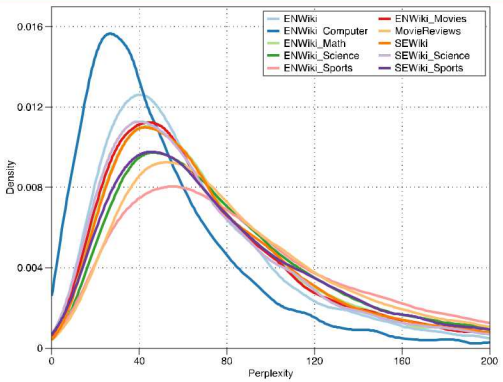
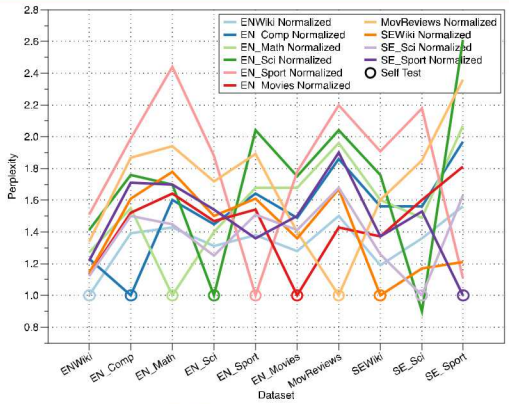
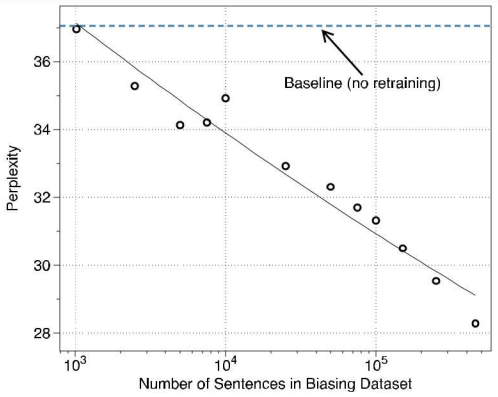
SAND2018-9120C







Efficient transfer learning for neural network language models

Jacek Skryzalin, Hamilton Link, Jeremy Wendt, Richard Field, and Samuel N. Richter

Sandia National Laboratories

Albuquerque, New Mexico, U.S.A.

This work was supported by the Laboratory Directed Research and Development program at Sandia National Laboratories, a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Language Modeling

Sandia National Laboratories

A language model associates probabilities to sequences of tokens.

- Language models are likelihood estimators that support many tasks…
  - Translation and automatic summarization
  - Bot detection
  - Parsing and entity resolution
  - Question answering and information retrieval
- … but they often require vast amounts of data to train
  - The “One Billion Word Benchmark”
  - The complete set of reviews from some large platform
- Can we still train a language model when our training set has a much more modest size?

Neural Network Language Modeling

Sandia National Laboratories

- NNLMs represent the state-of-the-art in language modeling
  - For next-token prediction, neural network language models outperform competing Kneser-Nay and sparse matrix factorization approaches
  - Neural networks also achieve state-of-the-art performance in machine translation and part-of-speech identification
- For next-token prediction, NNLMs require orders of magnitude less data than Kneser-Nay models to achieve similar results
  - Neural networks also take a great deal of care to train properly
  - NNLMs still require a corpus of millions of tokens to train

Neural Network Language Modeling

Sandia National Laboratories

A neural network language model (NNLM) decomposes language modeling into:

- A layer to translate between tokens and a vector space representation
  - Each token is associated to a ≈1000-dimensional vector
- A layer to model the transitions and flow of language
  - Recurrent neural networks (RNN)
  - Gated recurrent units (GRU)
  - Long short-term memory layers (LSTM)
- A layer to associate the hidden layer to a value to be predicted
  - The next token in a sequence
  - Entity resolution or disambiguation
  - Part-of-speech identification
  - Question answering applications

Transfer learning for NNLMs

Sandia National Laboratories

A neural network language model (NNLM) decomposes language modeling into:

- A layer to translate between tokens and a vector space representation ← Language-dependent, but not context-dependent
- A layer to model the transitions and flow of language ← Context-dependent
- A layer to associate the hidden layer to a value to be predicted ← Language-dependent, but not context-dependent

Solution: Learn parameters on an appropriate large corpus; refine the transition (middle) layer on a smaller dataset.

