Recurrent Neural Network Based Language Model

Tomas Mikolov, Martin Karafiat, Lukas Burget,
Jan “Honza” Cernock, Sanjeev Khudanpur

INTERSPEECH 2010

Presenter: Hiroaki Hayashi

Overview

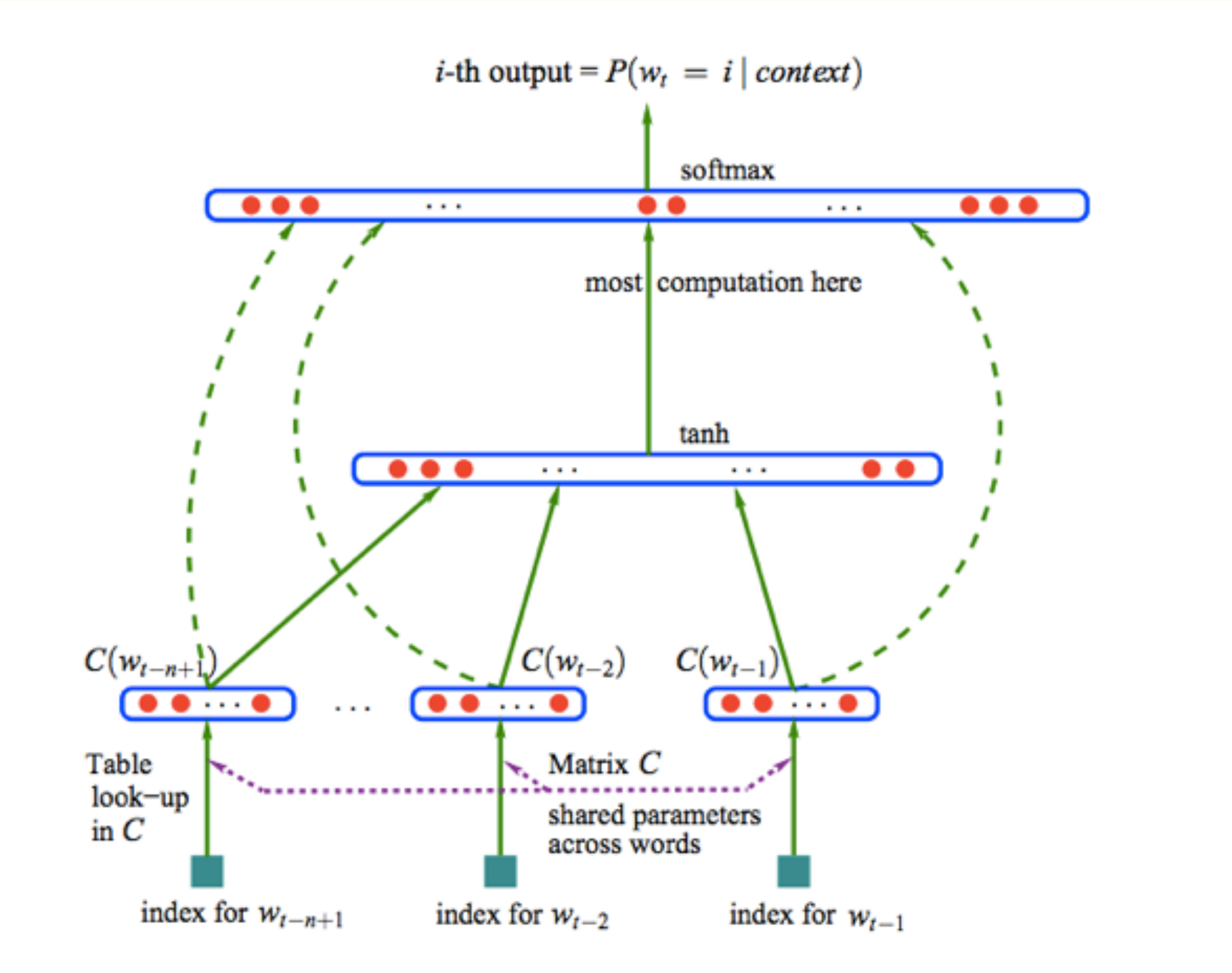
- Conventional “good” language models
  → Not applicable to practical tasks
  → Tiny improvements against each other
  ex. Cache, Class-based

- Neural probabilistic language model [Bengio et al, 03]

