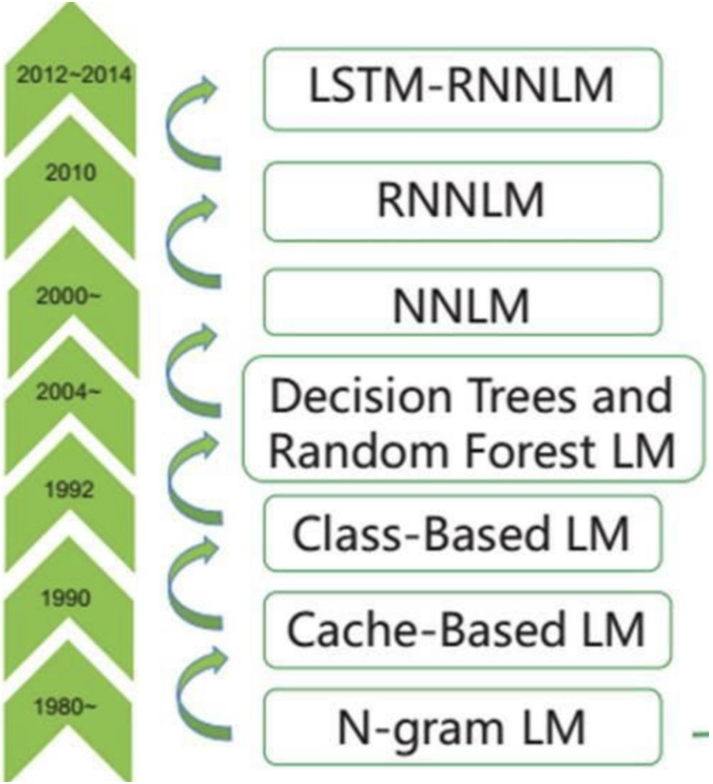
RNN based Language model

Akash Govindarajula & Jakub Wasylkowski

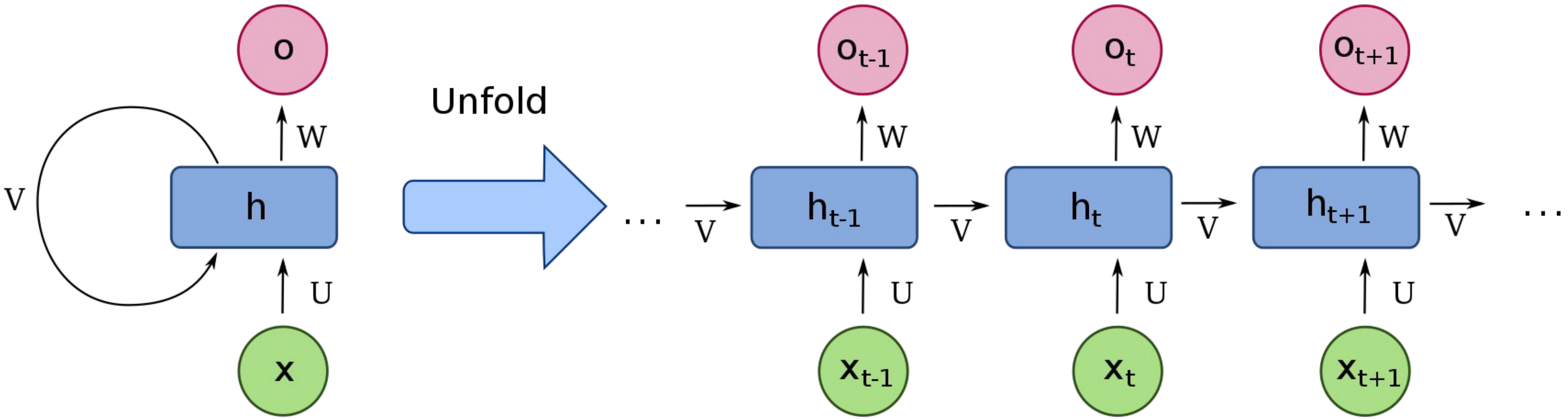
Contents

Intro

- Current language models
- Perplexity & experiments explored
- Connectionist LM's vs. n-grams
- Cache and class-based



RNN for Language Modelling



$$h_t = \sigma_h(W_h x_t + U_h h_{t-1} + b_h)$$
$$y_t = \sigma_y(W_y h_t + b_y)$$

Elman Network

- Simple RNN (3 layers only).
- No fixed context length.
- Context recycled many times.

Dynamic network

- Training continues through testing.
- Model is updated after processing testing data.
- Better domain adaptation.
- Lower perplexity

Optimization

- Only the size of the hidden layer is parametrized
- "Rare token" for vocabulary reduction

$$P(w_i(t+1)|w(t), h(t-1)) = \begin{cases} \frac{y_{rare}(t)}{C_{rare}} & \text{if } w_i(t+1) \text{ is rare} \\ y_i(t) & otherwise \end{cases}$$

"All Rare words have equal probability"

WSJ Experiments

- DARPA and Wall Street Journal ‘92 & ‘93 dataset of spoken / written data
- RNN’s trained on up to 6.4M words from NYT English Gigaword
- RNN combined with 5-gram Kneser-Ney (KN5) model.
- RNN 250 / 5 → RNN *size of hidden layer* / *rare token cutoff*

