Back-Off Language Model Compression

Boulos Harb, Ciprian Chelba, Jeffrey Dean, Sanjay Ghemawat

{harb,ciprianchelba,jeff,sanjay}@google.com

Google

Outline

- Motivation: Language Model (LM) Size Matters
- Integer Trie LM Representation
- Techniques for LM Compaction:
  - N-gram Map: Block Compression
  - Probabilities and Back-off Weights: Quantization and Block Compression
- Experiments
- Conclusions and Future Work

Google

How Big a Language Model?

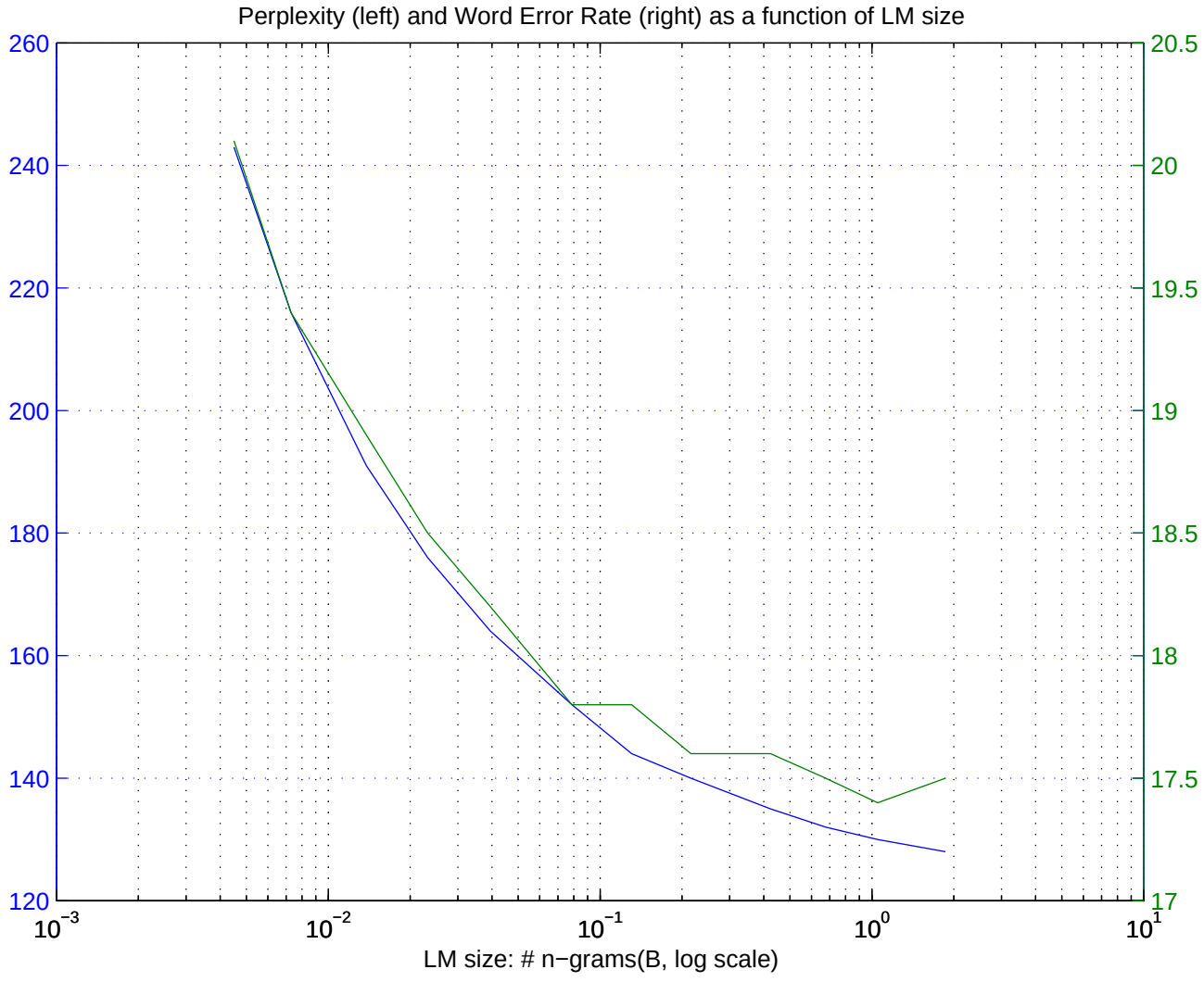
Typical Voicesearch LM training setup is *data rich*:

- vocabulary size: 1 million words, OoV rate 0.57%
- training data: 230 billion words from google.com query logs, after text normalization for ASR

Order	# n-grams	pruning	PPL	n-gram hit-ratios
3	15M	entropy	190	47/93/100
3	7.7B	1-1-1	132	97/99/100
5	12.7B	1-1-2-2-2	108	77/88/97/99/100

- A lot of *float* numbers along with n-grams!

Is Bigger 1st Pass LM Better? YES!



