Deep Recurrent Neural Network applied to Incomplete Data

Gwenaelle Cunha Sergio and Minho Lee
School of Electronics Engineering, Kyungpook National University

Motivation

- Text summarization with incomplete data
- In the real world, data is not always correct or complete.
- This work compares the effect that this type of "noisy" data has in models that have been trained with complete/correct data.
- We compare the Long Short-Term Memory (LSTM) model with the Multiple Timescale LSTM in a RNN Encoder-Decoder Framework
- We show that MTLSTM has better performance

Building dataset

- Term Frequency-Inverse Document Frequency (TF-IDF)
- Goal: to calculate the most relevant (important) words in a given document

$$Tf(d,t) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

Tf: frequency of appearance
d: document; *t*: # times term *t* appears in document *d*

$$Idf(c,t) = log_{10}\Big(\frac{N}{1 + Df(c,t)}\Big)$$

Idf: number of documents term *t* appears in
c: document (corpus); *t*: # times term *t* appears in corpus *c*;
log: tweak to avoid giving high scores to very rare words

$$Tf.Idf(c,d,t) = Tf(d,t)Idf(c,t)$$

Tf.Idf: term frquency-inverse document frequency
N: # total documents

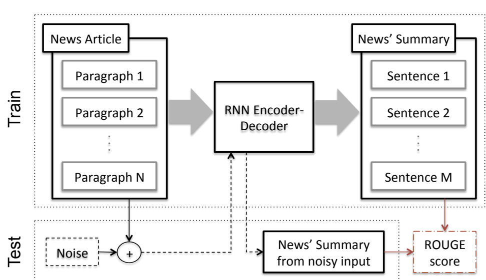
- Two datasets formed of incomplete sentences as input and complete sentences as output
  - 1. Missing relevant (important) words
  - 2. Missing common (non-important)
  - Variants of those dataset by changing the missing words ratio (from 0 to 20%)

Evaluation Method

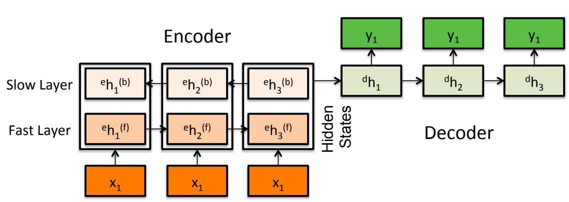
- Recall-Oriented Understudy for Gisting Evaluation (ROUGE), proposed by Lin (2004)
  - ROUGE-N (1 and 2): *N*-grams between generated and reference sentences
  - ROUGE-L: Longest Common Subsequence based statistics
- Package for automatic evaluation of summaries
- Test scores:
  - P-score: *precision*, # correct positive results / # all positive results returned by the model
  - R-score: *recall*, # correct positive results / # all relevant samples that should have been classified as positive.
  - F-Score: *accuracy*, considers both *P* and *R*.

Approach (1/2)

- Noise: missing words
- Compare effect of noise in model trained with complete (correct) data

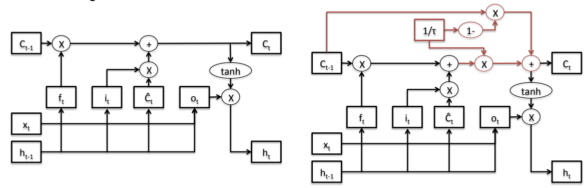


- Bi-directional LSTM RNN Encoder-Decoder



Approach (2/2)

- LSTM vs MTLSTM
- MTLSTM: same forget, input and output gates as LSTM + Multiple Timescales which differ depending on the layer. The deeper, the slower it is so as to mimic different levels of abstraction
- Model used: taus = {1, 0.9, 0.8}, 256 hidden units, 128 word embedding, 50000 vocabulary, 0.15 learning rate

