Temporal Hierarchies in Sequence to Sequence for Sentence Correction

Gwenaelle Cunha Sergio, Dennis Singh Moirangthem, Minho Lee

Artificial Brain Research Lab., Kyungpook National University, South Korea

WCCI 2018, 13-July-2018

Contents

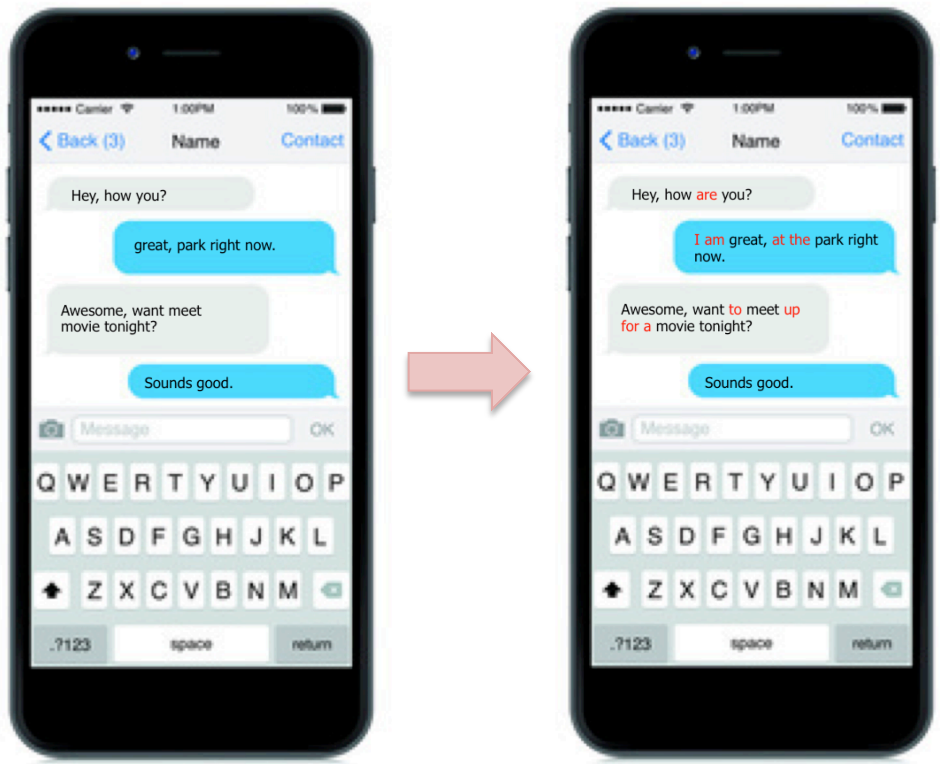
- **Introduction**
- **Proposed model**
- **Building the dataset**
- **Results**
- **Conclusion**

Motivation

Grammatical error correction (**GEC**) is *"the task of detecting and correcting grammatical errors in text written by non-native English writers"*.

Our problem

Noise in the form of missing words in the language domain



Statistical vs Neural

- Statistical vs Neural Machine Translation

- **SMT**: *"consists of components that are trained separately and combined during decoding"* (Koehn, 2010)
  - Usually built for specific error types (e.g. determiner or preposition errors)

- **NMT**: *"learns a single large neural network which inputs a sentence and outputs a translation, being able to correct erroneous word phrases and sentences that have not been seen in the training set more effectively"* (Luong, 2015).
  - Able to handle all error types simultaneously
  - Helpful due to the lack of error-annotated learner corpora for GEC
  - Able to generate new, original sentences
  - Seq2seq Encoder-Decoder approach

