Math Word Problem Generation with Multilingual Language Models

Problem

"How to *generate Math Word Problems (MWPs) in Sinhala, Tamil and English languages as an autoregressive manner*"

Background

- A Mathematical word problem (MWP) is a mathematical problem expressed in natural language and it requires problem-solving ability.
- It is a narrative with a specific topic that provides clues to the correct equation with numerical quantities and variables therein

Sinhala: බිල්ට බෝල *9* ක් ඇති අතර ජිම්ට බිල්ට වඩා *7* ක් අඩුවෙන් බෝල ඇත. ජිම්ට බෝල කොපමණ ප්‍රමාණයක් තිබේද*?*
English: *The sum of two numbers is 18 and their difference is 4, what are the two numbers?*

Motivation

- There is only a limited amount of research done for multilingual MWP generation.
- Existing research mainly focused on language-dependent techniques that rely on man-made templates for MWP question generation.
- Recent NLG research has shown very promising results with the use of pre-trained deep contextual embedding models s.a. GPT.

Generating MWPs is challenging because,

- Algebraic questions are normally deep and logical.
  Eg: *"The sum of two numbers is 78. if four times the smaller number is subtracted from the larger number. the result is 13. what are the two numbers"*
- Most of the MWPs contain numerical constraints and units.

Methodology

Model Selection

We choose the model variants that have roughly the same configurations. According to Huggingface, **GPT2-Medium**, **T5-base** and **BART-large** variants have approximately 300M model parameters. Therefore these were used for further experiments. For multilingual MWP generation, we selected **mT5** and **mBART** models.

Dataset

- Two types of MWPs - Simple MWPs and Algebraic MWPs[1].
- Extended this dataset using the Dolphin18K dataset and the allArith dataset
- Selected questions are similar or slightly higher in complexity compared to baseline[1]
- The extended dataset contains 4210 Algebraic MWPs and 3160 simple MWPs.

Simple MWP
Kamal found 7 balls but 4 were broken. How many unbroken balls did Kamal find?
Algebraic MWP
The sum of two numbers is twenty-three, and the larger number is five more than the smaller number. Find these numbers

Evaluation Matrices

- BLEU score
- ROUGE score
- Human Evaluation

- All the models considered in this research are trained using the Transformer architecture
- GPT, BART and T5 are pre-trained with English data & mBART and mT5 are pre-trained with data from multiple languages.
- GPT models are decoder based & T5, BART, mBART and mT5 are encode-decoder models.
- We follow the standard training procedure of GPT-2 model in training it & used the conditional generator option of BART/mBART and T5/mT5.
- In both these cases, the models generate the rest of the MWP for a given seed.

Our Experiments Show

- How the performance of mT5 and mBART varies depending on the language
- How the performance of the models varies depending on the amount of fine-tuning data
- How much information (size of the seed) should be provided to the model at the inference stage for it to generate a meaningful MWP
- How the context of an MWP affects the generation performance

Human Evaluation

- Measured the quality of the created dataset
- Measured the quality and correction of generated questions.
- We identified the types of errors in MWP generation.

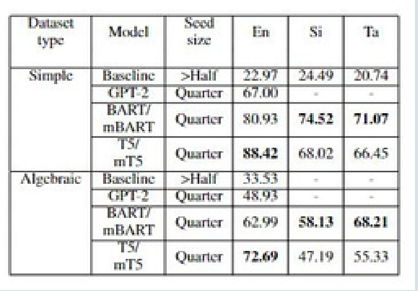
Contributions

- We created a benchmark dataset by extending the dataset created by Liyanage and Ranathunga (2020)[1] for MWP generation
- First to conduct an empirical analysis on the use of GPT, BART, T5, mBART, and mT5 for auto-regressive generation of MWPs and mBART and mT5 for the general task of auto-regressive text generation considering low-resource languages.

Result & Evaluation

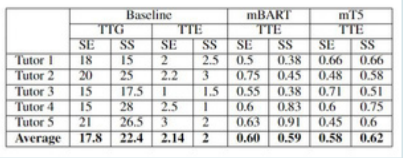
1. Pre-trained models vs Baseline
2. Effect of Fine-tuning Dataset Size
3. Effect of Pre-training Dataset Size
4. Effect of the Context of MWPs
5. Zero-shot MWP Generation
6. Effect of Seed Length
7. Human Evaluation

Baseline MWP results

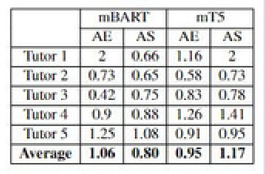


Dataset type	Model	Seed size	En	Si	Ta
Simple	Baseline	>Half	22.97	24.49	20.74
	GPT-2	Quarter	67.00	-	-
	BART/ mBART	Quarter	80.93	**74.52**	**71.07**
	T5/ mT5	Quarter	**88.42**	68.02	66.45
Algebraic	Baseline	>Half	33.53	-	-
	GPT-2	Quarter	48.93	-	-
	BART/ mBART	Quarter	62.99	**58.13**	**68.21**
	T5/ mT5	Quarter	**72.69**	47.19	55.33

