Language Model Algorithms

Kenneth Heafield

Carnegie Mellon, University of Edinburgh

March 7, 2013

MT is Expensive

- "Since decoding is very time-intensive" [Jehl et al, 2012]
- "Based on the amount of memory we can afford" [Wuebker et al, 2012]

MT is Expensive

- "Since decoding is very time-intensive" [Jehl et al, 2012]
- "Based on the amount of memory we can afford" [Wuebker et al, 2012]

Much of the computational cost is due to the language model.

Desiderata

- More data
- Less CPU time
- Less RAM
- High Quality
  - Search Accuracy
  - BLEU

Language Model Algorithms

Estimating Text → ARPA file with probabilities.

Querying ARPA file → efficient data structure.

Decoding Searching for high-scoring translations.

Estimating: The Problem

SRILM uses too much memory $\Longrightarrow$ limit on data size.
Also, annoying to compile.

Estimating: The Problem

SRILM uses too much memory $\Longrightarrow$ limit on data size.
Also, annoying to compile.

IRSTLM does not estimate modified Kneser-Ney models.
Also, segfaults.

Stupid Backoff [Brants et al, 2007]

1. Count n-grams offline

2. Compute pseudo-probabilities at runtime

Stupid Backoff [Brants et al, 2007]

1. Count n-grams offline

$$\text{count}(w_1^n)$$

2. Compute pseudo-probabilities at runtime

$$s(w_n|w_1^{n-1}) = \begin{cases} \frac{\text{count}(w_1^n)}{\text{count}(w_1^{n-1})} & \text{if count}(w_n) > 0 \\ 0.4s(w_n|w_2^{n-1}) & \text{if count}(w_n) = 0 \end{cases}$$

NB: s does not sum to 1.

Counting *n*-grams

<s> Australia is one of the few

5-**gram**	**Count**
<s> Australia is one of	1
Australia is one of the	1
is one of the few	1

Hash table?

Counting *n*-grams

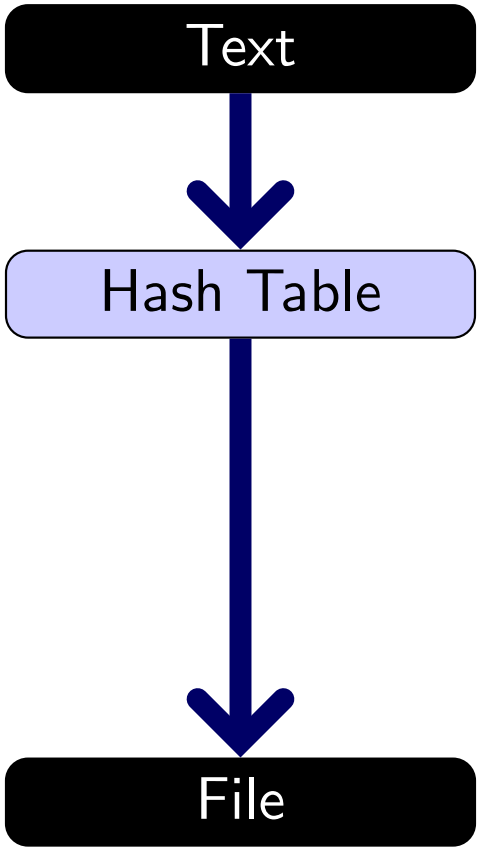
<s> Australia is one of the few

5-**gram**	**Count**
<s> Australia is one of	1
Australia is one of the	1
is one of the few	1