Preprocessing…

Sandia National Laboratories

- Convert all tokens to lower-case
- Retain only the most frequently (64k) occurring tokens
- Resolve how to split hyphenated tokens, contractions, and punctuation
- Resolve how to input data into the network (e.g., each sentence separately, a continuous stream of tokens, etc.)
- Many languages conjugate verbs and decline nouns; unless a truly large corpus is available, it might be prudent to separate the stem and ending of a word

Training…

Sandia National Laboratories

- Use a 2-layer LSTM network designed to predict the next token in a sentence
- For each token, the LSTM stores a 2,048-dimensional state
- Predicting 64k tokens can be onerous; use a sampled softmax loss function instead!
  - Sample 1k incorrect tokens to estimate the incorrect tokens' contribution to the loss
  - Sample tokens proportional to (unigram distribution)$^{0.4}$
- Regularize using 25% dropout between layers
- Train using the ADAM optimizer with a learning rate decreasing from 0.001 to 0.0001

Retraining…

- Freeze the values in the input and output layers
- Continue training on a smaller, specialized corpus
- Depending on how small the specialized corpus, it is still quite likely that the network will overfit to the training data
- Monitor performance (i.e., perplexity) on a validation set, and stop training once a local minimum is reached

Data

- Our general, large dataset is the text of English Wikipedia.
- We consider smaller dataset with different styles and/or topics

DATASET STATISTICS

Dataset	Type	Number of Words	Number of Sentences
ENWiki	Gen	1,597,148,670	65,548,135
ENWiki_Computer	Top	4,531,972	146,159
ENWiki_Math	Top	5,081,695	170,400
ENWiki_Movies	Top	1,763,730	63,594
ENWiki_Science	Top	17,774,526	572,702
ENWiki_Sport	Top	34,224,253	1,241,570
SEWiki	Sty	18,597,300	1,042,226
SEWiki_Science	S+T	179,457	8,212
SEWiki_Sport	S+T	390,283	19,768
MovieReviews	S+T	1,494,946	63,577
WSJ	S+T	1,199,206	48,288

Results

Sandia National Laboratories

- We trained a model for each dataset and evaluated each corpus against each model using perplexity
  - The inverse geometric mean of the probabilities of the correct tokens
  - Lower is better
- Training time was approximately 10,000 cpu-hours for the general language model; refining on smaller, more specialized data typically finished in under 100 cpu-hours
- Some datasets are naturally more complex than others; to better understand the interactions among the various models and corpora, normalize test perplexities by dividing by the perplexity attained by evaluating the dataset on its corresponding model

