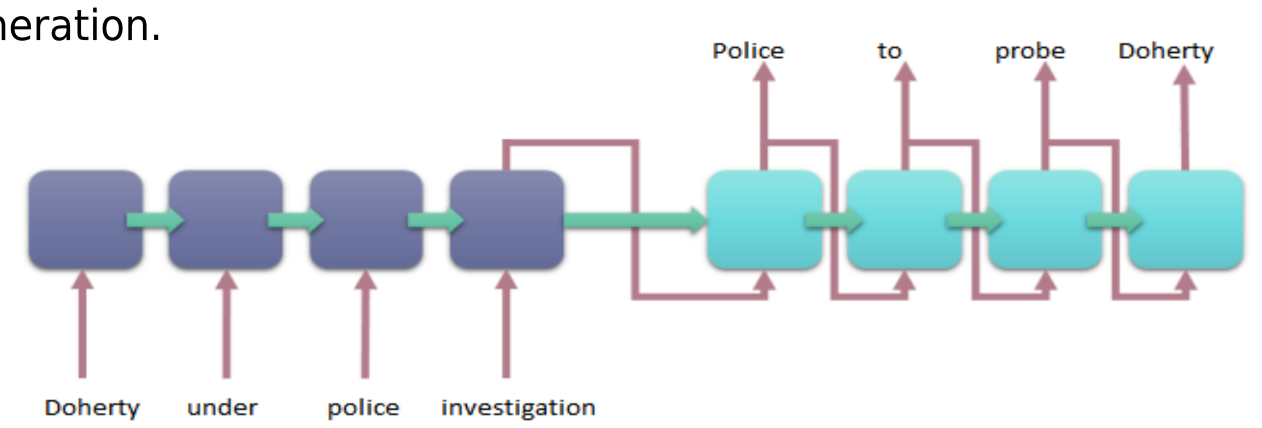
Neural Paraphrase Generation with Stacked Residual LSTM Networks

Aaditya Prakash[1,2], Sadid A. Hasan[2], Kathy Lee[2], Vivek Datla[2], Ashequl Qadir[2], Joey Liu[2], Oladimeji Farri[2]
[1]Brandeis University, Waltham, MA, USA
[2]Artificial Intelligence Laboratory, Philips Research North America, Cambridge, MA, USA
aprakash@brandeis.edu

Problem Definition

Paraphrasing, the task of expressing the same meaning in different possible ways, is an important subtask in various Natural Language Processing (NLP) applications such as question answering, information extraction, information retrieval, summarization and natural language generation.



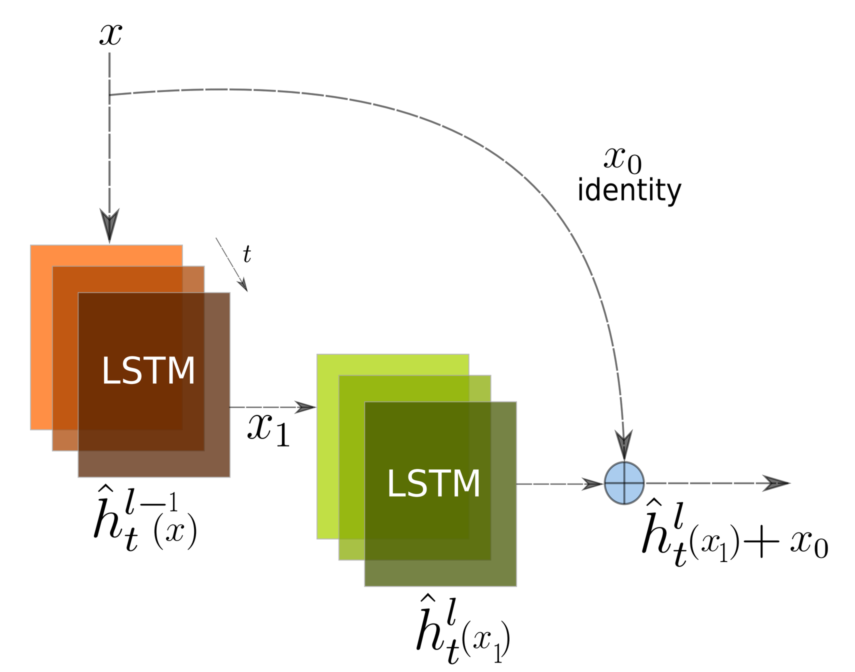
- Conventional paraphrase generation methods either leverage hand-written rules and thesauri-based alignments, or use statistical machine learning principles.

- To the best of our knowledge, this work is the first to explore deep learning models for paraphrase generation.

- Our primary contribution is a stacked residual LSTM network, where we add residual connections between LSTM layers. This allows for efficient training of deep LSTMs.

- Our model outperforms sequence to sequence, attention-based, and bi-directional LSTM models on BLEU, METEOR, TER, and greedy embedding.