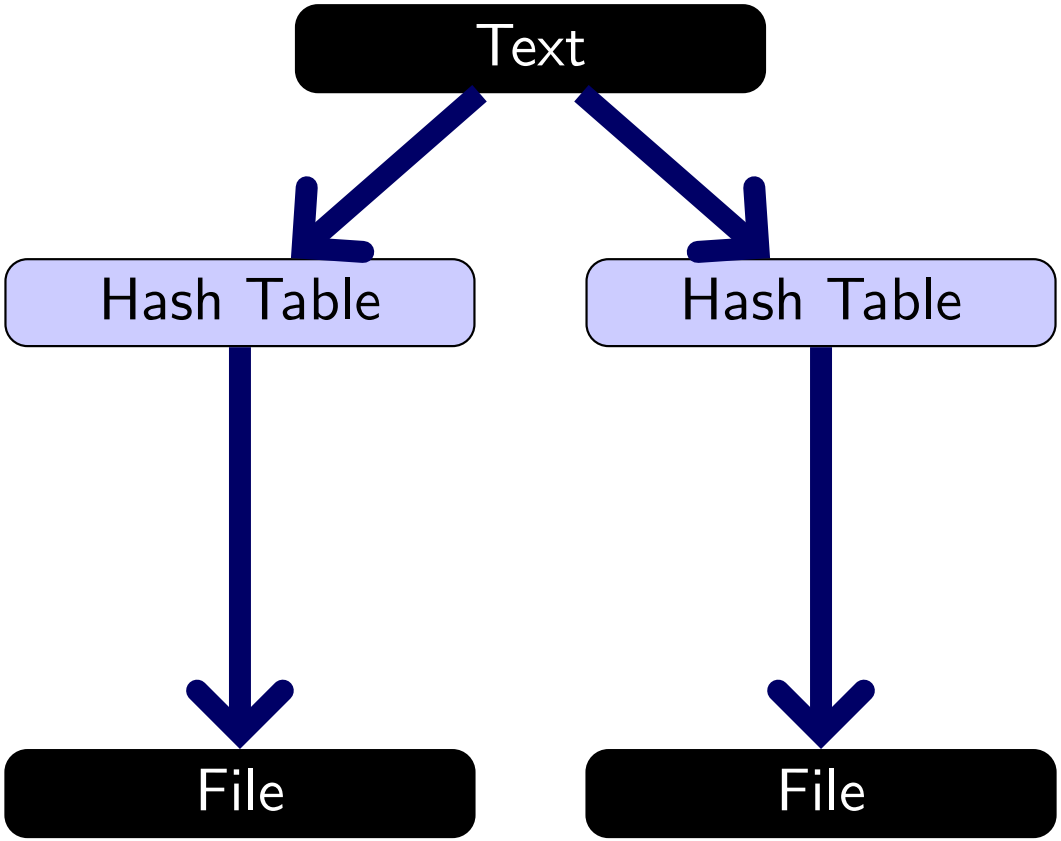
Hash table?

Runs out of RAM.

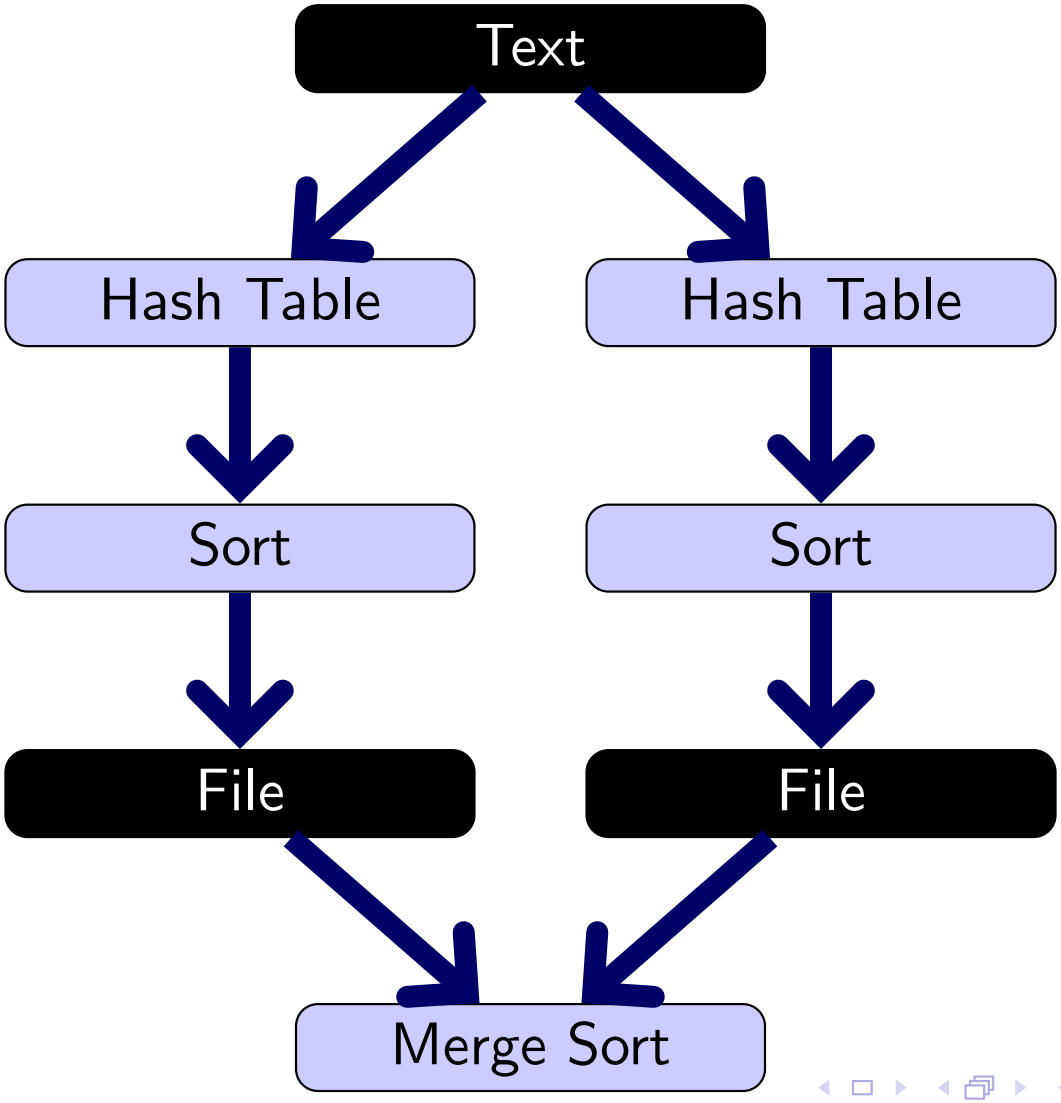
Spill to Disk When RAM Runs Out



Split Data



Split and Merge



Fanout: Merging More Than 2 Files

$\log_{\text{fanout}} |\text{data}|$ passes, each of which reads *and* writes $|\text{data}|$.