Results

Sandia National Laboratories

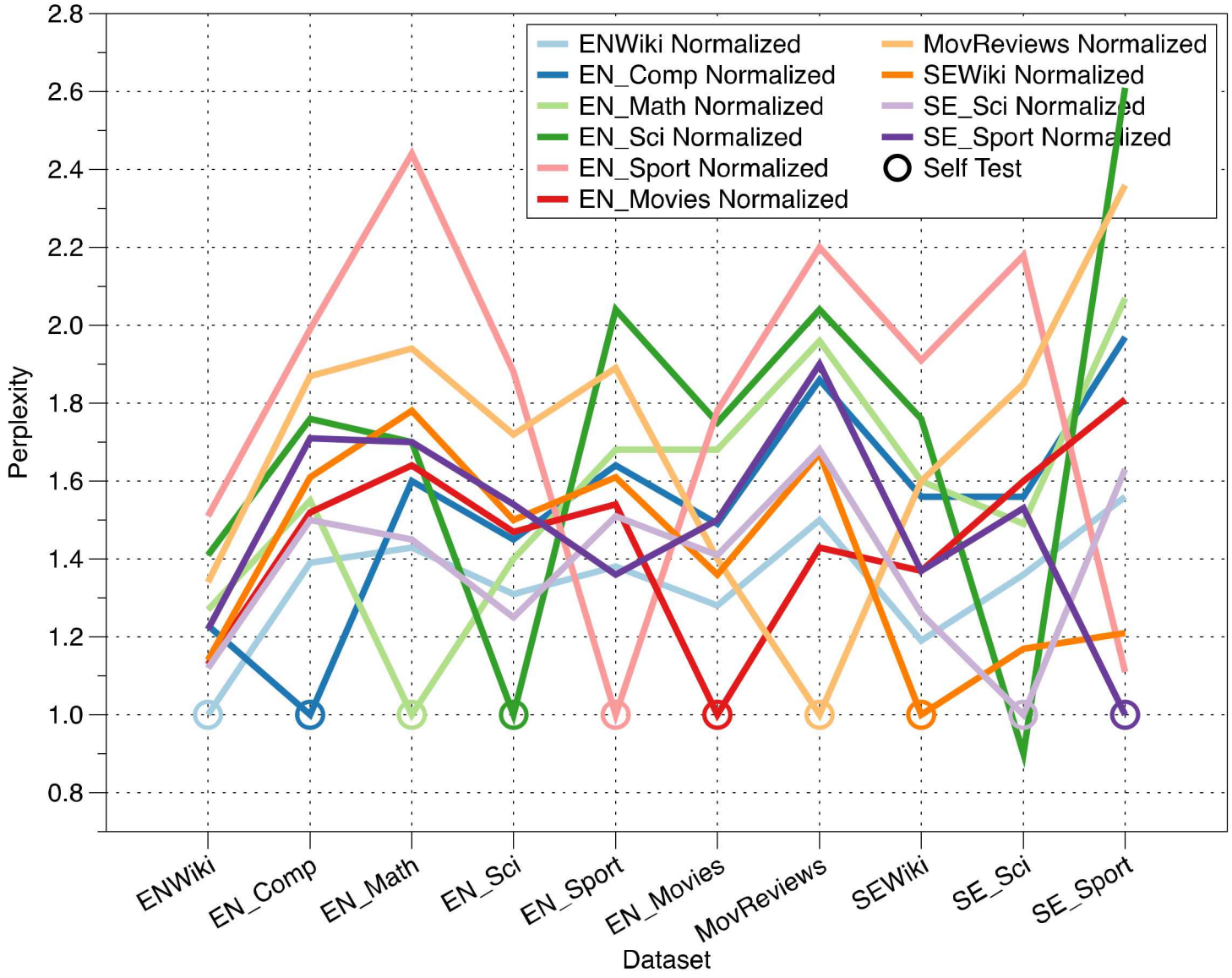
PERPLEXITY ATTAINED BY EVALUATING STANDARD ENGLISH MODELS ON STANDARD ENGLISH DATASETS

