Are NLP Models really able to Solve Simple MathWord Problems?

김현주

목차

01 Introduction

Math Word Problem (MWP)

PROBLEM:
Text: Jack had 8 pens and Mary had 5 pens. Jack gave 3 pens to Mary. How many pens does Jack have now?
Equation: 8 - 3 = 5

- 짧은 natural language narrative으로 이루어져 있는 수학적 문제
- 문제가 주어지면 계산할 수 있는 수식을 도출하는 task
- 자연어를 이해해 중요한 정보를 추출하고 동시에 수학적인 reasoning도 시행할 수 있어야 함

Problem Formulation

$$Problem : P = (w_1, \cdots, w_n)$$
$$Body : B = (w_1, \cdots, w_k)$$
$$Question : Q = (w_{k+1}, \cdots, w_n)$$

용어

- Grade Level
  - 난이도를 나타내는 표현
  - 비슷한 종류의 MWPs를 어느 학년 수준에서 가르치는지를 나타냄
- One-unknown arithmetic word problems
  - 숫자들과 하나 혹은 여러 개의 연산자(+, −,×,÷)로 이루어진 수식 도출
  - 단 하나의 미지수 존재
- Execution Accuracy
  - 도출된 수식을 계산했을 때 그 예측 값이 실제 값과 일치하는지 계산

Recent Works

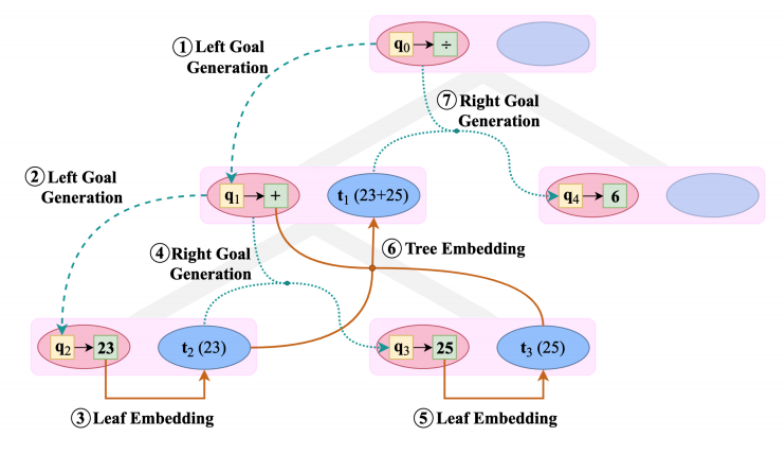
- 최근 MWPs를 해결하기 위한 다양한 datasets와 methods가 제시되고 있음
- 많은 논문에서 이전 datasets에는 결함이 있다고 생각하여 그것을 다루는 여러 datasets가 제시됨
  - MAWPS, Math23K, ASDiv, HMWP 등
- 그러나 아직까지 datasets의 어느 요소가 문제가 되는가에 대한 분석은 진행되지 않음

02 Existing Methods

Seq2Seq

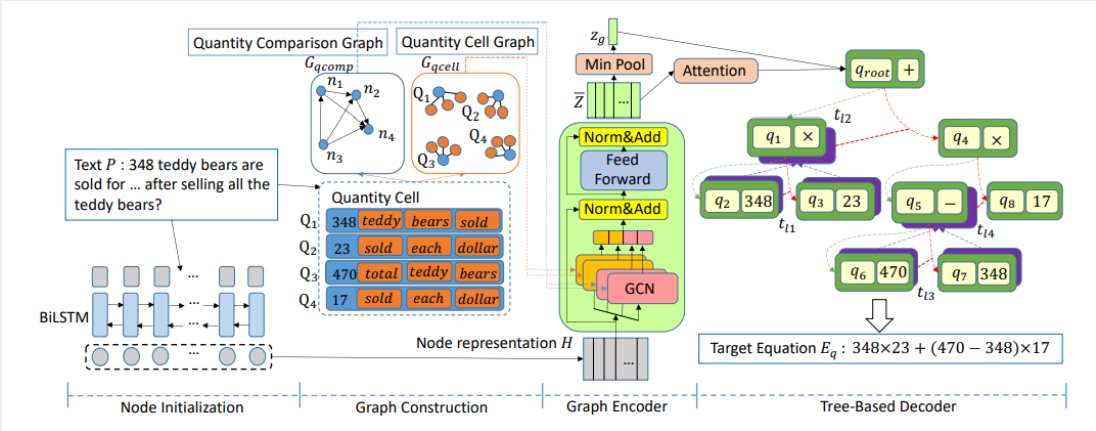
- Encoder: bidirectional LSTM
- Decoder: LSTM with attention

