Exploring the Limits of Language Model

Raghavendra R Bilgi

Feb 21, 2017

Automatic Speech Recognition (ASR)

- Main component of Voice Assistants
- Converts Speech to text (STT)
- Goal is to recognize as many words correctly as possible (low Word Error Rate (WER))

Fundamental Equation of Speech Recognition

$$W^* = \underset{W}{argmax}\ p(W/O;\Theta)$$

$$W^* = \underset{W}{argmax}\ p(O/W;\Theta_A)\ p(W;\Theta_L)$$

- $P(O/W)$ links state sequence to words
- $P(W)$ Assigns Probability (prior) on word sequences

$$P(W) = P(w_1, w_2, w_3, ..., w_N)$$
$$= \prod_{n=1}^{N} p(w_n/w_1, w_2, ..., w_{n-1})$$

- Use n-gram models
- Probability is conditioned on window of n previous words

$$Unigram : P(w_n)$$
$$Bigram : p(w_n/w_{n-1})$$
$$Trigram : p(w_n/w_{n-2}, w_{n-1})$$

Language Model (LM)

Advantages of n-gram language models

- Performance improvement with higher n-gram (more context)
- Faster score computation (Faster look up)
- Can be represed as a WFST (useful for speech)
- Can be easily adpated to specific domain

Limitations

- Data spartisity is an issue
- More data, Smoothing, interpolation, back-off's required
- Exponential increase in the size with n-gram, and requrie more RAM

Language Model in ASR

- Two pass decoding stratergy used
- Smaller LM to generate the lattice which can fit in GPU memory
- Unpruned (bigger lm) to rescore the lattice
- Selection of pruning and smoothing method and stratergy is critical
- Agressive LM pruning has effect with certain smoothing techniques
- Lower Beam can result in shallow lattice

Language Model in ASR

Variants of Language Model

- Class n-gram model
- Cache model
- Skip-Gram Model
- Maximum entropy model

Amazon Echo Study and Findings

WHAT TASKS HAVE ECHO OWNERS TRIED WITH ALEXA?

ECHO TASKS

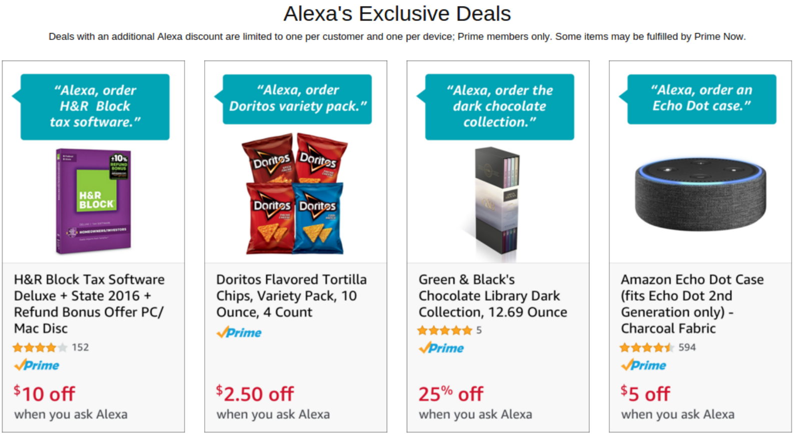
Tasks owners have tried at least once

Set a timer	84.9%	Add an item to your to-do list	32.7%
Play a song	82.4%	Buy something on Amazon Prime	32.1%
Read the news	66.0%	Control smart thermostat	30.2%
Set an alarm	64.2%	Play children's music	28.9%
Check the time	61.6%	Check or add an item to calendar	21.4%
Tell a joke	60.4%	Other	19.5%
Control smart lights	45.9%	Spell something	17.6%
Add item to shopping list	45.3%	Call an Uber	6.3%
Connect to paid music service	40.9%	Connect to phone via Bluetooth	3.5%
Provide the traffic	36.5%		

Survey respondents have tried an **AVERAGE OF EIGHT TASKS** from the above list.

