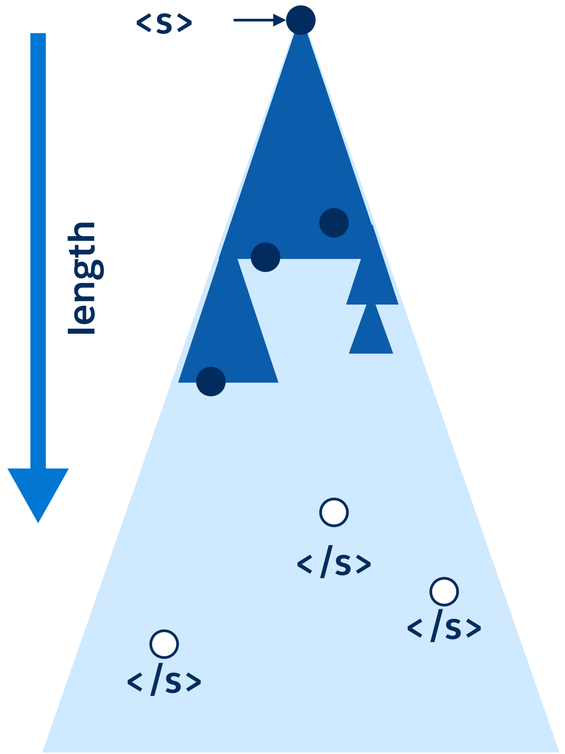
Best-*k* Search Algorithm

Our Approach

- Foundation: examining Best-first search (BFS) for text generation
- Our approach: Best-*k* search algorithm
  - Components: parallel exploration, heap pruning, temporal decay
- Experiments:
  - question generation, commonsense generation, summarization, translation.
- Results: great diversity, naturalness and quality.

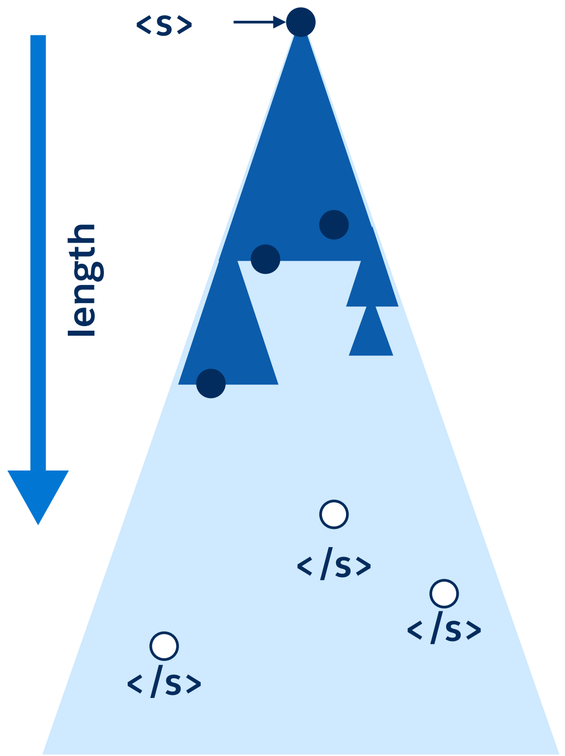
An efficient inference-only algorithm, no parameter, no training or tuning.

Lessons from Best-First Search

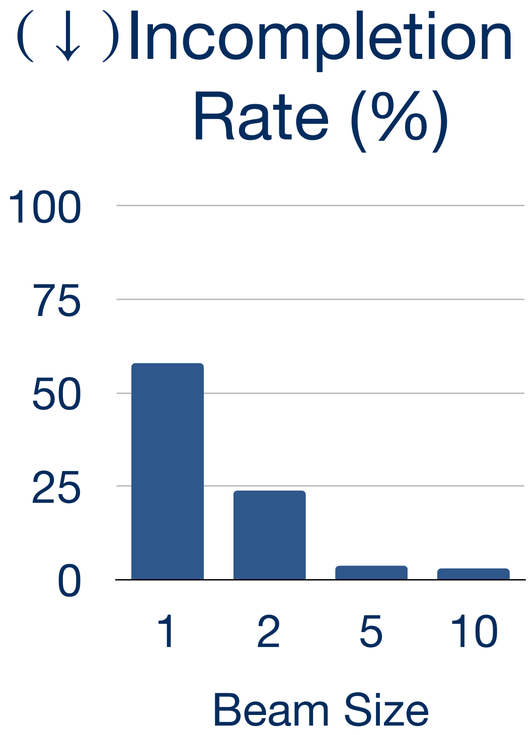


Incompletion = no </s> reached by the end of search

Lessons from Best-First Search

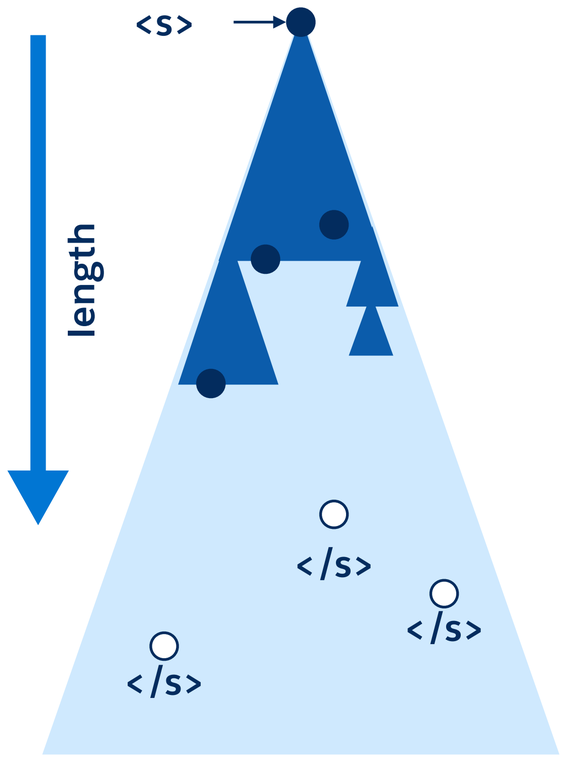


Incompletion = no </s> reached by the end of search

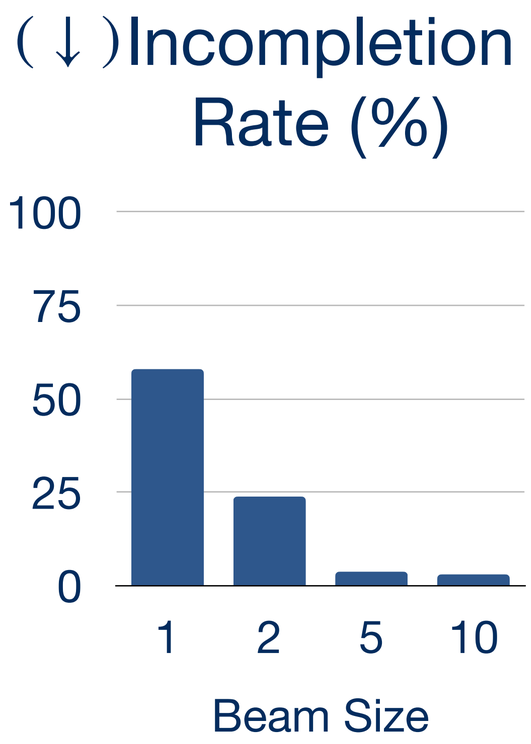


Insufficient completions.

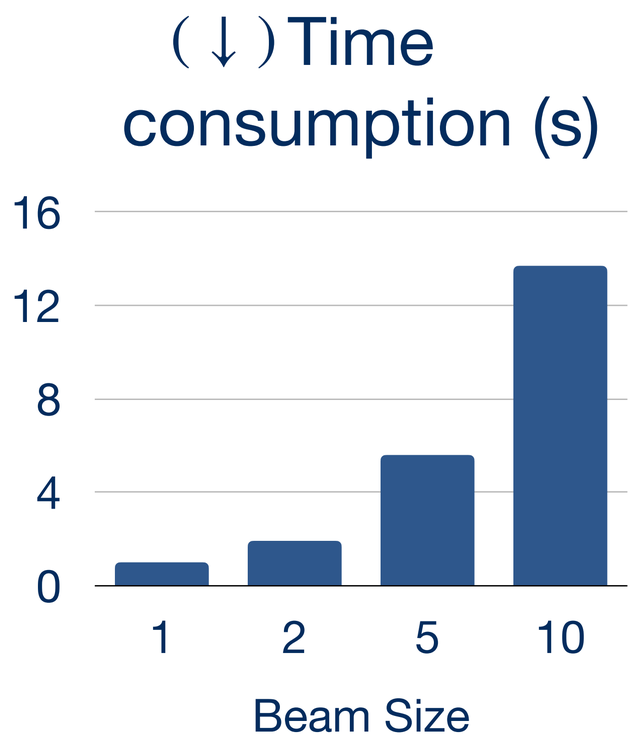
Lessons from Best-First Search



Incompletion = no </s> reached by the end of search



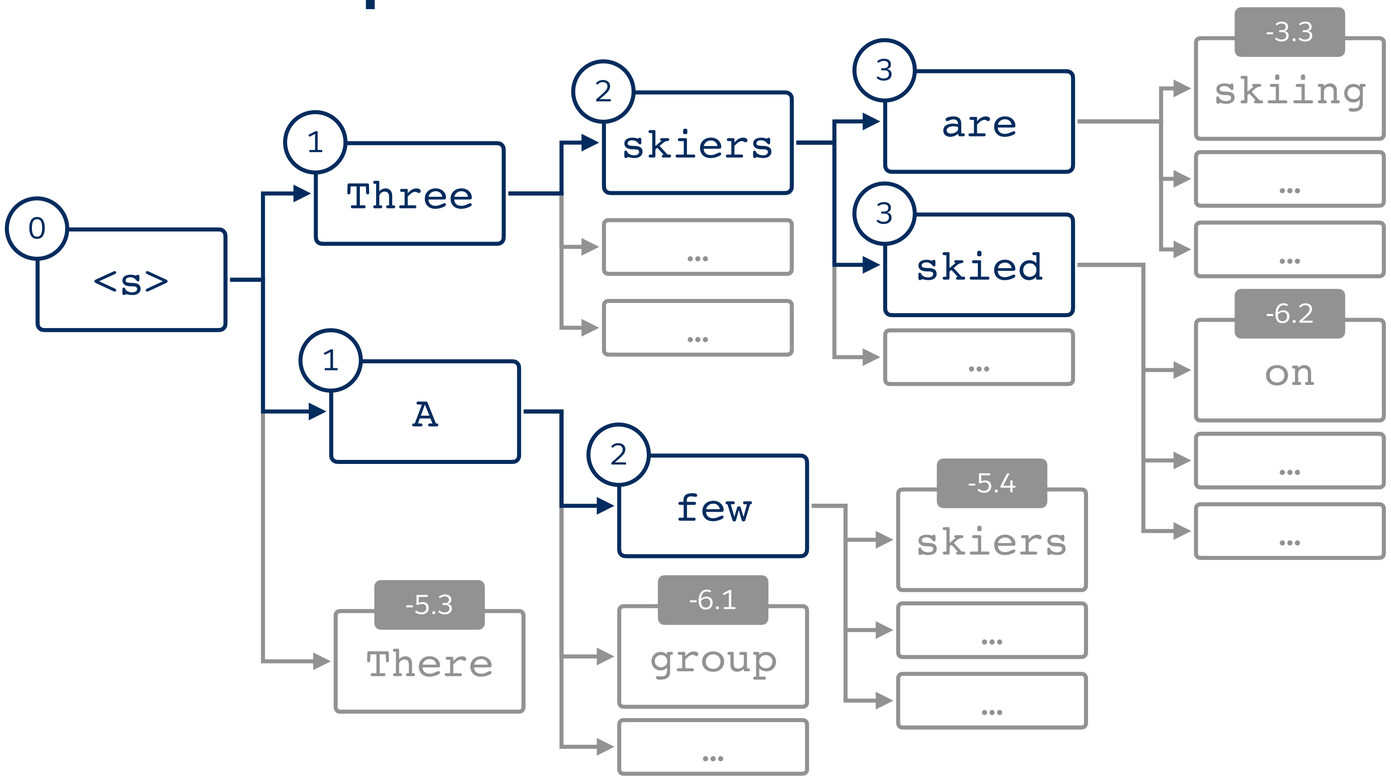
Insufficient completions.



Not efficient enough.

