

Temporal Hierarchies in Sequence to Sequence for Sentence Correction

Gwenaelle Cunha Sergio, Dennis Singh
Moirangthem, Minho Lee



WCCI 2018, 13-July-2018

Contents

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



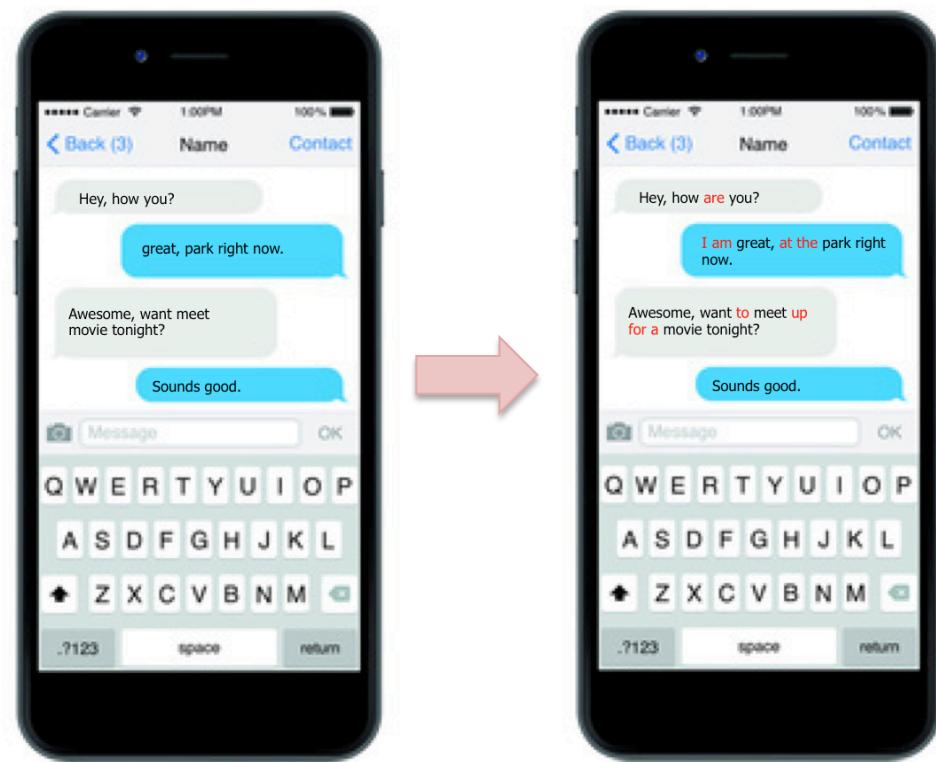
Motivation

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



Our problem

Treating missing data in the language domain, which can also be seen as a type of noise in that domain.



Machine Translation

- 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



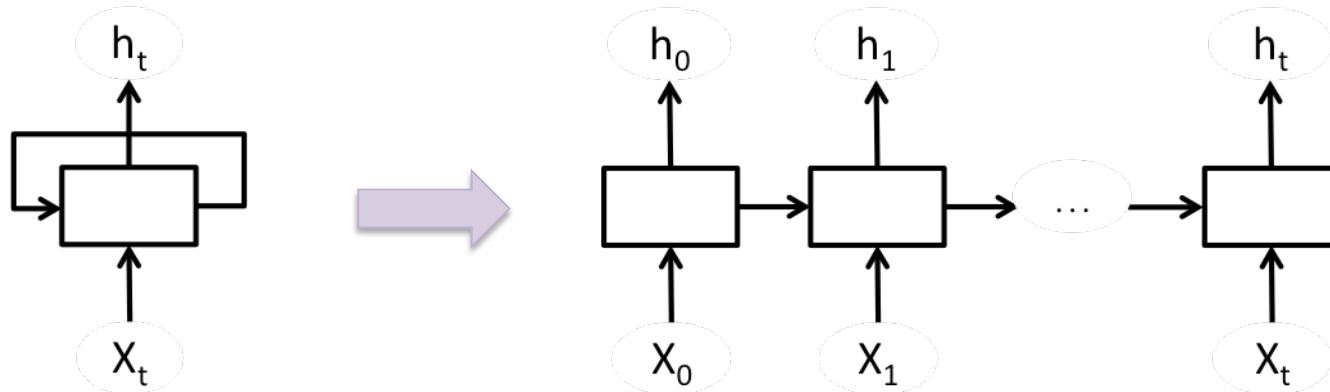
Limitations

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 compositionality of language

