Capitalization Normalization for Language Modeling with an Accurate and Efficient Hierarchical RNN Model

Hao Zhang, You-Chi Cheng, Shankar Kumar, W. Ronny Huang, Mingqing Chen, Rajiv Mathews

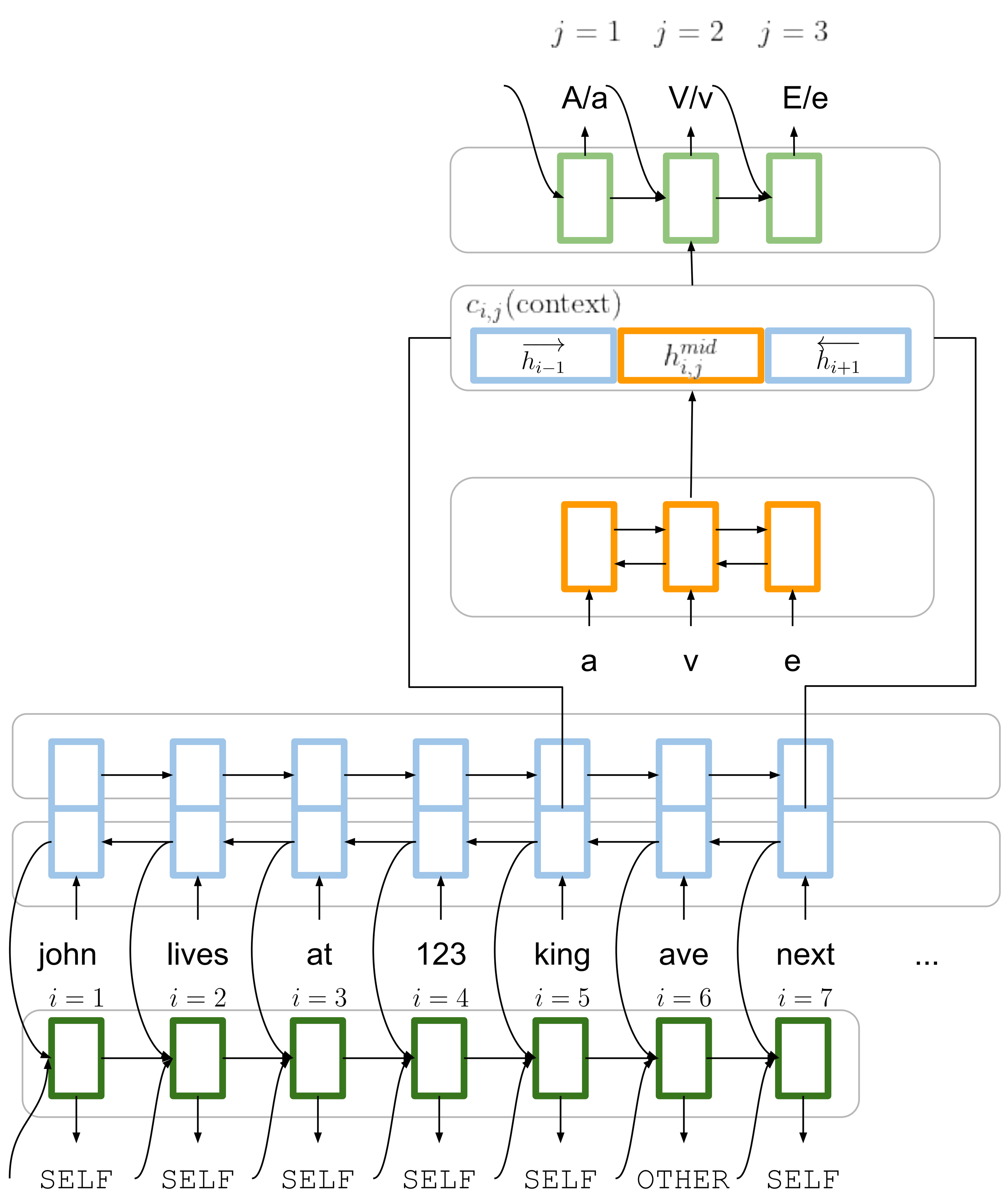
haozhang@google.com

Google Research

Abstract

Capitalization normalization (truecasing) is the task of restoring the correct case (uppercase or lowercase) of noisy text. We propose a fast, accurate and compact two-level hierarchical word-and-character-based recurrent neural network model. We use the truecaser to normalize user-generated text in a Federated Learning framework for language modeling. A case-aware language model trained on this normalized text achieves the same perplexity as a model trained on text with gold capitalization. In a real user A/B experiment, we demonstrate that the improvement translates to reduced prediction error rates in a virtual keyboard application. Similarly, in an ASR language model fusion experiment, we show reduction in uppercase character error rate and word error rate.

Word-and-Character-based Hierarchical RNN



Accuracy, Speed, and Model Size Comparison

System		*Precision*	*Recall*	*F1*	*Speed*	*# of params*
5-gram FST	unpruned, unoptimized	91.64	43.55	59.04	1.0x	10M
	pruned, optimized	**91.88**	41.19	56.88	88.0x	1M
char. RNN	small, 1-layer uni-, dec.	69.11	22.86	34.35	0.7x	230K
	small, 1-layer bi-, enc.&dec.	86.12	75.07	80.22	0.5x	400K
	large, 2-layer bi-, enc.&dec.	87.06	78.09	82.33	0.1x	8.4M
hier. RNN, student	small, 1-layer bi-, enc.&dec. $\times 2$	86.95	79.81	83.23	2.2x	1.3M
hier. RNN, teacher	large, 2-layer bi-, enc.&dec. $\times 2$	88.01	**82.60**	**85.22**	0.3x	19.2M

- FST models have high precision but low recall. Character-based RNN models are slow.
- Hierarchical RNN models have the best accuracy and speed trade-offs.

Case-aware Language Models

Capitalization Model	*Perplexity*
50% corrupt	59.41
25% corrupt	54.68
5-gram FST	51.74
hier. RNN	**51.60**
oracle	51.61

- Perplexities of RNN language models on LM1B using different capitalization normalization methods.

Case-aware Language Models in Speech Recognition

Model	*WER*	*UER*
5-gram FST	5.8	32.6
hier. RNN	**5.6**	**32.4**

- ASR LM fusion experiment results. The two systemsin comparison differ only in the capitalization normalization model used to pre-process the LM training data. UER stands for upper-case error rate.

Case-aware Language Models in Virtual Keyboard Applications

Model	*WMR*	*RAC*
5-gram FST	5.81%	2.91%
hier. RNN	5.78%	2.87%
Rel. Reduction	**[-0.92,-0.11]**%	**[-2.21, -0.69]**%

- Virtual keyboard A/B experiment results. *WMR* is the fraction of words modified or retyped. *RAC* is the auto-correction rejection rate. The last row shows the 95% confidence interval of the relative reductions.

Conclusions

Truecasing provides a factored solution to improve case-aware language modeling for applications such as ASR and text input in virtual keyboards. We propose a hierarchical word-and-character-based RNN model with the speed advantage of word-based models and accuracy advantage of character-based models. The model is efficient enough to be uploaded to mobile devices to train a language model using Federated Learning. The improvement is manifested in reduction of prediction error rates in a large-scale A/B experiment using a virtual keyboard and an ASR LM fusion experiment.