Unpacking the Algorithm

Parallel exploration

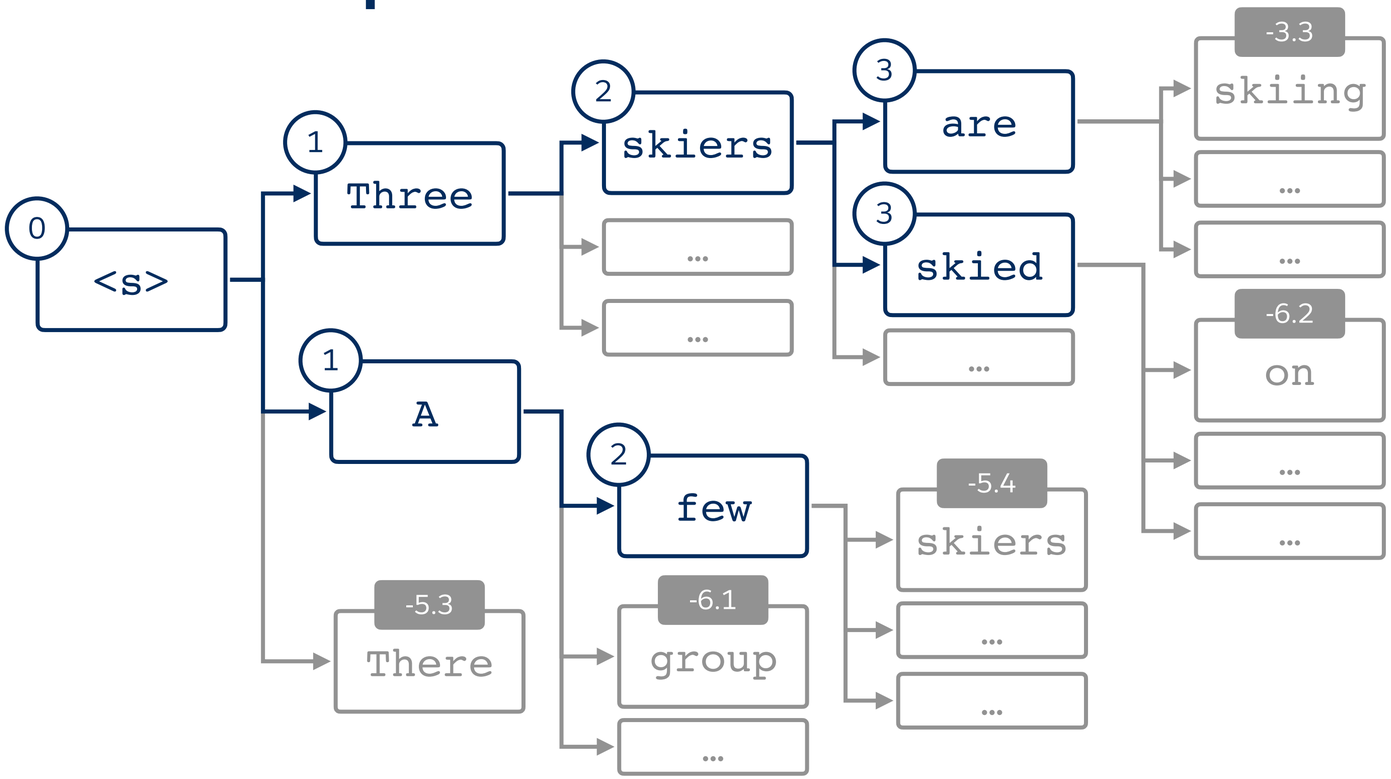


Heap Top

Example: $k = 2$,beam size $= 3$

Unpacking the Algorithm

Parallel exploration



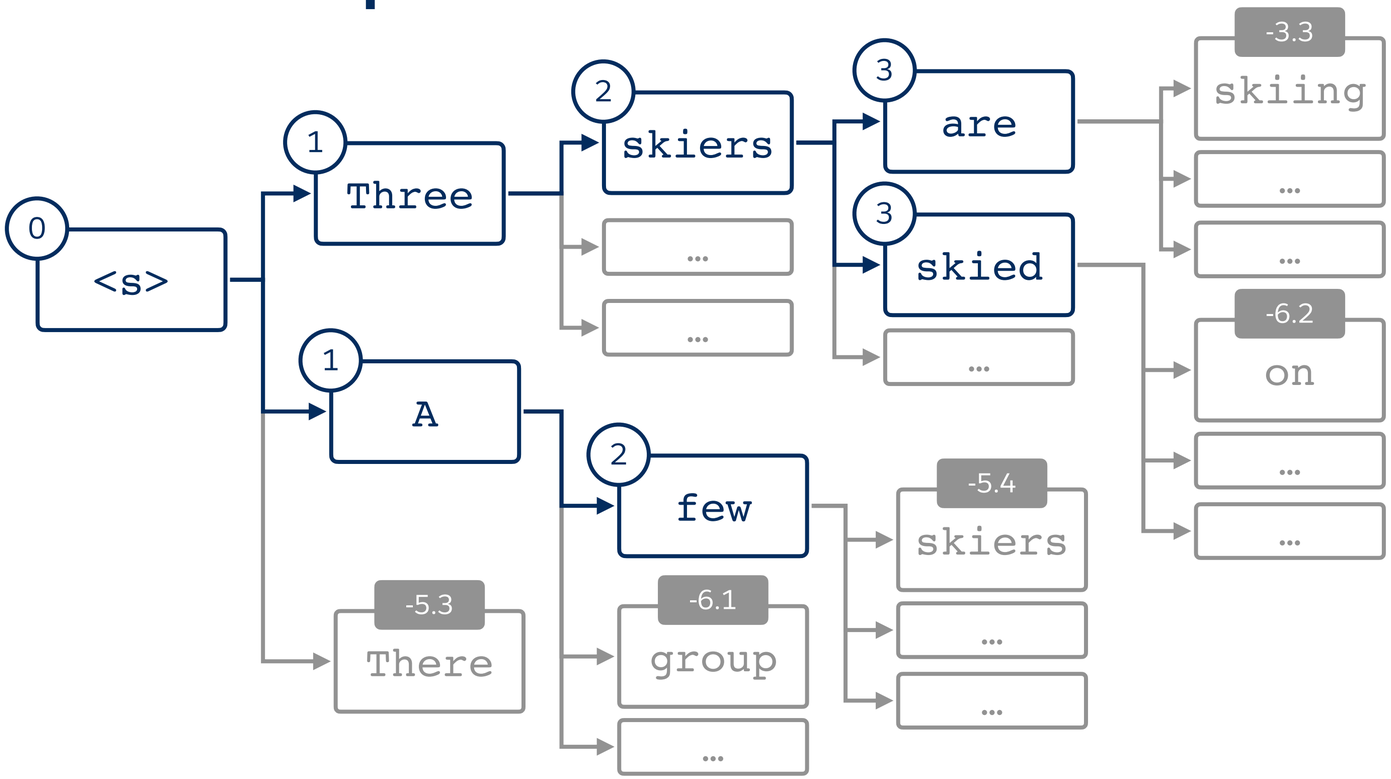
Best-first search:

PQ.heappop(1)

Example: $k = 2$,beam size $= 3$

Unpacking the Algorithm

Parallel exploration



Best-first search:

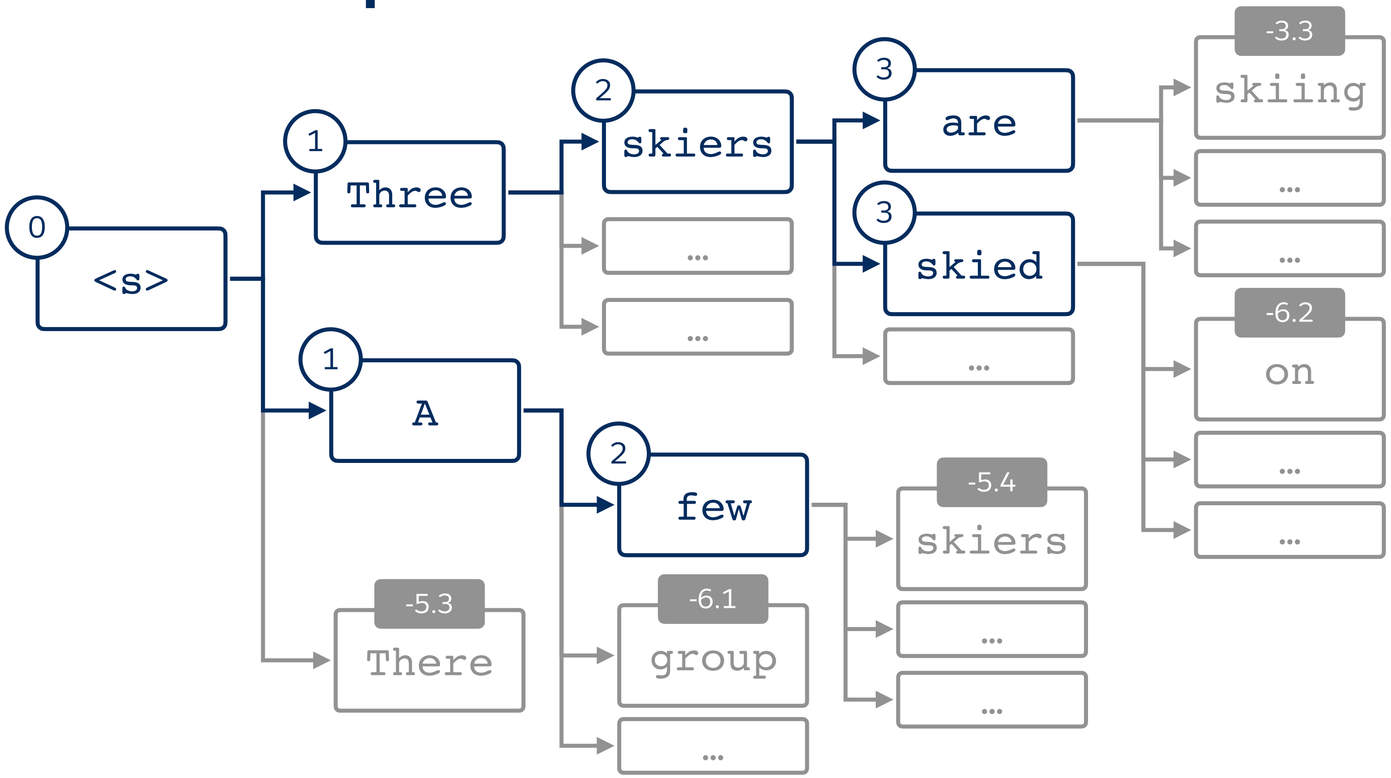
PQ.heappop(1)

Heap Top

Example: $k = 2$,beam size $= 3$

Unpacking the Algorithm

Parallel exploration



Best-first search:

PQ.heappop(1)

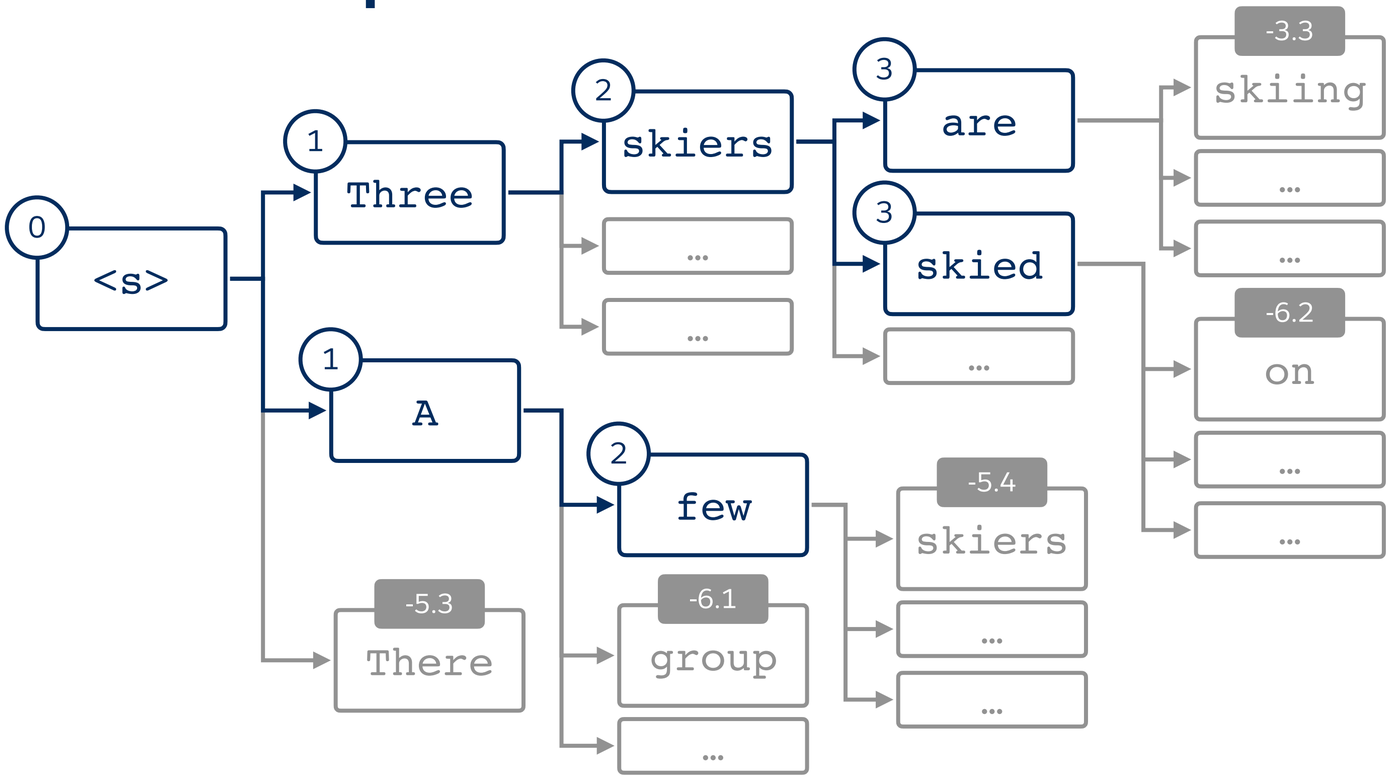
Best-*k* search:

PQ.heappop(k)

Example: $k = 2$,beam size $= 3$

Unpacking the Algorithm

Parallel exploration

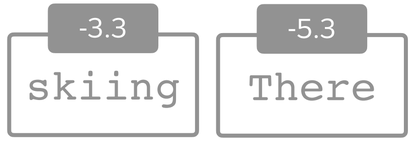


Best-first search:

PQ.heappop(1)

Best-*k* search:

PQ.heappop(k)

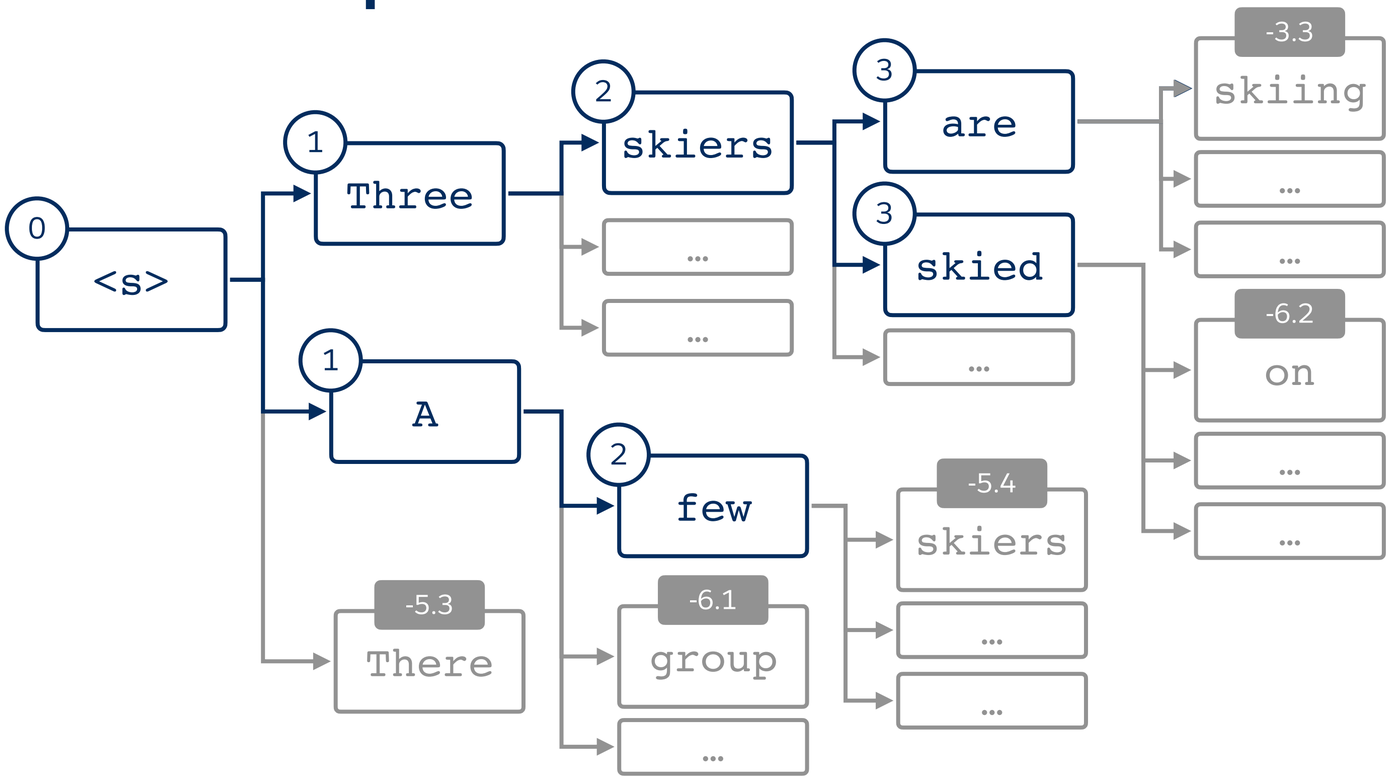


Heap Top

Example: $k = 2$,beam size $= 3$

Unpacking the Algorithm

Parallel exploration

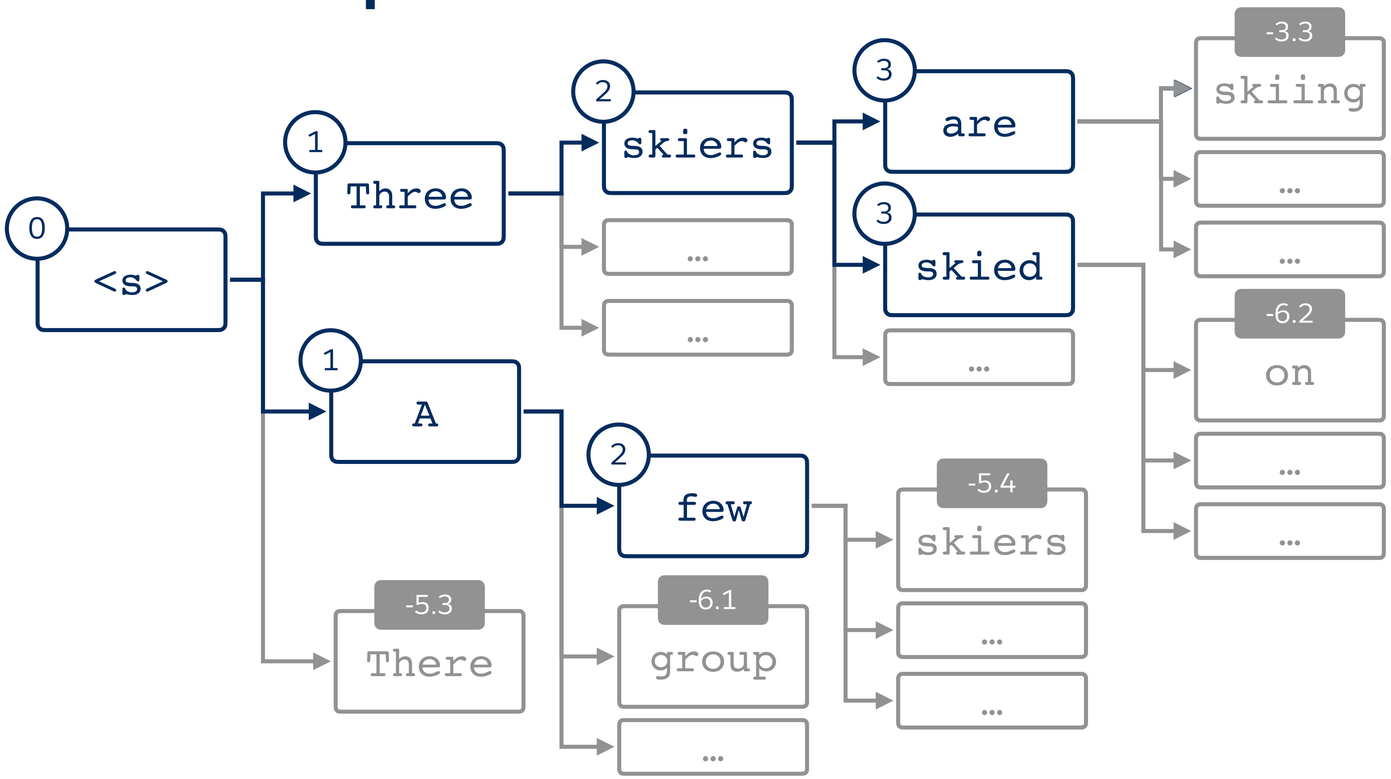


Heap Top

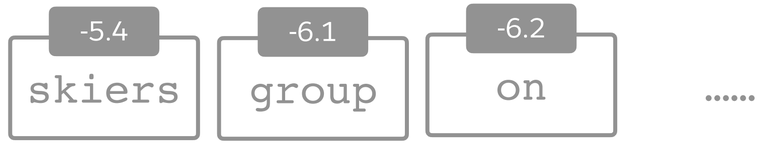
Example: $k = 2$,beam size $= 3$

Unpacking the Algorithm

Parallel exploration



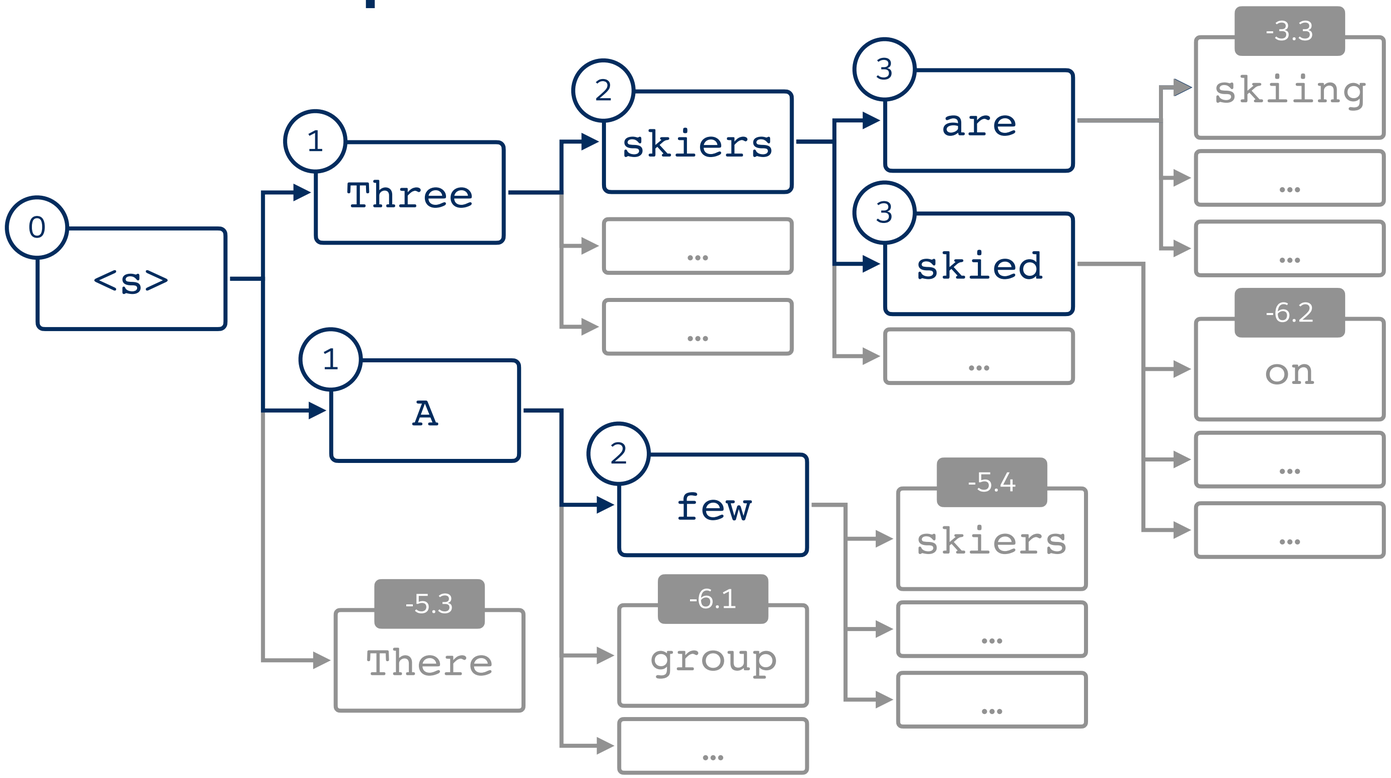
Heap Top



Example: $k = 2$,beam size $= 3$

Unpacking the Algorithm

Parallel exploration