Training corpus size

Model	# words	PPL	WER
KN5 LM	200K	336	16.4
KN5 LM + RNN 90/2	200K	271	15.4
KN5 LM	1M	287	15.1
KN5 LM + RNN 90/2	1M	225	14.0
KN5 LM	6.4M	221	13.5
KN5 LM + RNN 250/5	6.4M	156	11.7

Model combinations

	PPL		WER	
Model	RNN	RNN+KN	RNN	RNN+KN
KN5 - baseline	-	221	-	13.5
RNN 60/20	229	186	13.2	12.6
RNN 90/10	202	173	12.8	12.2
RNN 250/5	173	155	12.3	11.7
RNN 250/2	176	156	12.0	11.9
RNN 400/10	171	152	12.5	12.1
3xRNN static	151	143	11.6	11.3
3xRNN dynamic	128	121	11.3	11.1

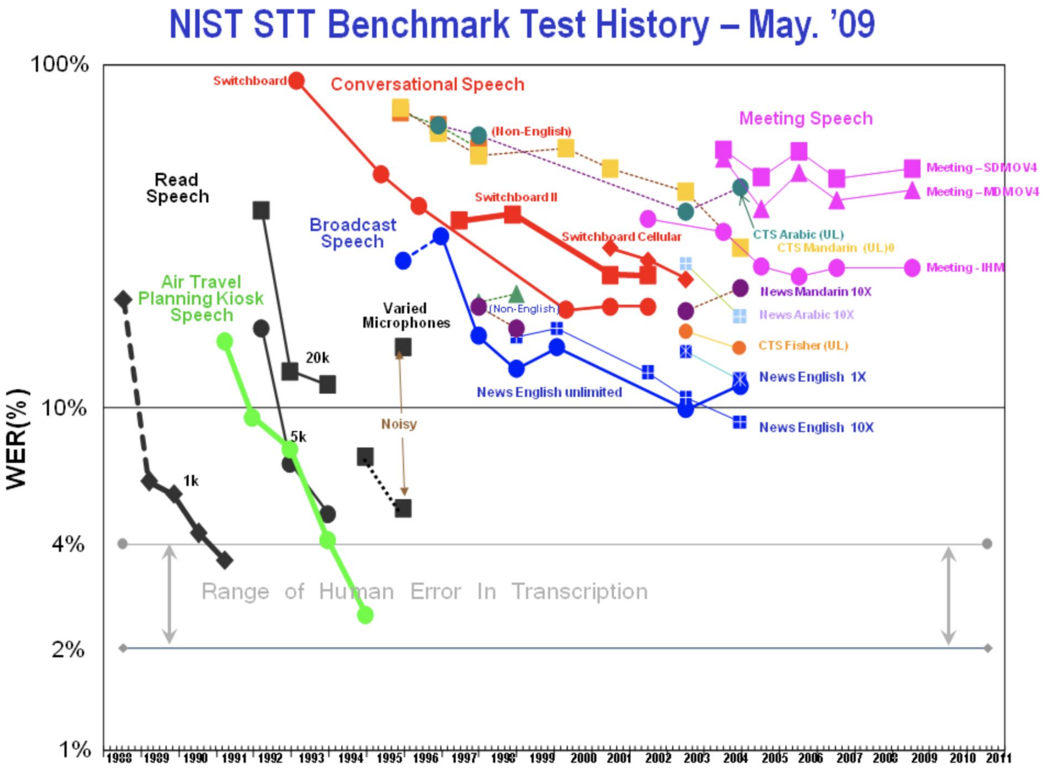
Vs. other models

Model	DEV WER	EVAL WER
Lattice 1 best	12.9	18.4
Baseline - KN5 (37M)	12.2	17.2
Discriminative LM [8] (37M)	11.5	16.9
Joint LM [9] (70M)	-	16.7
Static 3xRNN + KN5 (37M)	11.0	15.5
Dynamic 3xRNN + KN5 (37M)	10.7	16.3[4]

NIST RT05

- RT05 focused on the English Meeting Domain speech. There were four evaluation tasks: STT, MDE Speaker Diarization, MDE Speech Activity Detection, and MDE Source Localization.

- The cross site evaluation corpora included "conference" room meetings and "lecture" room meetings.



WER results

Model	WER static	WER dynamic
RT05 LM	24.5	-
RT09 LM - baseline	24.1	-
KN5 in-domain	25.7	-
RNN 500/10 in-domain	24.2	24.1
RNN 500/10 + RT09 LM	**23.3**	23.2
RNN 800/10 in-domain	24.3	23.8
RNN 800/10 + RT09 LM	23.4	23.1
RNN 1000/5 in-domain	24.2	23.7
RNN 1000/5 + RT09 LM	23.4	22.9
3xRNN + RT09 LM	**23.3**	**22.8**

Conclusion & Future

- Significantly outperforms SOTA's
  - 18% error rate reduction over 12% of backoff in WSJ experiments
  - In NIST experiments, over 5.4M words can outperform big backoff models
- Large perplexity improvements
- Explore more into backpropagation & on-line learning (BPTT)
  - Link: wildml.com/2015/10/recurrent-neural-networks-tutorial-part-3-backpropagation-through-time-and-vanishing-gradients/
  - Can be extended to applications in OCR and backoff model use-cases

Questions?