Jointly Learning to Align and Convert Graphemes to Phonemes with Neural Attention Models

Shubham Toshniwal, Karen Livescu
Toyota Technological Institute at Chicago (TTIC)

Abstract

- Most prior work on grapheme-to-phoneme (G2P) conversion requires explicit alignments for training [1, 2].
- Recent work using recurrent neural network (RNN) in an encoder-decoder fashion, requiring no alignment, has shown potential [3, 4].
- However, to date the best performing models still use explicit alignment [3, 4].
- We use the attention enabled encoder-decoder model and achieve state-of-the-art results on three standard data sets (CMUDict, Pronlex, and NetTalk).

Grapheme to Phoneme Conversion

- Problem: Convert a word, a sequence of characters/graphemes, to its pronunciation, a sequence of phonemes. For example,
  - *knife* $\longrightarrow$ [N AY F], *exit* $\longrightarrow$ [EH K S IH T]
- Motivation: Essential component of text-to-speech (TTS) and automatic speech recognition (ASR) systems for augmenting static pronouncing dictionaries.
- Challenges:
  - Output sequence can be shorter/longer than input sequence.
  - Grapheme pronunciation depends on its context.
  - Word pronunciation depends on its etymology.
- Performance metrics:
  - Word Error Rate (WER): $\mathbb{1}_{\boldsymbol{y} \neq \boldsymbol{y}_{\text{pred}}}$
  - Phoneme Error Rate (PER): $\frac{\text{Edit distance}(\boldsymbol{y}, \boldsymbol{y}_{\text{pred}})}{|\boldsymbol{y}|}$

Models

Global Attention

Uses the attention mechanism of [5], shown in Figure 1.

Local Attention

- Context vector c_t, used by attention, is calculated using a *localized* context window [p_t - D, p_t + D] centered at alignment position p_t.
- We consider 2 such variants proposed by [6]:
  - Monotonic Alignment (local-m): $p_t = t$
  - Predictive Alignment (local-p): $p_t = T_g \cdot \sigma(\boldsymbol{v_p}^T \tanh(\boldsymbol{W_p d_t}))$

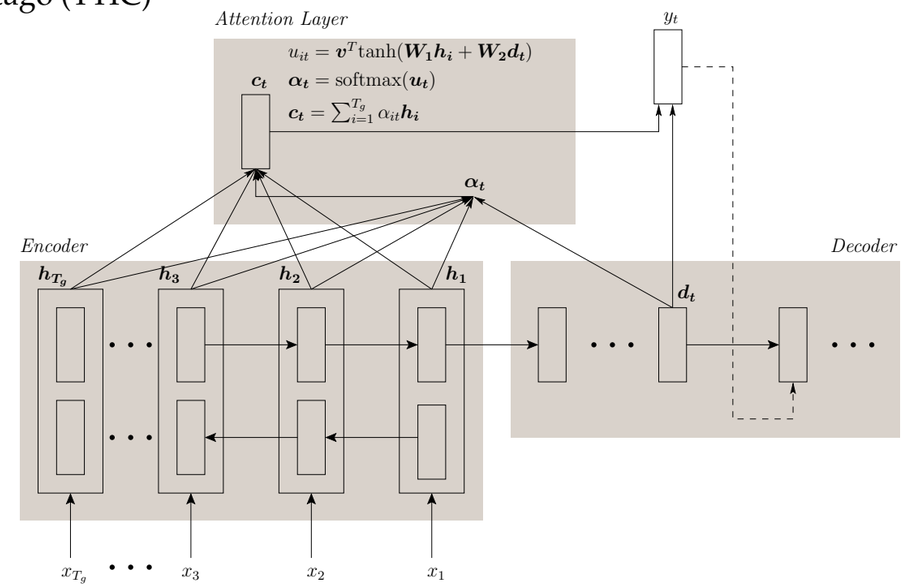
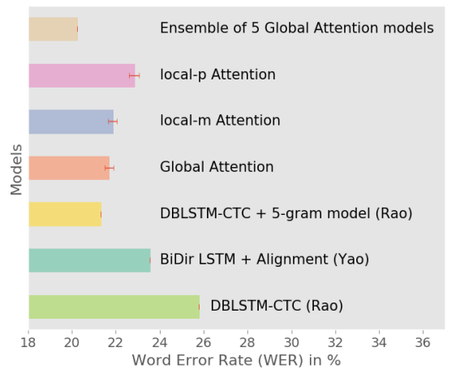
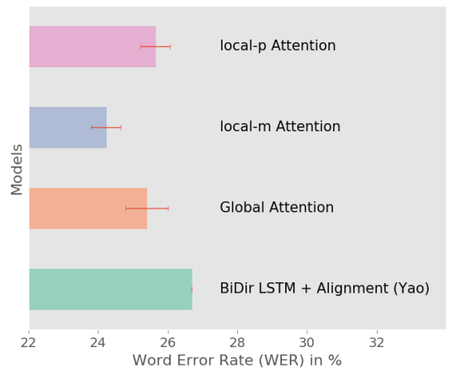