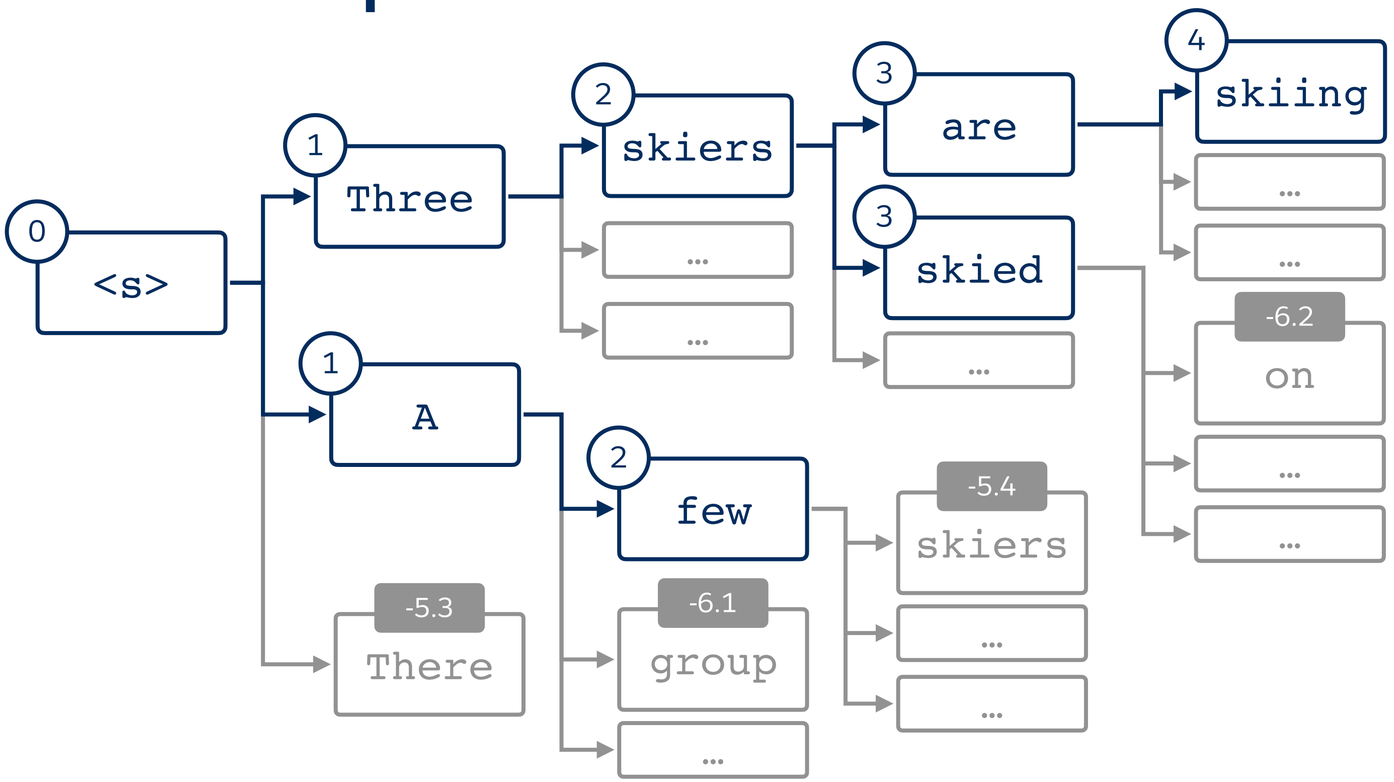


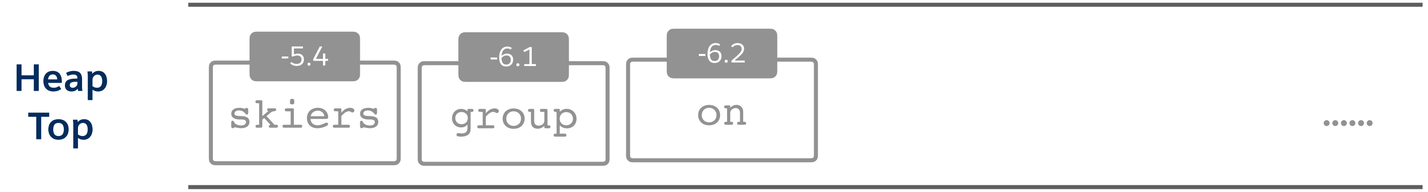
Heap Top

Example: $k = 2$,beam size $= 3$

Unpacking the Algorithm

Parallel exploration

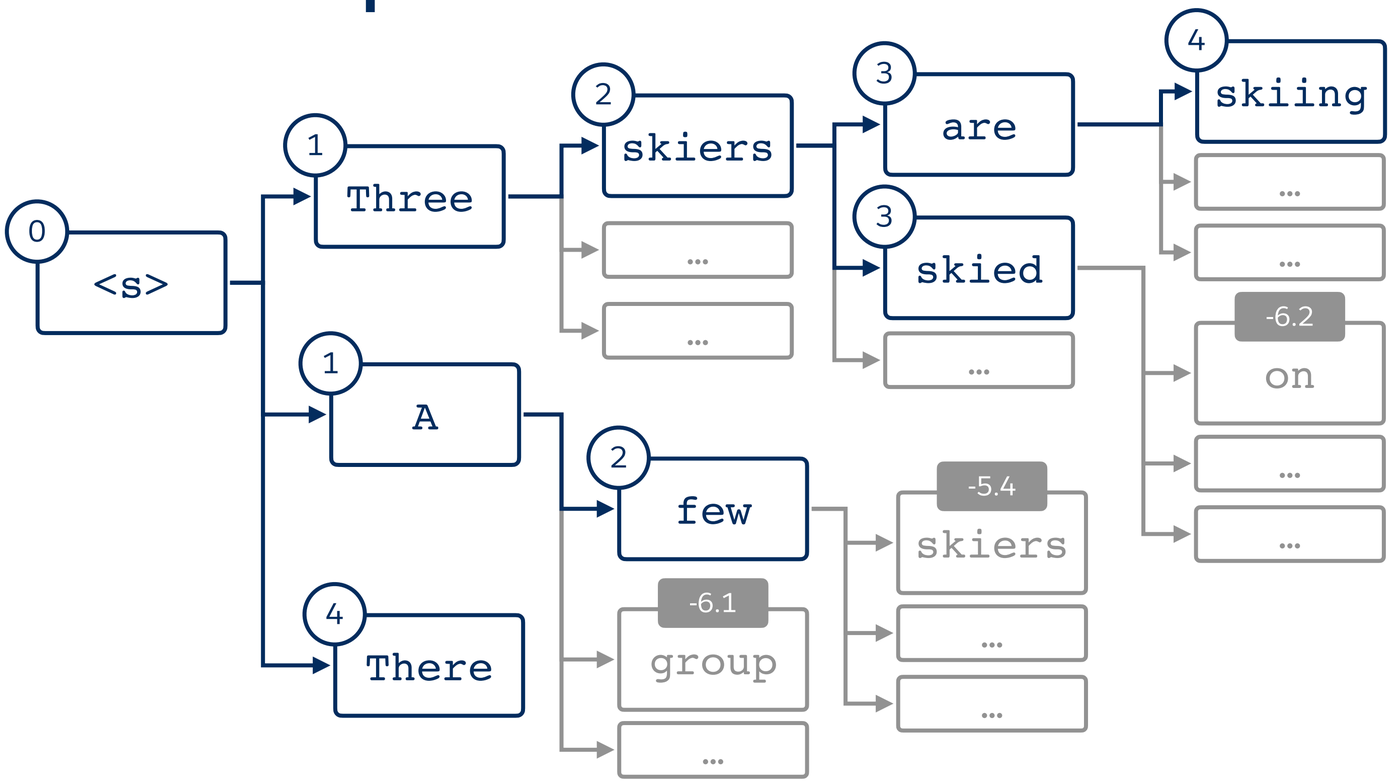




Example: $k = 2$,beam size $= 3$

Unpacking the Algorithm

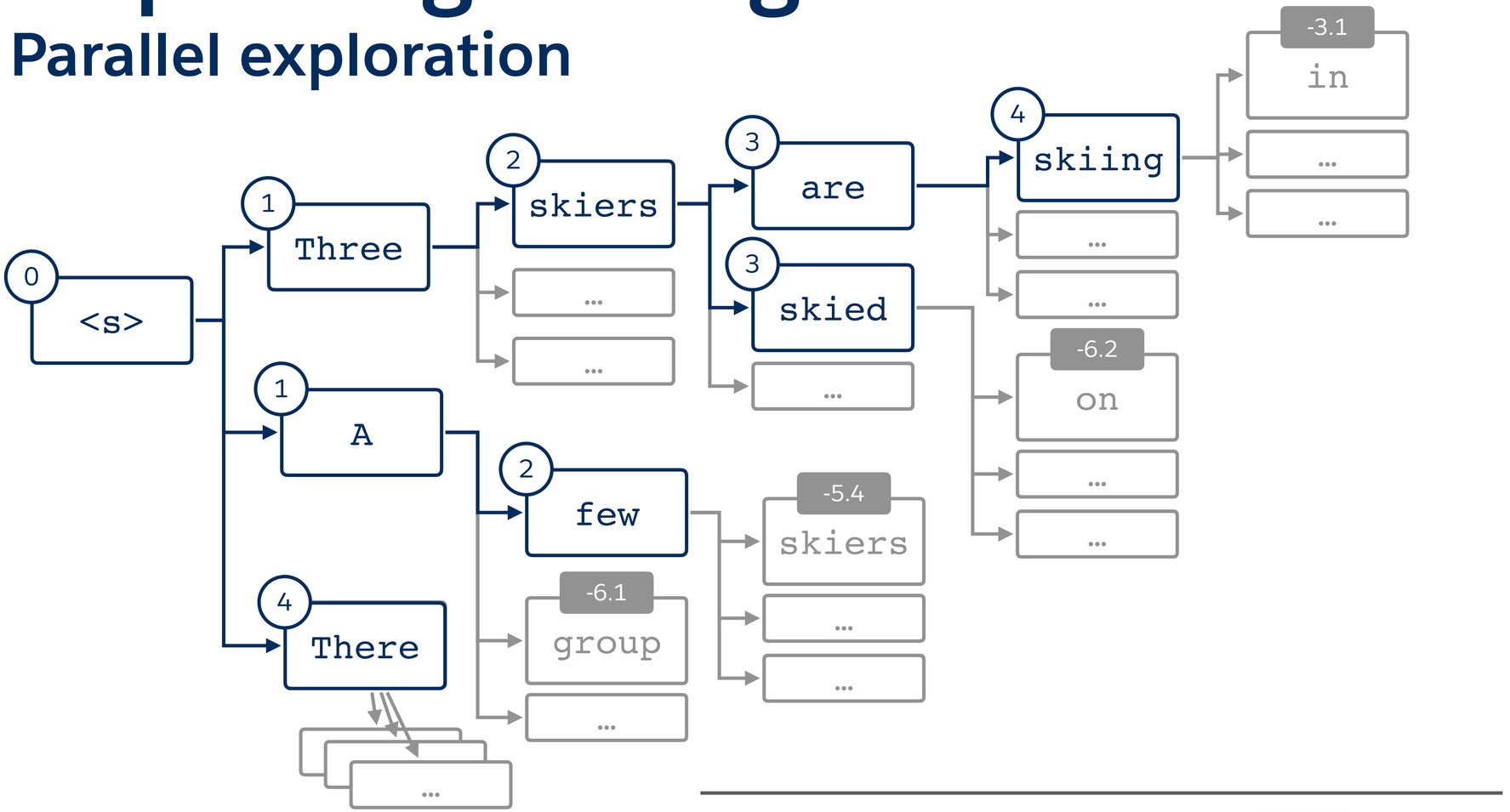
Parallel exploration



Example: $k = 2$,beam size $= 3$

Unpacking the Algorithm

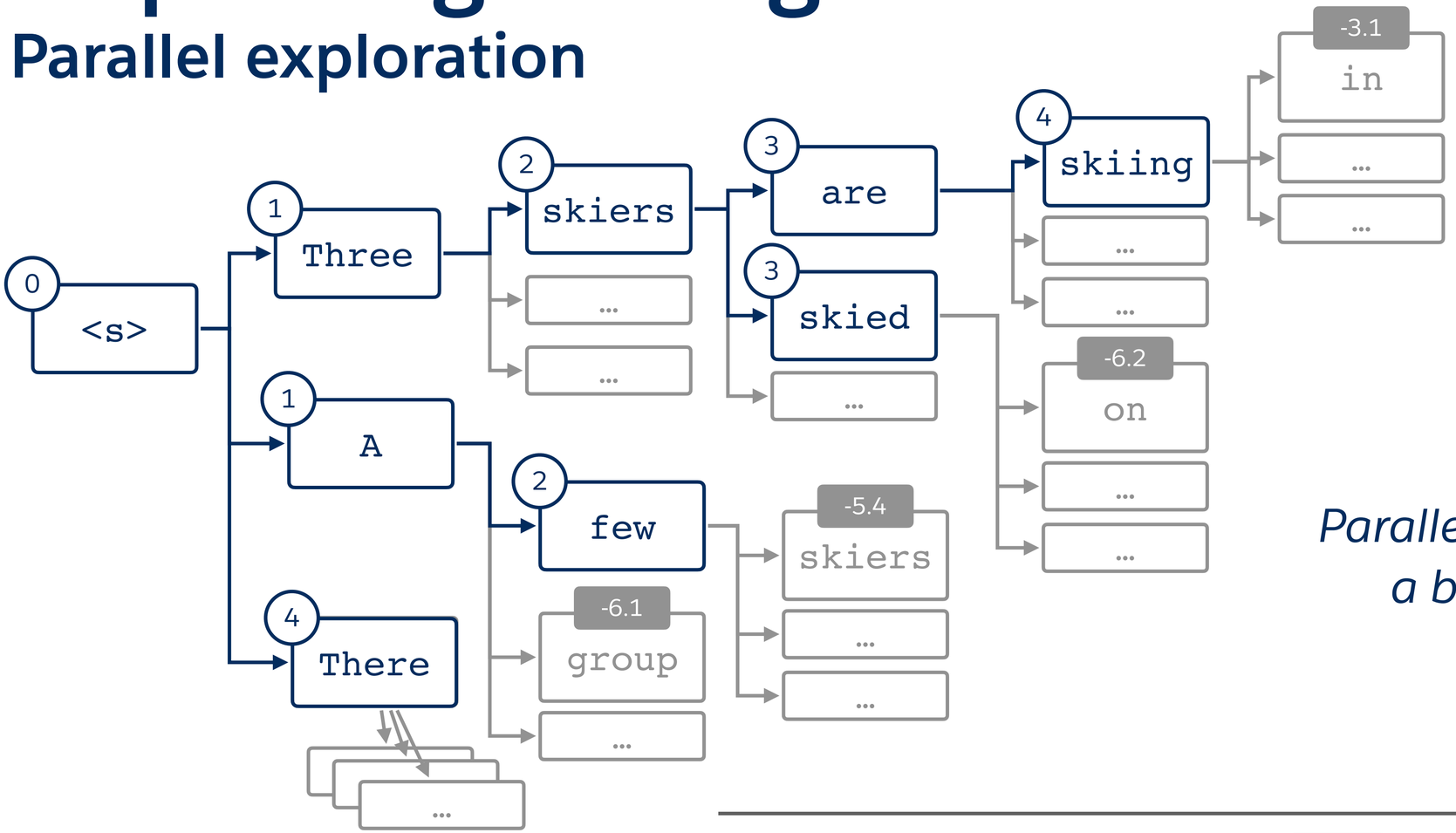
Parallel exploration



Example: $k = 2$,beam size $= 3$

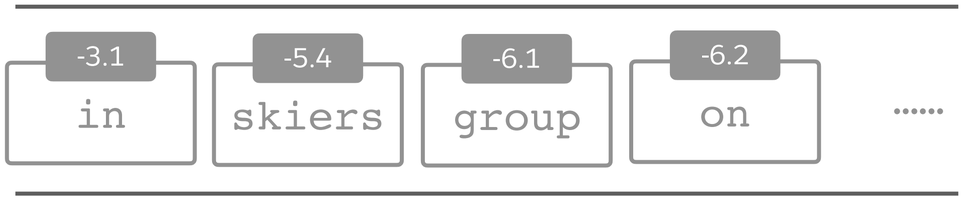
Unpacking the Algorithm

Parallel exploration



Parallel exploration: pop & expand a batch of nodes each time.