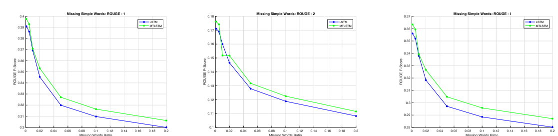


$$f_t = \sigma(W_{hf}h_{t-1} + W_{xf}x_t + b_f)$$
$$i_t = \sigma(W_{hi}h_{t-1} + W_{xi}x_t + b_i)$$
$$\tilde{C}_t = tanh(W_{hC}h_{t-1} + W_{xC}x_t + b_C)$$
$$C_t = f_tC_{t-1} + i_t\tilde{C}_t$$
$$o_t = \sigma(W_{ho}h_{t-1} + W_{xo}x_t + b_o)$$
$$h_t = o_t tanh(C_t)$$

$$C'_t = f_tC_{t-1} + i_t\tilde{C}_t$$
$$C_t = C'_t\frac{1}{\tau} + C_{t-1}\Big(1 - \frac{1}{\tau}\Big)$$

Results

- ROUGE F-Score Degradation Curve
- Simple Words



Excerpt of original input

[…]
The only thing crazier than a guy in snowbound Massachusetts boxing up the powdery white stuff and offering it for sale online?

People are actually buying it.

For $89, self-styled entrepreneur Kyle Waring will ship you 6 pounds of Boston-area snow in an insulated Styrofoam box -- enough for 10 to 15 snowballs, he says.
[…]

MT Output

Social media accounts claim filled more than 133 orders snow -- more than 30 tuesday alone , his busiest day yet. More than 45 total inches , boston has set record this winter snowiest month itshistory.

ROUGE-1	**0.3164**
ROUGE-2	**0.1225**
ROUGE-3	**0.2958**

Vanilla Output

Crazier massachusetts boxing up powdery white stuff offering it sale online? People are actually buying it. $ 89, self-styled entrepreneur kyle waring will ship you 6 pounds boston-area snow an insulated styrofoam box -- enough 10 15 snowballs, he says .

ROUGE-1	0.3097
ROUGE-2	0.1188
ROUGE-3	0.2885

Reference

A man in suburban boston is selling snow online to customers in warmer states. For $ 89 , he will ship 6 pounds of snow in an insulated styrofoam box.

- Complex Words

Excerpt of original input

[…]
The only thing crazier than a guy in snowbound Massachusetts boxing up the powdery white stuff and offering it for sale online?

People are actually buying it.

For $89, self-styled entrepreneur Kyle Waring will ship you 6 pounds of Boston-area snow in an insulated Styrofoam box -- enough for 10 to 15 snowballs, he says.
[…]

MT Output

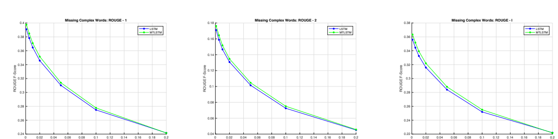
The only thing than a guy in massachusetts boxing up the white stuff and offering it for sale online. Most residents see the huge piles of choking their yards and sidewalks as a nuisance, but sawan opportunity.

ROUGE-1	**0.2777**
ROUGE-2	**0.0751**
ROUGE-3	**0.2551**

Vanilla Output

Kyle will you 6 pounds of - area in an -- enough for 10 to 15, he says. But not if you live in new england or surrounding. "But i 've thought about taking this idea and running with it for other seasonal items".

ROUGE-1	0.2748
ROUGE-2	0.0725
ROUGE-3	0.2519



Conclusion & Future works

- MTLSTM performs better than LSTM
- Better sentence generation

- Future works
  - Expand results by increasing the missing words ratio
  - Different tuning parameters and model structure (more layers, units per layer)

Acknowledgement

This work was partly supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government(MSIT) (R7124-16-0004, Development of Intelligent Interaction Technology Based on Context Awareness and Human Intention Understanding) (50%) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. NRF-2017R1E1A2A02023371) (50%)