Will recurrent neural network language models scale?

Tony Robinson
Tom Ash
Niranjani Prasad
David Mrva
Will Williams

Cantab Research Limited

Abstract

In “Up from trigrams! The struggle for improved language models” (1991) Fred Jelinek described the first use of trigrams in 1976 and then lamented “The surprising fact is that now, a full 15 years later, after all the solid progress in speech recognition, the trigram model remains fundamental”. Almost two decades later the situation was largely unchanged but in 2010 Tomas Mikolov presented the "Recurrent neural network based language model” (RRN LM). After many decades we now have a new means for language modelling which is clearly much better than the n-gram. Having actively pioneered the use of RNNs in the 80's and 90's the concern arises as to whether the RNNs will continue to outperform or whether there will be another “neural net winter”. This talk address the problem of whether RNN LMs will scale by looking at the scaling properties of n-grams, and then doing the same for RNN LMs. Scaling is considered in terms of LM words, number of parameters, processing power and memory. Preliminary results will be presented showing the largest reductions in perplexity reported so far, an analysis of the performance on frequent and rate words, results on the newly released 1-billion-word-language-modelling-benchmark and the impact on word error rates in a commercial LVCSR system. The talk concludes by justifying whether RNN LMs will scale with respect to the previously incumbent n-grams.

Past performance is no guarantee..

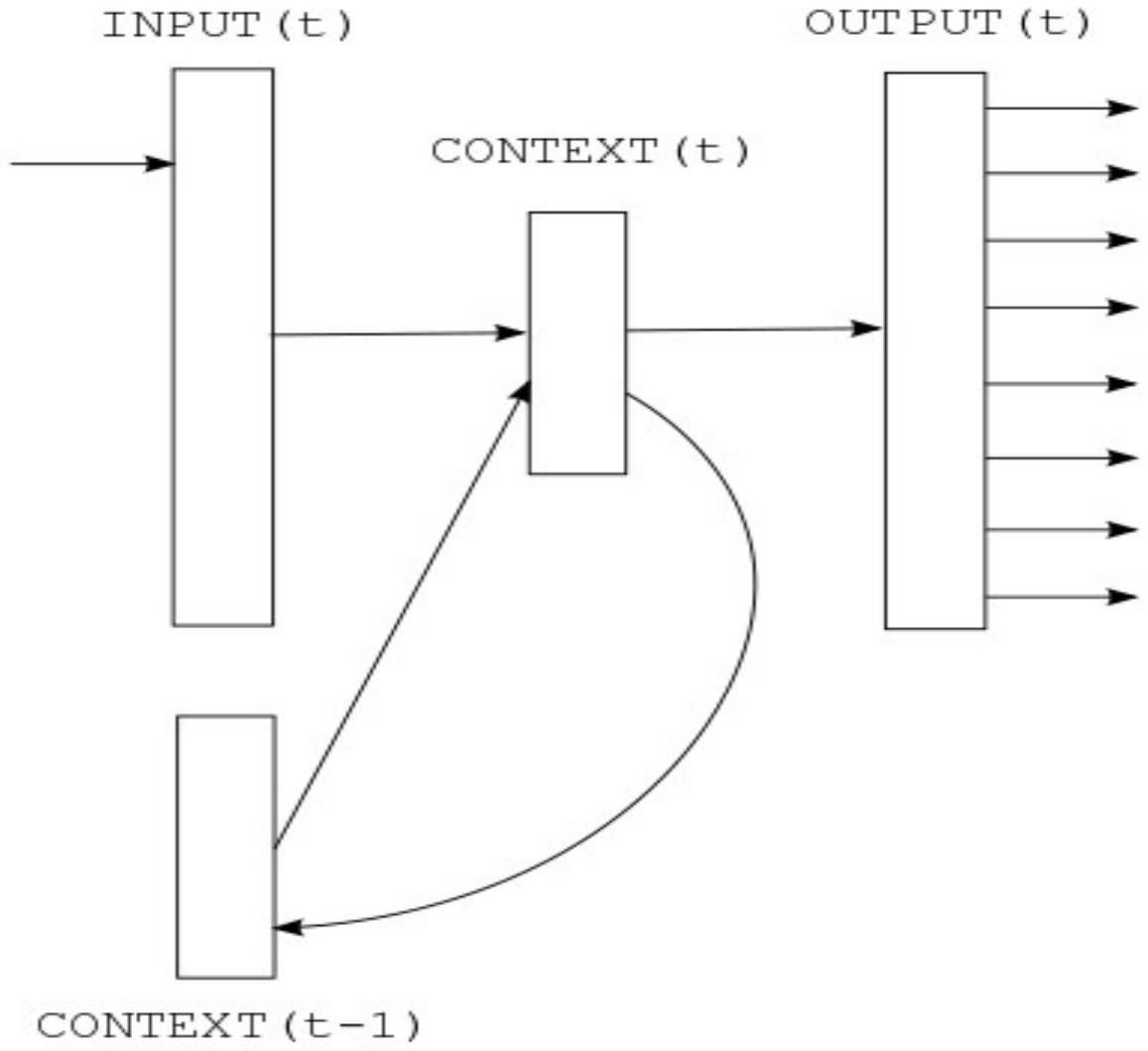
- Recurrent Neural Network (RNN) acoustic models once (briefly) offered the best performance on some tasks
- They didn't scale and Gaussian mixture model HMMs dominated
- RNN Language Models (LMs) now offer the best performance on large vocabulary tasks
- Will RNN LMs survive, i.e. will they scale with increasing LM data and CPU/RAM/HDD?

Background: n-grams

- n-grams dominated statistical language modelling from inception until 2013
- In the simplest form:

  $Pr(w_n|w_{n-1}, w_{n-2}) = c(w_n, w_{n-1}, w_{n-2})/c(w_{n-1}, w_{n-2})$
- In practice use discounting and backoff in the form of a modified Kneser Ney 5-gram (KN 5-gram)

