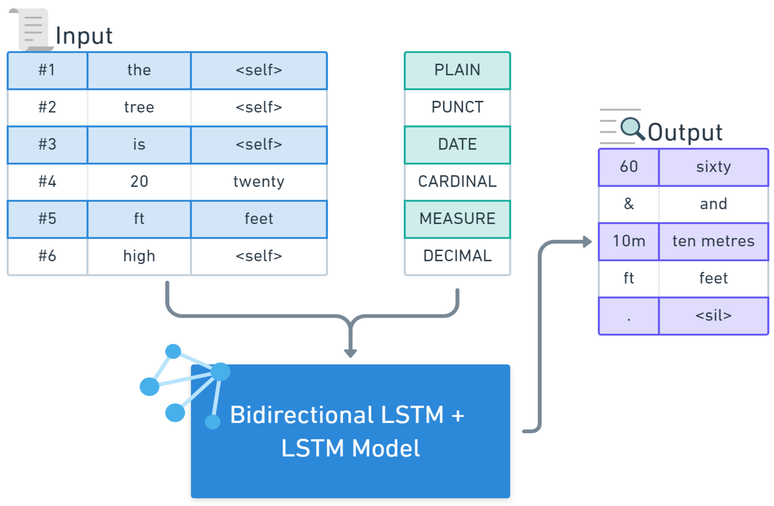
conTEXT – Normalization using LSTM

Simon Fraser University

Siddhant Singhal, Raghib Musarrat, Rahul Rajeev, Milbir Guram

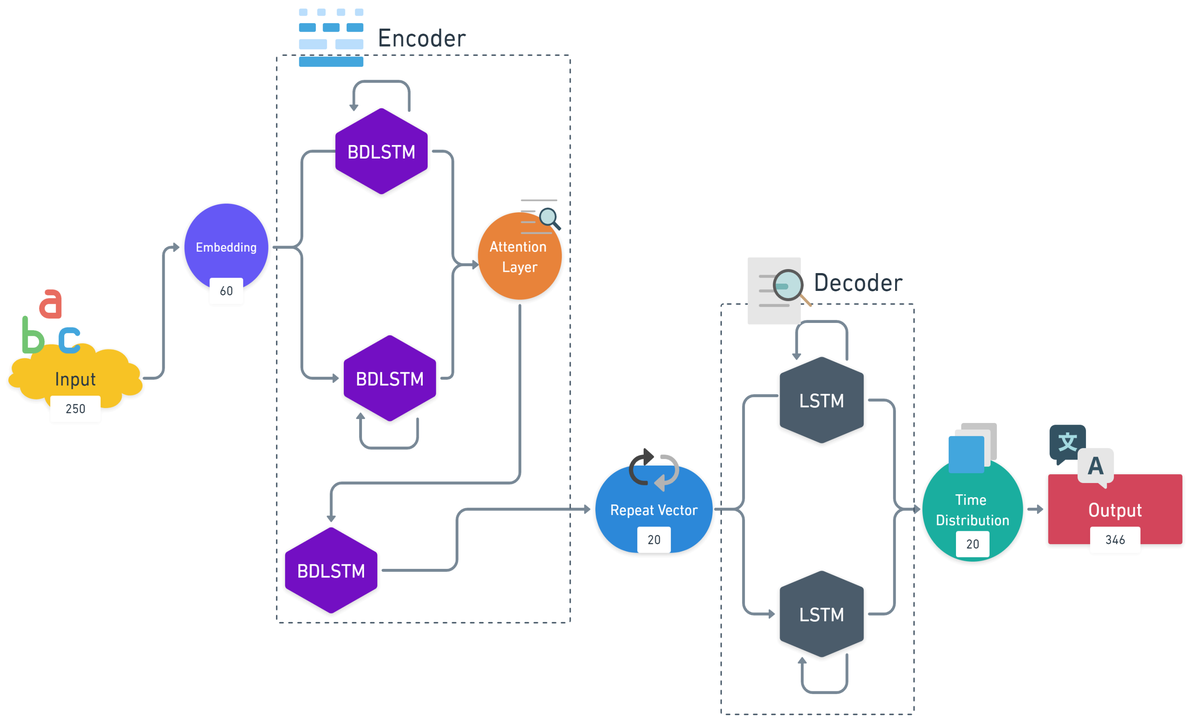
PROBLEM

The task of text normalization is performed in the following manner:
– **division** of the sentence input to the tokens and **inclusion** of the tokens to different entities
– understanding the **context** through the encoder layer and showing the relevant **representation** during the decoder layer



Contribution: Implementation of the LSTM with 3 bidirectional layers in the encoder + 2 LSTM layers for the decoder.
Comparative analysis of the accuracy obtained with our approach to the **Sproat et al. arXiv 2016** & **Kestrel TTS 2014**