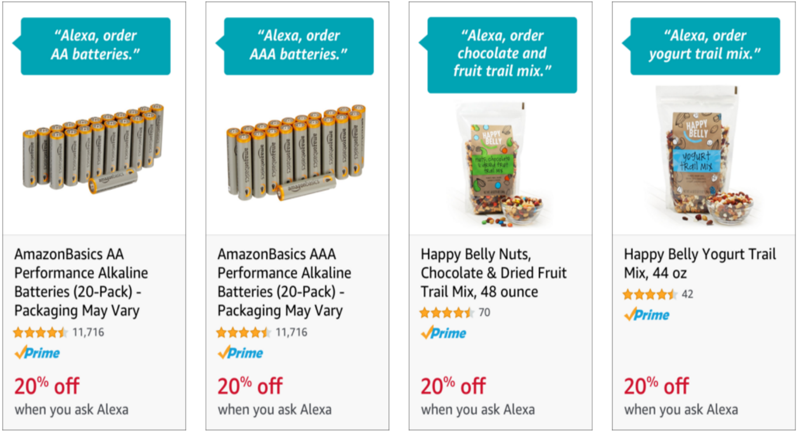
Source : Amazon Echo Study and Findings

Alexa Voice Shopping



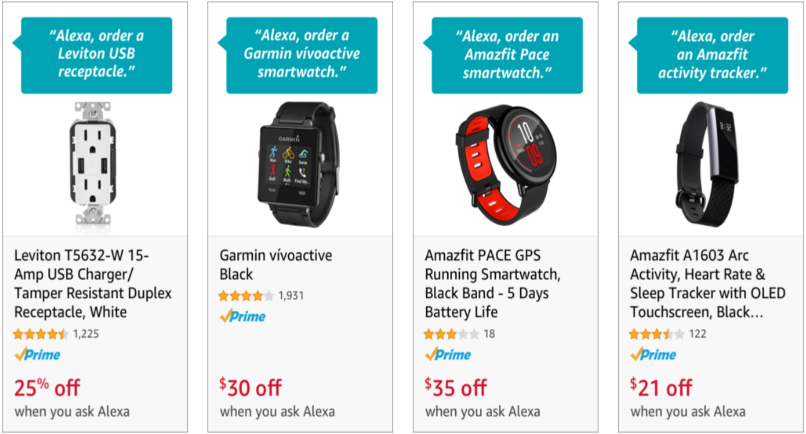
Source : Alexa Voice Shopping

Alexa Voice Shopping



Source : Alexa Voice Shopping

Alexa Voice Shopping



Source : Alexa Voice Shopping

Google Voice Shopping

Source : Shopping with Google Assistant, Feb 16, 2017

Alexa Voice Shopping

Amcrest IP2M-841 1080p dome surveillance camera

- Wifi security camera black
- Amcrest i. p. to m. security camera
- wifi dome surveillance
- Amcrest indoor don't surveillance camera
- Amcrest dumb surveillance camera
- Amcrest ip to him surveillance camera
- Echo crest eight four one security camera
- Amcrest ten eighty p wi fi security camera

Source : Amazon Echo Prime day Review

Alexa Voice Shopping

Amcrest IP2M-841 1080p dome surveillance camera

- Amcrest IP2M security camera $\rightarrow$ Amcrest i. p. to m. security camera
- Amcrest IP2M surveillance camera $\rightarrow$ Amcrest ip to him surveillance camera
- Amcrest 841 security camera $\rightarrow$ Echo crest eight four one security camera
- Amcrest dome surveillance camera $\rightarrow$ Amcrest dumb surveillance camera
- Amcrest 1080p wi-fi security camera $\rightarrow$ Amcrest ten eighty p wi fi security camera

Source : Amazon Echo Prime day Review

Alexa Voice Shopping

Noisy LM Training data

Kanvas Katha Women's Multi color Ballet Flats - 3 UK/India (36 EU)(KKFTOXDOCT00303)
Royal Son Rimless Rectangular Women Spectacle Frame (RS0650ER 50 Transparent)
Syska B22 15-Watt LED Bulb (Pack of 2, Cool Day Light)
FabHomeDecor Elzada Five Seater Sofa 3+2 (Black)
Goodway Pack Of 3 Junior Boys Graphic Tee C'mon Bro-Give Your- Lazy Boy Prints Combo
Butterflies Women's Wallet (Dark Pink) (BNS 2320 DPK)
Fila Unisex Relaxer III Red and Navy Sneakers - 7 UK/India (41 EU)
Vvoguish Full Sleeve Indigo Red Round Size -S-VVTOP928INDGMELRD-S
Skil 6513 JD 13mm Drill Kit with 15 Drill Bits
IDEE Round Sunglasses (IDS1986C2SG 49 Matte Black)

Recurrent Neural Network Language Model (RNN LM)