Human Evaluation

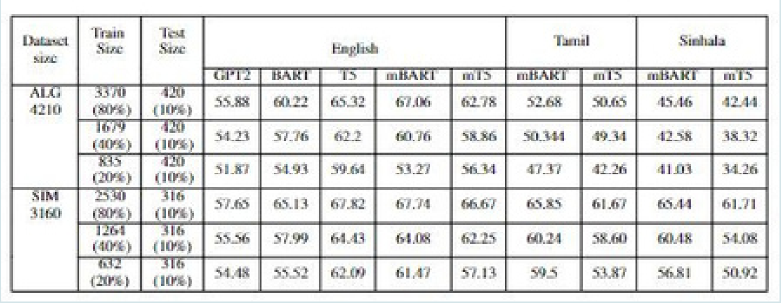


	Baseline				mBART		mT5	
	TTG		TTE		TTE		TTE	
	SE	SS	SE	SS	SE	SS	SE	SS
Tutor 1	18	15	2	2.5	0.5	0.38	0.66	0.66
Tutor 2	20	25	2.2	3	0.75	0.45	0.48	0.58
Tutor 3	15	17.5	1	1.5	0.55	0.38	0.71	0.51
Tutor 4	15	28	2.5	1	0.6	0.83	0.6	0.75
Tutor 5	21	26.5	3	2	0.63	0.91	0.45	0.6
Average	**17.8**	**22.4**	**2.14**	**2**	**0.60**	**0.59**	**0.58**	**0.62**



	mBART		mT5	
	AE	AS	AE	AS
Tutor 1	2	0.66	1.16	2
Tutor 2	0.73	0.65	0.58	0.73
Tutor 3	0.42	0.75	0.83	0.78
Tutor 4	0.9	0.88	1.26	1.41
Tutor 5	1.25	1.08	0.91	0.95
Average	**1.06**	**0.80**	**0.95**	**1.17**

Monolingual Results variation with train size



Dataset size	Train Size	Test Size	English					Tamil		Sinhala	
			GPT2	BART	T5	mBART	mT5	mBART	mT5	mBART	mT5
ALG 4210	3370 (80%)	420 (10%)	55.88	60.22	65.32	67.06	62.78	52.68	50.65	45.46	42.44
	1679 (40%)	420 (10%)	54.23	57.76	62.2	60.76	58.86	50.344	49.34	42.58	38.32
	835 (20%)	420 (10%)	51.87	54.93	59.64	53.27	56.34	47.37	42.26	41.03	34.26
SIM 3160	2530 (80%)	316 (10%)	57.65	65.13	67.82	67.74	66.67	65.85	61.67	65.44	61.71
	1264 (40%)	316 (10%)	55.56	57.99	64.43	64.08	62.25	60.24	58.60	60.48	54.08
	632 (20%)	316 (10%)	54.48	55.52	62.09	61.47	57.13	59.5	53.87	56.81	50.92

Examples

- T5 Generation

½ seed (input) - Twice the larger number is 3 more than five times the smaller number, the sum
Generated MWP - Twice the larger number is 3 more than five times the smaller number, the sum of the 2 numbers is 7 less than 3 times the larger number, Find the smaller number.

¼ seed (input) -Twice the larger number is 3 more than
Generated MWP - Twice the larger number is 3 more than five times the smaller number. The sum of the 2 numbers is 27. What are the two numbers?

- mBART Generation

¼ seed (input) - 150 ක
Generated questions- 150 ක එකතුවක් ඇති අඛණ්ඩ සංඛ්‍යා තුනක් සොයා ගන්න.

½ seed (input) - 150 ක එකතුවක් ඇති අඛණ්ඩ
Generated questions - 150 ක එකතුවක් ඇති අඛණ්ඩ ඉරට්ටේ සංඛ් යා දෙක කුමක්ද?

Conclusion

We evaluated several multilingual and monolingual pre-trained models for the task of MWP generation considering four factors - the amount of language-specific pre-trained data, amount of fine-tuning data, length of the seed, and type of the MWP. We also presented a multi-way parallel dataset for MWP evaluation, which includes two languages underrepresented in these pre-trained models

References

[1] Liyanage, V. and Ranathunga, S. Multi-lingual mathematical word problem generation using long short term memory networks with enhanced input features. In Proceedings of The 12th Language Resources and Evaluation Conference, pages 4709–4716, 2020

Authors

Kashyapa Niyarepola Dineth Athapaththu Savindu Ekanayake Surangika Ranathunga

Department of Computer Science and Engineering
University of Moratuwa
Sri Lanka