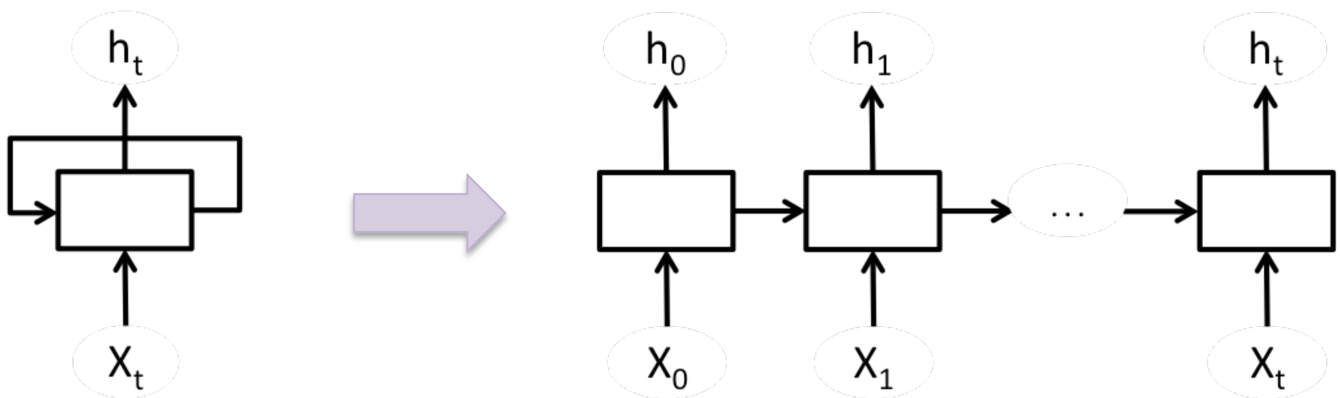
Limitations

Current models don't consider different levels of **compositionality** between words and sentences without dramatically increasing training time and memory usage.

Proposed Model

- Seq2seq + Multiple Timescale
- Used in Kim's (2016) work for abstractive summarization of scientific articles
- Temporal hierarchy concept in MTGRU performs well in language modeling tasks
- Handles long term dependency better with the help of the varying timescales to represent multiple compositionalities of language

Recurrent Neural Network



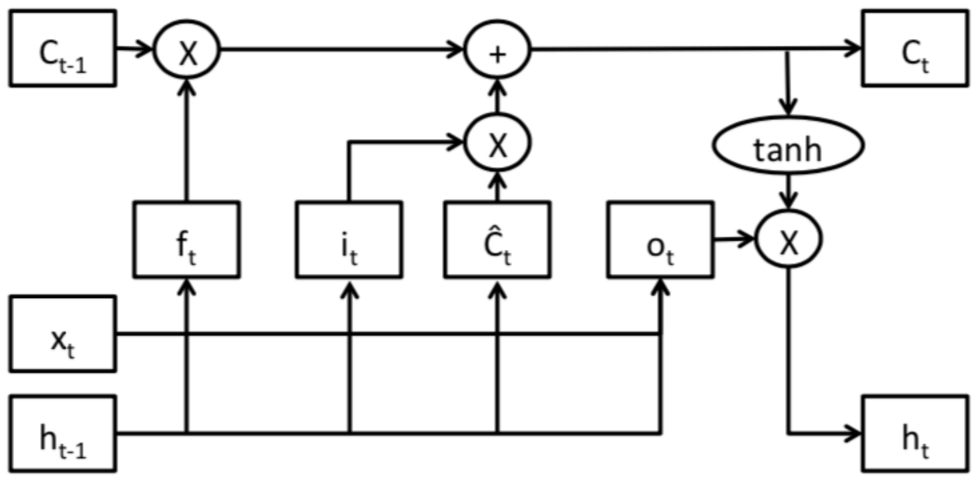
$$h_t = \sigma(W x_t + U h_{t-1})$$

- Limitations:
  - Learning temporal dependencies of long-term nature, such as longer sentences in language
  - Vanishing and exploding gradient problem