Model \ Dataset	*ENWiki*	*EN_Comp*	*EN_Math*	*EN_Sci*	*EN_Sport*	*EN_Movies*	*MovReviews*	*SEWiki*	*SE_Sci*	*SE_Sport*
ENWiki	40.8	54.5	34.8	37.0	26.3	48.3	89.4	35.5	36.2	17.9
ENWiki_Computer	50.1	39.2	38.9	41.1	31.4	56.3	110.6	46.5	41.5	22.6
ENWiki_Math	51.8	60.9	24.3	39.7	32.1	63.4	116.8	47.9	39.6	23.8
ENWiki_Science	57.7	68.9	41.4	28.3	38.9	66.1	121.4	52.6	23.9	30.0
ENWiki_Sport	61.6	78.0	59.4	53.2	19.1	67.4	130.6	57.1	58.1	12.8
ENWiki_Movies	46.1	59.5	39.9	41.7	29.4	37.8	85.2	40.8	42.5	20.8
MovieReviews	54.5	73.2	47.1	48.6	36.1	53.1	59.4	47.9	49.2	27.1
SEWiki	46.5	63.2	43.2	42.3	30.8	51.6	99.4	29.8	31.2	13.9
SEWiki_Science	45.7	58.9	35.4	35.4	28.8	53.5	100.0	37.6	26.6	18.7
SEWiki_Sport	49.7	66.9	41.3	43.7	26.0	56.7	112.8	40.9	40.7	11.5

PERPLEXITY, NORMALIZED BY COLUMN, ATTAINED BY EVALUATING STANDARD ENGLISH MODELS ON STANDARD ENGLISH DATASETS