$$\text{Small fanout} \implies \text{more passes.}$$
$$\text{Large fanout} \implies \text{more disk seeks.}$$

Compromise: 64 MB buffer per file; fanout $= (\text{RAM}/64\text{MB}) - 1$.

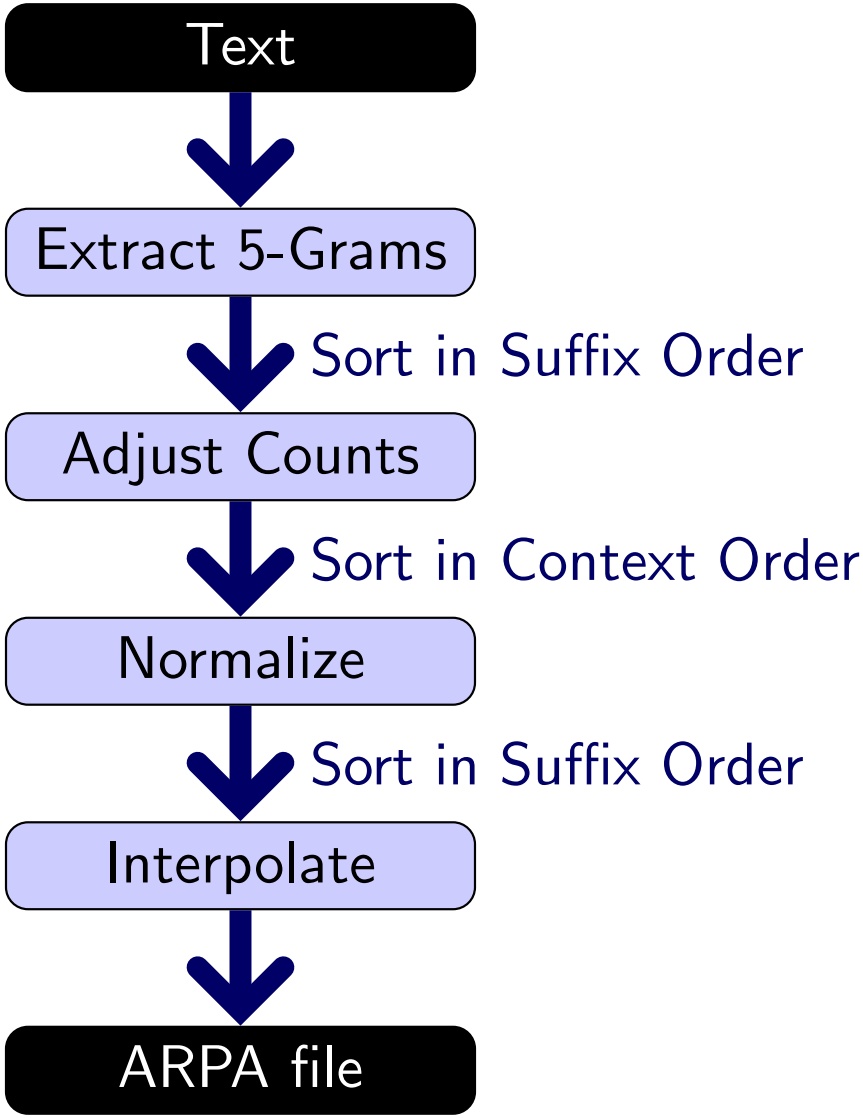
Optimizing Merge Sort

- Laziness: do the last merge as the file is being read
- Sharding: partition the data by key
- Combine counts during merge passes

Now we can estimate Stupid Backoff models.

What about modified Kneser-Ney?

Modified Kneser-Ney



Sorting Orders

Suffix Order

5	4	3	2	1
A	A	A	A	A
C	A	A	A	A
A	Y	A	B	A
A	Y	Z	B	A
A	A	A	A	Y
C	A	B	A	Z

Context Order

4	3	2	1	5
A	A	A	A	A
A	A	A	A	Y
C	A	A	A	A
C	A	B	A	Z
A	Y	A	B	A
A	Y	Z	B	A

Each step is a streaming algorithm in suffix or context order.

Adjust Counts

$$\text{adjusted}(w_1^n) = \begin{cases} \text{count}(w_1^n) & \text{if } n = 5 \text{ or } w_1 = \text{<s>} \\ |v : \text{count}(vw_1^n) > 0| & \text{otherwise} \end{cases}$$

Suffix order makes vw_1^n consecutive.