Long Short-Term Memory

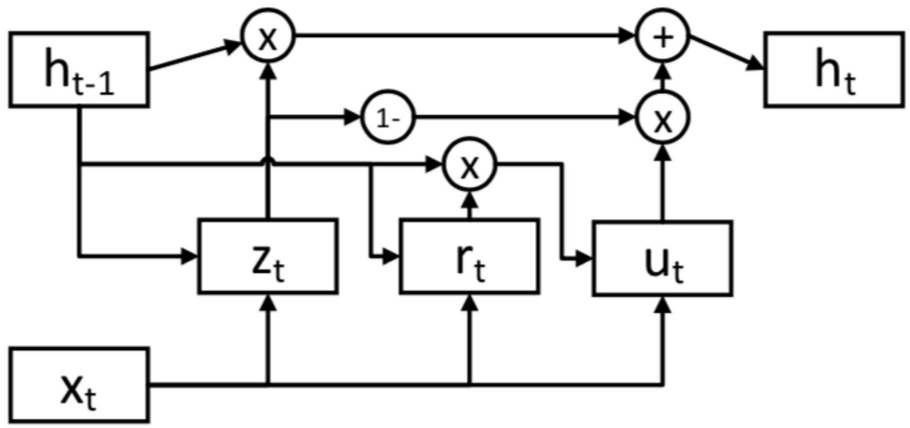
Gating mechanism to allow learning of longer-term dependencies

$$
\begin{aligned}
f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1}) \\
i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1}) \\
\tilde{C}_t &= tanh(W_{xC}x_t + W_{hC}h_{t-1}) \\
C_t &= f_t C_{t-1} + i_t \tilde{C}_t \\
o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1}) \\
h_t &= o_t tanh(C_t)
\end{aligned}
$$



Gated Recurrent Unit

Similar to LSTM requiring less memory due to the deletion of the output gate and separate memory cells



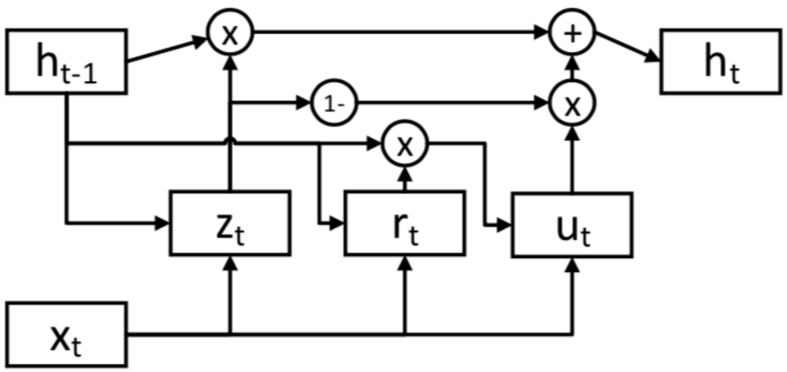
$$r_t = \sigma(W_{xr}x_t + W_{hr}h_{t-1})$$
$$z_t = \sigma(W_{xz}x_t + W_{hz}h_{t-1})$$
$$u_t = tanh(W_{xu}x_t + W_{hu}(r_t \odot h_{t-1}))$$
$$h_t = (1 - z_t)h_{t-1} + z_t u_t$$