- Efficiency ++
- Diversity +



Example: $k = 2$,beam size $= 3$

Unpacking the Algorithm

Temporal Decay

- Scoring function determines the order of exploration.
- Modify the objective by adding a *temporal decay* term:

$$\text{decay}(n.\text{time}, t) = -\kappa(t - n.\text{time})^{\beta}$$

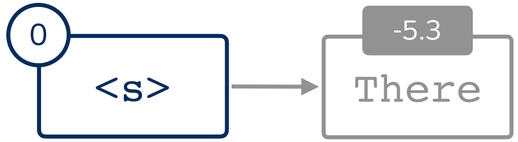
- where $\kappa > 0, \beta > 0$.
- Idea: more recently discovered nodes get a bonus!

Unpacking the Algorithm

Temporal Decay

$$\text{decay}(n.\text{time}, t) = -\kappa(t - n.\text{time})^{\beta}$$

- where $\kappa > 0, \beta > 0$.
- Example: we set $\kappa = \beta = 1$, current time $t = 4$

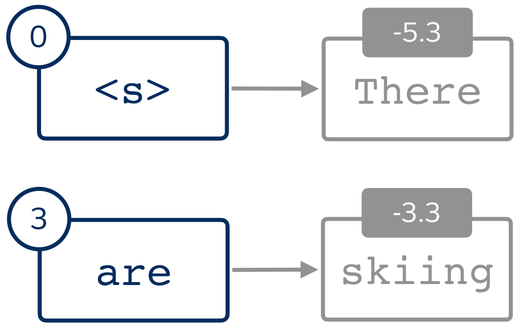


Unpacking the Algorithm

Temporal Decay

$$\text{decay}(n.\text{time}, t) = -\kappa(t - n.\text{time})^{\beta}$$

- where $\kappa > 0, \beta > 0$.
- Example: we set $\kappa = \beta = 1$, current time $t = 4$

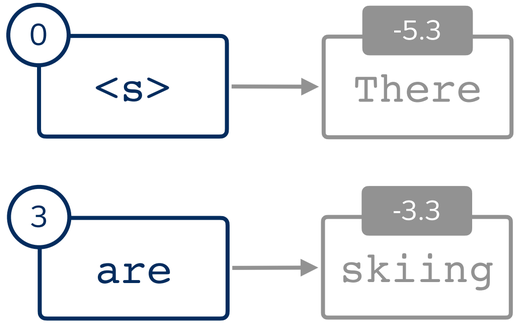


Unpacking the Algorithm

Temporal Decay

$$\text{decay}(n.\text{time}, t) = -\kappa(t - n.\text{time})^{\beta}$$

- where $\kappa > 0, \beta > 0$.
- Example: we set $\kappa = \beta = 1$, current time $t = 4$



$$= -\kappa(4 - 0)^{\beta} = -4$$

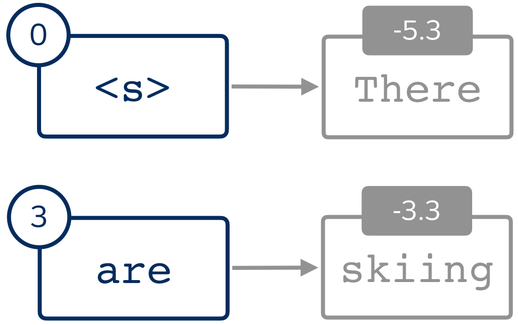
$$= -\kappa(4 - 3)^{\beta} = -1$$

Unpacking the Algorithm

Temporal Decay

$$\text{decay}(n.\text{time}, t) = -\kappa(t - n.\text{time})^{\beta}$$

- where $\kappa > 0, \beta > 0$.
- Example: we set $\kappa = \beta = 1$, current time $t = 4$



$$= -\kappa(4 - 0)^{\beta} = -4$$

$$= -\kappa(4 - 3)^{\beta} = -1$$ ★

Recent completion is more likely to continue due to temporal decay.

Unpacking the Algorithm

Heap pruning & model score

Heap pruning:

- Intuition: low prob nodes won't be visited anyway
- Truncate PQ to 500;
- Ignore tokens with prob below threshold $\gamma = 0.05$

Model score:

- Define a memoryless scoring function:

$$h(\mathbf{y}) = \log p_\theta(y_t | \mathbf{y}_{<t}, \mathbf{x})$$

Algorithm 2 Best-k Search with parallel exploration, heap pruning, and temporal decay.

Input: Generation model θ with vocabulary $\mathcal{V}$, search budget, $\mathcal{O}$ denotes open set (max priority queue). group size k. T is the number of explored steps; t is the time stamp.

Output: All completed paths P.

$\mathcal{O} \leftarrow \{\langle \infty, \text{BOS}, -1 \rangle\}, T \leftarrow 0, t \leftarrow 0.$
while $T <$ budget **do**
  $\mathcal{PQ} \leftarrow \emptyset$
  for $n \in \mathcal{O}$ **do**
    $\mathcal{PQ} \leftarrow \mathcal{PQ} + \langle n.\texttt{score} + \texttt{decay}(n.\texttt{time}, t), n \rangle$
  end for
  $g \leftarrow min(k, \mathcal{PQ}.\texttt{size()})$
  $\mathcal{H} \leftarrow \mathcal{PQ}.\texttt{heappop(g)}$ // $\mathcal{H}$ is the group of candidates to explore.
  $\mathcal{O} \leftarrow \mathcal{O} \setminus \mathcal{H}$
  for $\langle score, n \rangle \in \mathcal{H}$ **do**
    for $v \in \mathcal{V}$ **do**
      if is-complete$(n \circ v)$ **then**
        $P \leftarrow P \cup (n \circ v)$
        continue
      end if
      child $\leftarrow \langle h(n \circ v), v, t \rangle$
      $\mathcal{O} \leftarrow \mathcal{O} \cup$ child
    end for
  end for
  $\mathcal{O} \leftarrow \mathcal{O}.\texttt{prune()}$
  $T \leftarrow T + g$
  $t \leftarrow t + 1$
end while

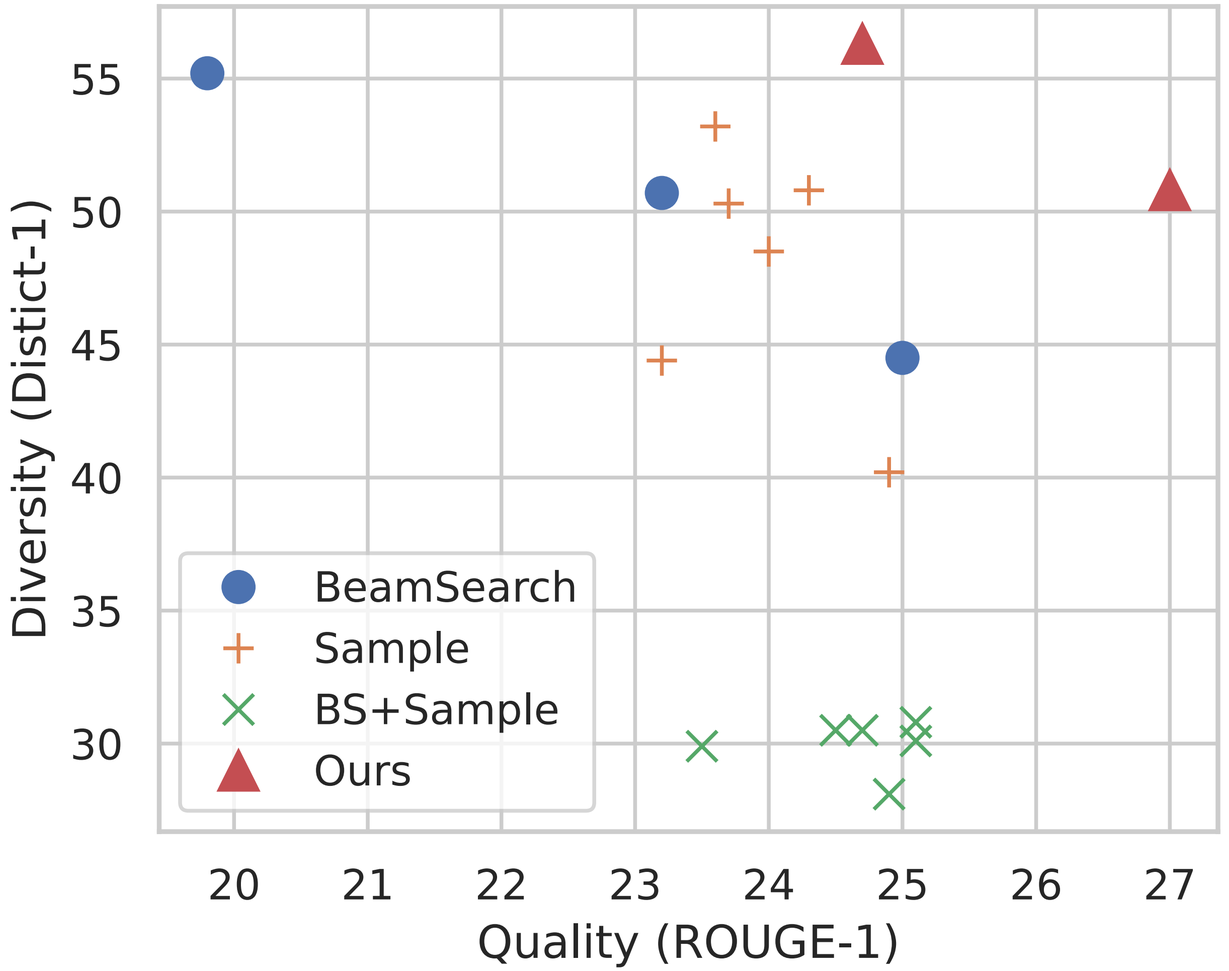
Best-k Search Algorithm
Please check the paper for detail.