Background: Recurrent Neural Nets

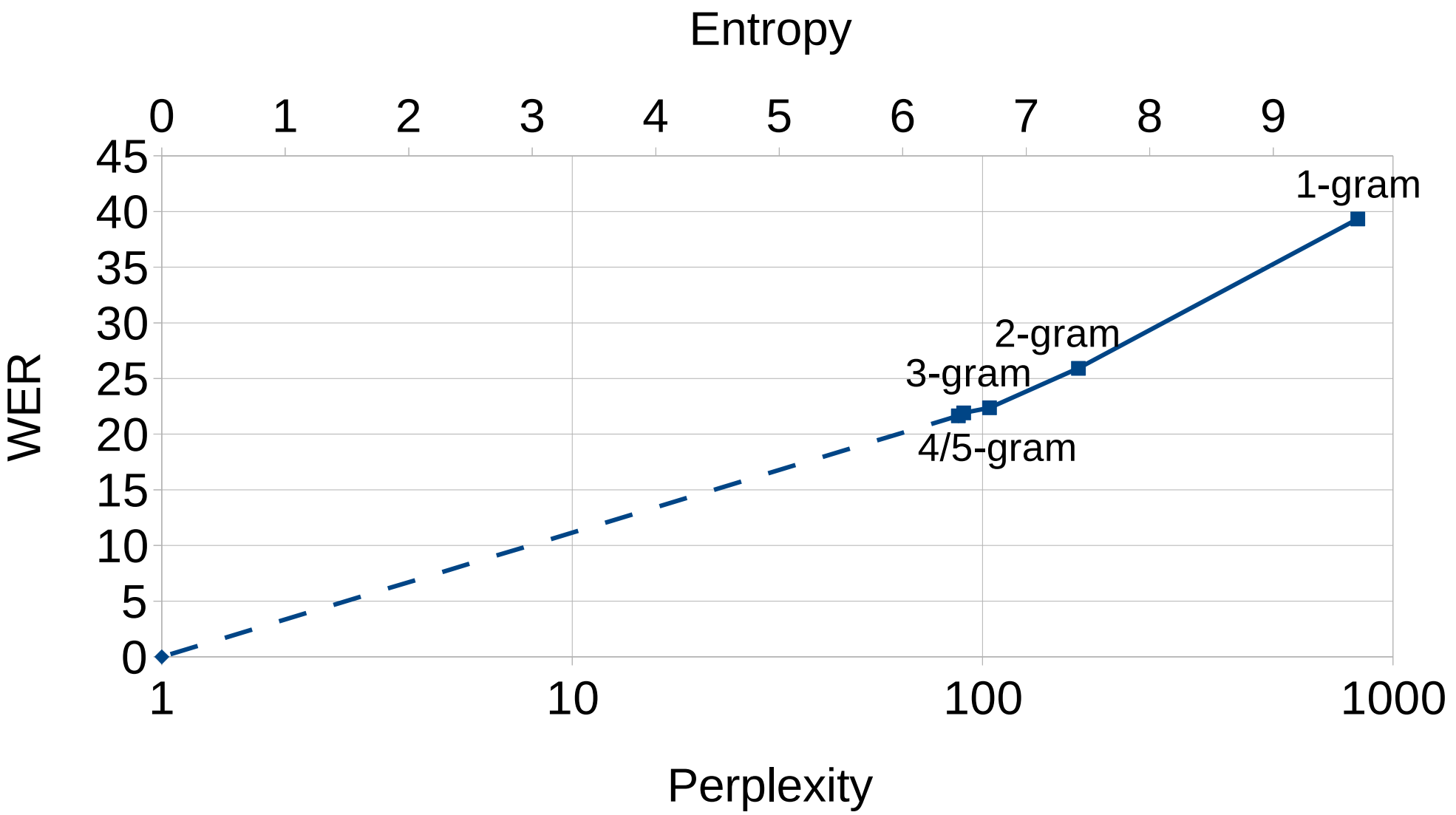


See PhD of Tomas Mikolov and papers/code at rnnlm.org

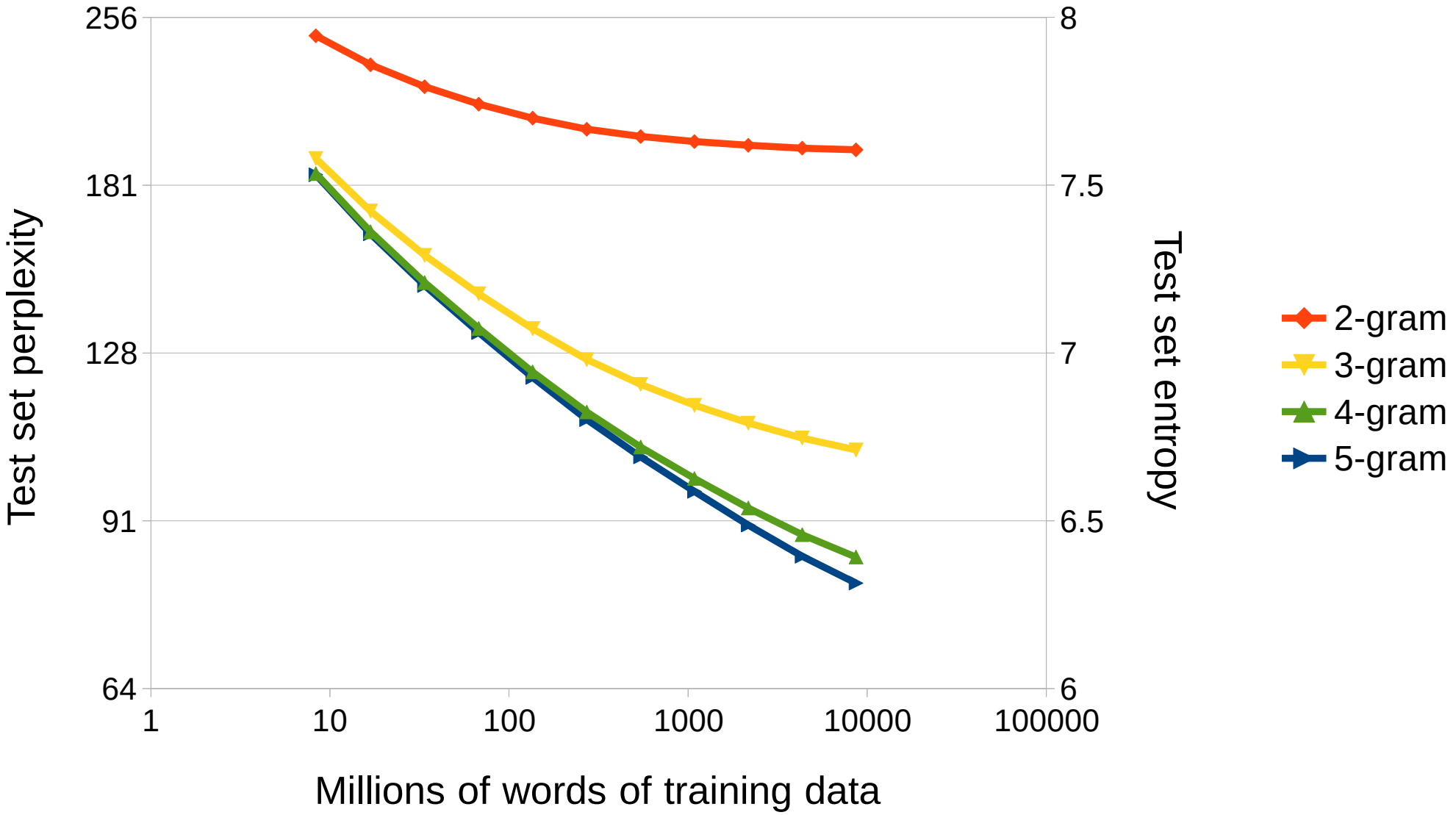
Performance is measured in bits

- LM performance is normally quoted as the perplexity on a held-out test set. Entropy (H) is far more useful
  - $H = 1/N \sum \log_2 Pr(Wn|Hn)$
  - $Perplexity = 2^H$
- Perplexity is the equivalent number of equally likely next words
- Entropy (log perplexity) is the natural measure
  - Used in Viterbi decoding
  - Correlates directly with real metrics such as word error rate
  - Produces straight line plots for the rest of the talk...