Recurrent Neural Network



$$h_t = \sigma(Wx_t + Uh_{t-1})$$

- 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

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1})$$

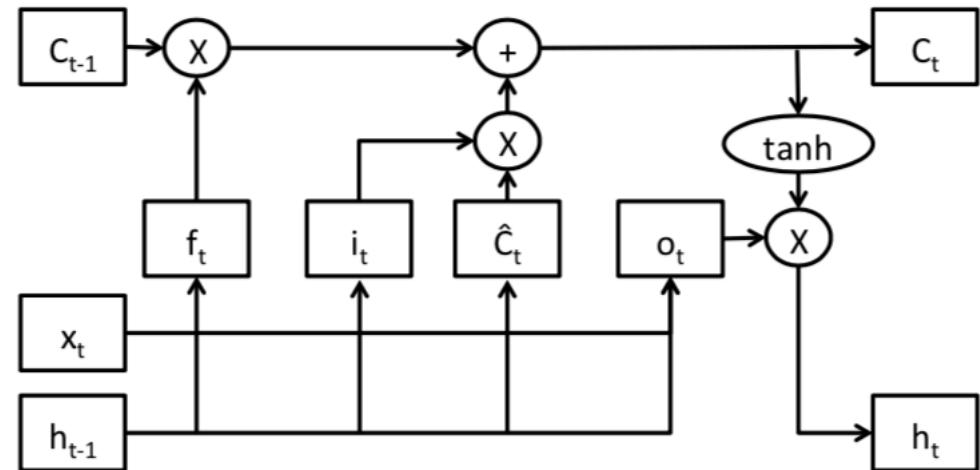
$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1})$$

$$\tilde{C}_t = \tanh(W_{xC}x_t + W_{hC}h_{t-1})$$

$$C_t = f_t C_{t-1} + i_t \tilde{C}_t$$

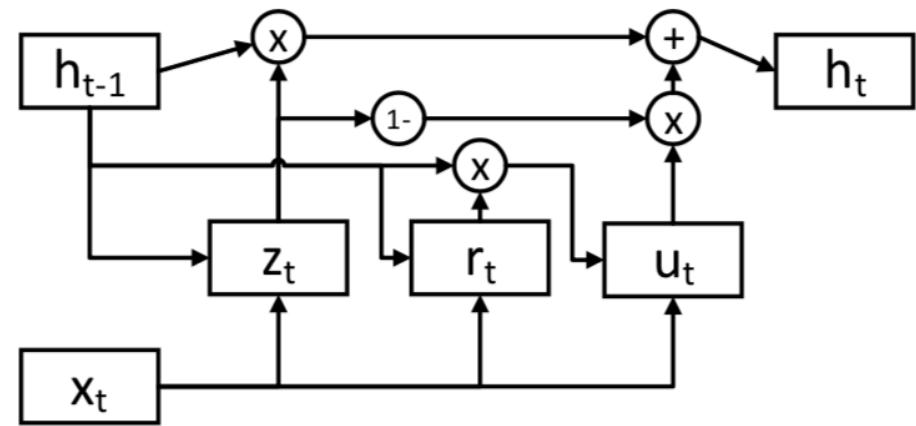
$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1})$$

