INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

Practice Makes a Solver Perfect: Data Augmentation for Math Word Problem Solvers

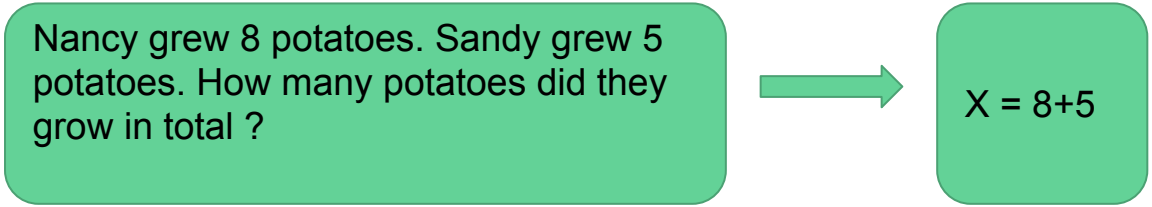
Authors: Vivek Kumar, Rishabh Maheshwary, Vikram Pudi
Affiliation: IIIT Hyderabad, India

What are Math Word Problems ?

- Math word problems are arithmetic problems involving text and numbers in combination to identify unknown values.
- Each word problem contains a textual narrative of a sequence of events involving numerical transactions followed by a question text to identify some unknown values.
- Each of the problems are similar to grade 6 - grade 8 level involving equations in one variable.

What are Math Word Problems ?

- Math word problems are arithmetic problems involving text and numbers in combination to identify unknown values.
- Each word problem contains a textual narrative of a sequence of events involving numerical transactions followed by a question text to identify some unknown values.
- Each of the problems are similar to grade 6 - grade 8 level involving equations in one variable.



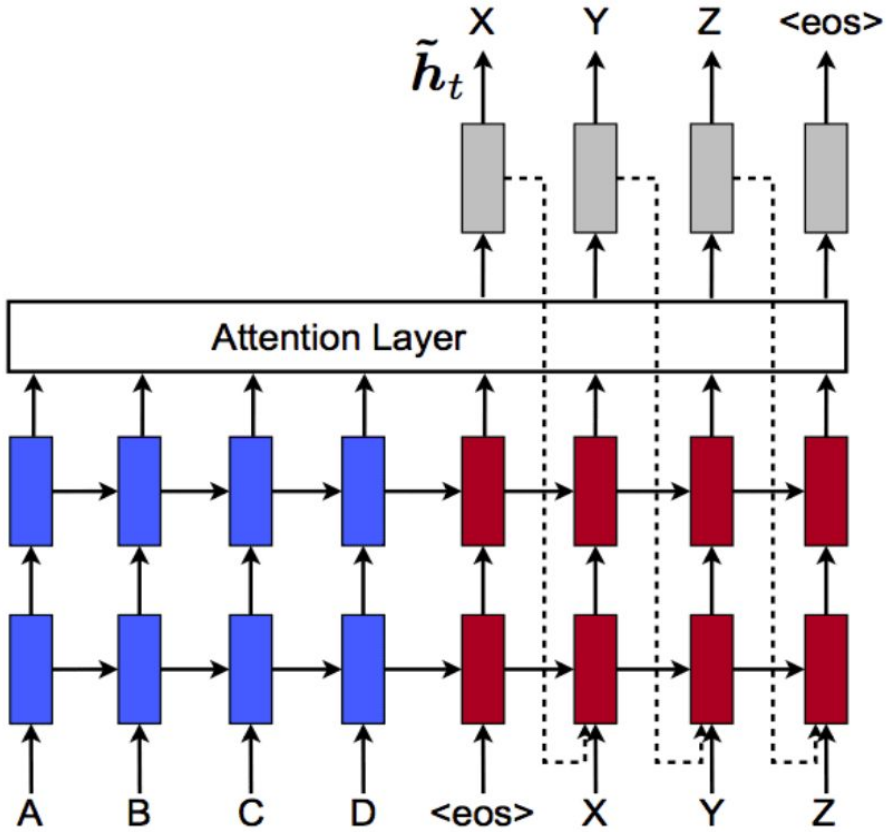
Evaluation Setup

Models:

- Seq2Seq with Attention
- GTS
- Graph2Tree

Datasets:

- MaWPS
- ASDiv-A
- SVAMP



INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

Evaluation Setup

Models:

- Seq2Seq with Attention
- GTS
- Graph2Tree

Datasets:

- MaWPS
- ASDiv-A
- SVAMP

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

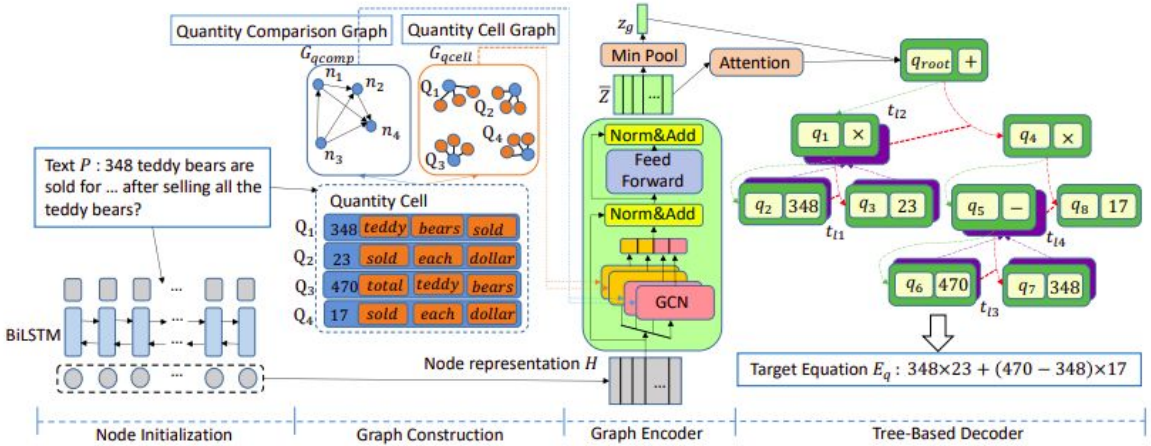
Evaluation Setup

Models:

- Seq2Seq with Attention
- GTS
- Graph2Tree

Datasets:

- MaWPS
- ASDiv-A
- SVAMP



Evaluation Setup

Models:

- Seq2Seq with Attention
- GTS
- Graph2Tree

Datasets:

- MaWPS
- ASDiv-A
- SVAMP

- MaWPS contains 2,373 English math word problems.
- It is one of the largest available english language dataset for MWP solving.
- All the problems are linear equations in one variable.

Evaluation Setup

Models:

- Seq2Seq with Attention
- GTS
- Graph2Tree

Datasets:

- MaWPS
- ASDiv-A
- SVAMP

- ASDIV-A consists of 1,213 math word problems.
- It also includes annotations for equation, problem type and grade level.
- The diversity of problems is larger than MaWPS.

Evaluation Setup

Models:

- Seq2Seq with Attention
- GTS
- Graph2Tree

Datasets:

- MaWPS
- ASDiv-A
- SVAMP

- SVAMP is a challenge set designed manually by altering the problem statements in MaWPS and ASDiv-A.
- These alterations are made to test the robustness of existing solvers.
- SVAMP consists of around 1000 simple math word problems.

Drawbacks of existing MWP Solvers

We conduct following experiments to identify the drawbacks of existing MWP solvers:

- Dropping Question text
- Randomly shuffling sequence of sentences
- Random word deletion
- Random word reordering

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

Drawbacks of existing MWP Solvers

We conduct following experiments to showcase the drawbacks of existing MWP solvers:

- Dropping Question text
- Randomly shuffling sequence of sentences
- Random word deletion
- Random word reordering

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

Drawbacks of existing MWP Solvers

We conduct following experiments to showcase the drawbacks of existing MWP solvers:

- Dropping Question text
- Randomly shuffling sequence of sentences
- Random word deletion
- Random word reordering