Integer Trie LM Representation

- 1-1 mapping between n-grams and dense integer range using *integer trie*:
  - 2 vectors that concatenate, for each n-gram context:
    - cummulative diversity count
    - list of future words
- look-up time: $\mathcal{O}((n-1) \cdot \log(V))$, in practice much smaller
- once n-gram key is identified, lookup probability and back-off weight in 2 separate arrays

Google

Integer Trie LM Compaction

Sequence of entries in vectors is far from memoryless.

N-gram Map:

- block compression for both diversity and word vectors
  - `GroupVar`: variable integer length per block
  - `RandomAccess`: fixed integer length per block
  - `CompressedArray`: a version of Huffman coding enhanced with simple operators

Probabilities and Back-off Weights:

- linear quantization to 1 byte
- block compression of 4 byte bundles cast to `int`

Google

Experiments

Google Search by Voice LM:

- : 3-gram LM, 13.5 million n-grams
- 1.0/8.2/4.3 million 1/2/3-grams, respectively

We measure:

- storage: representation rate, no. bytes/n-gram
- speed (relative to uncompressed): computed PPL on unseen test data

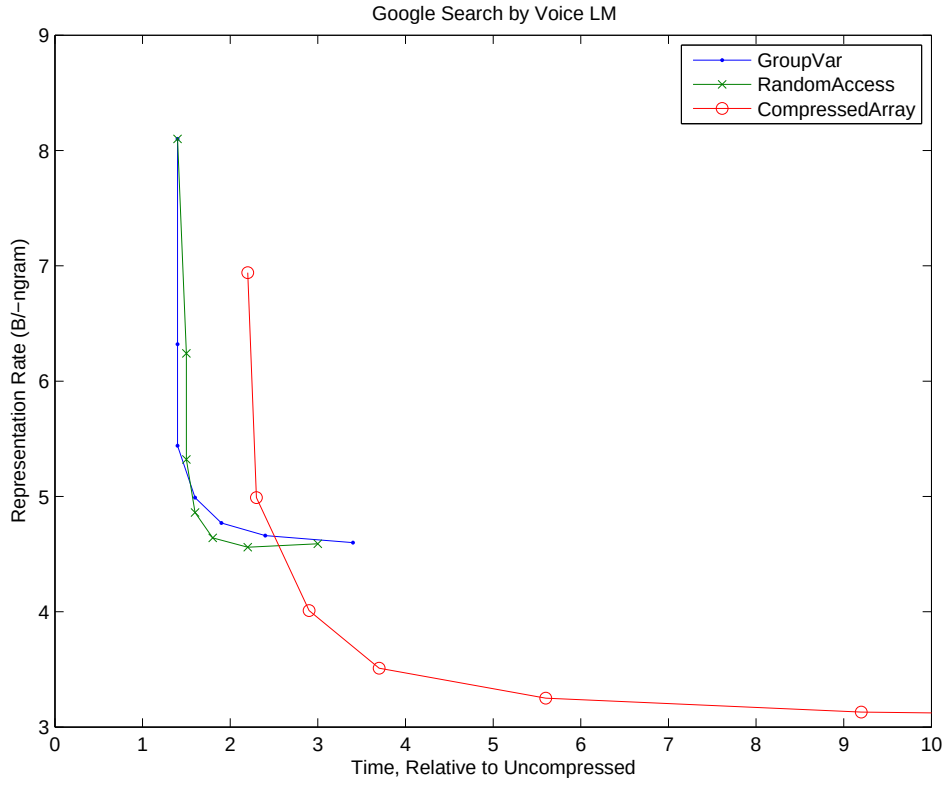
Google

LM Representation Rate vs. Speed

Compression Technique	Block Length	Relative Time	Bytes per n-gram
None	—	1.0	13.2
Quantized	—	1.0	8.1
CMU 24b, Quantized	—	1.0	5.8
GroupVar	8	1.4	6.3
	64	1.9	4.8
	256	3.4	4.6
RandomAccess	8	1.5	6.2
	64	1.8	4.6
	256	3.0	4.6
CompressedArray	8	2.3	5.0
	64	5.6	3.2
	256	16.4	3.1
+ logprob/bow arrays	256	19.0	2.6

Google

LM Representation Rate vs. Speed



- 1 billion 3-grams: 4GB of RAM @acceptable lookup speed

Conclusions

- can achieve 2.6 bytes/n-gram representation rate if speed is not a concern
- 4 bytes/n-gram at reasonable speed
- 1st pass LM using 1 billion n-grams is feasible, with excellent results in WER:
  - 10% rel. reduction in WER over 13.5 million n-gram LM baseline

Google

Future Work

- Integrate with reachable composition decoder at real-time factor close to 1.0:
  - Allauzen, Riley, Schalkwyk: A Generalized Composition Algorithm for Weighted Finite-State Transducers
- Scale up to 10 billion n-grams (40-60GB)?

Google