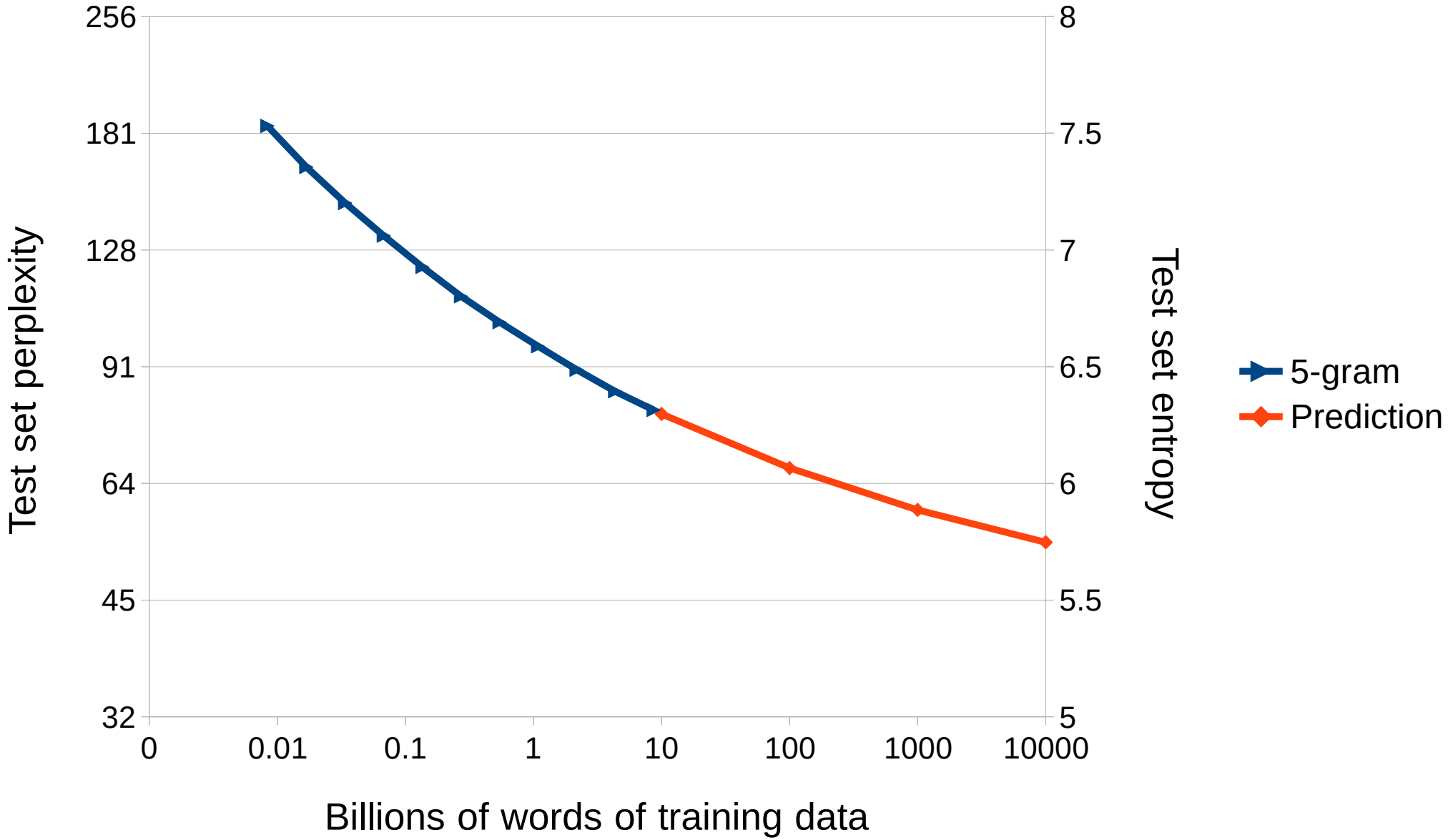
LM entropy and ASR WER



Question: Do n-grams scale?