Experiments

- Tasks: Question generation (QuoRef, DROP, SQuAD), Commensense Generation, Machine Translation (WMT14 EnFr & EnDe), Text Summarization (XSum)
- Models: corresponding SOTA-ish model available off-the-shelf
  - QG: MixQG; CommenGen: T5-fine-tuned-cg
  - Summarization: BART-large-xsum; MT: mBART50
- Baselines:Beam Search, Diverse Beam Search, Sampling (Nucleus sampling, Typical Sampling), Beam Search + Sampling (BeamNuc, BeamTyp)

Measuring Quality & Diversity

Question generation on QuoRef

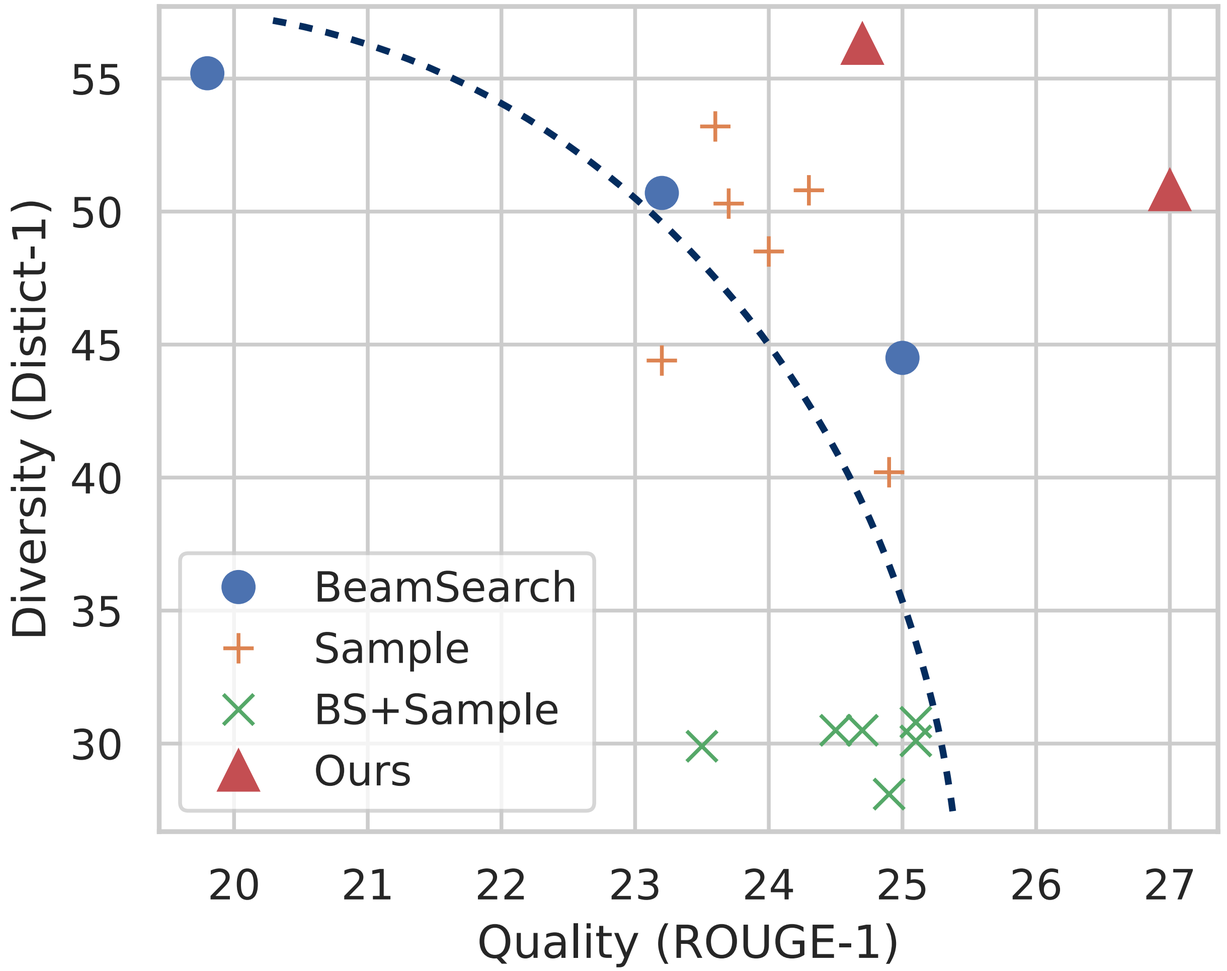


Measuring Quality & Diversity

Question generation on QuoRef

Invisible curve of quality and diversity for existing methods;

- Tradeoff (Zhang+'21)



Zhang et al., Trading off diversity and quality in natural language generation.

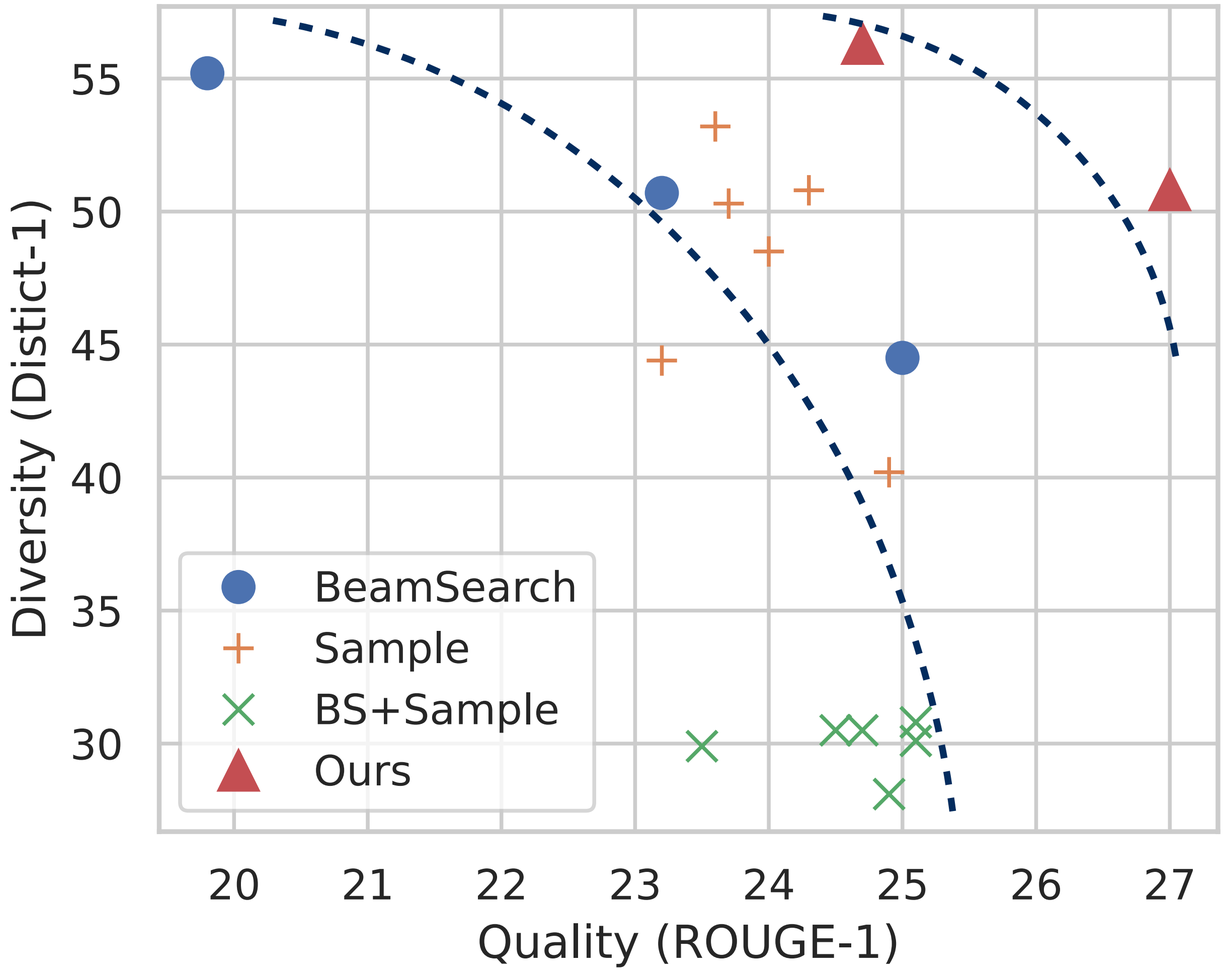
Measuring Quality & Diversity

Question generation on QuoRef

Invisible curve of quality and diversity for existing methods;

- Tradeoff (Zhang+'21)

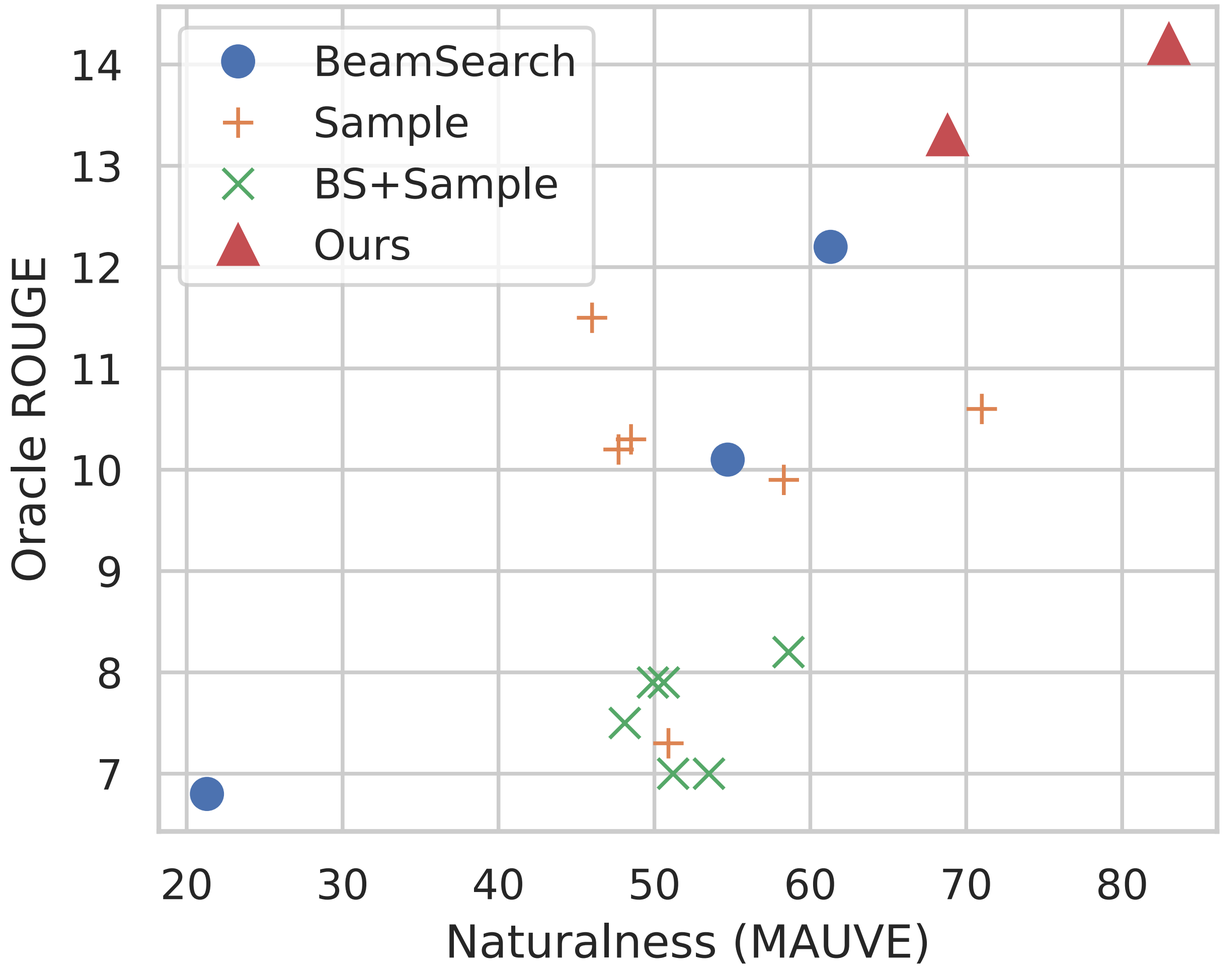
Our approach pushes the line forward by a significant margin.



Zhang et al., Trading off diversity and quality in natural language generation.

Measuring Oracle ROUGE & Naturalness

Question generation on QuoRef



Measuring Oracle ROUGE & Naturalness

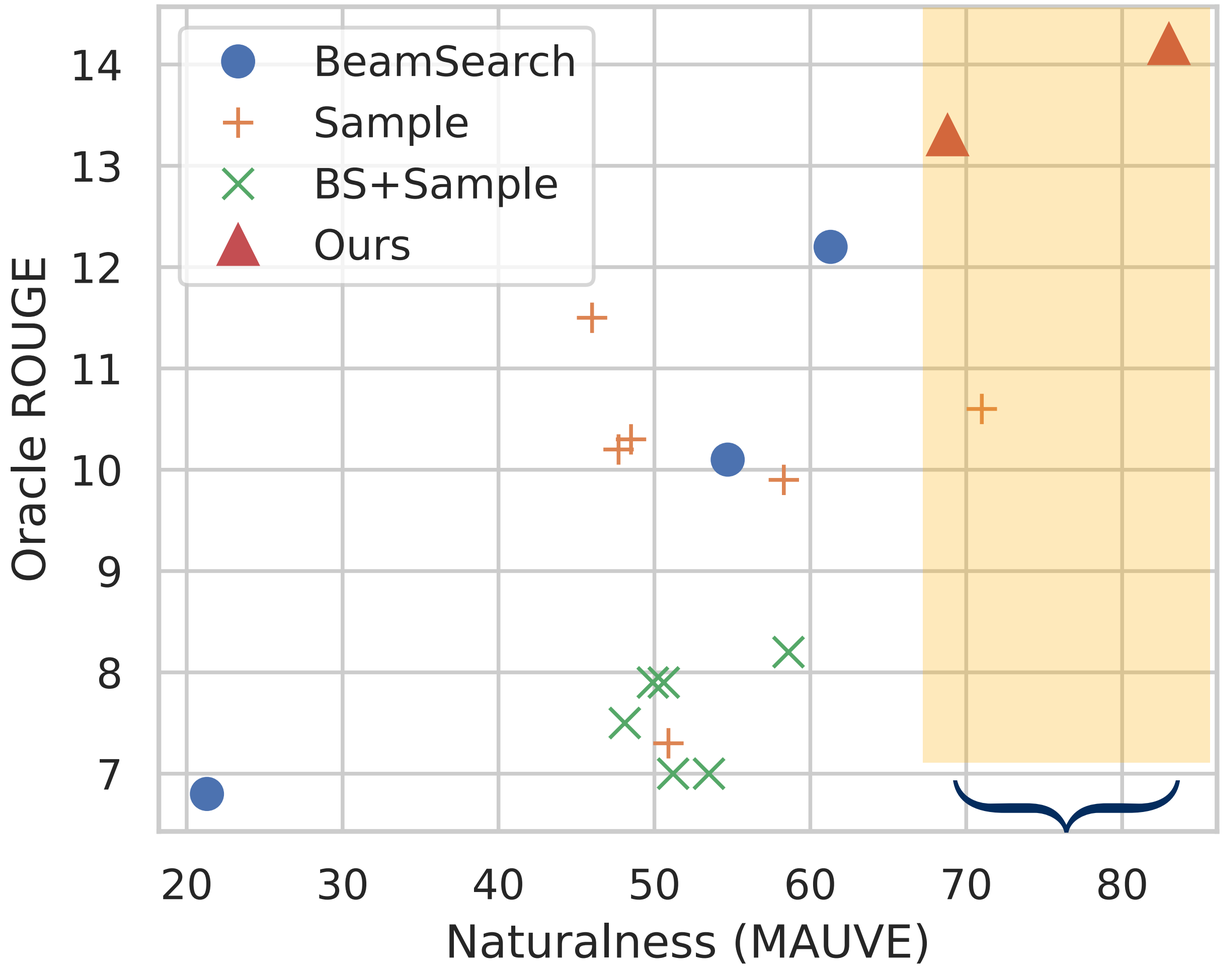
Question generation on QuoRef

Oracle R2: **14.2** vs. 11.5 ($\text{Nuc}_{0.9}$)

