

Investigating Numeracy of a Text-to-Text Transfer Model

EMNLP 2021

Kuntal Kumar Pal, Chitta Baral

Introduction

- Current transformer-based LM shows SOTA performance on almost all of the conventional NLP tasks.
- But do they understand texts where numbers are the integral part and can change the meaning of the text ?
- Are they able to
 - Understand numbers in multiple surface forms
 - Understand its values based on its context.
 - Compare their values with others
 - Rearrange them based on their values.

Goal

- Here we test numeracy of a text-to-text transfer transformer (T5) model with four basic tasks.
 - Numeration, Magnitude order prediction, List min-max, List sorting
- We test three versions of T5 models
 - T5-SM (Small), T5-BS (Base), T5-LG (Large)
- Experimental Settings
 - Interpolation
 - Training on a certain number range and testing on same number range
 - Extrapolation
 - Training on a certain number range and testing on different number ranges

Numeracy Tests

What is three thousand four hundred seventy three ?	3473 (FULL) 3 4 7 3 (SPLIT)
Hain Celestial Q1 adjusted earnings per share \$ <MASK>	0.40 (0)
Shop Black Friday in Orange County, where sales begin Nov <MASK>	28 (2)
Which is minimum in value among 49 3 55 51 25 62 45 4 42 39 ?	3
Which is maximum in value among 49 3 55 51 25 62 45 4 42 39 ?	62
Sort in ascending order : 49 3 55 51 25 62 45 4 42 39	3 4 25 39 42 45 49 51 55 62
Sort in descending order : 49 3 55 51 25 62 45 4 42 39	62 55 51 49 45 42 39 25 4 3

 NUMERATION  MAGNITUDE  MIN-MAX  SORTING

Results - Numeration

- Data : Templates
- Number Representation matters
 - Split representation better than Full
- Fewer shot setting
 - Large model maintains performance better than base or small models
- Extrapolation setting
 - None of the model performs well

# TRAIN →		4.9K		1.3K		0.9K	
TP	Model	IN	EX	IN	EX	IN	EX
FL	T5-SM	45.31	0.08	1.90	0.01	0.33	0.00
	T5-BS	92.16	1.03	66.47	0.45	37.20	0.42
	T5-LG	98.06	1.91	89.49	1.96	79.48	1.58
SP	T5-SM	69.67	39.35	26.89	1.10	0.23	0.01
	T5-BS	99.50	11.31	81.21	22.44	73.61	31.06
	T5-LG	100.00	10.05	99.97	7.35	91.59	12.92

Table 1: **Numeration** EM scores w/ split (SP) and w/o split (FL) representation on 4.9K, 1.3K, 0.9K train-data in Interpolation (IN) and Extrapolation (EX) settings.

What is three thousand four hundred seventy three ?

3473 (FULL)
3 4 7 3 (SPLIT)

Results - Magnitude Order Prediction

- Data : Numeracy600K
 - Article Titles (AT)
 - Stock Market Comments (MC)
- Interpolation performance 5% better than previous SOTA
- Extrapolation performance is quite good
 - 25% better than prior SOTA

Datasets →	AT		MC	
	μ F1	mF1	μ F1	mF1
LR	62.49	30.81	71.25	60.80
CNN	69.27	35.96	77.17	58.49
GRU	70.92	38.43	78.25	58.08
BiGRU	71.49	39.94	<u>80.16</u>	62.74
CRNN	69.50	36.15	78.00	<u>64.62</u>
CNN-capsule	63.11	29.41	75.89	59.22
GRU-capsule	70.73	33.57	77.36	64.71
BiGRU-capsule	71.49	34.18	77.97	64.34
BiLSTM-DICE	75.56	46.80	-	-
T5-SM	69.87	31.36	66.11	34.68
T5-BS	<u>78.06</u>	40.04	72.22	47.44
T5-LG	81.40	<u>44.64</u>	80.29	59.16

Table 3: **Magnitude Order Prediction** for Market Comments (MC) and Article Titles (AT) datasets of numeracy600K in micro-F1 (μ F1) and macro-F1 (mF1). Best score is in bold and second-best is underlined.

Train on →	AT		MC	
	μ F1	mF1	μ F1	mF1
BiGRU	25.59	10.58	31.38	11.08
T5-SM	28.88	12.04	37.35	10.81
T5-BS	35.53	14.48	31.51	12.25
T5-LG	50.18	21.24	38.43	12.32

Table 4: **Cross Domain** (Extrapolation) Tests of Order Prediction. Train on MC, test on AT and vice-versa.