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

Drawbacks of existing MWP Solvers

We conduct following experiments to showcase the drawbacks of existing MWP solvers:

- Dropping Question text
- Randomly shuffling sequence of sentences
- Random word deletion
- Random word reordering

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

Dataset	Eval Type	Seq2Seq	GTS	Graph2Tree
MaWPS	True	84.6	87.5	88.7
	WD	**80.2**	**81.5**	**77.3**
	QR	**77.4**	**82.0**	**80.2**
	SS	**77.0**	**60.4**	**66.4**
	WR	**54.9**	**34.8**	**39.3**
ASDiv-A	True	70.6	80.3	82.7
	WD	**60.2**	**61.3**	**56.7**
	QR	**58.7**	**52.4**	**54.1**
	SS	**56.2**	**59.3**	**60.7**
	WR	**47.1**	**32.3**	**34.6**

WD,QR,SS,WR represent word deletion, question reordering, sentence shuffling and word reordering respectively.

- There is a moderate drop in the accuracy scores for WD, QR and SS transformations.
- A large drop is observed for Word reordering transformation.
- The reduction in accuracy scores is relatively very less than expected.
- We infer that solvers pick word patterns and keywords from the problem statement.

Why Data Augmentation ?

- Constructing large datasets which are annotated, labeled and have MWPs of similar difficulty level is a very expensive and tedious task.
- Humans learn to solve MWPs by going through a variety of similar examples and slowly become capable enough to tackle variations of similar difficulty levels.

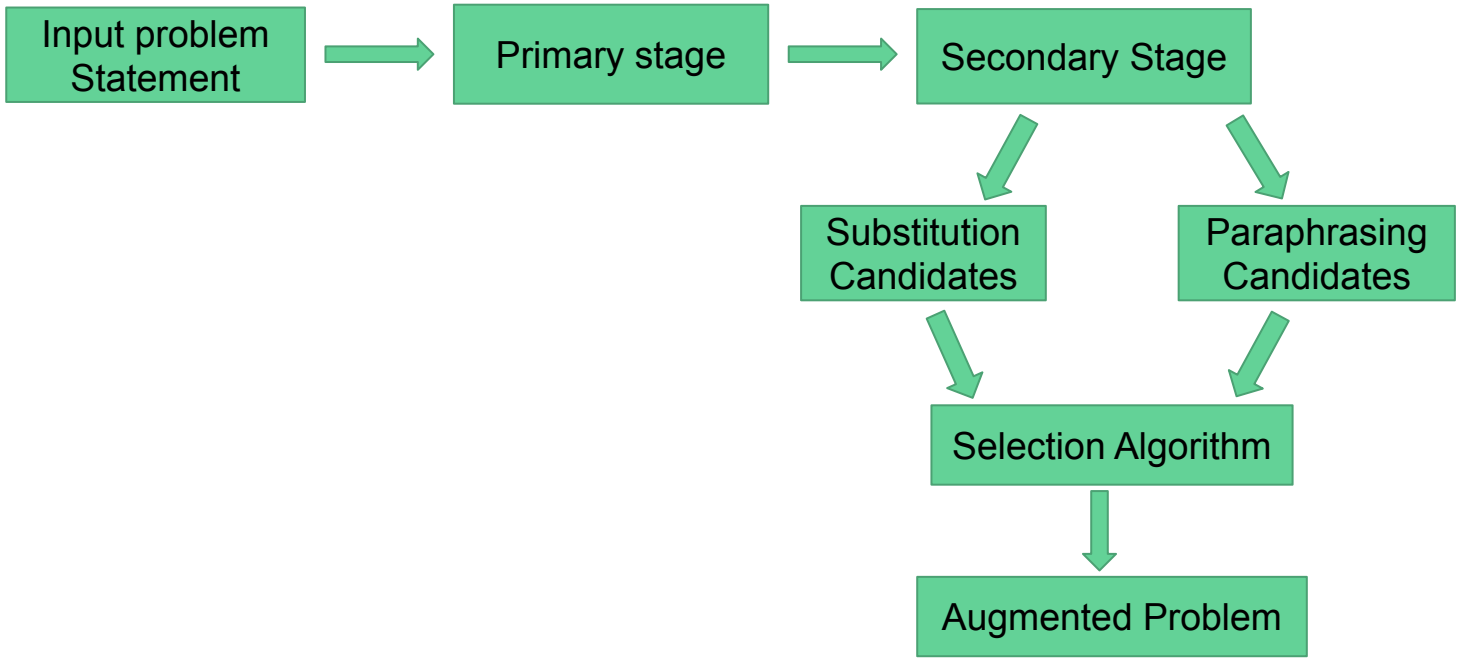
Following conditions must be satisfied by augmented examples:

- Preserve the equation labels.
- Same numeric values and relationships with their entities.
- Sequence of events should be preserved.
- Semantically similar to the original counterpart.

Proposed Approach

- Based on the experimental results, and previous works by Kumar et al 2021, Patel et al 2020, we infer that the existing solvers are not robust.
- One of the key reasons for lack of robustness is unavailability of a large dataset.
- To improve the robustness of existing solvers on available datasets we propose data augmentation methods.
- We further propose candidate selection algorithm which ensures diversity in the augmented problem statements.
- To validate our method, we evaluate on SVAMP challenge set and conduct human evaluation apart from the extensive experimentation.

Outline of Proposed Approach

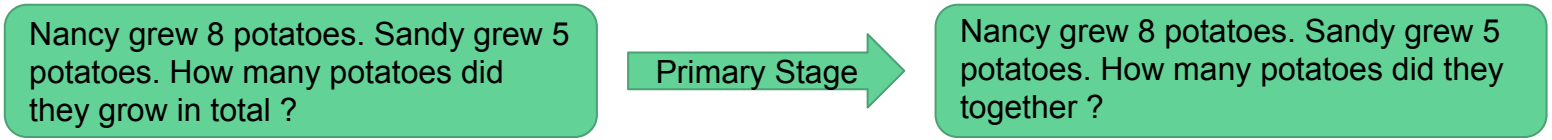


Primary Stage

- In the primary stage we generate a set of base candidates from problem statement P by inducing variations in the question text Q.
- We use T5 paraphrasing model to generate top 5 variations of Q per problem.
- Key motivation is to ensure that each augmentation of a given problem has a different question text.
- This step is aimed to empower the solver to focus more on the question text.

Primary Stage

- In the primary stage we generate a set of base candidates from problem statement P by inducing variations in the question text Q.
- We use T5 paraphrasing model to generate top 5 variations of Q per problem.
- Key motivation is to ensure that each augmentation of a given problem has a different question text.
- This step is aimed to empower the solver to focus more on the question text.



Secondary Stage

- In the secondary stage we work on all the base candidates to generate augmentation candidates for a problem statement.

Following methods are proposed:

- Paraphrasing Methods
  1. Round Trip Translations
  2. Problem Reordering
- Substitution Methods
  1. Fill Masking
  2. Named Entity replacement
  3. Synonym replacement

Round Trip Translations

- Problems are translated from their original language to foreign languages and then translated back to the original language.
- The motivation is to utilize the different structural constructs and linguistic variations present in other languages.

Challenges:

1. Numerical quantities are fragile to translations and their order and representation may change.
2. Back-translation is known to diverge uncontrollably (Tan et al., 2019) for multiple round trips. This can lead semantic variance.

Round Trip Translations

To overcome these challenges:

1. To preserve numerical quantities, we replace them with special symbols and keep a map to restore numerical quantities.
2. To control divergence, we worked with languages that have structural constructs similar with English.

English - Russian - English: Word order does not affect the syntactic structure of a sentence in Russian, making it a good choice for single hop.

English - German - French - English: Languages similar to English for multiple round trips to maintain both semantic in-variance and induce minor alterations in the paraphrases.