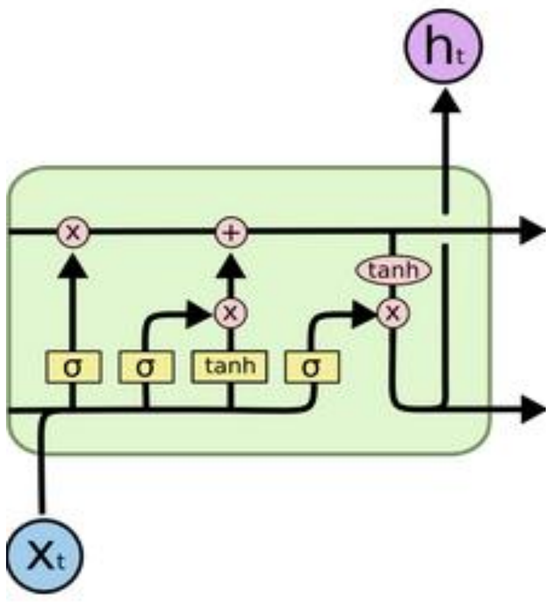
MODEL FORMULATION



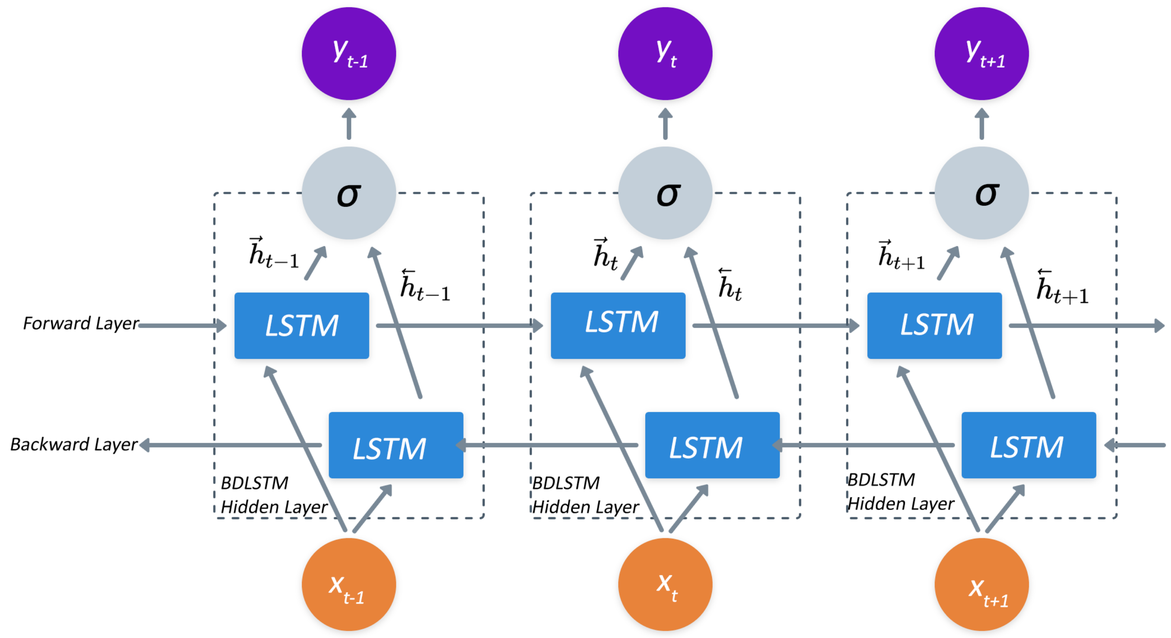
Input takes into consideration 250 distinct character values, the **output** results after normalization results in 346 different words.
In order to create **fixed-length sequences,** the input padding sequence and output padding sequence to the max. length is 60 and 20 units resp.

BIDIRECTIONAL LSTM

The difference between LSTM and bi-directional LSTM is that not only does it take past inputs(x(t-1)), but also takes into account future inputs(x(t+1)). This helps to understand the context of the input(x(t)) at time(t):



$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right)$$
$$h_t = o_t * \tanh\left(C_t\right)$$
$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$



The bi-directional LSTM has 2 layers - **'forward'** and **'backward'** which lead to the additional constraints repeated twice that makes it further difficult for implementation.
Here, the forward layer output sequence **h(right)** is calculated using input sequence from time **T-1 to T-n**, whereas the backward layer output sequence **h(left)** is calculated in the opposite manner, i.e. using input sequence from **T-n to T-1.** Therefor, **y(t) = sigmoid[h(right), h(left)]**

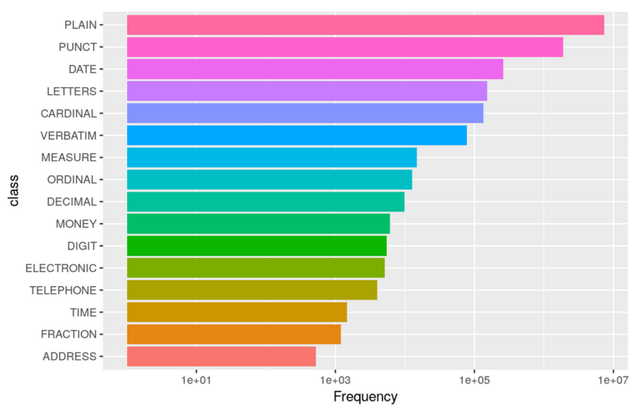
EXPERIMENTS

Dataset: Google Dataset – Text Normalization **Paper:** RNN approaches to Text Normalization

Comparative analysis of the normalization done through various models:

	Sproact et al.	conTEXT	Kestrel TTS	XGBoost
Accuracy	99.6	**98.2**	91.3	97.4
C.D.	–	**1.366**	8.264	1.977

EDA of the dataset and Qualitative example of the result obtained on the model -



	sentence_id	token_id	before	after
65	4	1	2	two
81	5	5	3,400	thousand thousand five hundred
85	5	9	10,200 ft	hundred hundred million dollars
92	5	16	7,000	two thousand
100	6	5	1895	eighteen ninety five
101	6	6	-	to
102	6	7	1945	nineteen forty seven
106	6	11	W.	w
113	7	0	Pgs	p g