Hain Celestial Q1 adjusted earnings per share \$ <MASK>	0.40 (0)
Shop Black Friday in Orange County, where sales begin Nov <MASK>	28 (2)

Results - List Min-Max

- Data : Templates
- Interpolation
 - T5-BS and T5-LG perform over 80%
 - T5-SM performance decreases with increase in range and series length.
- Extrapolation
 - For small series lengths performance increases with increase in variations of numbers
 - With increase in series length the performance decreases considerably

		LIST MINIMUM						LIST MAXIMUM					
# ELEMENTS		3		5		10		3		5		10	
Range	Model	IN	EX	IN	EX	IN	EX	IN	EX	IN	EX	IN	EX
< 99	T5-SM	90.5	0.6	86.5	0.1	65.9	0.0	80.4	0.5	71.6	0.3	74.7	0.1
	T5-BS	96.2	33.9	99.1	13.0	98.2	2.8	92.3	22.7	96.8	6.0	90.4	1.1
	T5-LG	100.0	22.2	99.4	2.8	100.0	0.5	100.0	29.6	100.0	13.6	100.0	2.0
< 999	T5-SM	72.6	41.8	55.5	22.2	49.9	9.7	65.3	38.4	54.8	17.5	40.0	5.2
	T5-BS	91.5	67.2	92.1	42.6	80.4	27.1	89.1	65.3	90.8	47.2	88.3	25.0
	T5-LG	98.3	70.1	96.1	49.3	87.4	34.7	96.1	61.2	97.8	58.7	95.2	35.3
< 9999	T5-SM	59.1	44.7	43.5	30.4	30.7	17.1	51.2	47.0	36.0	27.0	20.9	11.1
	T5-BS	89.6	68.8	86.9	53.8	85.4	38.1	87.1	58.6	83.1	43.4	81.6	29.9
	T5-LG	97.1	81.3	93.7	71.8	94.0	58.2	96.2	84.9	94.9	76.4	94.9	59.1

Which is minimum in value among 49 3 55 51 25 62 45 4 42 39 ?

3

Which is maximum in value among 49 3 55 51 25 62 45 4 42 39 ?

62

Results - List Sorting

- Data : Templates
- T5-SM - No correct extrapolation result for series lengths 10 and <10% success in length 5
- Increase in sequence length :
 - Performance degrades considerably both in interpolation and extrapolation settings
- Increase in range :
 - T5-LG extrapolation performance increases

# ELEMENTS		LIST-SORT ASCENDING						LIST-SORT DESCENDING					
		3		5		10		3		5		10	
Range	Model	IN	EX	IN	EX	IN	EX	IN	EX	IN	EX	IN	EX
< 99	T5-SM	54.0	12.4	7.6	0.0	0.0	0.0	56.0	12.6	5.9	0.4	0.0	0.0
	T5-BS	80.6	12.2	87.2	0.0	0.4	0.0	84.3	12.9	75.5	0.0	6.2	0.0
	T5-LG	100.0	5.8	99.9	0.0	69.7	0.1	100.0	13.1	96.6	0.1	57.6	0.1
< 999	T5-SM	32.6	15.1	1.4	0.6	0.0	0.0	38.0	22.3	3.4	1.3	0.0	0.0
	T5-BS	74.7	45.7	64.0	8.0	12.5	0.0	73.1	42.0	62.6	9.6	16.8	0.1
	T5-LG	95.1	64.2	91.8	16.8	61.9	1.7	94.7	63.5	92.5	25.7	61.2	1.6
< 9999	T5-SM	23.4	17.1	1.0	0.1	0.0	0.0	30.4	21.2	0.7	0.4	0.0	0.0
	T5-BS	63.1	45.5	51.1	12.7	15.0	0.2	59.8	43.9	51.4	12.4	14.3	0.3
	T5-LG	94.5	76.0	87.4	43.2	74.6	12.6	94.2	76.1	86.1	44.4	75.6	11.9

Sort in ascending order : 49 3 55 51 25 62 45 4 42 39

3 4 25 39 42 45 49 51 55 62

Sort in descending order : 49 3 55 51 25 62 45 4 42 39

62 55 51 49 45 42 39 25 4 3

Error Analysis

What is five hundred thousand three ?

Label : 500003 **Predicted** : 5003

Numeration : Missing keyword hundred

Nonprofit Homefront America Receives \$ < MASK > from Walmart Foundation

Label : 10000 (5) **Predicted** : 100000 (6)

Magnitude Order Prediction : Predicted an extra zero

Which is minimum in value among 92473 52823 52746 68801 69389 54929

96584 81316 57345 92317 ?

Label : 52746 **Predicted** : 52723

List-Min-max : Found the second minimum element

Sort in descending order : 594 598 632 600 633 630 560 574 634 599

Label : 634 633 632 630 600 599 598 594 574 560

Predicted : 634 633 632 630 599 598 594 574 560 600

List-Min-max : Missed order of one element 600

Conclusion

- All the models show promise in learning numeracy in interpolation setting.
- The smaller models parameter space limits its learning ability in all four tasks
- Overall, none of the models were able to learn well on extrapolation samples showing that the inherent rules of numeracy is difficult to learn
- We also find more variation in numbers (increasing number ranges) leads to better performance in extrapolation settings

Thank You !!!

Contacts : Kuntal Kumar Pal, Chitta Baral

Paper : <https://arxiv.org/pdf/2109.04672.pdf>

Code : <https://github.com/kuntalkumarpal/T5Numeracy>

Email : {kkpal, chitta}@asu.edu