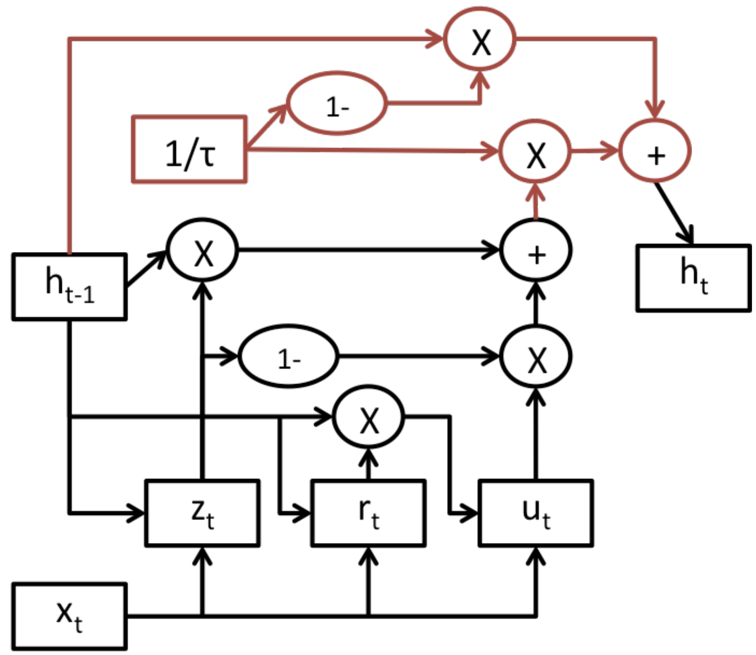
Multiple Timescale GRU

- Aim:
  - Incorporate the temporal hierarchy structure to the GRU so as to enable it to handle multiple levels of compositionality, similar to how human brain organizes itself to a temporal hierarchical structure to handle language.

- GRU vs MTGRU
  - Introduction of another gating unit (timescale constant $1/\tau$) that modulates the mixture of past and current hidden states in the MTGRU
  - The other gates remain unchanged

Multiple Timescale GRU





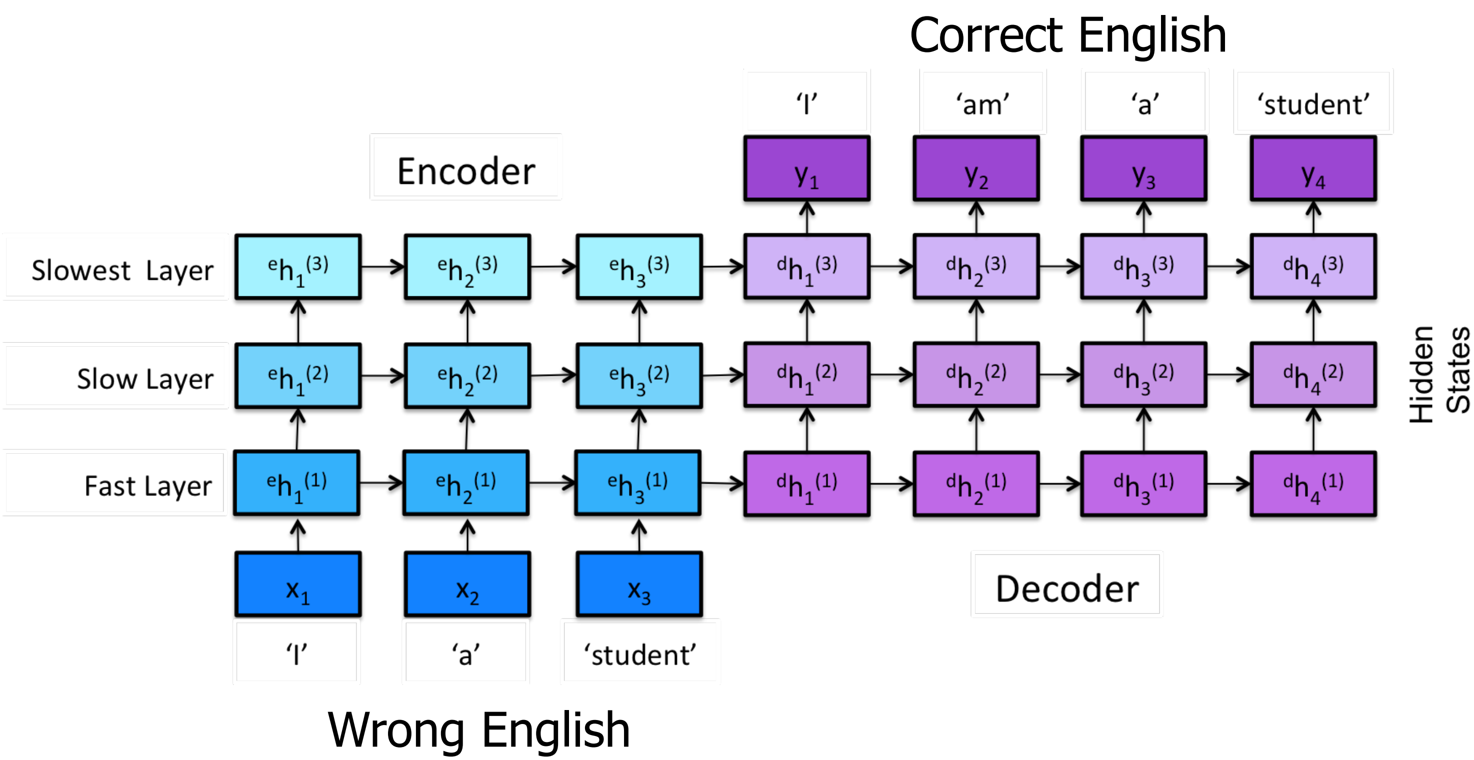
$$r_t = \sigma(W_{xr}x_t + W_{hr}h_{t-1})$$
$$z_t = \sigma(W_{xz}x_t + W_{hz}h_{t-1})$$
$$u_t = tanh(W_{xu}x_t + W_{hu}(r_t \odot h_{t-1}))$$