Normalize

$$\text{normalized}(w_1^n) = \frac{\text{adjusted}(w_1^n) - \text{discount}(\text{adjusted}(w_1^n))}{\sum_v \text{adjusted}(w_1^{n-1} v)}$$

Context order makes $w_1^{n-1} v$ consecutive.
The denominator is computed by a thread that reads ahead.

Interpolate

$$p(w_n|w_1^{n-1}) = \text{normalized}(w_1^n) + \text{backoff}(w_1^{n-1})p(w_n|w_2^{n-1})$$

Suffix order means $p(w_n|w_2^{n-1})$ and $p(w_n|w_1^{n-1})$ are close.

Estimation Results

Implz	132 billion tokens, 4.3 days, one machine (140 GB RAM), lossless
Google 2007	31 billion tokens, 2 days, 400 machines, pruned singleton words

Querying: The Problem

The LM is queried ≈2.7 million times per sentence translated.

$\Longrightarrow$ Represent the LM to make queries fast.

Querying: The Problem

The LM is queried ≈2.7 million times per sentence translated.
$\Longrightarrow$ Represent the LM to make queries fast.

The University of Edinburgh owns a machine with 1 TB RAM.
$\Longrightarrow$ Find more training data!

Example Language Model

Unigrams

Words	log p	log b
<s>	-∞	-2.0
iran	-4.1	-0.8
is	-2.5	-1.4
one	-3.3	-0.9
of	-2.5	-1.1

Bigrams

Words	log p	log b
<s> iran	-3.3	-1.2
iran is	-1.7	-0.4
is one	-2.0	-0.9
one of	-1.4	-0.6

Trigrams

Words	log p
<s> iran is	-1.1
iran is one	-2.0
is one of	-0.3

Example Queries

Unigrams

Words	log p	log b
<s>	-∞	-2.0
iran	-4.1	-0.8
is	-2.5	-1.4
one	-3.3	-0.9
of	-2.5	-1.1

Bigrams

Words	log p	log b
<s> iran	-3.3	-1.2
iran is	-1.7	-0.4
is one	-2.0	-0.9
one of	-1.4	-0.6

Trigrams

Words	log p
<s> iran is	-1.1
iran is one	-2.0
is one of	-0.3

Query: <s> iran is

$\log p(\text{is} \mid \text{<s> iran}) = -1.1$

Query: iran is of

$$\begin{array}{ll} \log p(\text{of}) & -2.5 \\ \log \text{backoff}(\text{is}) & -1.4 \\ \log \text{backoff}(\text{iran is}) & + -0.4 \\ \hline \log p(\text{of} \mid \text{iran is}) & = -4.3 \end{array}$$

Lookup I: Giant Hash Table

Hash every n-gram to a 64-bit integer. Ignore collisions.
Store *n*-grams in custom linear probing hash tables.

Fastest.

Lookup II: Minimal Perfect Hash [Talbot and Brants, 2008]

Function from seen *n*-grams to an integers:

Perfect Each seen *n*-gram has a *unique* integer.

Minimal Integers are in $[0, |\text{n-grams}|)$.

Use the integer to index an array of probability and backoff.

Lookup II: Minimal Perfect Hash [Talbot and Brants, 2008]

Function from seen n-grams to an integers:

Perfect Each seen n-gram has a *unique* integer.

Minimal Integers are in $[0, |\text{n-grams}|)$.

Use the integer to index an array of probability and backoff.

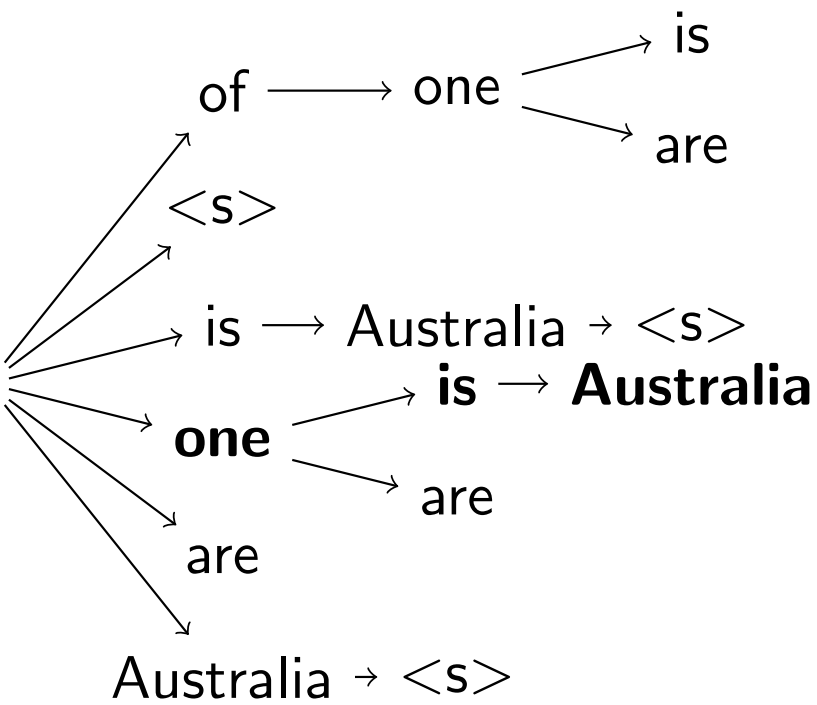
False Positives

Problem Unseen n-grams collide with seen n-grams.