- Recurrence allows for unbounded context
- RNN Model compactly represents world knowledge
- Impressive Perplexity improvemnts
- No more feature engineering, model learns to extract latent features

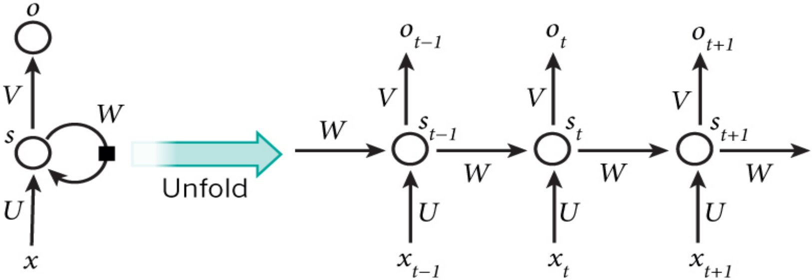


Figure : A recurrent neural network and the unfolding in time of the computation involved in its forward computation : Source Nature

RNN LM

$$h_t = f(W_t h_{t-1} + U_t x_t)$$
$$y_t = softmax(h_t V_t)$$

Table 2. *Comparison of different neural network architectures on Penn Corpus (1M words) and Switchboard (4M words).*

	Penn Corpus		Switchboard	
Model	NN	NN+KN	NN	NN+KN
KN5 (baseline)	-	141	-	92.9
feedforward NN	141	118	85.1	77.5
RNN trained by BP	137	113	81.3	75.4
RNN trained by BPTT	123	106	77.5	72.5

Figure : RNN LM Perplexity [Mikolov et al. 2010]

Distributional Representation of Words

- Word meaning defined in terms of vectors
- Vectors are learned such that, words with similar context are close in vector space
- CBOW, Skip-Gram to learn the parameters

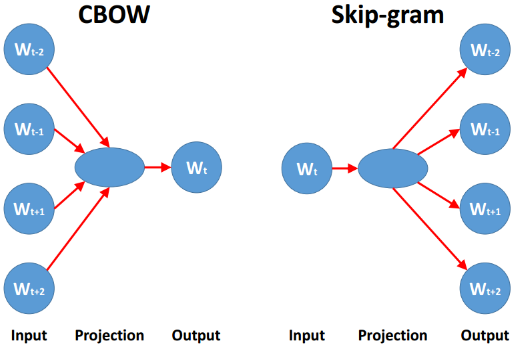


Figure : CBOW and Skip-Gram Models

Sequence to Sequence model

$$h_t = f(W_t h_{t-1} + U_t x_t)$$
$$y_t = softmax(h_t V_t)$$
$$p(y_1, y_2, ..., y_{T'}/x_1, x_2, ..., x_T) = \prod_{t=1}^{T'} p(y_t/h_t, y_1, ..., y_{t-1})$$

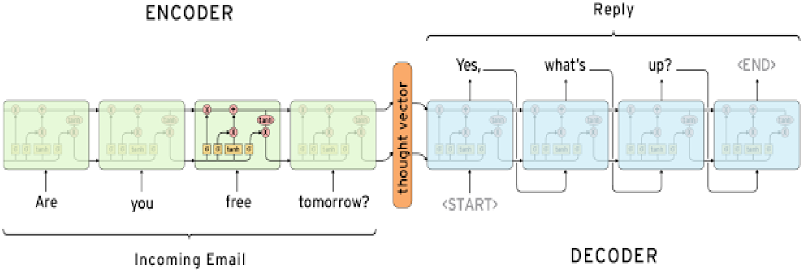


Figure : Sequence to Sequence Model

Sequence to Sequence model with Attention

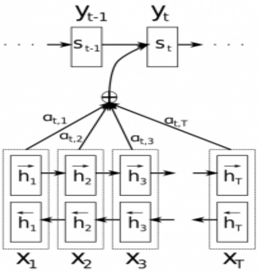


Figure : Sequence to Sequence Model with Attention

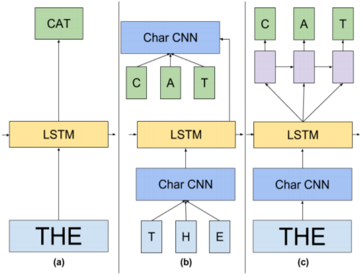
$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}$$