$$h_t = (1 - z_t)h_{t-1} + z_t u_t$$

$$h_t = ((1 - z_t)h_{t-1} + z_t u_t)\frac{1}{\tau} + (1 - \frac{1}{\tau})h_{t-1}$$

Seq2Seq

- MTGRU Encoder-Decoder



Dataset

- Difficult to find
  - Create own dataset: Subset of WMT'15 English-to-French dataset
  - Use the correct English sentences as output and do pre-processing on that to obtain the wrong sentences data
- Original data
  - 20GB of disk space
  - 22,000,000 sentences of various lengths
- Increase the problem complexity
  - Longer sentences: 15 to 20 words
  - Total: 3,000,000 sentences (**target** dataset)

Dataset

- Target: 3,000,000 sentences of correct English
- **Input**: modification of the target data using Python's Natural Language Toolkit (NLTK)
  - Allows for part-of-speech tagging, or POS-tagging, of words.
  - Delete words in following group of tags (tagset)

Tag	Meaning	Example
CC	coordinating conjunction	*and*
DT	determiner	*the*
IN	preposition/subordinating conjunction	*in, of, for, like*
LS	list marker	*1)*
TO	to	*go 'to' the store*
UH	interjection	*errrrrrrrm*

Experimental Setup

- Models: 4 seq2seq models
- 3 model variations: 2, 3 and 4 layers
- Timescale constants ($1/\tau$)
  - 2 layers (1, 0.999), 3 layers (1, 0.999, 0.998), 4 layers (1, 0.999, 0.998, 0.997)
  - After many experiments, we found that our model is highly sensi tive to larger timescales.
  - Reason: sentences have a maximum length of 20 words
  - Larger timescales are effective for handling longer term dependencies in paragraphs in summarization tasks

Experimental Setup

- Models hyper-parameters:
  - Buckets: [(10,15),(10,20),(15,20),(20,20)]
  - Units per layer: 1024 hidden units
  - Batch size: 64
  - Learning rate: 0.5 with decay factor of 0.99
  - Embedding size: 512
  - Vocabulary size: 40,000
  - Training: Nvidia Titan X GPU

Model	**Perplexity**	
	Train	*Test*
Vanilla RNN	892	8,521
LSTM	1.53	2.14
GRU	1.41	2.03
MTGRU	1.43	2.11

Evaluation Metric

- **BLEU**: BiLingual Evaluation Understudy (Papineni et al., 2002)
- Automatic machine translation evaluation method
- Range: 0 to 1
- Quick, language-independent, and correlates highly with human evaluation.
- 4 types of BLEU-n : n-gram precision

Results: 2,3,4-layer models

Model	**Score (BLEU-n)**			
	1-gram	*2-gram*	*3-gram*	*4-gram*
Vanilla RNN	0.0319	0.00133	$4.77e-09$	$9.35e-12$
LSTM	0.343	0.204	0.128	0.0819
GRU	0.392	0.239	0.152	0.0986
MTGRU	**0.415**	**0.270**	**0.183**	**0.127**