Solution Store an f-bit signature $\implies$ 2^{-f} false positives.

Lowest memory but has false positives.

Lookup III: Reverse Trie



Most popular format. Exact, middle memory usage.

Optimizing Storage

Bit-Level Packing

Use only as many bits as needed.

Chop Bits [Raj and Whittaker, 2003]

In a sorted array, encode bits by the offset where they roll over.

Quantization [Whittaker and Raj, 2001] [IRSTLM]

Cluter floats into 2^q bins then store q bits/float.

Storing Less

Filtering

Remove *n*-grams that will not be queried during decoding.

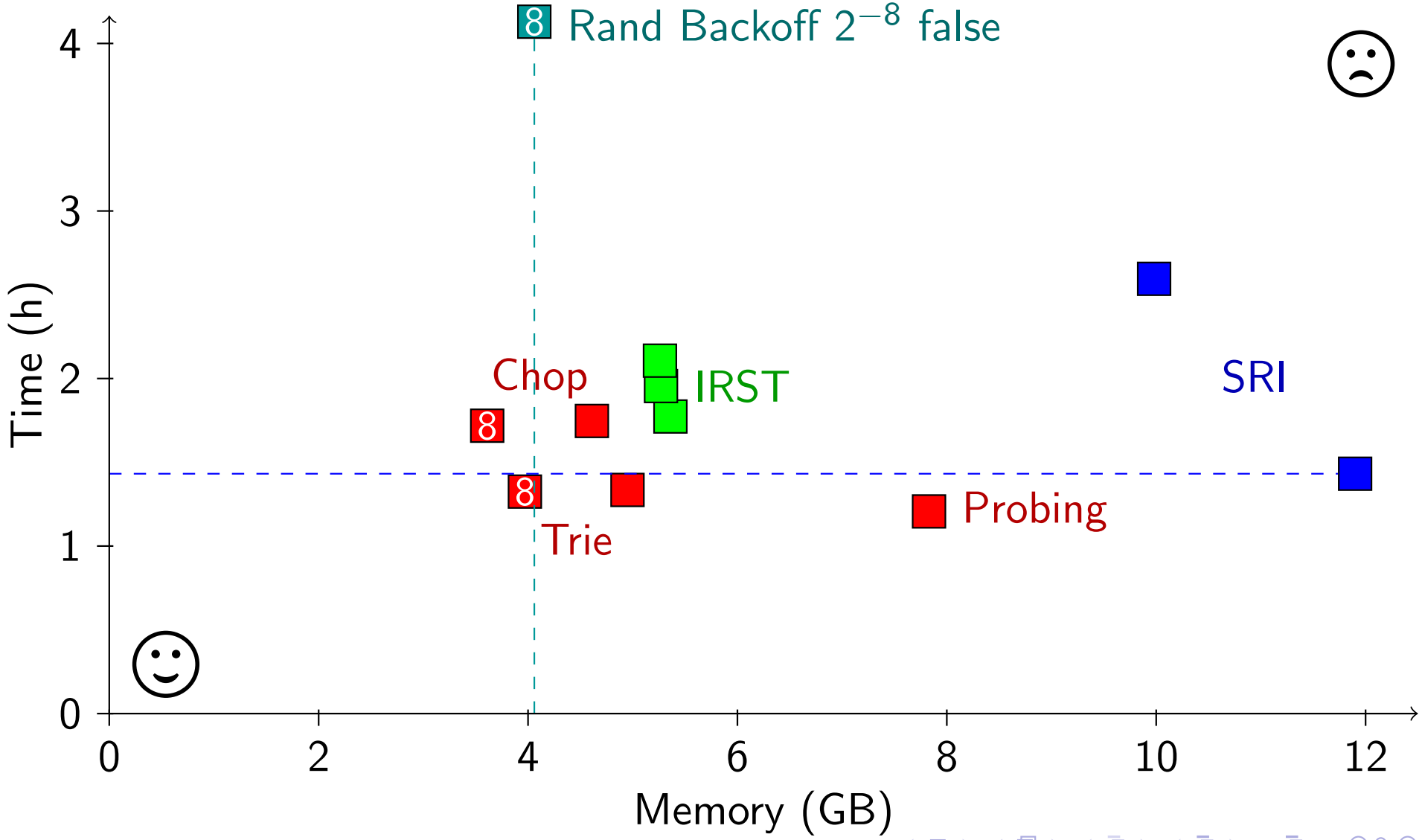
Pruning

Remove low-count *n*-grams.