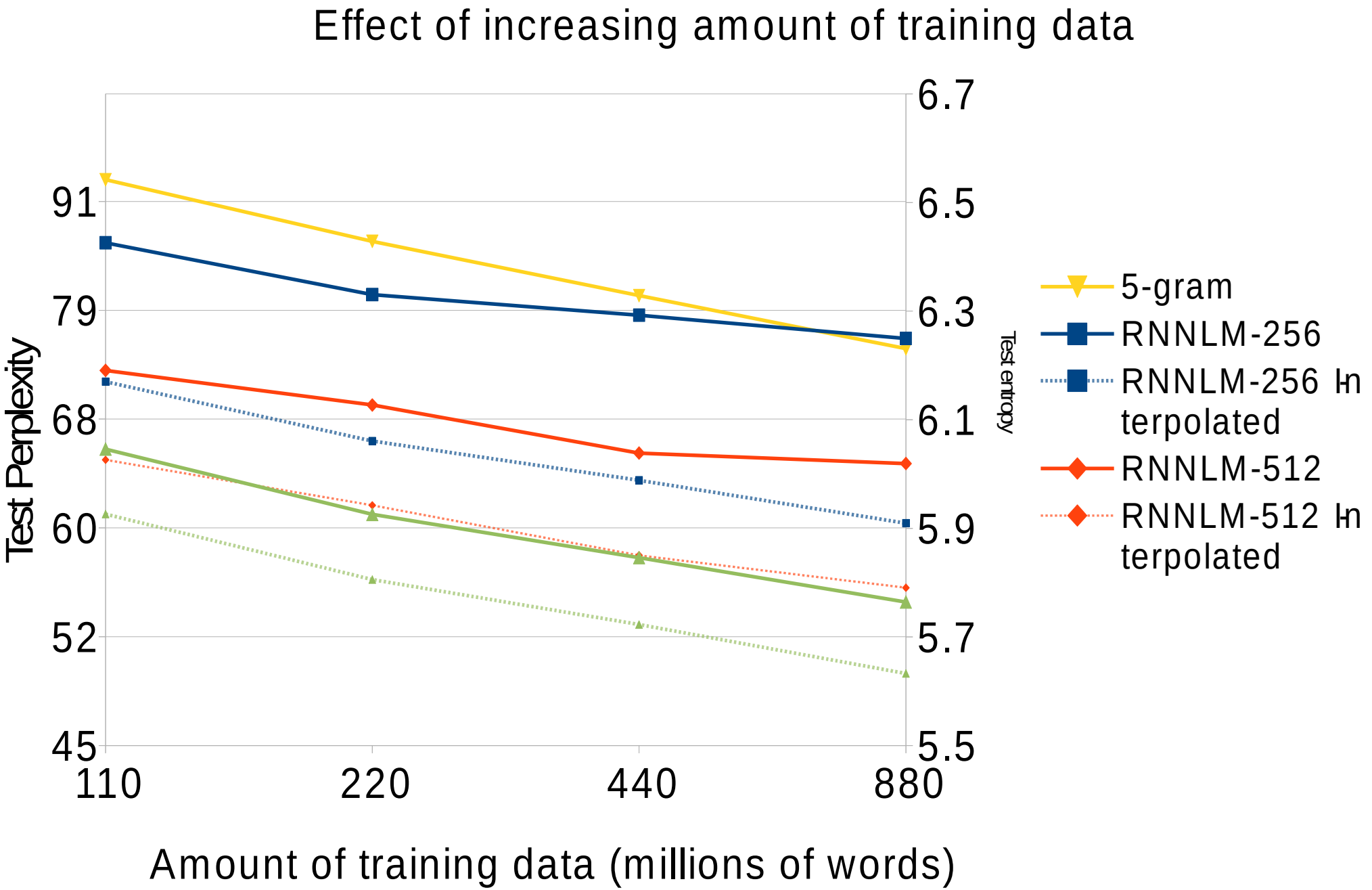
Extrapolating...



So do KN 5-gram scale?

- The asymptote of the previous graph is about perplexity 35 – but it's **very** hard to get close
- At 9bn words the KN 5-gram takes up 362 Gbyte in ARPA format (in 12 hours) and 69 Gbyte in KenLM format (in another 12 hours) – already impractically large for current commercial ASR
- Each order of magnitude more training data above 100bn words will decrease the entropy by less than 3%

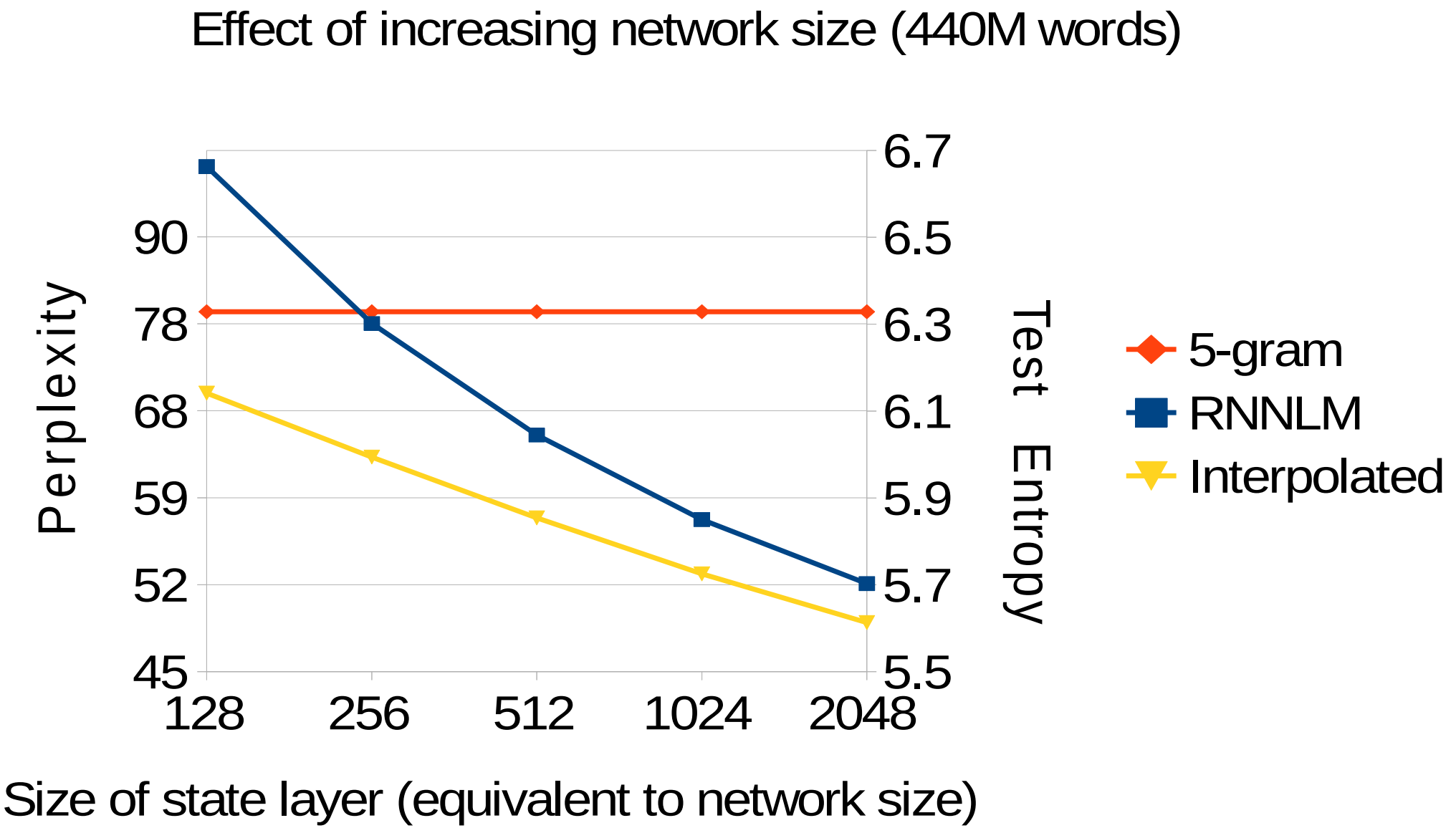
Q: Do RNN LMs scale to 1bn words?