$$e_{ij} = a(s_{i-1}, h_j)$$

Source : Neural Machine Translation by jointly learning to align and transalate

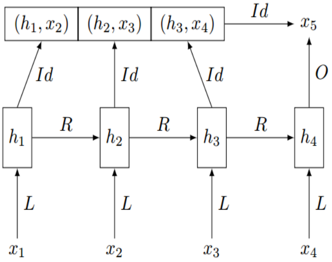
RNN LM with CNN Softmax



MODEL	TEST PERPLEXITY	NUMBER OF PARAMS [BILLIONS]
SIGMOID-RNN-2048 (JI ET AL., 2015A)	68.3	4.1
INTERPOLATED KN 5-GRAM, 1.1B N-GRAMS (CHELBA ET AL., 2013)	67.6	1.76
SPARSE NON-NEGATIVE MATRIX LM (SHAZEER ET AL., 2015)	52.9	33
RNN-1024 + MAXENT 9-GRAM FEATURES (CHELBA ET AL., 2013)	51.3	20
LSTM-512-512	54.1	0.82
LSTM-1024-512	48.2	0.82
LSTM-2048-512	43.7	0.83
LSTM-8192-2048 (NO DROPOUT)	37.9	3.3
LSTM-8192-2048 (50% DROPOUT)	32.2	3.3
2-LAYER LSTM-8192-1024 (BIG LSTM)	30.6	1.8
BIG LSTM+CNN INPUTS	**30.0**	**1.04**
BIG LSTM+CNN INPUTS + CNN SOFTMAX	39.8	**0.29**
BIG LSTM+CNN INPUTS + CNN SOFTMAX + 128-DIM CORRECTION	35.8	**0.39**
BIG LSTM+CNN INPUTS + CHAR LSTM PREDICTIONS	47.9	**0.23**

Figure : Exploring Limits of Language Model, [Jozefowiez et al. 2016]

Neural Cache Model



Model	Test PPL
RNN+LSA+KN5+cache (Mikolov & Zweig, 2012)	90.3
LSTM (Zaremba et al., 2014)	78.4
Variational LSTM (Gal & Ghahramani, 2015)	73.4
Recurrent Highway Network (Zilly et al., 2016)	66.0
Pointer Sentinel LSTM (Merity et al., 2016)	70.9
LSTM (our implem.)	82.3
Neural cache model	72.1

Figure : Neural LM with Continuous Cache, [Edouard Grave et al. 2017]

Ensemble of Language Model

Model	Test Perplexity
Large Ensemble (Chelba et al., 2013)	43.8
RNN+KN-5 (Williams et al., 2015)	42.4
RNN+KN-5 (Ji et al., 2015a)	42.0
RNN+SNM10-skip (Shazeer et al., 2015)	41.3
Large Ensemble (Shazeer et al., 2015)	41.0
Our 10 best LSTM models (equal weights)	26.3
Our 10 best LSTM models (optimal weights)	26.1
10 LSTMs + KN-5 (equal weights)	25.3
10 LSTMs + KN-5 (optimal weights)	25.1
10 LSTMs + SNM10-skip (Shazeer et al., 2015)	**23.7**

Figure : Ensemble of LM, [Jozefowiez et al. 2016]

N-Gram LM or RNN LM

- RNN LMs are very popular results in lower perplexity
- However, they are not easy to adapt, cannot scale to to several million word dataset like n-grams
- RNN LM can't be compiled into an FST but can rescore the word lattice.
- Primary domains of Voice-Assistants use short utterances
- Ensemble of these to seems to be best middle gound
- n-gram model with context or domain knowledge is much better than single model


2nd
car
per
IIK
new
29g
Age
War
S.S
Ram
Mat
Joe
Aux
Ltr
7th
NFL
5th
PCs
LCD
6Ft
Pan
SLR
Wow
Jio
2gb
COD
Gym
Pet
Dad


Red
Grey
Set
Pink
Watch
Size
Jeans
Sticker
Light
Dial
Lancer
Piece
Tan
Shirt
USB
Half
Men
Bag
Table
Full
Top
Slim
Suede
Year
One
Face
Wine
Fast
Card
Oak
Cream
India
Lime
Bed
kids
Free
Plus

N-Gram LM or RNN LM

- RNN LMs are very popular results in lower perplexity
- However, they are not easy to adapt, cannot scale to to several million word dataset like n-grams
- Primary domains of Voice-Assistants use short utterances
- Ensemble of these to seems to be best middle gound
- n-gram model with context or domain knowledge is much better than single model
- Lot of domain specific clean data