Encode Less Than Probability and Backoff

- Stupid backoff has one value: count.
- Collapse probability and backoff (requires more CPU time).

Moses Benchmarks: Single Threaded



Querying Takeaways

- Optimizing the LM optimizes the decoder.
- Approximations can have negligible impact on quality
  27.29 BLEU $\rightarrow$ 27.09 BLEU by quantizing to 4 bits.
- There is no single best data structure.

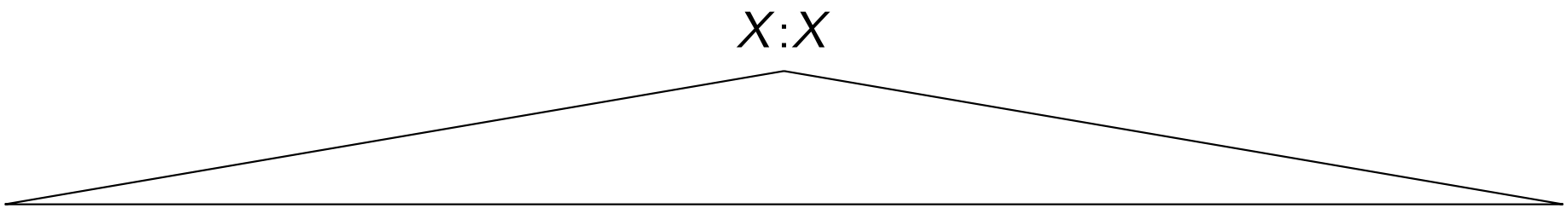
Decoding

March 5: Cube Pruning
Can we do better?

Recall: Cube Pruning

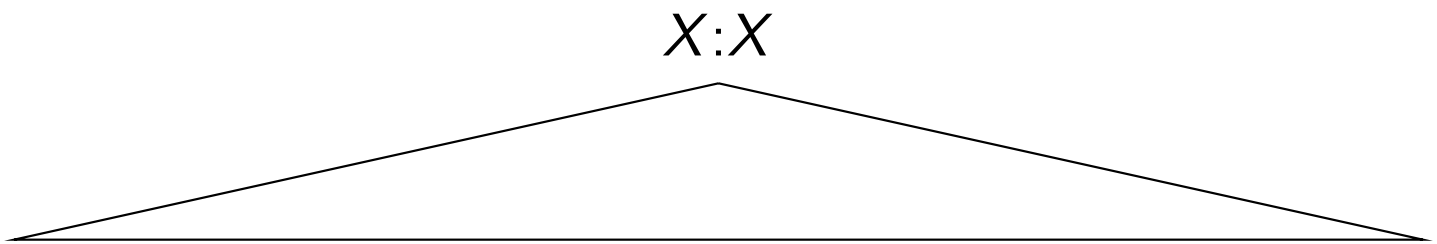
- Each constituent is visited in bottom-up (topological) order.
- Generate a fixed number of hypothese per constituent.
- Estimated probabilities for words on the edge.

A Beam



Left State		**Right State**
countries that ⊣	maintain diplomatic	relations ties ⊢ with North Korea .
countries that have ⊣	an embassy in	⊢ DPR Korea .
country ⊣	that maintains some diplomatic ties	⊢ in North Korea .
nations which has ⊣	some diplomatic ties	⊢ with DPR Korea .
country ⊣	that maintains some diplomatic ties	⊢ with DPR Korea .

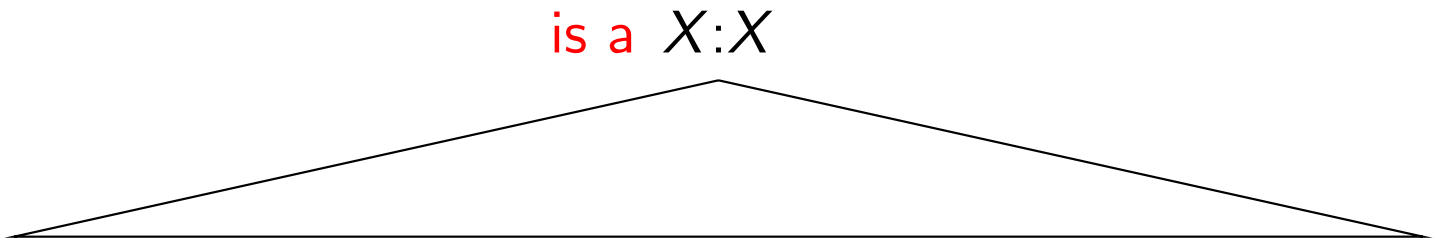
A Beam With State



Left State		Right State	Score
(countries that	⊣ ◇ ⊢	with North Korea .)	-2
(nations which has	⊣ ◇ ⊢	with DPR Korea .)	-4
(countries that have	⊣ ◇ ⊢	DPR Korea .)	-5
(country	⊣ ◇ ⊢	in North Korea .)	-8
(country	⊣ ◇ ⊢	with DPR Korea .)	-9

◇ denotes words omitted by state.