Problem Reordering

- We reorder the problem such that question text is at the start of problem statement.
- To preserve the semantic and syntactic meaning of problem statement we use filler phrases like ’Given that’ and ’If-then’.
- To make these paraphrases more fluent, we use NER and co-reference resolution to replace the occurrences of pronouns with their corresponding references.
- This method is better than random shuffling of sentences as it preserves the sequence of events in the problem statement.

The schools debate team had 4 boys and 6 girls on it. If they were split into groups of 2, how many groups could they make ?

Round Trip Translation

The school discussion group consisted of 4 boys and 6 girls. If they are divided into groups of 2, How many groups could they have created ?

Lucy has an aquarium with 5 fish. She wants to buy 1 more fish. How many fish would Lucy have then ?

Problem Reordering

If lucy has an aquarium with 5 fish and she wants to buy 1 more fish then how many fish would lucy have ?

Secondary Stage

- In the secondary stage we work on all the base candidates to generate augmentation candidates for a problem statement.

Following methods are proposed:

- Paraphrasing Methods
  1. Round Trip Translations
  2. Problem Reordering
- Substitution Methods
  1. Fill Masking
  2. Named Entity replacement
  3. Synonym replacement

Fill Masking

- We model the challenge of generating candidates as a masked language modelling problem.
- Instead of randomly choosing words for masking, we use part of speech tags to focus on nouns and adjectives, preferably in the vicinity of numerical quantities.
- These words are then replaced with MASK tokens and then passed through a masked language model.

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

Named Entity Replacement

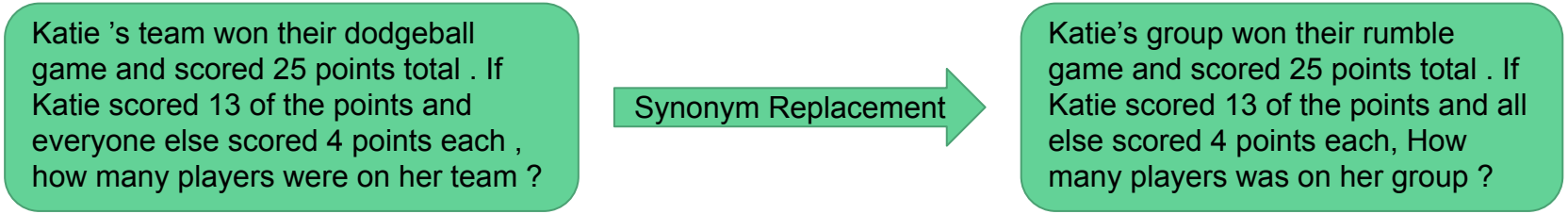
- Named entities play a crucial role in stating the problem statement, but the solution equations do not change on altering these entities.
- We identify and replace named entities like person, place and organizations with their corresponding substitutes.

Sally found 7 seashells , Tom found 12 seashells , and Jessica found 5 seashells on the beach . How many seashells did they find together ?

Edd found 7 seashells , Alan found 12 seashells , and Royal found 5 seashells on the beach . How many seashells were found together ?

Synonym Replacement

- After stop-word removal, we select keywords randomly for substitution.
- Here we use Glove embeddings (Pennington et al., 2014) to find the top k candidates that are close synonyms of the keywords.
- To ensure syntactic correctness in candidates, we maintain the part of speech tags for the substitute candidates.



Algorithm 1 MWP Candidate Selection Algorithm

Requires: $\mathcal{M}$ is augmentation method, $\mathcal{S}$ is similarity model, $\mathcal{F}$ is solver model, $\mathcal{L}$ is Loss function.

Input: Problem text $\mathcal{P}$

Output: Augmented Text $\mathcal{P}^*$

$\mathcal{E}_P \leftarrow \mathcal{F}(\mathcal{P})$

$Candidates \leftarrow \mathcal{M}(\mathcal{P})$

for C_j *in* $Candidates$: **do**

    $S_j \leftarrow \mathcal{S}(C_j, \mathcal{P})$

    $L_j \leftarrow (\mathcal{L}(C_j) - \mathcal{L}(P))/\mathcal{L}(P)$

    $CandidateScore.add(S_j * L_j)$

$\mathcal{P}^* = \arg\max_{C_j} CandidateScore(C_j)$

end

Candidate Selection Algorithm

- Potential candidates are filtered to select the best possible candidate using selection algorithm.
- Select candidates on which the solver does not perform well and which are similar to the original problem statement.
- We evaluate all candidates loss values and select the candidate with the maximum mean normalized loss and similarity score.
- Negative log likelihood and Sentence-BERT are used to compute loss values and similarity scores respectively.