$$h_t = o_t \tanh(C_t)$$



Gated Recurrent Unit

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



$$r_t = \sigma(W_{xr}x_t + W_{hr}h_{t-1})$$

$$z_t = \sigma(W_{xz}x_t + W_{hz}h_{t-1})$$

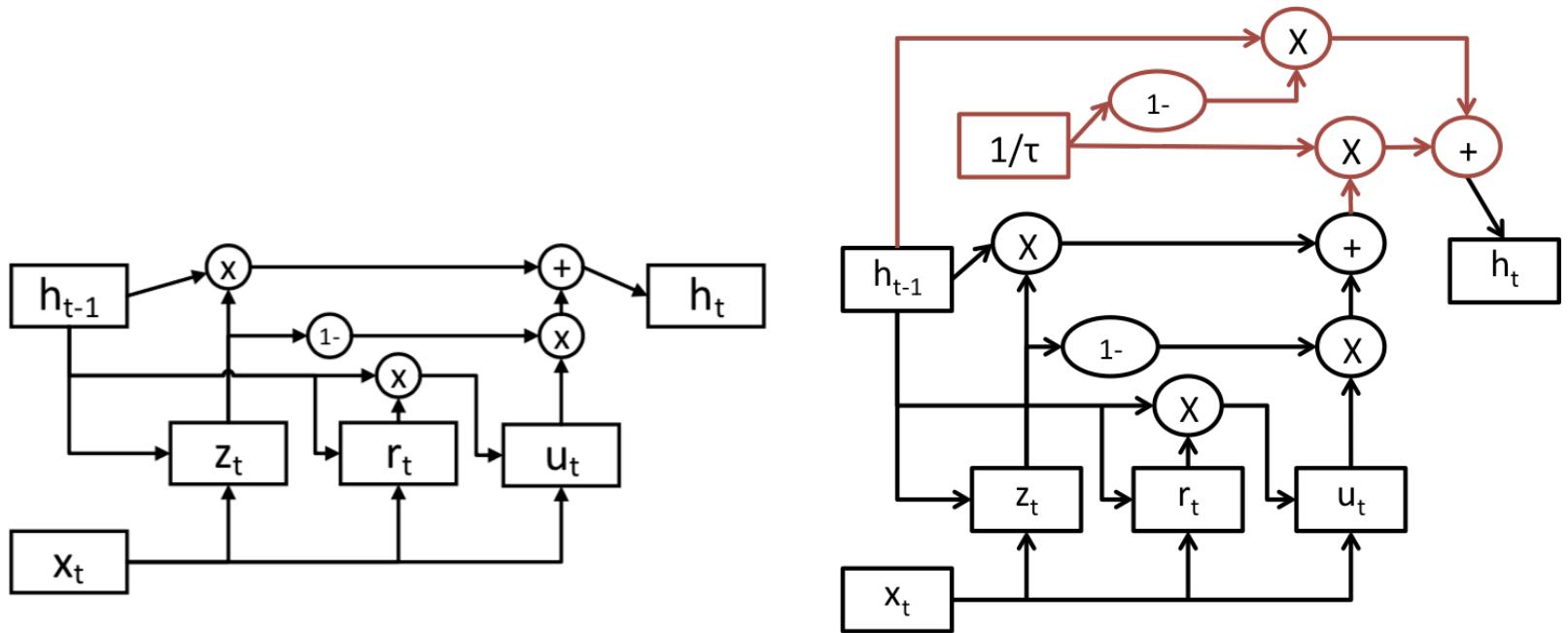
$$u_t = \tanh(W_{xu}x_t + W_{hu}(r_t \odot h_{t-1}))$$