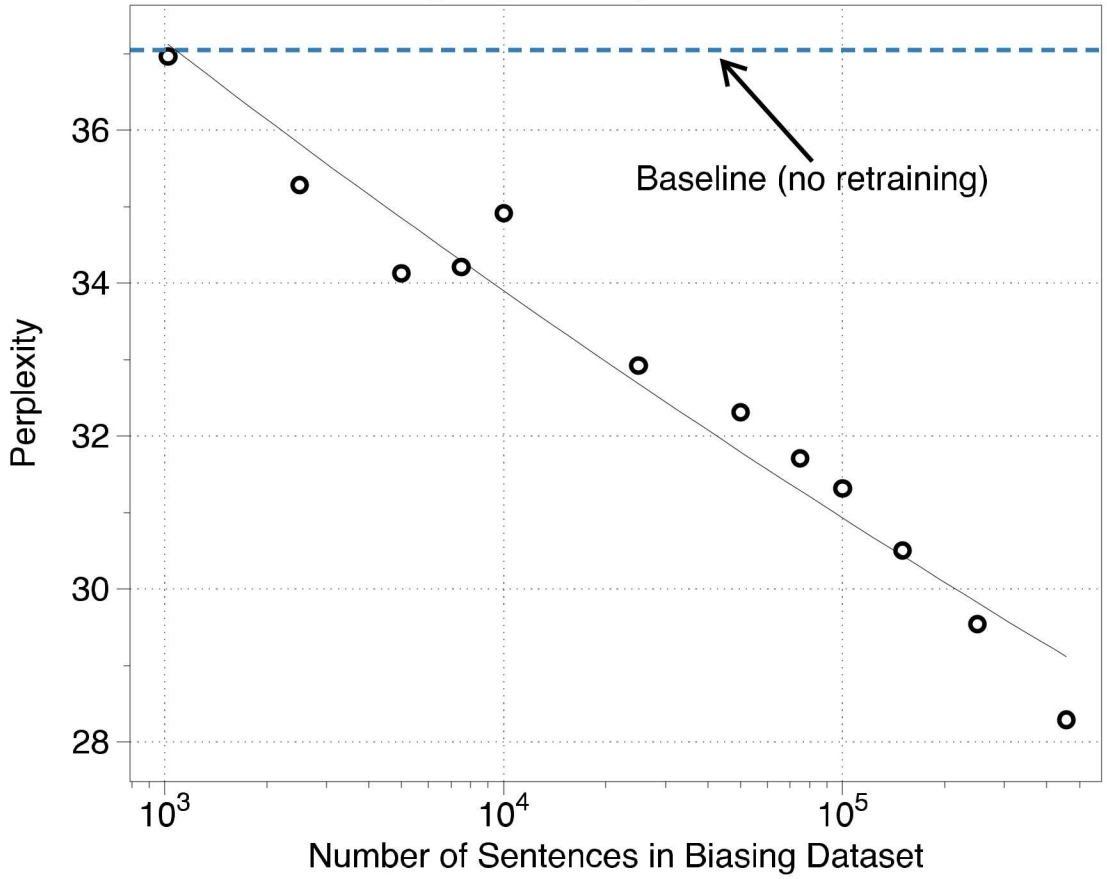
Model \ Dataset	*ENWiki*	*EN_Comp*	*EN_Math*	*EN_Sci*	*EN_Sport*	*EN_Movies*	*MovReviews*	*SEWiki*	*SE_Sci*	*SE_Sport*
ENWiki	1	1.39	1.43	1.31	1.38	1.28	1.50	1.19	1.36	1.56
ENWiki_Computer	1.23	1	1.60	1.45	1.64	1.49	1.86	1.56	1.56	1.97
ENWiki_Math	1.27	1.55	1	1.40	1.68	1.68	1.96	1.60	1.49	2.07
ENWiki_Science	1.41	1.76	1.70	1	2.04	1.75	2.04	1.76	0.90	2.61
ENWiki_Sport	1.51	1.99	2.44	1.88	1	1.78	2.20	1.91	2.18	1.11
ENWiki_Movies	1.13	1.52	1.64	1.47	1.54	1	1.43	1.37	1.60	1.81
MovieReviews	1.34	1.87	1.94	1.72	1.89	1.40	1	1.60	1.85	2.36
SEWiki	1.14	1.61	1.78	1.50	1.61	1.36	1.67	1	1.17	1.21
SEWiki_Science	1.12	1.50	1.45	1.25	1.51	1.41	1.68	1.26	1	1.63
SEWiki_Sport	1.22	1.71	1.70	1.54	1.36	1.50	1.90	1.37	1.53	1