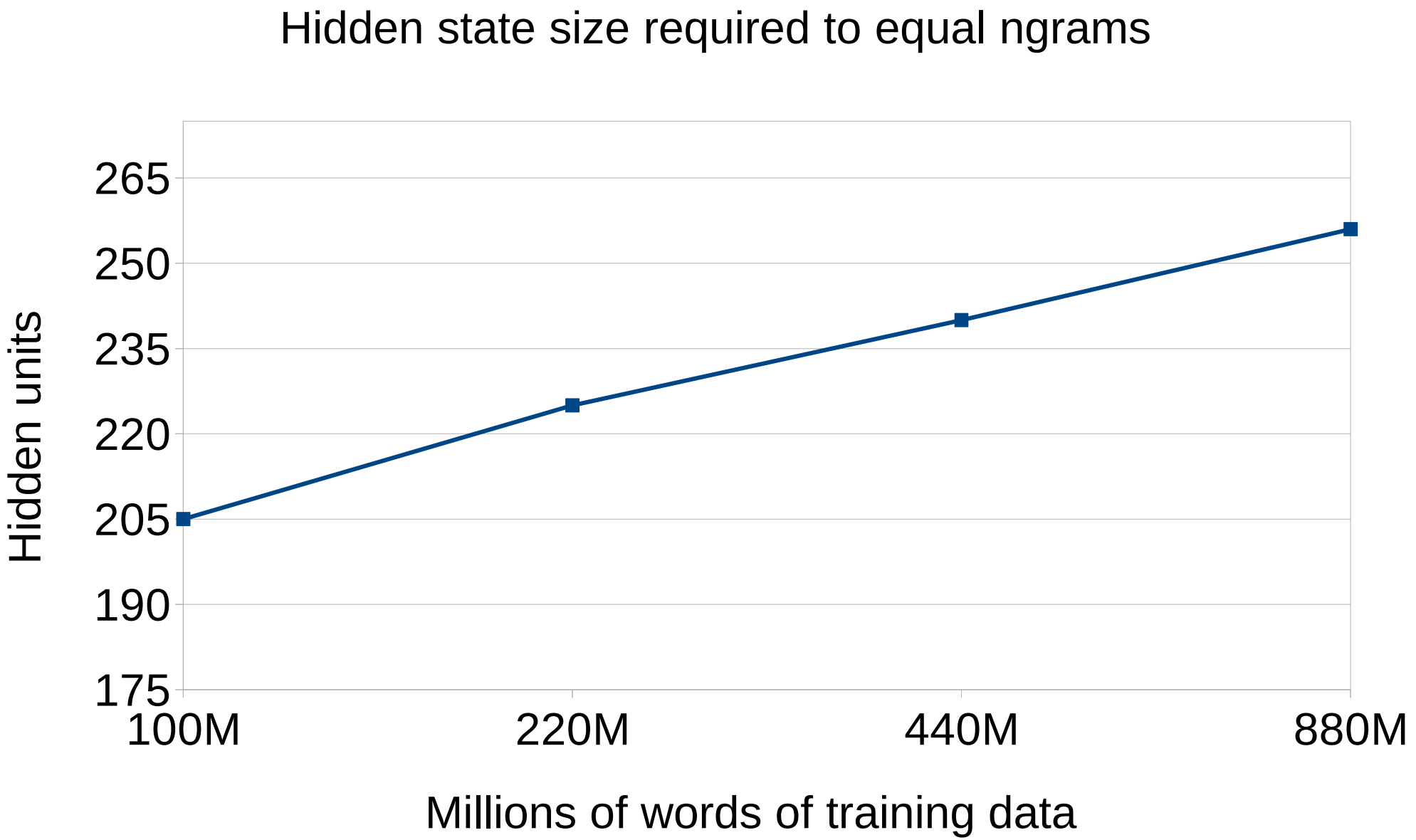
Summary so far...

- Right now RNN LMs significantly out perform grams
- RNN LMs need increasing number of parameters to maintain their lead over n-gram
- But n-grams take up massively more space for modest entropy improvements so which one scales best...

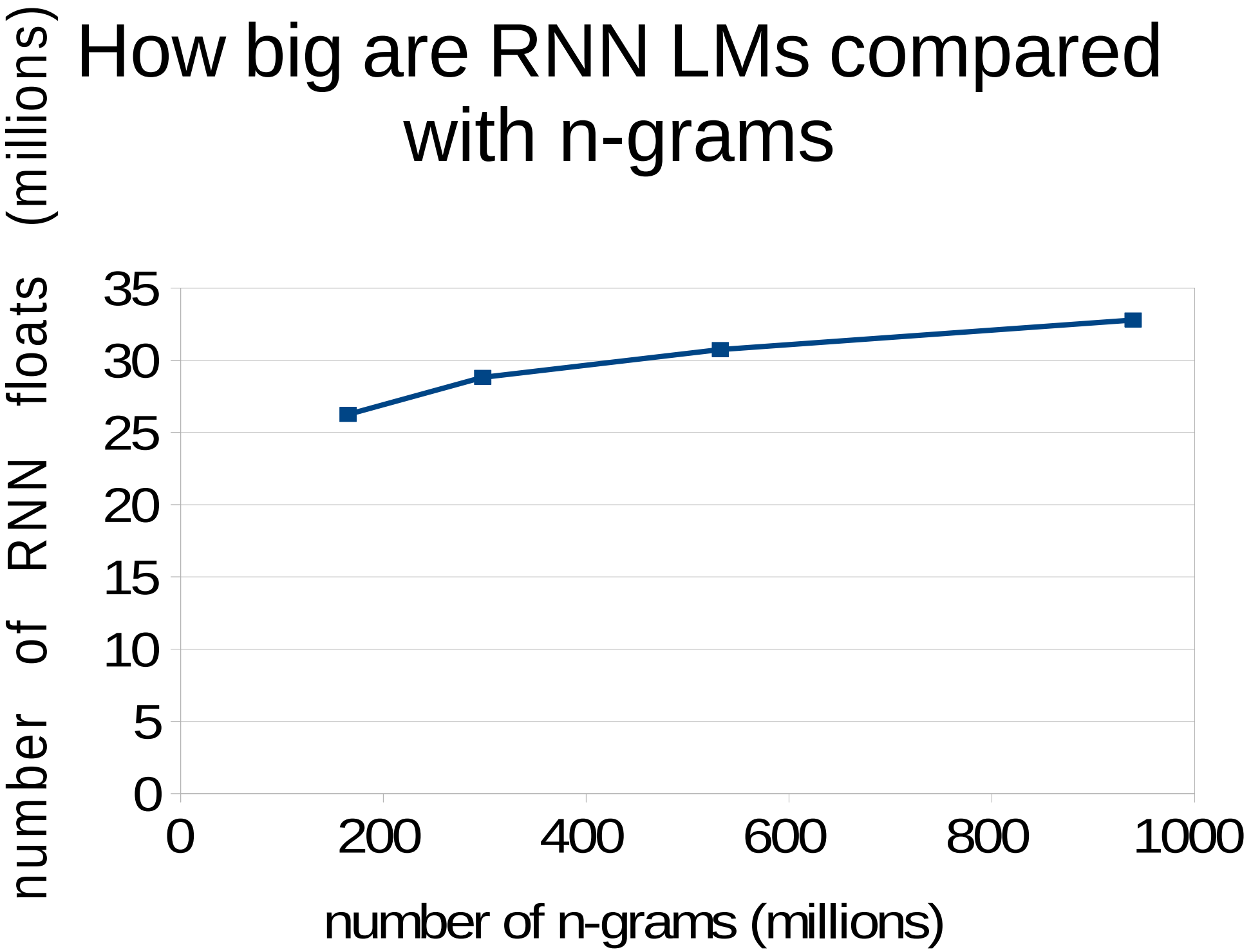
By how much can RNN LMs outperform n-grams



