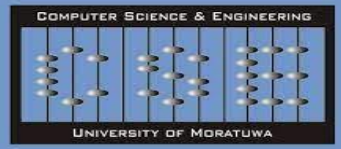

INLG 2022

# Math Word Problem Generation with Multilingual Language Models

**Kashyapa Niyarepola** **Savindu Ekanayaka** **Dineth Athapaththu** **Surangika Ranathunga**

Department of Computer Science and Engineering
University of Moratuwa, Sri Lanka



# Math Word Problems (MWPs)

- A Mathematical word problem (MWP) is a mathematical problem expressed in natural language and it requires problem solving ability.

- MWP should provide clues to the correct equation with numerical quantities and variables.

  e.g. The sum of two numbers is 18 and their difference is 4, what are the two numbers?

# Why MWP?

- Algebraic MWPs are a major component in elementary Mathematics, and provide the problem-solving ability that is essential for every student.
- However, most students face difficulties in solving MWPs related to algebra[9].
- The most effective way to overcome this problem is to provide the students with lots of algebraic problems and simple multiple MWP to work on.
- MWPs Generation is a time-consuming task for teachers and tutors.

# Research Objectives

- Evaluate the performance of monolingual and multilingual pre-trained language models for auto-regressive MWP generation under different constraints
  - ❏ The amount of language specific pre-trained data
  - ❏ Amount of fine-tuning data
  - ❏ Length of the seed
  - ❏ Type of the MWP

- Create a multi-parallel dataset (English, Tamil, Sinhala) of MWPs

# Previous Work

- Question rewriting[13]
- Template-based generation[5]
- Text generation with Neural Networks[14]
    - **Multilingual elementary level MWPs generation using character-level LSTMs[1]**

# Dataset

| Dataset | Size | Languages | Example |
|---|---|---|---|
| Simple MWPs[2] | 3160 | English, Sinhala and Tamil | Kamal has 16 marbles and Nimal has 12 less marbles than Kamal. How many marbles does Nimal have? |
| Algebraic MWPs[2] | 4210 | English, Sinhala and Tamil | The sum of two numbers is 38. their difference is 12. what are the two numbers? |

# Dataset Evaluation

| Dataset type | Avg. Num. of words per question | Avg. Num. of characters per question |
|---|---|---|
| English Simple (ES) | 15 | 54 |
| English Algebraic (EA) | 14 | 62 |
| Sinhala Simple (SS) | 19 | 61 |
| Sinhala Algebraic (SA) | 17 | 59 |
| Tamil Simple (TS) | 13 | 49 |
| Tamil Algebraic (TA) | 16 | 57 |

- 200 translated MWPs along with the original English questions were selected for the evaluation.
- Rated the translated version with respect to adequacy and fluency

| Data set | Rank | | | | | |
|---|---|---|---|---|---|---|
| | 0-10 | 11-29 | 30-50 | 51-69 | 70-90 | 91-100 |
| SS | 0% | 1.6% | 3% | 6.3% | 22.6% | 66% |
| SA | 0% | 0% | 0.3% | 2.6% | 12.6% | 84.3% |
| TS | 0% | 1% | 4% | 8.3% | 27.6% | 59% |
| TA | 7% | 12% | 6.3% | 6% | 11.3% | 57% |

# Methodology

| Tasks | Models | Description |
|---|---|---|
| Pre-trained models vs Baseline[2] | GPT2 , T5, BART, mT5 and mBART | Determined whether fine-tuning the pre-trained models is better than the selected RNN baseline[2]. |
| Effect of Fine-tuning Dataset Size | GPT2 , T5, BART, mT5 and mBART | Analyzed how the quality of the results varies with different fine-tuning dataset sizes. |

**Models were fine-tuned for 20 epochs with 16 batch size and 1e-4 learning rate.**

# Methodology

| Tasks | Models | Description |
| --- | --- | --- |
| Effect of Pre-training Dataset Size | GPT2 , T5, BART, mT5 and mBART | Observed the results and analyzed with pre-trained dataset size from each models. |
| Effect of the Context of MWPs | mT5 and mBART | Trained the models with One dataset type, and tested with the other. |

**Models were fine-tuned for 20 epochs with 16 batch size and 1e-4 learning rate.**

# Methodology

| Tasks | Models | Description |
| --- | --- | --- |
| Zero-shot MWP Generation | mT5 and mBART | Analyzed how small amounts of fine-tuning data, affect for text generation. |
| Effect of Seed Length | mT5 and mBART | Analyzed how the quality of the results varies with different fine-tuning dataset sizes. |