Results


ENWiki Normalized
EN_Comp Normalized
EN_Math Normalized
EN_Sci Normalized
EN_Sport Normalized
EN_Movies Normalized
MovReviews Normalized
SEWiki Normalized
SE_Sci Normalized
SE_Sport Normalized
Self Test
Perplexity
2.8
2.6
2.4
2.2
2.0
1.8
1.6
1.4
1.2
1.0
0.8
ENWiki
EN_Comp
EN_Math
EN_Sci
EN_Sport
EN_Movies
MovReviews
SEWiki
SE_Sci
SE_Sport
Dataset

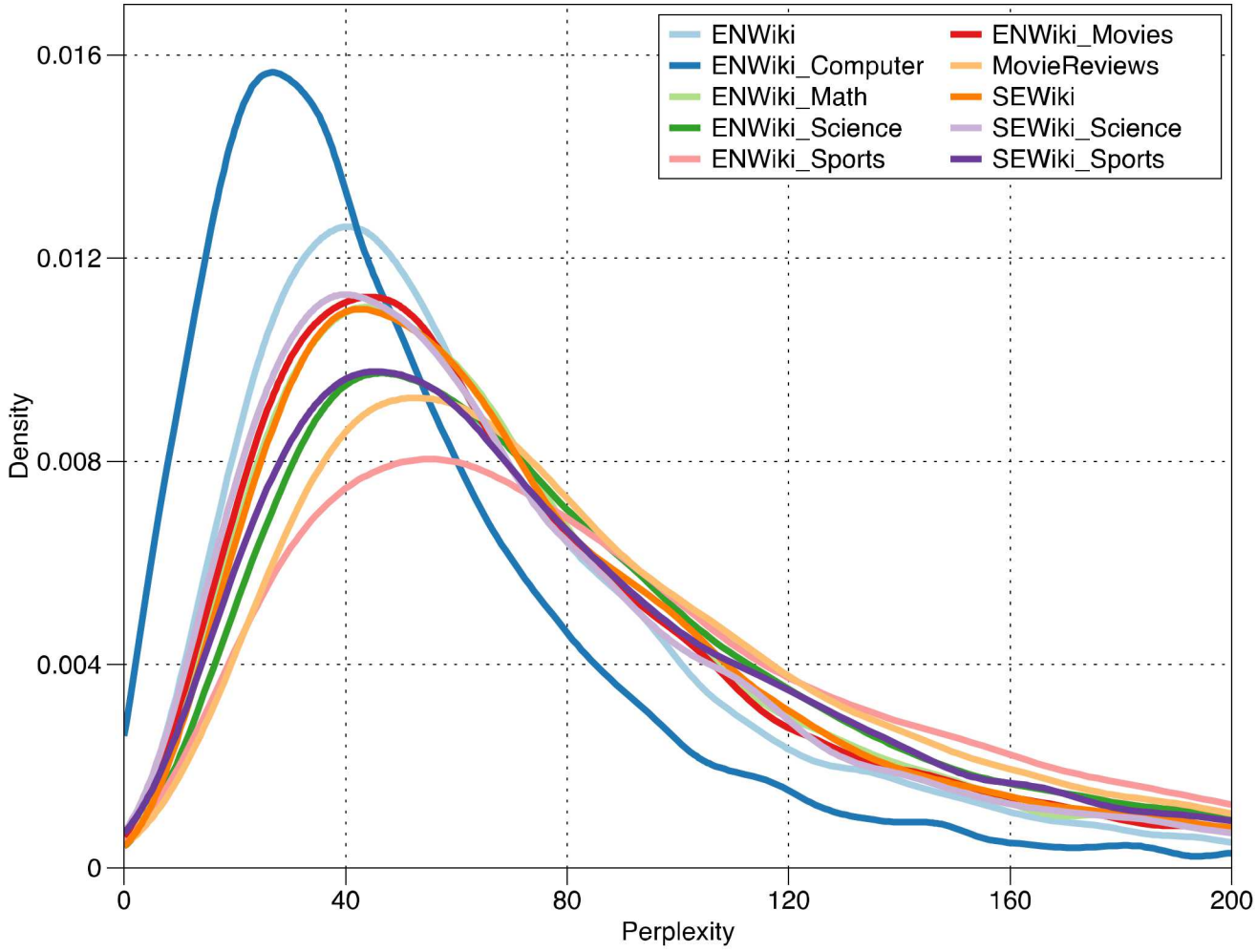
Results

- Performance on models biased on training subsets of ENWiki_Science of various sizes.
- Power law relationship: $\text{perplexity} = 48.9 \times \text{numSentences}^{-0.040}$



Results

- Perplexity distribution for sentences from ENWiki_Computer



Generated sentences – MovieReviews

- Is there a mission to be able to save our environment?
- There are both men and women and rule both sexes, while each member of the class is given eight <unk> once lost in an environment from <unk> to trolls or aliens
- However, Martin Scorsese's disappointing "the lantern of fools" bit proves the regular rotation can really be a bit more daunting
- The wonder-riddled body guns and safe thrills are such satisfying crowd names...
- Edward island has featured a automated software storage system

Generated sentences – ENWiki_Math

- In algebraic geometry, a space product, also known as a open set, is a topological space a for every subdivision of space that is compact, which is generally considered the regular topology
- <unk> and <unk> both argue that change can be objectively better than self-improvement (affecting content on the cognition of perceptual thinking), which explains conceptual actions to atypical applications in computational statistics
- We can view and explain changes in the vanishing bodies that are apparent in transcendental (or finitetime) geometry
- Graduates' successful range of technology can be generated using various experiments, such as data processing, equipment, conservation strategies, statistics, intellectual output and book-keeping

Conclusions

Sandia National Laboratories

- Neural network language models can be trained to model small training datasets
- The refinement process takes a small fraction of the time typically used to train NNLMs
- Models show promise for the use of topic and style identification, although more work is required
- Context-aware models might be able to understand more nuanced aspects of language, including humor and sarcasm

Questions?

Sandia National Laboratories