- Recurrent neural network based language model

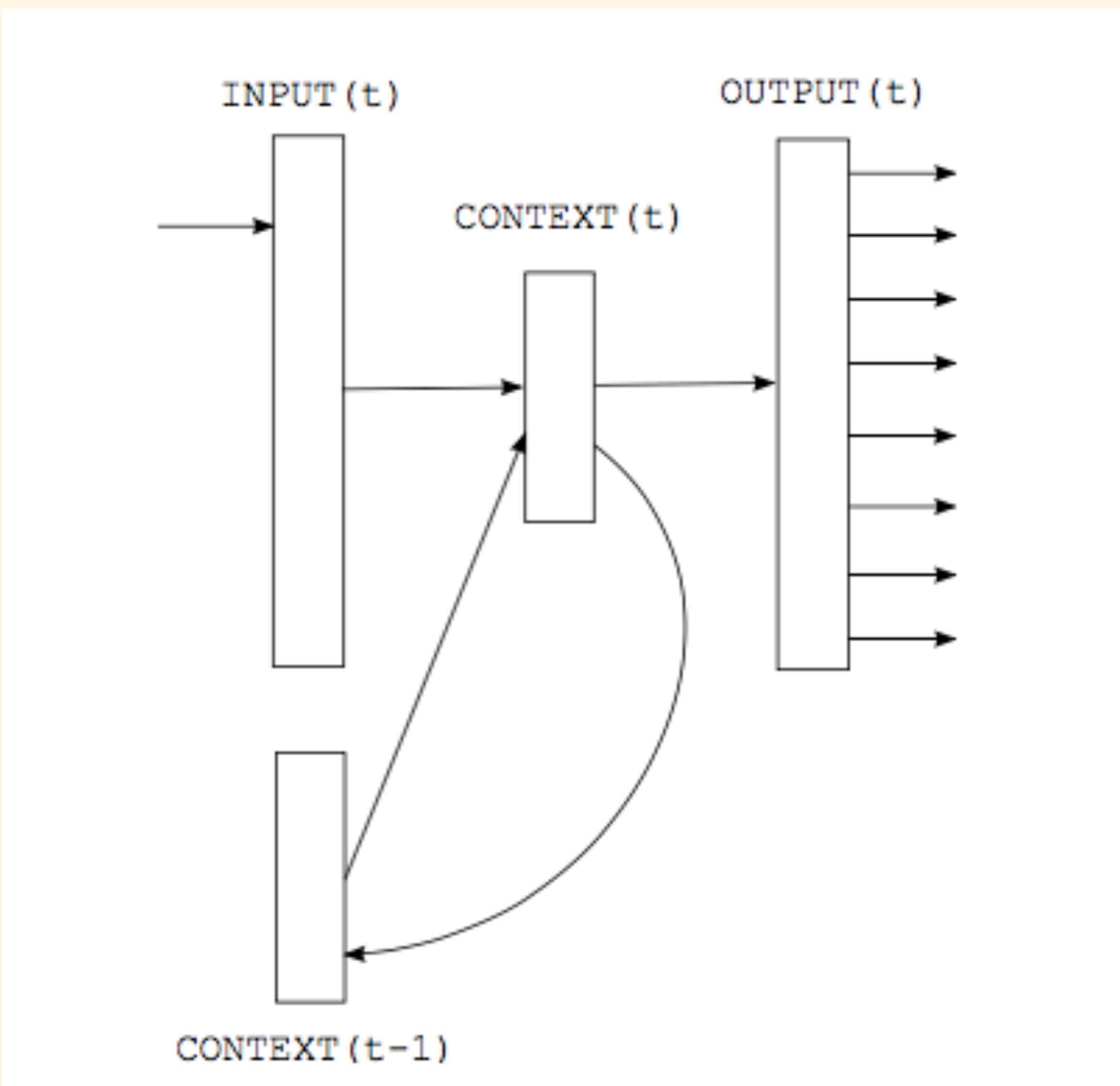
Feedforward NN LM

- Fixed size of previous contexts (N-gram)