A Beam With State



Left State		Right State	Score
is a (countries that	⊣ ◇ ⊢	with North Korea .)	-2
is a (nations which has	⊣ ◇ ⊢	with DPR Korea .)	-4
is a (countries that have	⊣ ◇ ⊢	DPR Korea .)	-5
(country	⊣ ◇ ⊢	in North Korea .)	-8
(country	⊣ ◇ ⊢	with DPR Korea .)	-9

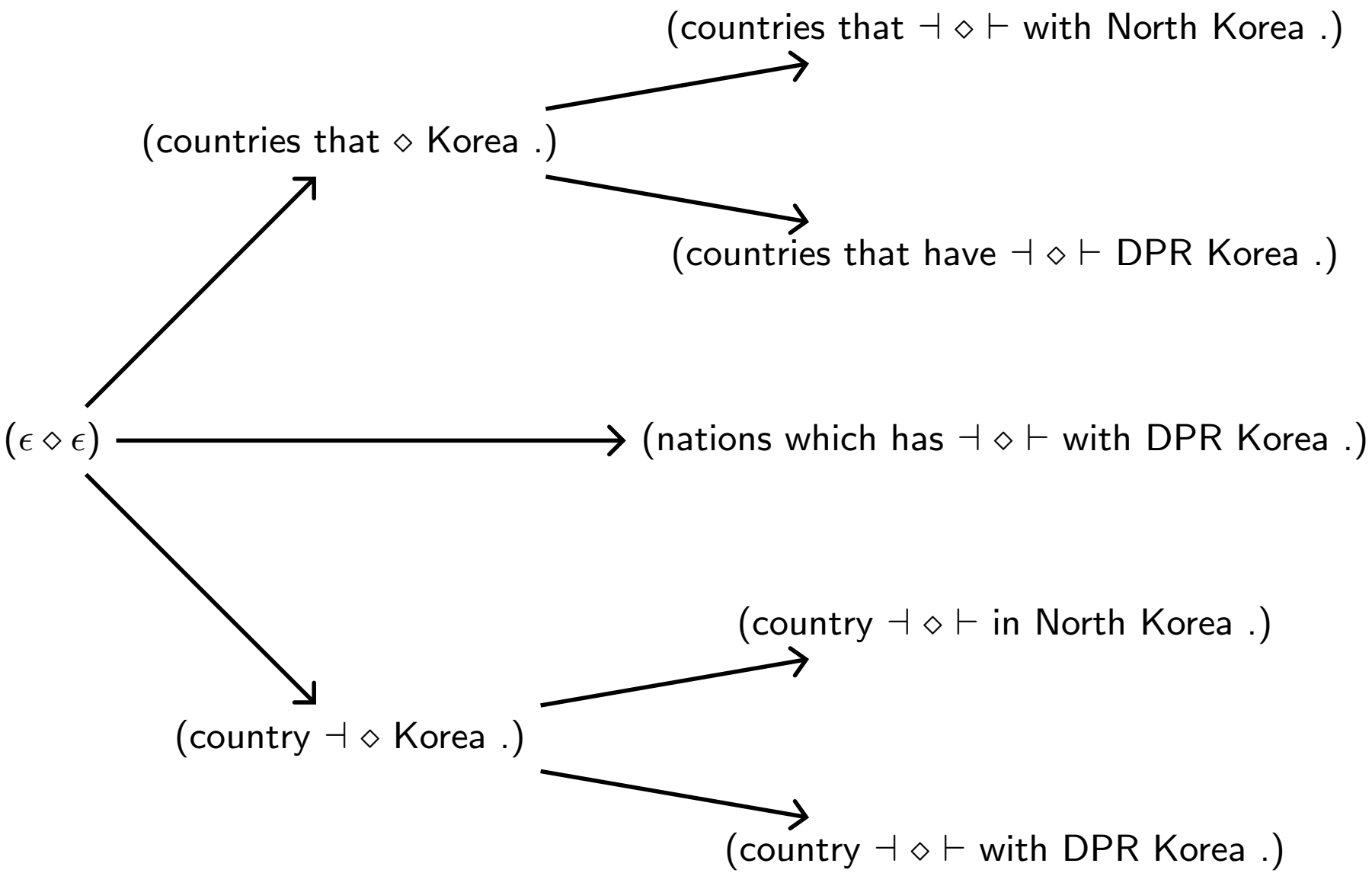
Rule "is a X:X"

Cube pruning tests variants of the same bad idea.

High Level Idea of Incremental Search

Group hypotheses by common outer words.
Score outer words first, see how well they do.

Make a tree


(countries that ⊣ ◇ ⊢ with North Korea .)
(countries that ◇ Korea .)
(countries that have ⊣ ◇ ⊢ DPR Korea .)
(ϵ ◇ ϵ)
(nations which has ⊣ ◇ ⊢ with DPR Korea .)
(country ⊣ ◇ ⊢ in North Korea .)
(country ⊣ ◇ Korea .)
(country ⊣ ◇ ⊢ with DPR Korea .)