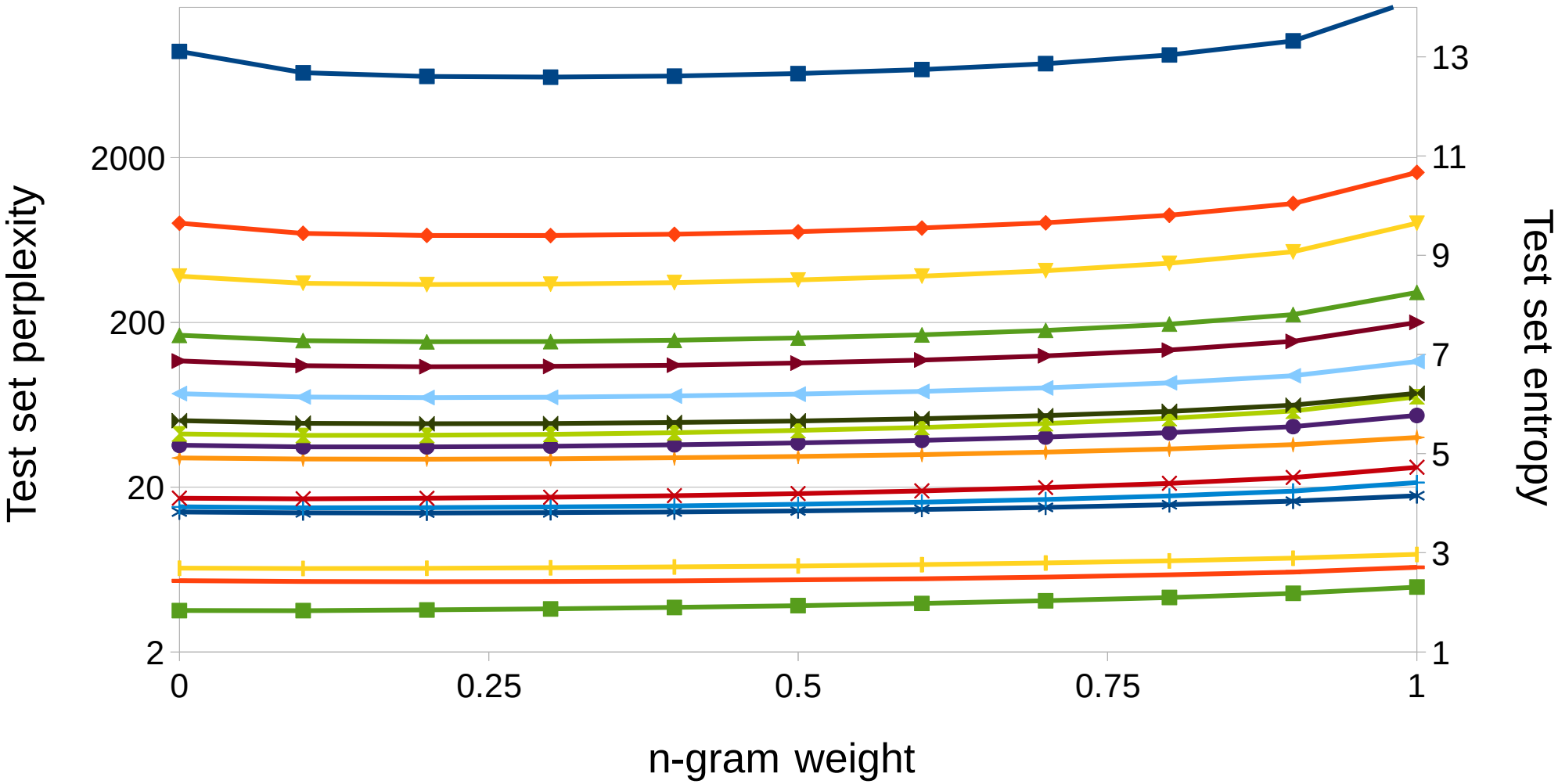
How many parameters do RNN LMs need to equal n-grams?