Results

Dataset	Problem Size	Vocabulary Size
MaWPS	2,373	2,632
ASDiv-A	1,213	2,893
Paraphrase	5,909	3,832
Substitution	6,647	3,923
Combined-MaWPS	**10,634**	**5,626**
Combined-ASDiv	**5,312**	**6,109**

Statistics of augmented dataset compared with MaWPS and ASDiv-A.

Results

Dataset	Evaluation Type	Seq2Seq	GTS	G2T
MaWPS	True	84.6	87.5	88.7
	Paraphrasing	**88.3**	**90.4**	**92.6**
	Substitution	**89.2**	**89.7**	**91.7**
	Combined	**91.3**	**92.6**	**93.5**
ASDiv-A	True	70.6	80.3	82.7
	Paraphrasing	**75.6**	**84.2**	**83.6**
	Substitution	**73.2**	**83.3**	**84.1**
	Combined	**78.2**	**85.9**	**86.3**

Result of augmentation methods. True is original dataset, Combined is combination of paraphrasing and substitution. G2T represents Graph2Tree solver.

Problem 1: Ricardo was making baggies of cookies with 5 cookies in each bag. If he had 7 chocolate chip cookies and 3 oatmeal cookies, how many baggies could he make ?
Solution Equation: X = (7+3)/5
Pre Augmentation Equation: X = (7/3)/3
Post Augmentation Equation: X = (7+3)/5

Problem 2: For halloween Destiny bought 9 pieces of candy. She ate 3 pieces the first night and then her sister gave her 2 more pieces. How many pieces of candy does Destiny have now ?
Solution Equation: X = 9-3+2
Pre Augmentation Equation: X = ((9+3-3
Post Augmentation Equation: X = (9+3-2)

Problem 3 : Audrey needs 6 cartons of berries to make a berry cobbler. She already has 2 cartons of strawberries and 3 cartons of blueberries. How many more cartons of berries should Audrey buy ?
Solution Equation: X = 6-2-3
Pre Augmentation Equation: X = (6-(2)+3)
Post Augmentation Equation: X = 6-(2+3)

Examples illustrating equation results before and after training on the full augmented dataset.

Ablation Studies

Method	Eval Type	Seq2Seq	GTS	Graph2Tree
RSA	True	84.6	87.5	88.7
	Paraphrasing	**85.3**	**88.1**	**89.2**
	Substitution	**86.8**	**87.3**	**87.9**
	Combined	**87.0**	**89.2**	**89.5**
CSA	True	84.6	87.5	88.7
	Paraphrasing	**88.3**	**90.4**	**92.6**
	Substitution	**89.2**	**89.7**	**91.7**
	Combined	**91.3**	**92.6**	**93.5**