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

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



$$r_t = \sigma(W_{xr}x_t + W_{hr}h_{t-1})$$

$$z_t = \sigma(W_{xz}x_t + W_{hz}h_{t-1})$$

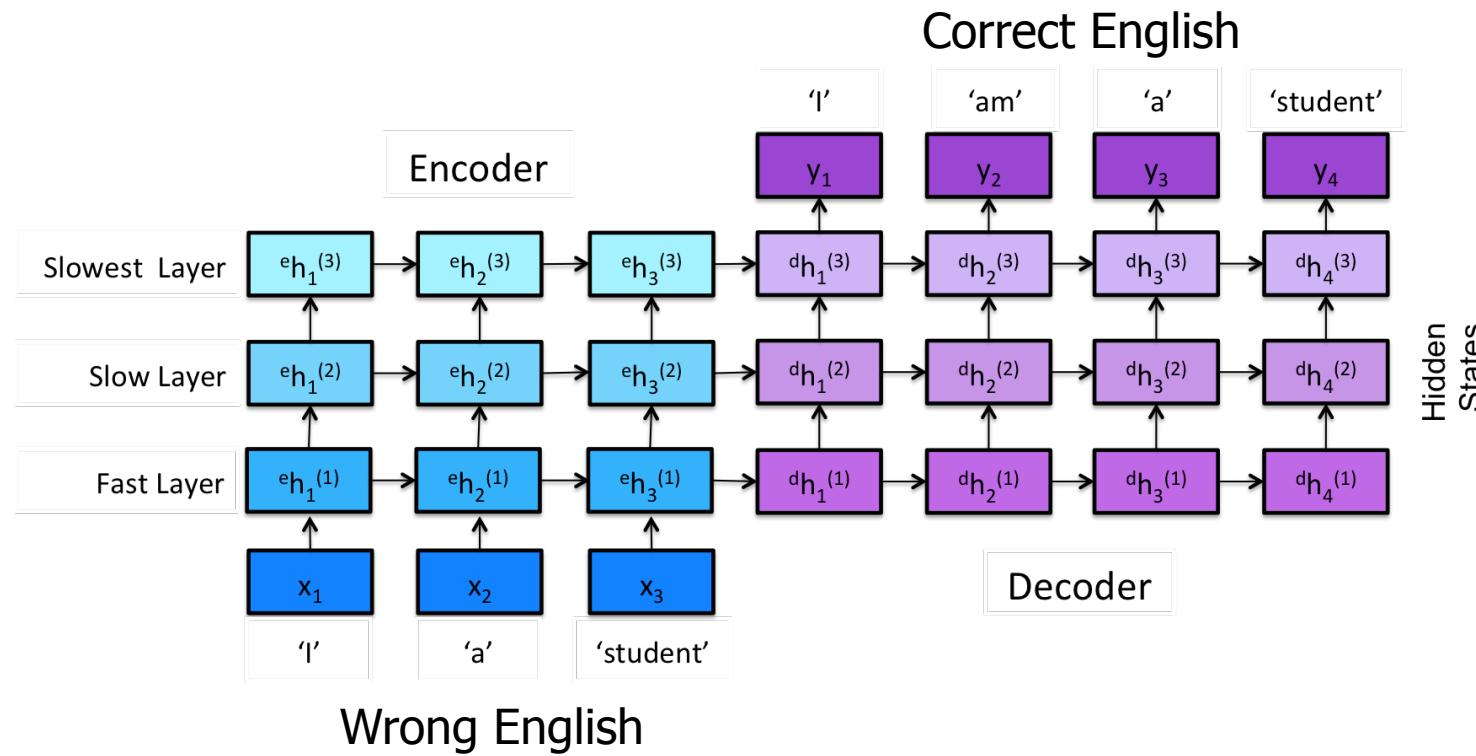
$$u_t = \tanh(W_{xu}x_t + W_{hu}(r_t \odot h_{t-1}))$$