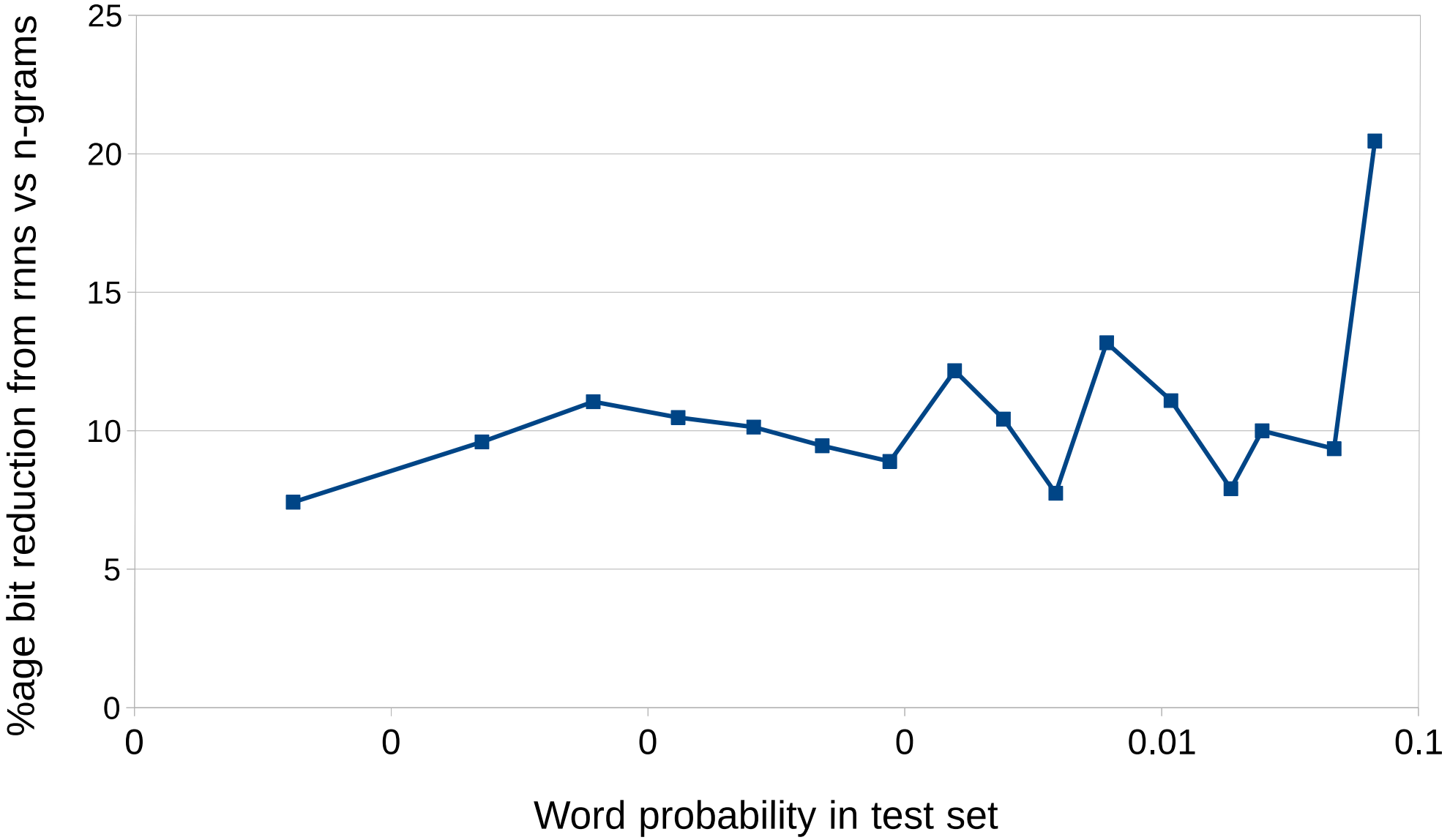
How big are RNN LMs compared with n-grams



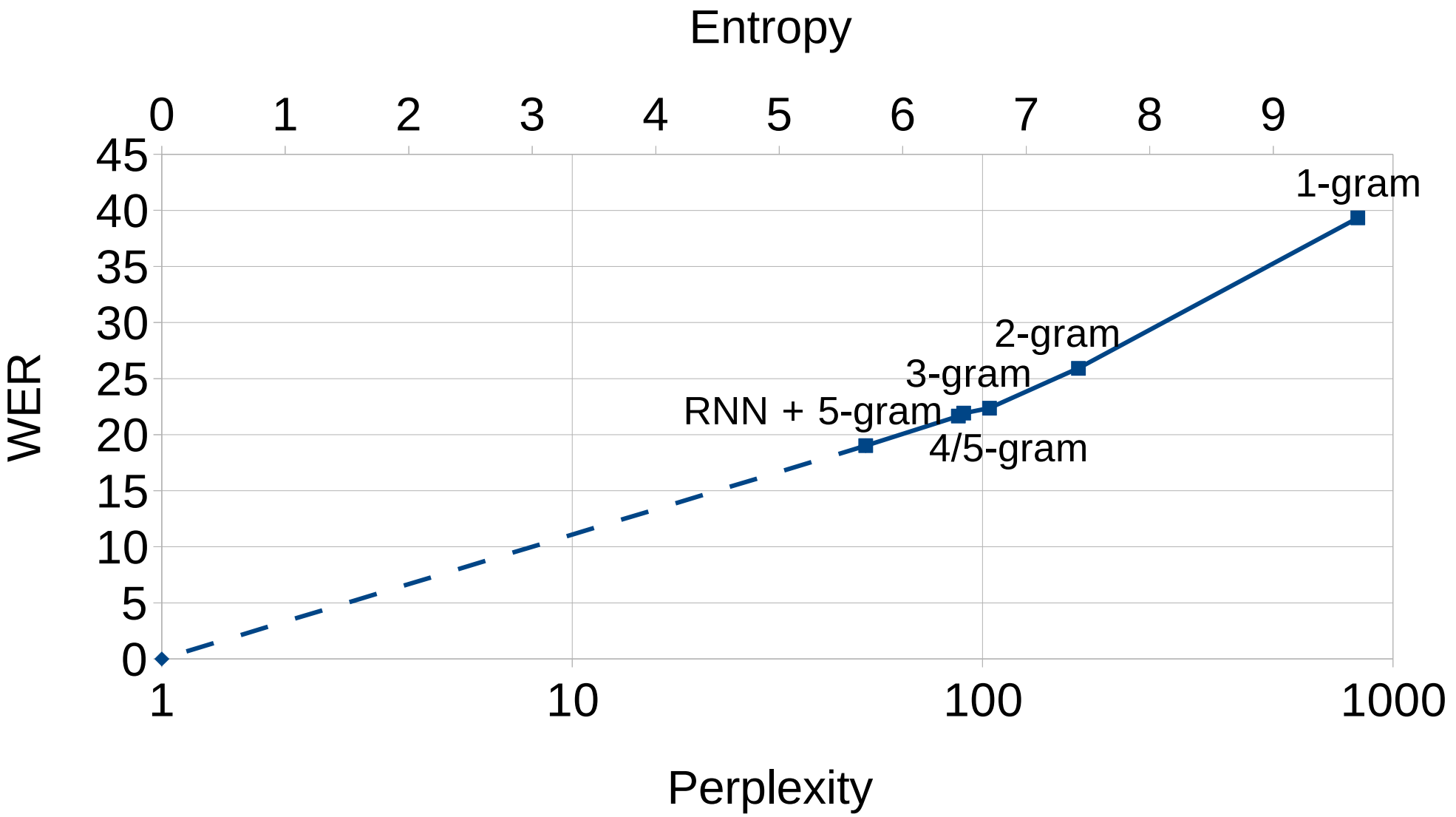
Does the advantage come from frequent or rare words?



Or simpler...