Use the Tree

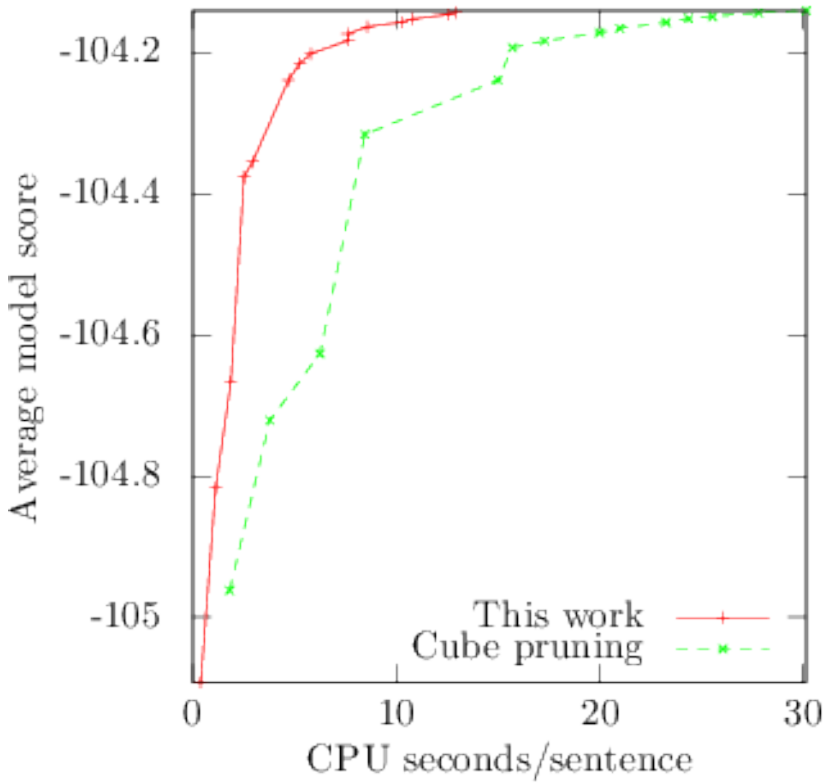
Try "in (country $\dashv \diamond \vdash$ Korea .)" once and see if you like it.

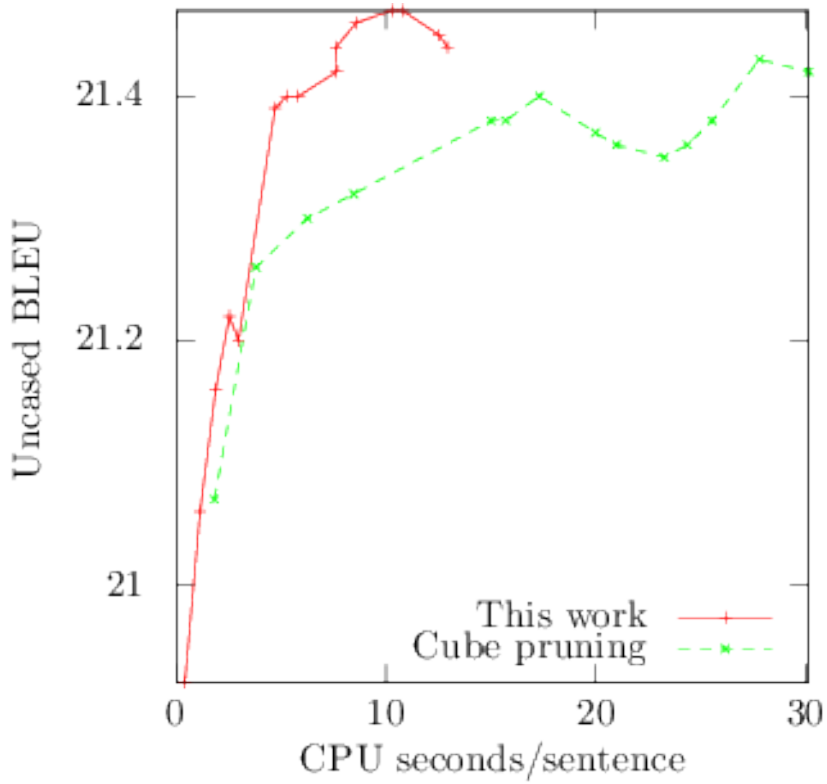
Use the Tree

Try "in (country $\dashv \diamond \vdash$ Korea .)" once and see if you like it.

Pop top candiate off the priority queue, expand children, push.

Results


-104.2
-104.4
-104.6
-104.8
-105
Average model score
0
10
20
30
CPU seconds/sentence
This work
Cube pruning


21.4
21.2
21
Uncased BLEU
0
10
20
30
CPU seconds/sentence
This work
Cube pruning