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

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

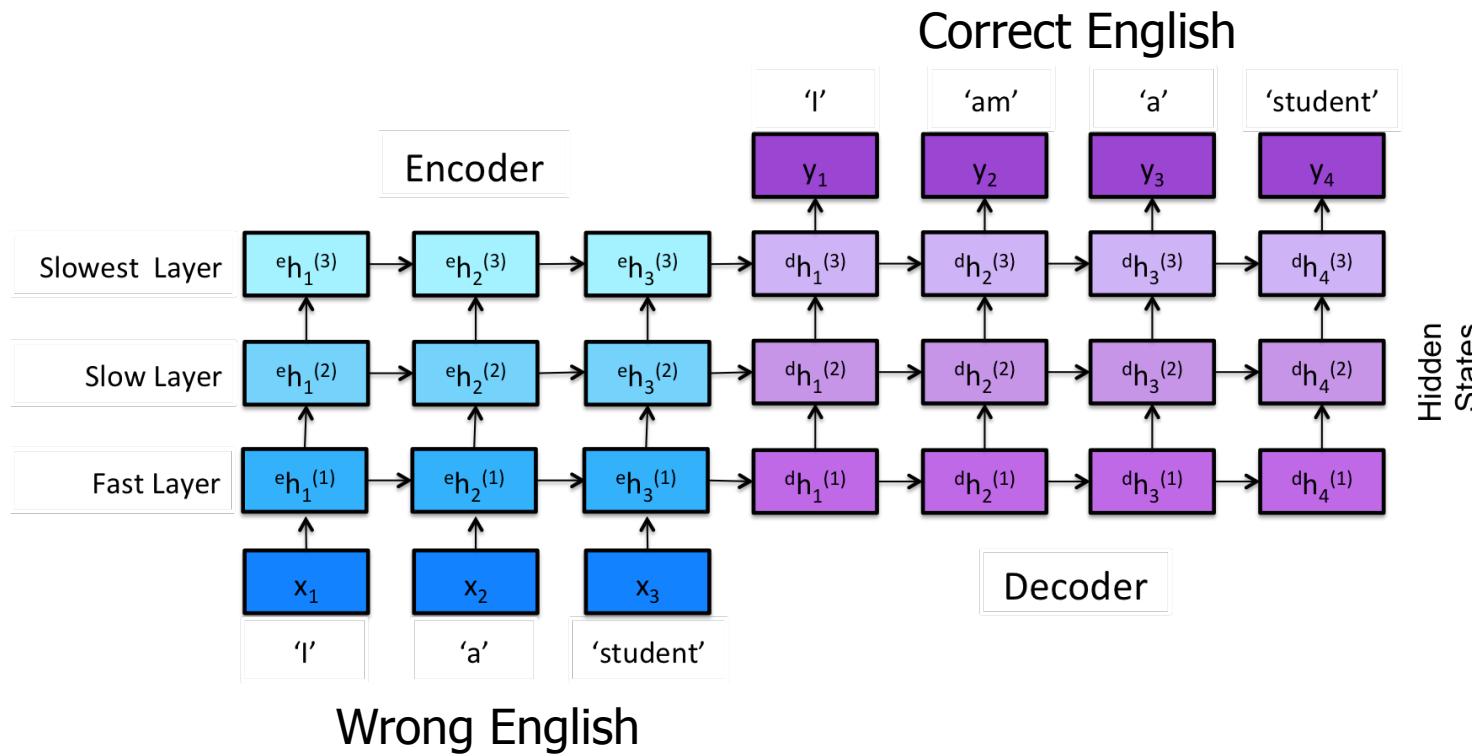
Seq2Seq

Model *encodes* wrong sentences so that the *decoder* is essentially a language model conditioned on the outputs of the encoder.