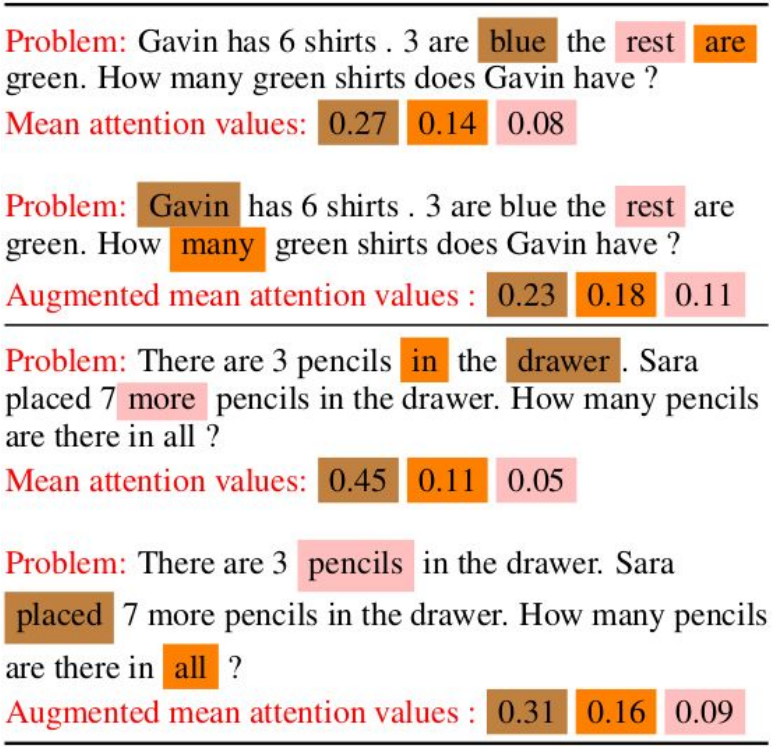
Ablation Study for Random Selection Algorithm (RSA) and Candidate Selection Algorithm (CSA).

Ablation Studies

Augmentation	Seq2Seq	GTS	Graph2Tree
True	84.6	87.5	88.7
RRT	**86.5**	89.1	**91.6**
PR	85.9	88.4	90.7
FM	84.8	87.2	89.1
SR	85.2	**90.1**	91.2
NER	86.1	88.3	89.7

Result of Ablation study for each augmentation method. True represents unaugmented MaWPS dataset, RRT, PR, FM, SR, NER represents round trip translations, problem reordering, fill masking,synonym replacement and named entity replacement respectively.

Attention Visualization



Examples illustrating distribution of top three attention weights before and after training on the full augmented dataset.

- To illustrate the effectiveness of our augmentation techniques, we show the distribution of attention weights for models trained on the augmented dataset.
- Before augmentation the focus of the solver is limited to a fixed region around numerical quantities and it does not pay heed to the question text.
- After training on the augmented dataset the solver has a better distribution of attention weights, the weights are not localised and and the model is also able to pay attention on the question text.

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

SVAMP Challenge Set

Augmentation	Seq2Seq	GTS	Graph2Tree
True	37.5	39.6	41.2
MaWPS(P+S)	39.2	40.1	42.3
ASDiv-A(P+S)	37.8	40.4	42.1
Combined	**40.2**	**41.3**	**43.8**

Result of augmentations on SVAMP Challenge Set. P and S represent paraphrasing and substitution methods. Combined represents augmented MaWPS and ASDiv-A. True is combined MaWPS and ASDiv-A.

BERT Embeddings

Augmentation Method	MaWPS		ASDiv-A	
	Scratch	BERT	Scratch	BERT
True	77.2	**84.6**	53.2	**70.6**
Paraphrasing	79.8	**88.3**	58.1	**75.6**
Substitution	81.3	**89.2**	57.3	**73.2**
Combined	82.7	**91.3**	60.4	**78.2**

Performance comparison of baseline model trained from scratch and trained using BERT embeddings. True represents unaugmented dataset.

Human Evaluation

Evaluation Criteria	MaWPS		ASDiv-A	
	Para	Sub	Para	Sub
Preserves Equation	92.3%	89.5%	93.6%	90.1%
Preserves Numbers	88.4%	91.2%	87.3%	90.3%
Semantic Similarity	0.96	0.89	0.91	0.87
Syntactic Similarity	4.67	4.36	4.59	4.33

Human Evaluation scores on augmented dataset. Para and Sub represents paraphrasing and substitution methods respectively.

Future Works

- Future works could focus on developing techniques to generate data artificially and making robust MWP solvers.
- WSM 8k is a recent dataset in this direction, however to develop robust solvers we must scale these datasets artificially to larger size.
- For more details, please find our repository at https://github.com/kevivk/MWP-Augmentation or reach out at vivek.k@research.iiit.ac.in