great optimality → great algorithm

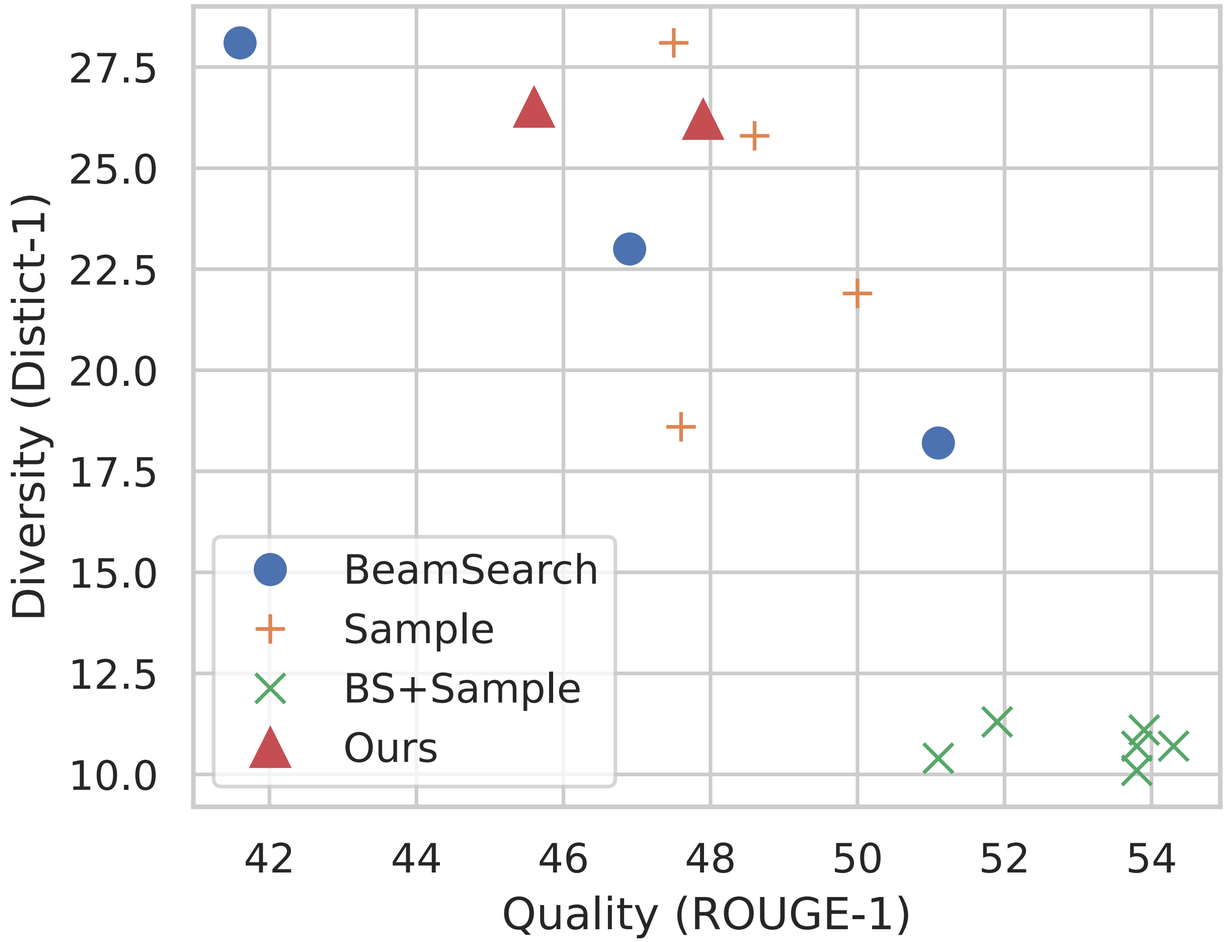
MAUVE: **83.0** vs. 71.0 ($\text{Typ}_{0.5}$)

great naturalness



Measuring Quality & Diversity

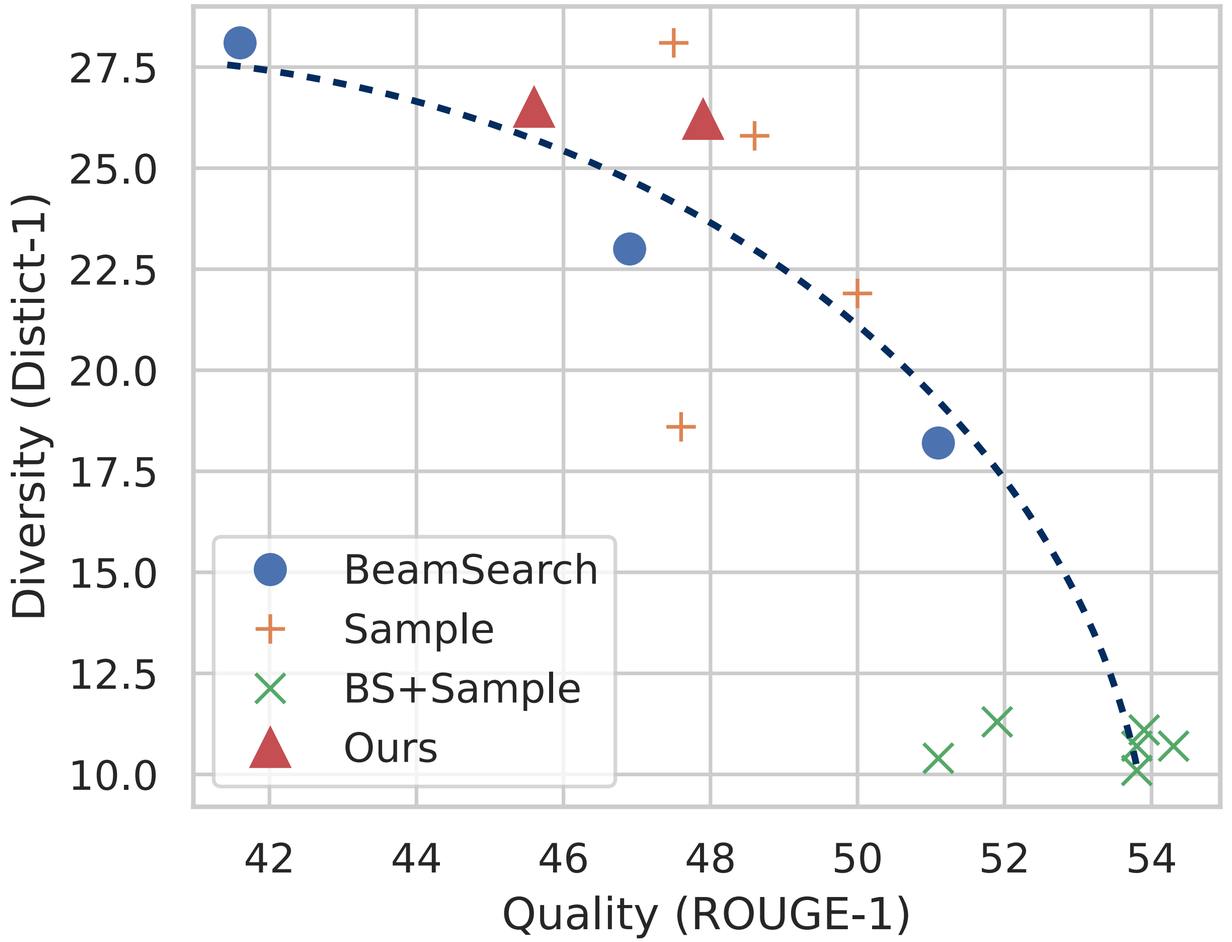
Question generation on SQuAD



Measuring Quality & Diversity

Question generation on SQuAD

For a dataset that our methods don’t perform super well …



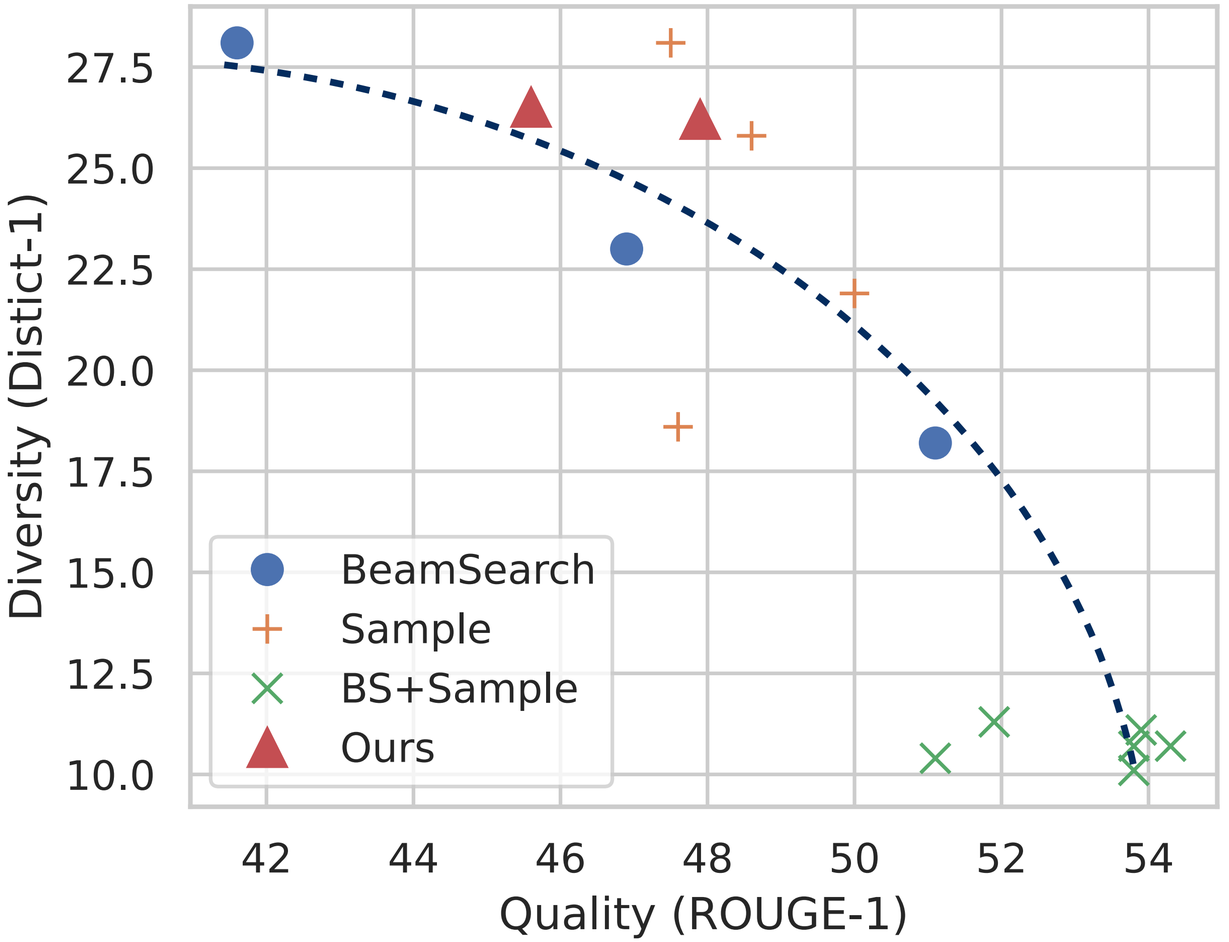
Zhang et al., Trading off diversity and quality in natural language generation.

Measuring Quality & Diversity

Question generation on SQuAD

For a dataset that our methods don't perform super well ...

Our approach is still competitive with best methods in-class.