Stacked Residual LSTM



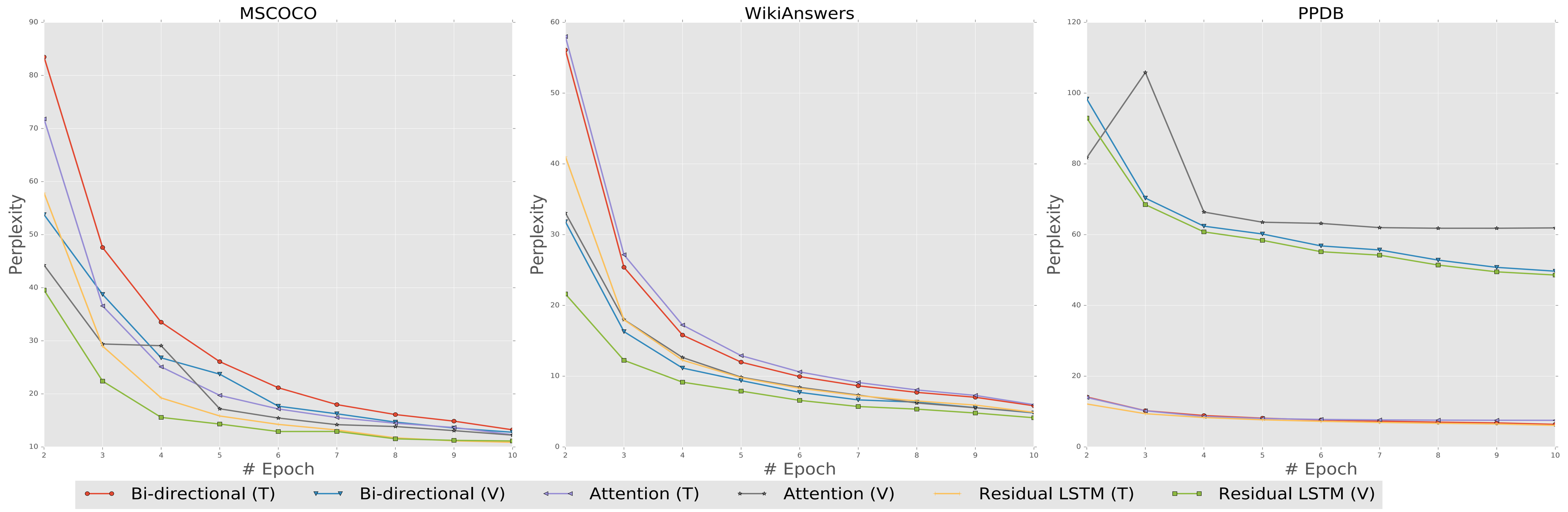
$$\hat{h}_t^{(l)} = f_h^l(h_t^{(l-1)}, h_{t-1}^{(l)}) + x_{l-n}$$

- h_t^l hidden state at layer l at time step t
- x input to layer $i+1$
- n^i # layers to skip between residual connection. Figure shows n=2
- Addition of residual connection does not add any learnable parameters

Training and Results

	PPDB	WikiAnswers	MSCOCO
Source	south eastern	what be the symbol of magnesium sulphate	a small kitten is sitting in a bowl
Reference	the eastern part	chemical formulum for magnesium sulphate	a cat is curled up in a bowl
Generated	south east	do magnesium sulphate have a formulum	a cat that is sitting on a bowl
Source	organized	what be the bigggest galaxy know to man	an old couple at the beach during the day
Reference	managed	how many galaxy be there in you known universe	two people sitting on dock looking at the ocean
Generated	arranged	about how many galaxy do the universe contain	a couple standing on top of a sandy beach
Source	counselling	what do the ph of acid range to	a little baby is sitting on a huge motorcycle
Reference	be kept informed	a acid have ph range of what	a little boy sitting alone on a motorcycle
Generated	consultations	how do acid affect ph	a baby sitting on top of a motorcycle

Example paraphrases generated using the 4-layer Residual LSTM with beam size 5.



