Recurrent Neural Network-based Language Models with Variation in Net Topology, Language, and Granularity

Tzu-Hsuan Yang, Tzu-Hsuan Tseng, and Chia-Ping Chen

Department of Computer Science and Engineering,
National Sun Yat-sen University

IALP 2016 @NCKU, Tainan, Taiwan

Overview

- Introduction
- Data
- Model Architecture
- Experiments
  - Comparison of Models and Databases
  - Model Complexity and Perplexity
  - Comparison of Granularity
- Conclusion

Introduction

- Language model (LM)
  - What is language model ?
  - Applications
  - Well-known LMs

- Major goals
  - Compare RNN-based LMs
  - Difference between character-based and word-based LMs in Chinese

Data

- Text databases in our experiments
  - Penn Tree Bank (PTB)
  - AMI meeting corpus (AMI)
  - Academia Sinica Balanced Corpus (ASBC)

- ASBC-r – a change of ASBC
  - Replace lower frequency tokens
  - Similar vocabulary size to PTB and AMI

Data

Databases	Vocabulary Size	Number of Words	
PTB	9999	Train	887521
		Validation	70390
		Test	78669
AMI	11883	Train	802824
		Validation	94953
		Test	89666
ASBC	49933	Train	4013468
		Validation	403482
		Test	411090
ASBC-r	10041	Train	4013468
		Validation	403482
		Test	411090

Table 1: Statistics of databases

Model Architecture

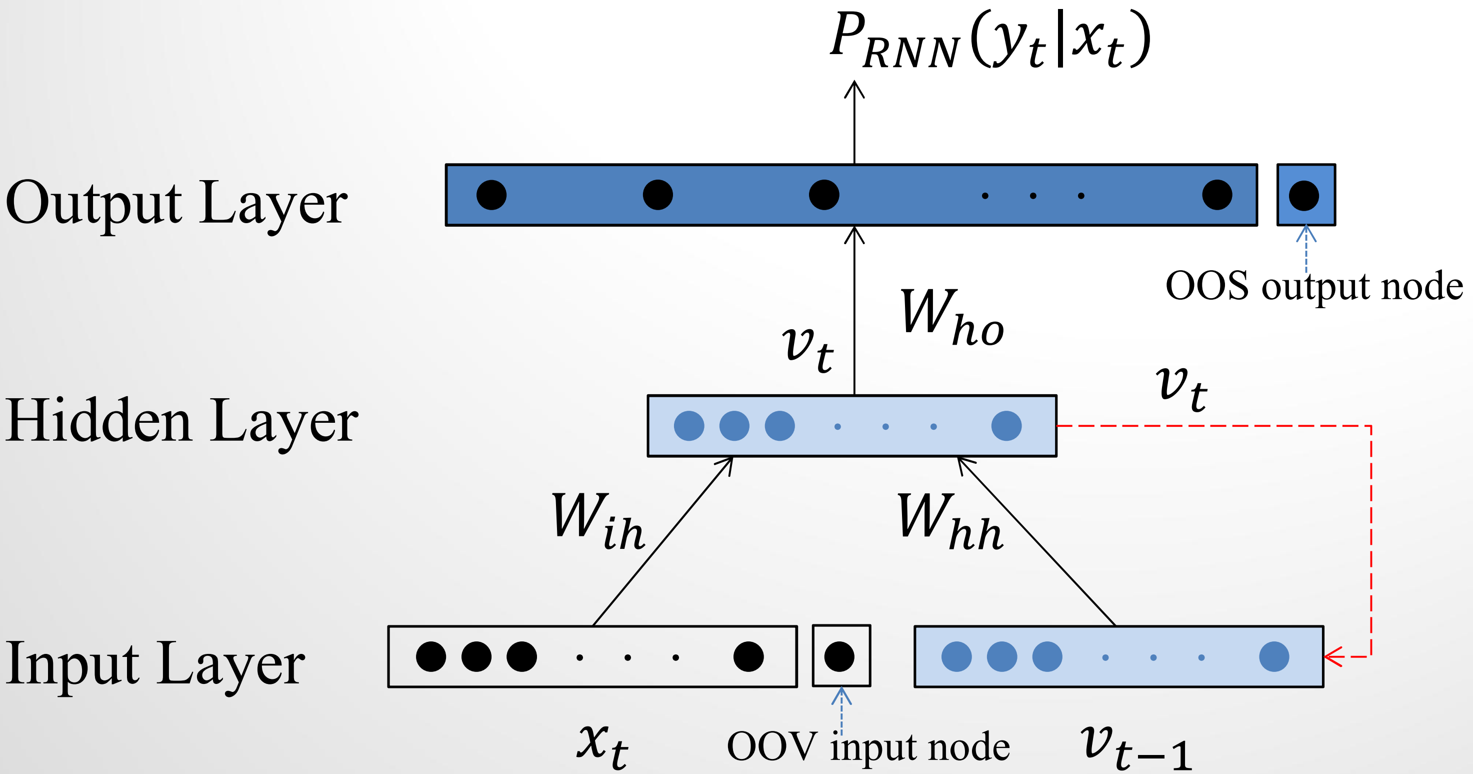

Figure 1: Basic RNN Architecture

Model Architecture

- OOV words in evaluation data
  - Some OOV words in AMI
  - Treat `unk` as OOV words in PTB and ASBC