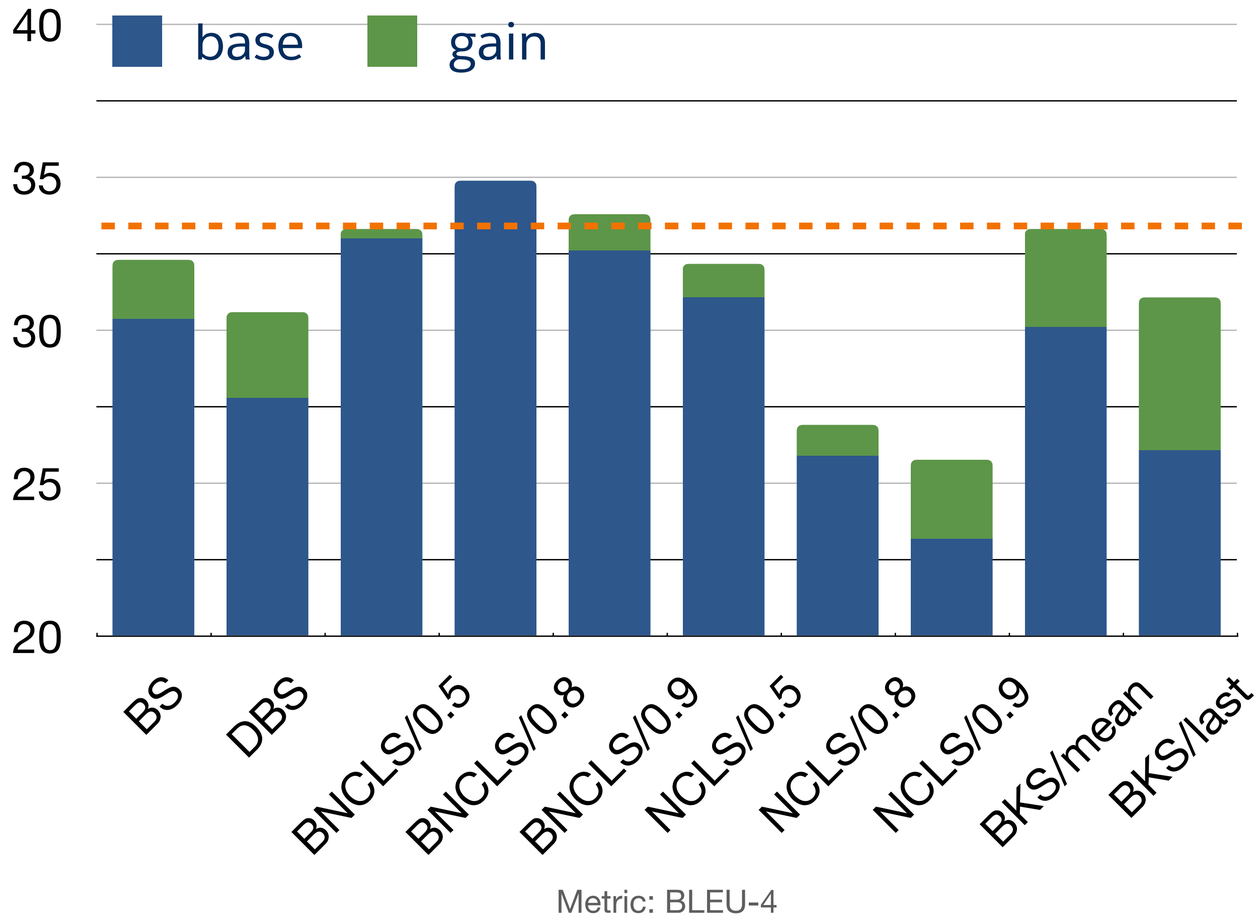


Zhang et al., Trading off diversity and quality in natural language generation.

Application: Reranking Diverse Outputs

Machine Translation

- Our approach outputs many (30+) hypotheses
  - baseline: 2 ~ 10
- Reranking: **30.1 → 33.3**
  - Reference-free COMET-QE
  - Best: BNcls/0.8: 34.9
- Higher than most baselines!



Sampling	BKS_{last}	BKS_{mean}
NCLS$_{0.8}$	What city is the fifth largest?	What city in OR is the fifth largest in OR?
What is the fifth largest city in OR?	What city is the fifth-largest city in the State?	What city is the fifth largest city in OR?
What is the fifth largest city in OR?	What is the 5th largest city in OR?	What city is the fifth largest city in the State?
What is the fifth largest city in OR?	What is the fifth largest city in OR?	What city is the fifth largest in OR?
What is the fifth largest city in OR?	What is the fifth largest city in the State of OR?	What city is the fifth largest in the state?
What is the fifth-largest city in OR?	What is the fifth largest city in the State?	What city is the fifth largest?
What is the fifth-largest city in OR?	What is the fifth-largest city in the State?	What city is the fifth-largest in the State?
What is the fifth-largest city in the State of OR?	Which city is the fifth largest city in OR?	What is the fifth largest city in OR?
What is the fifth-largest city in the State of OR?	Which city is the fifth largest city?	What is the fifth largest city in the State?
Which city in OR is the county seat of Washington County?	Which is the fifth largest city?	Which city in OR has the largest population?
TYP$_{0.5}$	Which OR city is the fifth largest in the state?	Which city in OR hosts Intel?
What is the fifth largest city in OR?	Which OR city is the fifth largest?	Which city in OR is known as the Silicon Forest?
What is the fifth largest city in OR?	Which OR town is home to Intel?	Which city in OR is the fifth largest in OR?
What is the fifth largest city in OR?	Which OR town is home to the tech company Intel?	Which city in OR is the fifth largest in the state?
What is the fifth largest city in OR?	Which OR town is known as the Silicon Forest?	Which city is the fifth largest city in OR?
What is the fifth largest city in OR?	Which OR town is the fifth largest city in the state?	Which city is the fifth largest city?
What is the fifth-largest city in OR?	Which OR town is the fifth largest city?	Which city is the fifth largest in the state?
What is the fifth-largest city in OR?	Which OR town is the fifth largest in size?	Which city is the fifth largest?
What is the fifth-largest city in OR?	Which OR town is the fifth largest in the state?	Which OR city is the county seat of Washington County?
Which city in OR is the county seat of Washington County?	Which OR town is the fifth largest?	Which OR city is the fifth largest in size?
Which city is the county seat of Washington County?		Which OR city is the fifth largest?

Input (Ans || Context): Hillsboro || Hillsboro is the fifth-largest city in the State of Oregon and is the county seat of Washington County. Lying in the Tualatin Valley on the west side of the Portland metropolitan area, the city hosts many high-technology companies, such as Intel, that comprise what has become known as the Silicon Forest. At the 2010 Census, the city's population was 91,611.For thousands of years before the arrival of European-American settlers, the Atfalati tribe of the Kalapuya lived in ...
Reference Question: What city is Intel located in?

Sampling	BKS_{last}	BKS_{mean}
NCLS$_{0.8}$	What city is the fifth largest?	What city in OR is the fifth largest in OR?
What is the fifth largest city in OR?	What city is the fifth-largest city in the State?	What city is the fifth largest city in OR?
What is the fifth largest city in OR?	What is the 5th largest city in OR?	What city is the fifth largest city in the State?
What is the fifth largest city in OR?	What is the fifth largest city in OR?	What city is the fifth largest in OR?
What is the fifth largest city in OR?	What is the fifth largest city in the State of OR?	What city is the fifth largest in the state?
What is the fifth largest city in OR?	What is the fifth largest city in the State?	What city is the fifth largest?
What is the fifth-largest city in OR?	What is the fifth-largest city in the State?	What city is the fifth-largest in the State?
What is the fifth-largest city in OR?	Which city is the fifth largest city in OR?	What is the fifth largest city in OR?
What is the fifth-largest city in the State of OR?	Which city is the fifth largest city?	What is the fifth largest city in the State?
What is the fifth-largest city in the State of OR?	Which is the fifth largest city?	Which city in OR has the largest population?
Which city in OR is the county seat of Washington County?	Which OR city is the fifth largest in the state?	Which city in OR hosts Intel?
TYP$_{0.5}$	Which OR city is the fifth largest?	Which city in OR is known as the Silicon Forest?
What is the fifth largest city in OR?	Which OR town is home to Intel?	Which city in OR is the fifth largest in OR?
What is the fifth largest city in OR?	Which OR town is home to the tech company Intel?	Which city in OR is the fifth largest in the state?
What is the fifth largest city in OR?	Which OR town is known as the Silicon Forest?	Which city is the fifth largest city in OR?
What is the fifth largest city in OR?	Which OR town is the fifth largest city in the state?	Which city is the fifth largest city?
What is the fifth largest city in OR?	Which OR town is the fifth largest city?	Which city is the fifth largest in the state?
What is the fifth-largest city in OR?	Which OR town is the fifth largest in size?	Which city is the fifth largest?
What is the fifth-largest city in OR?	Which OR town is the fifth largest in the state?	Which OR city is the county seat of Washington County?
What is the fifth-largest city in OR?	Which OR town is the fifth largest?	Which OR city is the fifth largest in size?
Which city in OR is the county seat of Washington County?		Which OR city is the fifth largest?
Which city is the county seat of Washington County?		

Input (Ans || Context): Hillsboro || Hillsboro is the fifth-largest city in the State of Oregon and is the county seat of Washington County. Lying in the Tualatin Valley on the west side of the Portland metropolitan area, the city hosts many high-technology companies, such as Intel, that comprise what has become known as the Silicon Forest. At the 2010 Census, the city's population was 91,611.For thousands of years before the arrival of European-American settlers, the Atfalati tribe of the Kalapuya lived in ...
Reference Question: What city is Intel located in?

Sampling	BKS_{last}	BKS_{mean}
$\text{NCLS}_{0.8}$	What city is the fifth largest?	What city in OR is the fifth largest in OR?
What is the fifth largest city in OR?	What city is the fifth-largest city in the State?	What city is the fifth largest city in OR?
What is the fifth largest city in OR?	What is the 5th largest city in OR?	What city is the fifth largest city in the State?
What is the fifth largest city in OR?	What is the fifth largest city in OR?	What city is the fifth largest in OR?
What is the fifth largest city in OR?	What is the fifth largest city in the State of OR?	What city is the fifth largest in the state?
What is the fifth largest city in OR?	What is the fifth largest city in the State?	What city is the fifth largest?
What is the fifth-largest city in OR?	What is the fifth-largest city in the State?	What city is the fifth-largest in the State?
What is the fifth-largest city in OR?	Which city is the fifth largest city in OR?	What is the fifth largest city in OR?
What is the fifth-largest city in the State of OR?	Which city is the fifth largest city?	What is the fifth largest city in the State?
What is the fifth-largest city in the State of OR?	Which is the fifth largest city?	Which city in OR has the largest population?
Which city in OR is the county seat of Washington County?	Which OR city is the fifth largest in the state?	Which city in OR hosts Intel?
$\text{TYP}_{0.5}$	Which OR city is the fifth largest?	Which city in OR is known as the Silicon Forest?
What is the fifth largest city in OR?	Which OR town is home to Intel?	Which city in OR is the fifth largest in OR?
What is the fifth largest city in OR?	Which OR town is home to the tech company Intel?	Which city in OR is the fifth largest in the state?
What is the fifth largest city in OR?	Which OR town is known as the Silicon Forest?	Which city is the fifth largest city in OR?
What is the fifth largest city in OR?	Which OR town is the fifth largest city in the state?	Which city is the fifth largest city?
What is the fifth largest city in OR?	Which OR town is the fifth largest city?	Which city is the fifth largest in the state?
What is the fifth-largest city in OR?	Which OR town is the fifth largest in size?	Which city is the fifth largest?
What is the fifth-largest city in OR?	Which OR town is the fifth largest in the state?	Which OR city is the county seat of Washington County?
What is the fifth-largest city in OR?	Which OR town is the fifth largest?	Which OR city is the fifth largest in size?
Which city in OR is the county seat of Washington County?		Which OR city is the fifth largest?
Which city is the county seat of Washington County?		

Input (Ans || Context): Hillsboro || Hillsboro is the fifth-largest city in the State of Oregon and is the county seat of Washington County. Lying in the Tualatin Valley on the west side of the Portland metropolitan area, the city hosts many high-technology companies, such as Intel, that comprise what has become known as the Silicon Forest. At the 2010 Census, the city's population was 91,611.For thousands of years before the arrival of European-American settlers, the Atfalati tribe of the Kalapuya lived in ...
Reference Question: What city is Intel located in?

Takeaways

Best-*k* Search Algorithm for Neural Text Generation

- Best-*k* Search, a novel decoding algorithm for text generation based on best-first search.
  - Including parallel exploration, temporal decay and heap pruning.
- Experiments: four tasks and six datasets
- Result: natural and diverse text while maintaining high quality.
- **The algorithm is orthogonal to sampling methods and it is parameter-free, lightweight, efficient, and easy to use.**

Structural Decoding in Neural Text Generation

Conclusion

- Massive-scale Decoding for Text Generation using Lattices
  - best-first search + path recombination
  - finds massive number of outputs and stores them in lattice structure
- Best-*k* Search Algorithm for Neural Text Generation
  - best-*k* search algorithm, simple yet effective
  - diversity, naturalness and quality covered; application: reranking

Email: jiacheng.xu@salesforce.com