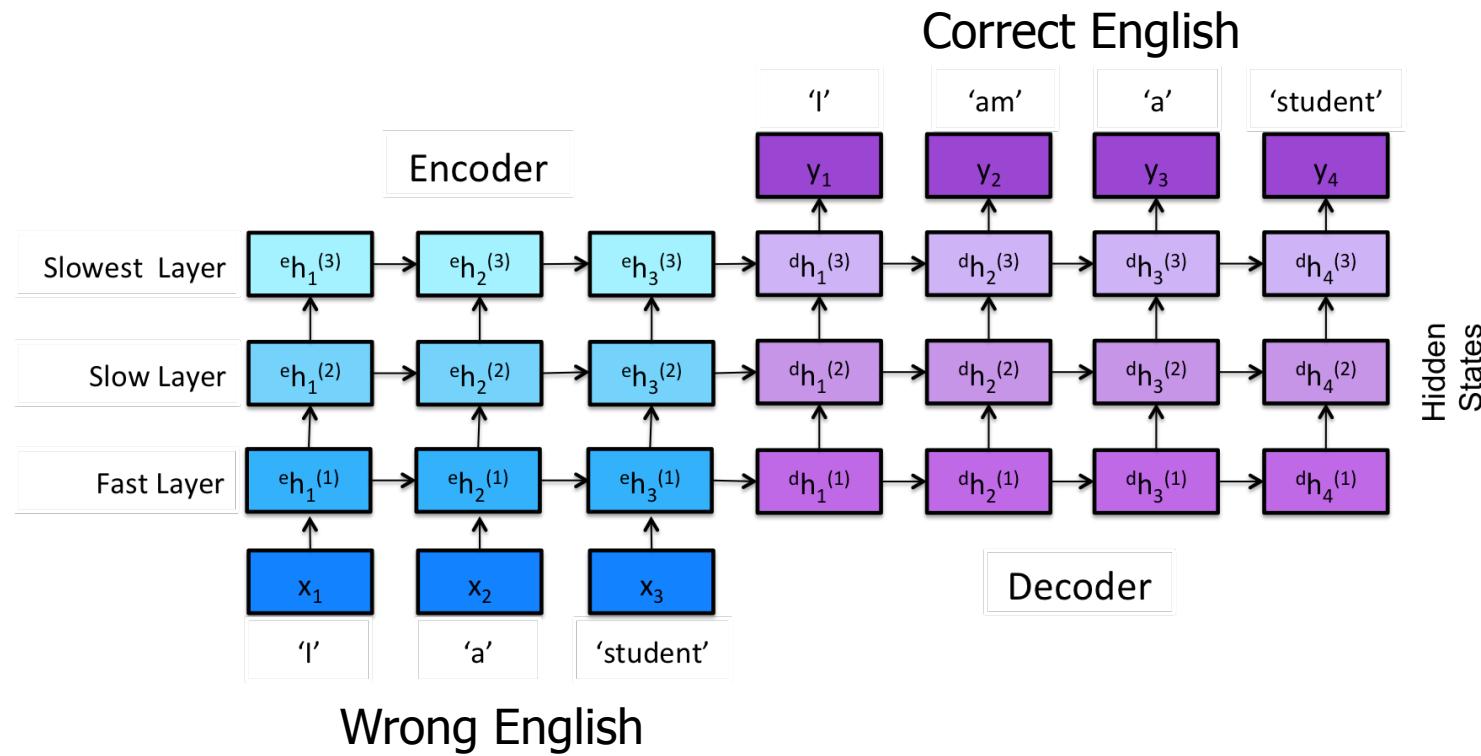
Seq2Seq

The different timescales in each layers enables the model to achieve *different levels of representation* and the deeper layers should encapsulate more *abstract features* of the input.



Seq2Seq

With the help of the encoder and decoder framework, our model can produce corrected sentences even though the input texts contain noise in the form of missing words.



Dataset

- Difficult to find
 - Create own dataset: Subset of WMT'15 English-to-French dataset
 - Use the correct English sentences as output
 - Pre-process output 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

- **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	<i>and</i>
DT	determiner	<i>the</i>
IN	preposition/subordinating conjunction	<i>in, of, for, like</i>
LS	list marker	<i>1)</i>
TO	to	<i>go ‘to’ the store</i>
UH	interjection	<i>errrrrrrrrm</i>

- Train 80% and test 20%

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 sensitive to larger timescales.
 - Reason: sentences have a maximum length of 20 words
 - Larger timescales are effective for handling longer term dependencies such as 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	
	<i>Train</i>	<i>Test</i>
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: BLEU

Model	Score (BLEU-n)			
	<i>1-gram</i>	<i>2-gram</i>	<i>3-gram</i>	<i>4-gram</i>
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)			
	<i>1-gram</i>	<i>2-gram</i>	<i>3-gram</i>	<i>4-gram</i>
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)			
	<i>1-gram</i>	<i>2-gram</i>	<i>3-gram</i>	<i>4-gram</i>
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 <u>assigned</u> assigned 0 less less did not receive an <u>assigned</u> were eliminated <u>for</u> review.

Example 2

Input	It must be reviewed by the European Parliament and the Council by ## June ####.
RNN	The
LSTM	It must be reviewed <u>by the</u> European Parliament <u>and the</u> Council on ## June ####.
GRU	It must be addressed <u>by the</u> European Parliament <u>and the</u> Council <u>of</u> June ## June ####.
MTGRU	It must be reviewed <u>by the</u> European Council <u>of the</u> Council <u>of</u> ## June ####.

Caption
— Correct additions of words
--- Unnecessary or incorrect substitutions
Placeholder for a number

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 ####.

Caption
— Correct additions of words
--- Unnecessary or incorrect substitutions
Placeholder for a number

Results: 4-layer models

Example 1

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

Example 2

Input	It must be reviewed <u>by</u> the European Parliament <u>and</u> the Council <u>by</u> ## June ####.
RNN	The in in on in of in of in of in of in of in
LSTM	It must be reviewed <u>by</u> the European Parliament <u>and</u> the Council <u>on</u> the ## June ####.
GRU	It must be reviewed <u>by</u> the European Parliament <u>and</u> <u>of</u> the Council <u>on</u> ## June ####.
MTGRU	It must be reviewed <u>by</u> the European Parliament <u>and</u> Council <u>on</u> <u>of</u> ## June ####.

Caption
— Correct additions of words
-- Unnecessary or incorrect substitutions
Placeholder for a number

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
gwna.cs@gmail.com