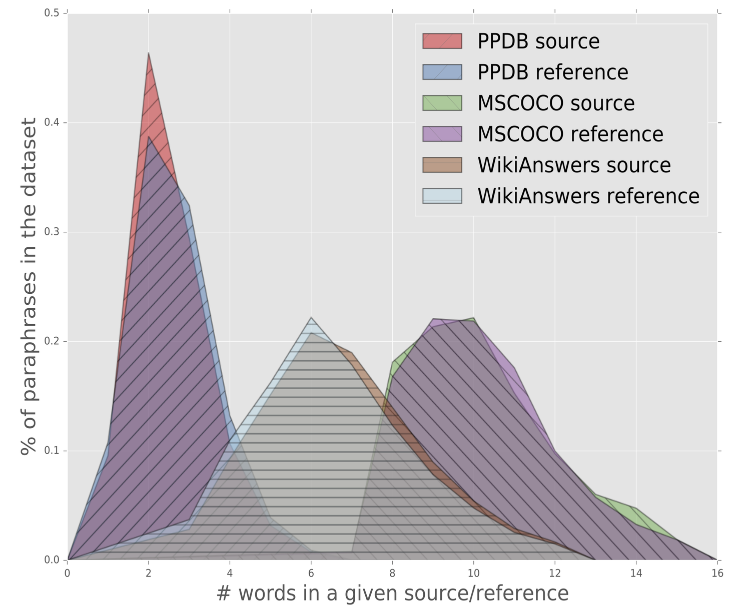
Dataset and Models

Dataset	Training	Test	Vocabulary Size
PPDB	4,826,492	20,000	38,279
WikiAnswers	4,826,492	20,000	50,000
MSCOCO	331,163	20,000	30,332

Dataset details

Models	Reference
Sequence to Sequence	(Sutskever et al., 2014)
With Attention	(Bahdanau et al., 2015)
Bi-directional LSTM	(Graves et al., 2013)
Residual LSTM	Our proposed model

Models



- Words were represented as one-hot vector

- Each model trained for 10 epochs

- Dropout of 50% was applied on LSTM layers

- Number of LSTM units was 512 for all models and all layers

- Training time (Titan X, Theano) : 36 hours for WikiAnswers and PPDB; : 14 hours for MSCOCO

- Perplexity was used as training loss

- Beam search used for generating samples

Analysis

- Scores on various metrics vary across the datasets due to differences in sentence length and vocabulary.

- PPDB contains very short phrases and does not score well with metrics like BLEU and METEOR, which penalizes short phrases.

- Deeper LSTM always leads to better performance.

- Larger beam size always improves performance but only marginally

- Models exploit the dataset '*bias*'. For example an OBJECT is mostly paraphrased with an OBJECT (eg. bowl, motorcycle). Shorter phrases generate shorter paraphrases.

- Perplexity does not incorporate reward "diversity", thus a better metric for paraphrase training and evaluation is required.

- Residual LSTM layers is useful for paraphrase generation, but it may not perform well for machine translation because not every word in a source sequence needs to be substituted for paraphrasing.

- Our work can be used to augment the datasets like MSCOCO used for image caption generation.

- We plan to explore memory networks to improve paraphrase generation.

Evaluation

		Beam size = 5				Beam size = 10			
#Layers	Model	BLEU↑	METEOR↑	Emb Greedy↑	TER↓	BLEU↑	METEOR↑	Emb Greedy↑	TER↓
	PPDB								
2	Sequence to Sequence	12.5	21.3	32.55	82.9	12.9	20.5	32.65	83.0
	With Attention	13.0	21.2	32.95	82.2	13.8	20.6	32.29	81.9
4	Sequence to Sequence	18.3	**23.5**	33.18	82.7	18.8	**23.5**	33.78	82.1
	Bi-directional	19.2	23.1	34.39	77.5	19.7	23.2	34.56	84.4
	With Attention	19.9	23.2	34.71	83.8	20.2	22.9	**34.90**	77.1
	Residual LSTM	**20.3**	23.1	**34.77**	**77.1**	**21.2**	23.0	34.78	**77.0**
	WikiAnswers								
2	Sequence to Sequence	19.2	26.1	62.65	35.1	19.5	26.2	62.95	34.8
	With Attention	21.2	22.9	63.22	37.1	21.2	23.0	63.50	37.0
4	Sequence to Sequence	33.2	29.6	73.17	28.3	33.5	29.6	73.19	28.3
	Bi-directional	34.0	30.8	73.80	27.3	34.3	30.7	73.95	27.0
	With Attention	34.7	31.2	73.45	27.1	34.9	31.2	73.50	27.1
	Residual LSTM	**37.0**	**32.2**	**75.13**	**27.0**	**37.2**	**32.2**	**75.19**	**26.8**
	MSCOCO								
2	Sequence to Sequence	15.9	14.8	54.11	66.9	16.5	15.4	55.81	67.1
	With Attention	17.5	16.6	58.92	63.9	18.6	16.8	59.26	63.0
4	Sequence to Sequence	28.2	23.0	67.22	56.7	28.9	23.2	67.10	56.3
	Bi-directional	32.6	24.5	68.62	53.8	32.8	24.9	68.91	53.7
	With Attention	33.1	25.4	69.10	54.3	33.4	25.2	**69.34**	53.8
	Residual LSTM	**36.7**	**27.3**	**69.69**	**52.3**	**37.0**	**27.0**	69.21	**51.6**

Evaluation results on PPDB, WikiAnswers, and MSCOCO (Best results are in **bold**).