Figure 1: A global attention encoder-decoder model reading the input sequence $x_1, \cdots, x_{T_g}$ and outputting the sequence $y_1, \cdots, y_t, \cdots$

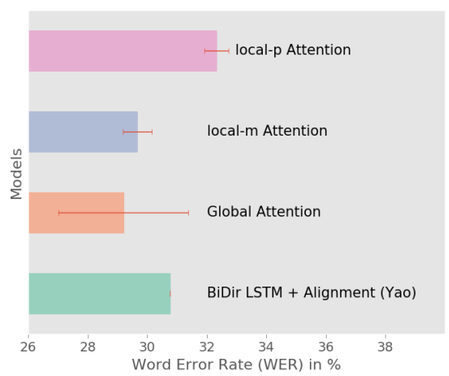
Results



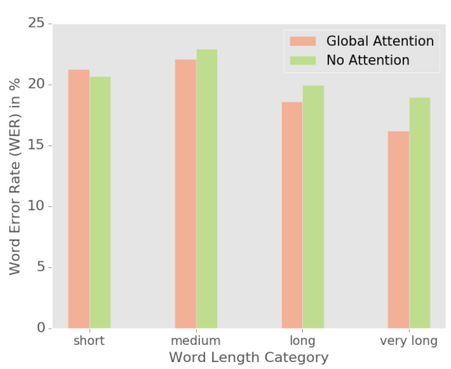
(a) CMUDict



(b) Pronlex



(c) NetTalk

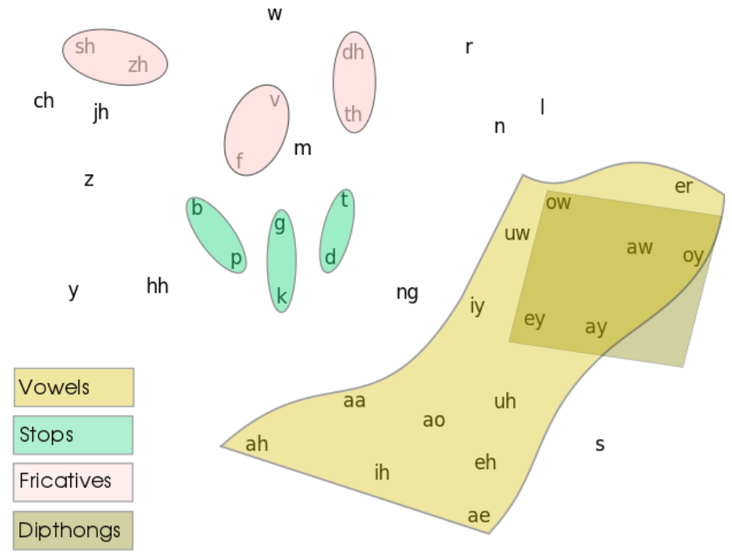


(d) WER as a function of word length

Error Analysis

	Foreign Origin Names	Abbreviations
Word	QUIXOTE (*Spanish*)	BLVD
Ground Truth	K IY HH OW T IY	B UH L AH V AA R D
Prediction	K W IH K S OW T	B L AH D
Word	MACIOCE (*Italian*)	JNA
Ground Truth	M AA CH OW CH IY	JH EY EH N EY
Prediction	M AH S IY OW S	N AH
	Wrong Ground Truth	**Under/Over Conversion**
Word	STACIE	KITTIWAKE
Ground Truth	S T AE K IY	K IH T AH W EY K
Prediction	S T EY S IY	K IH T AH W W K K
Word	COMMERICAL	LASTS
Ground Truth	K AH M ER SH AH L	L AE S T S
Prediction	K AH M EH R AH K AH L	L AE S

Phoneme Embedding Visualization



References

[1] Stanley F. Chen. Conditional and joint models for grapheme-to-phoneme conversion. 2003.

[2] Maximilian Bisani and Hermann Ney. Joint-sequence models for grapheme-to-phoneme conversion. 2008.

[3] Kanishka Rao et al. Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks. 2015.

[4] Kaisheng Yao and Geoffrey Zweig. Sequence-to-sequence neural net models for grapheme-to-phoneme conversion. 2015.

[5] Oriol Vinyals et al. Grammar as a foreign language. 2015.

[6] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. 2015.