- Interpolate with trigram model with same training data as RNN LM

Model Architecture

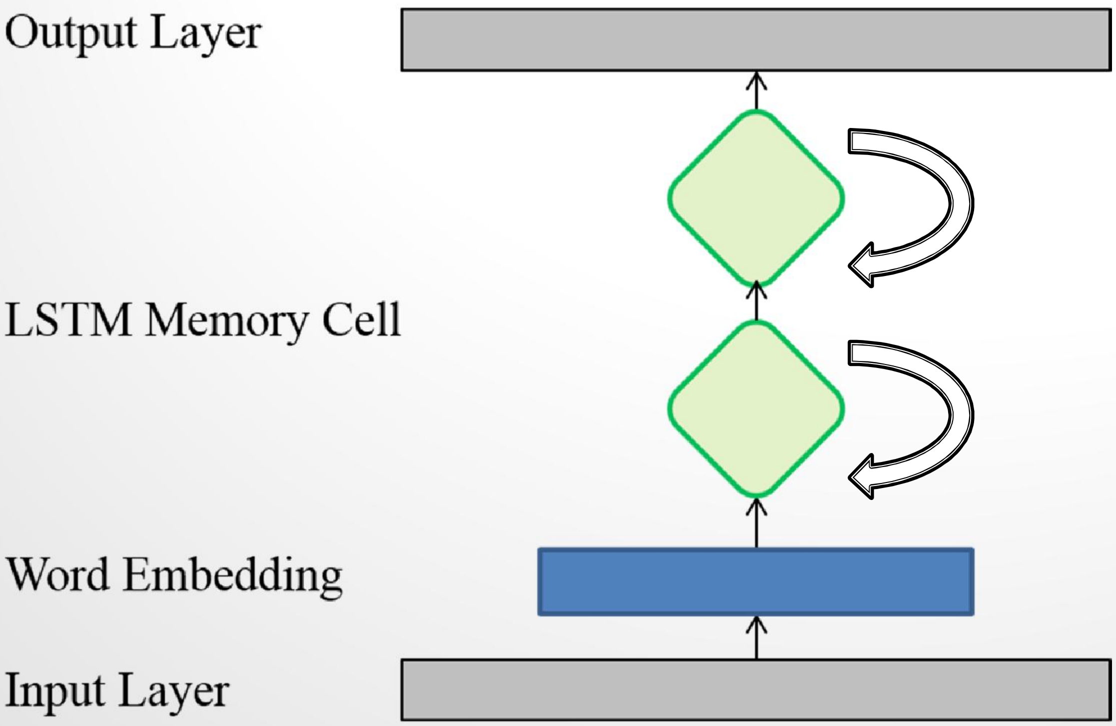

Figure 2: LSTM Architecture

Comparison of Models and Databases

- The perplexity in ASBC
  - LSTM is higher than RNN

- The perplexity in ASBC-r
  - LSTM is lower than RNN

Databases	RNN	LSTM
PTB	120.9	116.8
AMI	74.6	72.5
ASBC	306.7	317.1
ASBC-r	140.6	136.6

Table 2: RNN vs LSTM

Comparison of Models and Databases

- The increasing vocabulary size
  → lots of parameters
  → the LSTM model over-fitting

Databases	RNN	LSTM
PTB	120.9	116.8
AMI	74.6	72.5
ASBC	306.7	317.1
ASBC-r	140.6	136.6

Table 2: RNN vs LSTM

Comparison of Models and Databases

- The text in ASBC is more diverse than PTB and AMI
  - Even if the training set in ASBC-r is larger

Databases	RNN	LSTM
PTB	120.9	116.8
AMI	74.6	72.5
ASBC	306.7	317.1
ASBC-r	140.6	136.6

Table 2: RNN vs LSTM

Model Complexity and Perplexity

- All models are trained on ASBC
- Only change the hidden layer size

Hidden Size	RNN	LSTM
50	329.6	377.5
100	316.8	334.9
150	310.4	319.8
200	306.7	317.1
250	304.9	318.7
300	304.8	326.9

Table 3: Perplexities of ASBC with Different Hidden Size

Model Complexity and Perplexity

- Improvement in perplexity until the size up to 200
  - Too many parameters results in over-fitting

Hidden Size	RNN	LSTM
50	329.6	377.5
100	316.8	334.9
150	310.4	319.8
200	306.7	317.1
250	304.9	318.7
300	304.8	326.9

Table 3: Perplexities of ASBC with Different Hidden Size

Variations in ASBC-r

- Three variations of Chinese sentences

Variations	Example	Vocabulary Size	Number of Words	
Word-based	心 中 非常 著急	10041	Train	4013468
			Validation	403682
			Test	411090
Char-based	心 中 非 常 著 急	5633	Train	6470216
			Validation	650251
			Test	669229
Char-based with sp	心 sp 中 sp 非 常 sp 著 急	5634	Train	9901083
			Validation	986278
			Test	993493

Table 4: Statistics and Sample Sentences of Three Variations in ASBC-r

Comparison of Granularity

- The perplexity of character-based LM is lower
  - But the probability of the corpus is smaller

Variations	RNN	LSTM
Word-based	140.6	136.6
Char-based	60.4	60.5
Char-based with sp	17.5	15.4

Table 5: Perplexities of Three Variations in ASBC-r

Conclusion

- LSTM-based LM achieve lower perplexity than basic RNN
- The difference in diversity of the databases
- Larger model complexity will result in over-fitting
- The likelihood of the character-based corpus is smaller