**Models were fine-tuned for 20 epochs with 16 batch size and 1e-4 learning rate.**

# Methodology

| Tasks | Models | Description |
| --- | --- | --- |
| Human Evaluation | - | Analyzed the types of errors in MWP generation and Identified the actual utility of the generated questions<br><br>whether it is more effective for a tutor to correct a generated question, rather than generating a question from scratch |

# Results & Evaluation

# Pre-trained models vs Baseline

| 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 |

All our models outperform the baseline by a significant margin (even when using just the quarter seed)

# Effect of Fine-tuning Dataset 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 |

# Effect of Fine-tuning Dataset Size - Discussion

The performance of all the models drop when the fine-tuning dataset size drops

**English text generation,**

- Both sequence-to-sequence models outperform GPT-2.
- T5 outperforms BART.
- mBART and mT5 lag behind their monolingual counterparts.

**For multilingual models,**

- mBART outperforms mT5 in all the cases except for one case

# Effect of Pre-training Dataset Size

| Model | | English | Tamil | Sinhala |
|---|---|---|---|---|
| BART | Storage(GB) | 160 | - | - |
| T5 | Storage(GB) | 700 | - | - |
| mT5 | Token(B) | 2733 | 3.4 | 0.8 |
| | Pages(M) | 3,067 | 3.5 | 0.5 |
| mBART | Token(B) | 55.61 | 0.595 | 0.243 |
| | Storage(GiB) | 300.8 | 12.2 | 3.6 |

- Always English has the highest result, followed by Tamil, and then Sinhala.
- This could be due to the amount of pre-trained data in the models

# Effect of the Context of MWPs

| Train ID | Train Size | Test ID | Test Size | mBART | mT5 |
|---|---|---|---|---|---|
| SA | 1679 (40%) | SS | 1580 (50%) | 32.39 | 29.23 |
| SS | 1264 (40%) | SA | 2088 (50%) | 27.01 | 17.87 |
| TA | 1679 (40%) | TS | 1580 (50%) | 35.27 | 33.44 |
| TS | 1264 (40%) | TA | 2088 (50%) | 32.12 | 27.75 |

Substantial drop in the results, when the models are fine-tuned with the other dataset

# Zero-shot MWP Generation

| Test Dataset | Train Size | Test Size | mBART | mT5 |
|---|---|---|---|---|
| ES | 0 | 986 | 5.96 | 0.05 |
| EA | 0 | 1175 | 8.50 | 0.42 |
| SS | 0 | 986 | 6.37 | 0.01 |
| SA | 0 | 1175 | 7.50 | 0.03 |
| TS | 0 | 986 | 4.57 | 0.02 |
| TA | 0 | 1175 | 6.54 | 0.03 |
| ES | 100 | 986 | 23.24 | 4.30 |
| EA | 100 | 1175 | 34.50 | 3.93 |
| SS | 100 | 986 | 52.72 | 5.42 |
| SA | 100 | 1175 | 18.21 | 2.36 |
| TS | 100 | 986 | 48.86 | 2.87 |
| TA | 100 | 1175 | 39.95 | 0.60 |

- All the models fail on zero-shot text generation.
- Generated sentences are not questions but more like stories.
- The performance increases by a significant margin, When fine-tuned with just 100 data samples.
- A decent result can be expected,Even with a very small training dataset.

# Effect of Seed Length

| Seed size | SS | TS | SA | TA |
|---|---|---|---|---|
| 10% | 48.9 | 45.48 | 30.19 | 36.77 |
| 20% | 58.25 | 57.74 | 39.91 | 45.82 |
| 30% | 65.47 | 65.02 | 47.38 | 54.21 |
| 40% | 71.51 | 72.39 | 53.85 | 62.5 |

The quality of the generated text goes up when the seed length increases. Even 10% of the seed is enough to provide an acceptable result.

# Human Evaluation

| | | | | | mBART | | mT5 | |
|---|---|---|---|---|---|---|---|---|
| | TTG | | TTE | | TTE | | TTE | |
| | SE | SS | SE | SS | SE | SS | SE | SS |
| T1 | 18 | 15 | 2 | 2.5 | 0.5 | 0.38 | 0.66 | 0.66 |
| T2 | 20 | 25 | 2.2 | 3 | 0.75 | 0.45 | 0.48 | 0.58 |
| T3 | 15 | 17.5 | 1 | 1.5 | 0.55 | 0.38 | 0.71 | 0.51 |
| T4 | 15 | 28 | 2.5 | 1 | 0.6 | 0.83 | 0.6 | 0.75 |
| T5 | 21 | 26.5 | 3 | 2 | 0.63 | 0.91 | 0.45 | 0.6 |
| **Av** | **17.8** | **22.4** | **2.14** | **2** | **0.60** | **0.59** | **0.58** | **0.62** |