%age bit reduction from rnns vs n-grams
25
20
15
10
5
0
0
0
0
0
0.01
0.1
Word probability in test set

How does this improve ASR WER



1-billion-word-language-modelling-benchmark

Read all about it (MT task, 770k vocab):

https://code.google.com/p/1-billion-word-language-modeling-benchma

Cantab models have 64,000 word vocab and back off to the n-gram

System	# params [millions]	Perplexity
Google: KN 5-gram	1,760	67.6
Google: Recurrent NN1024 + MaxEnt 9-gram (rnn1024)	20,000	51.3
Google: 6% KN 61% rnn1024 13% rnn512 20% SBO	68,890	43.8
Cantab: 2048 nstate	266	65.2
Cantab: 2048 nstate + KN 5-gram	2,026	46.8
Cantab: 3520 nstate (unfinished)	463	61.0
Cantab: 3520 nstate (unfinished) + KN 5-gram	2,223	45.2
Cantab: Frankenstein 2048+3*3520+4096+5-gram	3,956	43.4

RNN LMs for Machine Translation

- The University of Cambridge Russian-English System at WMT13. J. Pino, A. Waite, T. Xiao, A. de Gispert, F. Flego, and W. Byrne. Proceedings of the Eighth Workshop on Statistical Machine Translation. 2013.
- 1000-best list rescoring
- Very preliminary, unfinished, results

System	tune BLEU	san BLEU	test BLEU
5-gram provided by CUED (6bn grams)	33.88	25.95	32.77
5-gram lower case 1bn word LM (1bn grams)	33.78	25.70	32.27
1bn word LM with 3520 nstate RNN LM	34.78	26.64	33.59

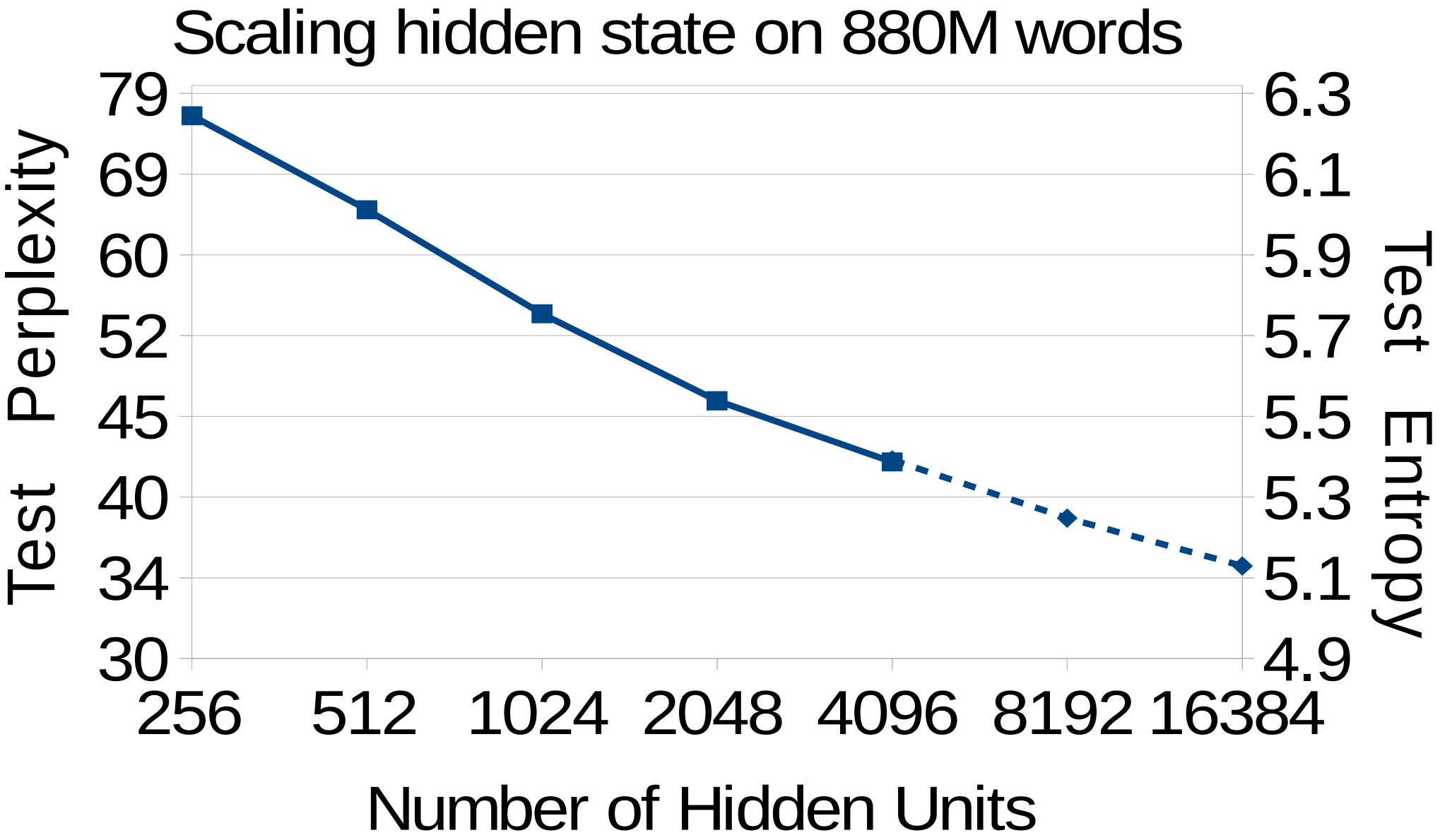
Example n-gram ramblings...

the outcome of the referendum in september will be that scotland back in nineteen eighty three in minneapolis yet from the white house </s> well first let me make a couple points </s> that is a very extreme makeovers that's keeping international golf after kosovo's controversial then controversial high-protein low-carb diet anyway </s> expand on that one constant good with extremism </s>

Example RNN + n-gram ramblings...

the outcome of the referendum in september will be that scotland is iraqi </s> and i believe that </s> there was a deal being reached on this </s> which choice </s> which countries do not comprehend the question remains a huge issue for their membership here </s> of sarkozy </s> he says it is unlikely he'll be quite clear </s> there are very substantial media campaign because one thing we're on in prime minister </s> they're not against it </s>

Could we do better?



So will RNN LMs scale?

- Already achieve perplexity 43.3 on 880m wor
- The equivalent KN 5-gram would:
  - need to be trained on a quadrillion (10^{15}) words
  - Take 1000 TB
- For smallish data (1bn words) the n-gram contributes very little
- Currently limited by GPU speed – 4096 state units take 1.5 weeks to train

Conclusion

- entropy is more useful than perplexity
- A KN 5-gram asymptotes and we think we can reach that limit with a RNN LM (by increasing size and algorithmic improvements)
- RNN LMs scale much better than n-grams

tonyr@cantabResearch.com

(Cantab is hiring!)