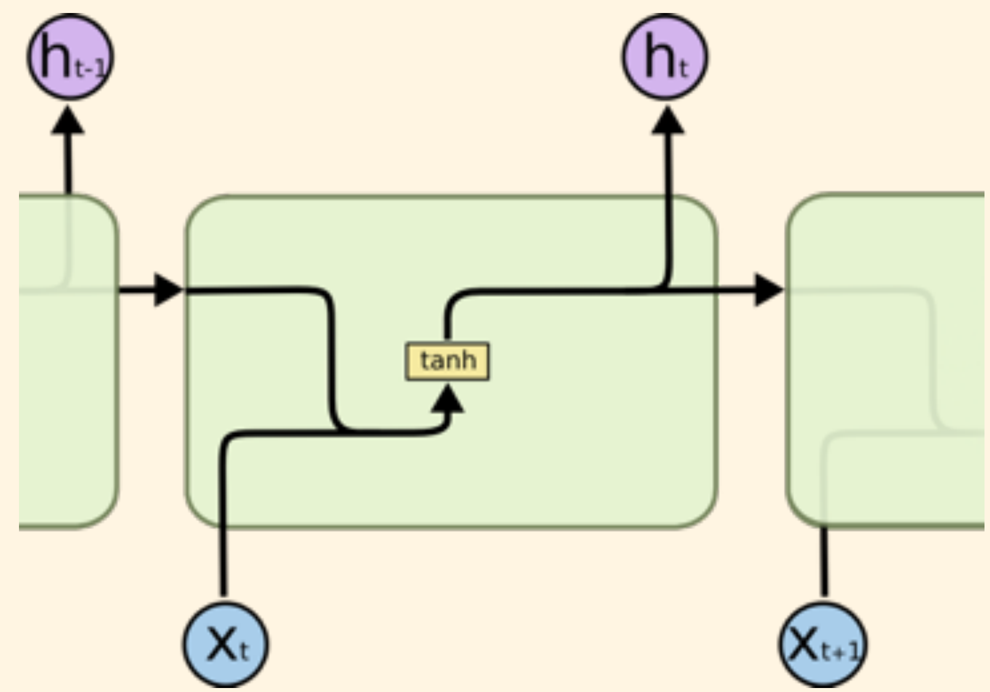


Recurrent NN LM

- Arbitrary-length contexts



Unfold



Model equation

$$x(t) = w(t) + s(t-1) \tag{1}$$

$$s_j(t) = f\left(\sum_i x_i(t) u_{ji}\right) \tag{2}$$

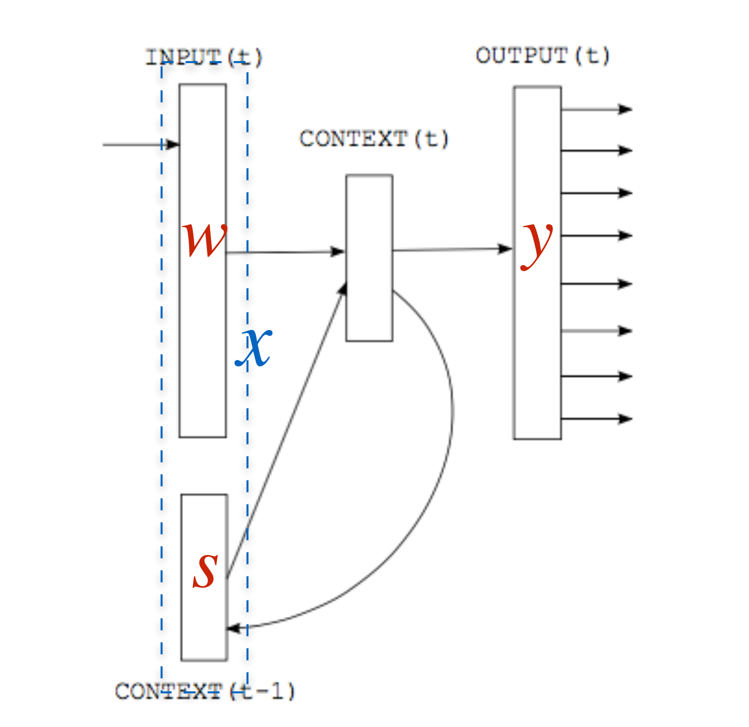
$$y_k(t) = g\left(\sum_j s_j(t) v_{kj}\right) \tag{3}$$

where $f(z)$ is sigmoid activation function:

$$f(z) = \frac{1}{1 + e^{-z}} \tag{4}$$

and $g(z)$ is softmax function:

$$g(z_m) = \frac{e^{z_m}}{\sum_k e^{z_k}} \tag{5}$$



Model Details

- Treating less-frequent words as `<rare>`
  → Uniformly distributed probabilities

- Setting $s(0)$ to be "small values"
  → Not crucial when the data is large

- Size of hidden layer reflects amount of training data

Experiments I - WSJ

- Linear interpolation:
  (.75) * KN5 + (.25) * RNN

- Smaller perplexity, Less error rate

Table 1: *Performance of models on WSJ DEV set when increasing size of training data.*

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

Experiments II - RNN params

- Dynamic model:
  Continue learning parameters from the test data

Table 2: *Comparison of various configurations of RNN LMs and combinations with backoff models while using 6.4M words in training data (WSJ DEV).*

	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

Experiments III - Data size

- RNN: 5.4M
  back-off: 1.3G

Table 4: *Comparison of very large back-off LMs and RNN LMs trained only on limited in-domain data (5.4M words).*

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

- Arbitrary-length context from the past
- Outperformance on various tasks with less data
- Need of improvement on capturing truly long context
  → LSTM..?

References

- Mikolov, Tomas, et al. "Recurrent neural network based language model." INTERSPEECH. Vol. 2. 2010.

- Bengio, Yoshua, et al. "Neural probabilistic language models." Innovations in Machine Learning. Springer Berlin Heidelberg, 2006. 137-186.

- http://colah.github.io/posts/2015-08-Understanding-LSTMs