| Errors% | mBART | | | | mT5 | | | |
|---|---|---|---|---|---|---|---|---|
| | SE | AE | SS | AS | SE | AE | SS | AS |
| Co-reference | 4 | 4 | 6 | 4 | 8 | 2 | 6 | 2 |
| Unit | 4 | 1 | 1 | 1 | 2 | 1 | 1 | 1 |
| Spelling | 0 | 0 | 4 | 2 | 2 | 0 | 0 | 2 |
| Grammar | 16 | 12 | 16 | 10 | 8 | 10 | 14 | 10 |
| math constraint % | 12 | 38 | 22 | 30 | 14 | 22 | 24 | 32 |

# Generated Questions - Algebraic

| Model | Generated MWPs |
| --- | --- |
| Reference | The sum of two numbers is 56, their difference is 22,Find the larger number. |
| Baseline | The sum of two numbers is 12. their differenct are the two consecutive integers if the sum of the second integers is 10. |

# Generated Questions - Algebraic

| Model | Generated MWPs |
| --- | --- |
| GPT2 | The sum of two numbers is 76, the second is 8 more than 3 times first, what are these 2 numbers? |
| BART | The sum of two numbers is 60. three times the smaller number minus twice the larger number is 56. Find the larger number. |
| T5 | The sum of two numbers is 91. the larger number is 1 more than 4 times the smaller number. Find the numbers? |

# Generated Questions - Algebraic

| Model | Generated MWPs |
| --- | --- |
| mBART (Sinhala) | 150 ක එකතුවක් ඇති අඛණ්ඩ සංඛ් යා තුනක් සොයා ගන්න. |
| mBART (Tamil) | சேஹனும் சரித்தும் ஒரு தொழிற்சாலை வைத்திருக்கிறார்கள். செஹான் 600 பங்குகளையும், நிமல் 1600 பங்குகளையும் வாங்கினர். நிமலை விட கமல் எத்தனை குறைவான பங்குகளை வாங்கினார்? |

# Conclusion

- 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
    - type of the MWP

- Presented a multi-way parallel dataset for MWP evaluation, which includes two languages underrepresented in these pre-trained models

# THANK YOU!

# References

[1] Vijini Liyanage and Surangika Ranathunga. “A multi-language platform for generating algebraic mathematical word problems”. In: 2019 14th Conference on Industrial and Information Systems (ICIIS). IEEE. 2019, pp. 332–337.

[2] Vijini Liyanage and Surangika Ranathunga. “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. 2020, pp. 4709–4716.

[3] Alec Radford et al. “Language models are unsupervised multitask learners”. In: OpenAI blog 1.8 (2019), p. 9.

[4] Colin Raffel et al. “Exploring the limits of transfer learning with a unified text-to-text transformer”. In: arXiv preprint arXiv:1910.10683 (2019).

# References cont.

[5] Qingyu Zhou and Danqing Huang. “Towards generating math word problems from equations and topics”. In: Proceedings of the 12th International Conference on Natural Language Generation. 2019, pp. 494–503.

[6] Kazemi, Ashkan, et al. "Extractive and Abstractive Explanations for Fact-Checking and Evaluation of News." arXiv preprint arXiv:2104.12918 (2021).

[7] Guan, Jian, et al. "A knowledge-enhanced pretraining model for common sense story generation." Transactions of the Association for Computational Linguistics 8 (2020): 93-108.

[8] Hu, Jinyi, and Maosong Sun. "Generating Major Types of Chinese Classical Poetry in a Uniformed Framework." arXiv preprint arXiv:2003.11528 (2020).

# References cont.

[9] M. J. Hosseini, H. Hajishirzi, O. Etzioni, and N. Kushman, Learning to Solve Arithmetic Word Problems with Verb Categorization, Proc. 2014 Conference. Empirical. Methods Natural. Language. Processing., pp. 523-533, 2014.

[10] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.arXiv preprint arXiv:1910.13461

[11] Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. arXiv preprint arXiv:2008.00401

# References cont.

[12] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale,Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. arXiv preprint arXiv:2010.11934.

[13] Rik Koncel-Kedziorski, Ioannis Konstas, Luke Zettle moyer, and Hannaneh Hajishirzi. 2016. A theme rewriting approach for generating algebra word problems. arXiv preprint arXiv:1610.06210.

[14] Qingyu Zhou and Danqing Huang. 2019. Towards generating math word problems from equations and topics. In Proceedings of the 12th International Conference on Natural Language Generation, pages 494503.