Model	**Score (BLEU-n)**			
	1-gram	*2-gram*	*3-gram*	*4-gram*
Vanilla RNN	0.0634	0.00259	$7.22e^{-09}$	$1.22e^{-11}$
LSTM	0.430	0.302	0.223	0.169
GRU	0.469	0.336	0.251	0.192
MTGRU	**0.484**	**0.350**	**0.263**	**0.202**

Model	**Score (BLEU-n)**			
	1-gram	*2-gram*	*3-gram*	*4-gram*
Vanilla RNN	0.0682	0.00281	$7.84e-09$	$1.33e-11$
LSTM	0.449	0.328	0.249	0.194
GRU	0.454	0.325	0.243	0.186
MTGRU	**0.489**	**0.353**	**0.266**	**0.203**

Results: 2-layer models

	Example 1
Input	Studies assigned ~~a~~ 0 or less did not receive ~~a~~ rating ~~and~~ were eliminated ~~from the~~ review.
RNN	to that . 0000
LSTM	Two confirmed 0 and the less did not receive the test were to the review effect.
GRU	The Studies for the age of 0 did not receive the matter were clear into these review.
MTGRU	The assigned assigned 0 less less did not receive an assigned were eliminated for review.

	Example 2
Input	It must be reviewed ~~by the~~ European Parliament ~~and the~~ Council ~~by~~ ## June ####.
RNN	The
LSTM	It must be reviewed by the European Parliament and the Council on ## June ####.
GRU	It must be addressed by the European Parliament and the Council of June ## June ####.
MTGRU	It must be reviewed by the European Council of the Council of ## June ####.

Results: 3-layer models

	Example 1
Input	Studies assigned ~~a~~ 0 or less did not receive ~~a~~ rating ~~and~~ were eliminated ~~from the~~ review.
RNN	. of . and and an . and . of . and . of . and . of . and
LSTM	Studies assigned 0 and less did not receive the rating or were raised under the review.
GRU	Studies assigned to 0 but less did not receive a rating or rating were eliminated.
MTGRU	Studies assigned 0 or less did not receive a rating but were eliminated from the review.

	Example 2
Input	It must be reviewed ~~by the~~ European Parliament ~~and the~~ Council ~~by~~ ## June ####.
RNN	and of . and that an . and . of . and . of . and . of . and
LSTM	It must be reviewed by the European Parliament of the Council on ## June ####.
GRU	It must be reviewed by the Parliament by Parliament and the European on ## June ####.
MTGRU	It must be reviewed by the European Parliament and the Council of ## June ####.

Results: 4-layer models

	Example 1
Input	Studies assigned ~~a~~ 0 or less did not receive ~~a~~ rating ~~and~~ were eliminated ~~from the~~ review.
RNN	The in of in of in of in of in of in of in of in of in of in
LSTM	Studies assigned that 0 less did not receive a rating and although apparently evaluated during review.
GRU	Studies assigned 0 but less did not receive any rating and were eliminated in the review.
MTGRU	The Recommendations assigned 0 or less did not receive a rating and were eliminated during the review.

	Example 2
Input	It must be reviewed ~~by the~~ European Parliament ~~and the~~ Council ~~by~~ ## June ####.
RNN	The in in in on in of in of in of in of in of in of in of in
LSTM	It must be reviewed by the European Parliament and the Council on the ## June ####.
GRU	It must be reviewed by the European Parliament and of the Council on ## June ####.
MTGRU	It must be reviewed by the European Parliament and Council on of ## June ####.

Conclusion

- Sentence correction model using a MT seq2seq architecture to handle incorrect data in the language domain
- Capable of better abstraction of the input data
- Model can handle longer sequences by representing multiple compositions of language in the data
- Dataset: modification of WMT'15
- MTGRU 3-layer model outperforms RNN, GRU and LSTM in BLEU-n evaluation and is comparable to the 4-layer model without the need of the additional layer complexity

Future Works

- Generalize model for other tasks
  - Specific grammatical errors
  - Misspelled words
  - Intrinsically wrong sentence structures including switching of nouns and verbs.

Thank you!

Gwenaelle Cunha Sergio

gwena.cs@gmail.com