GTS



- Encoder: LSTM
- Decoder: tree-based

Graph2Tree



- Encoder: graph-based
- Decoder: tree-based

03 Existing Datasets

- Math23K, HMWP : 중국어로 이루어져 있음
- Dolphin18K : 너무 어려운 문제 유형들로 이루어져 있음

MAWPS


```
[
    {
      "iIndex": 1,
      "sQuestion": "Joan found 70.0 seashells on the beach . She gave Sam some of her seashells .
      She has 27.0 seashells . How many seashells did she give to Sam ?",
      "lEquations": [
        "X=(70.0-27.0)"
      ],
      "lSolutions": [
        43.0
      ]
    }
]
```


ASDiv-A


```
<Problem ID="nluds-0001" Grade="1" Source="http://www.k5learning.com">
    <Body>Seven red apples and two green apples are in the basket.</Body>
    <Question>How many apples are in the basket?</Question>
    <Solution-Type>Addition</Solution-Type>
    <Answer>9 (apples)</Answer>
    <Formula>7+2=9</Formula>
</Problem>
```


Model	MAWPS	ASDiv-A
Seq2Seq (S)	79.7	55.5
Seq2Seq (R)	86.7	76.9
GTS (S) (Xie and Sun, 2019)	82.6	71.4
GTS (R)	88.5	81.2
Graph2Tree (S) (Zhang et al., 2020)	83.7	77.4
Graph2Tree (R)	**88.7**	**82.2**
Majority Template Baseline[2]	17.7	21.2

Table 2: 5-fold cross-validation accuracies (↑) of baseline models on datasets. (R) means that the model is provided with RoBERTa pretrained embeddings while (S) means that the model is trained from scratch.

03 Existing Datasets - Problems

Question-removed MWPs

Model	MAWPS	ASDiv-A
Seq2Seq	77.4	58.7
GTS	76.2	60.7
Graph2Tree	**77.7**	**64.4**

Table 3: 5-fold cross-validation accuracies (↑) of baseline models on Question-removed datasets.

	MAWPS		ASDiv-A	
Model	*Easy*	*Hard*	*Easy*	*Hard*
Seq2Seq	86.8	86.7	91.3	56.1
GTS	92.6	71.7	91.6	65.3
Graph2Tree	93.4	71.0	92.8	63.3

Table 4: Results of baseline models on the *Easy* and *Hard* test sets.

- Test set에서 Question 부분을 제거한 후 model을 평가함
- Question이 없음에도 불구하고 정답률이 꽤 높은 것을 알 수 있음
- MWPs의 body에 output equation과 직접적으로 연관될 수 있는 일정 패턴이 있는 것으로 판단됨
- 완전한 정보가 없더라도 정답을 맞출 수 있는 요소가 datasets에 존재함을 보여줌

- **Easy** : question이 없어도 정답을 맞추는 문제
- **Hard** : question 없이는 정답을 맞추지 못하는 문제
- Hard question도 어느정도 맞추는 모습을 보이기는 하지만 easy question이 높은 정답률의 큰 영향을 끼치는 것으로 보임

- SOTA methods가 MWPs를 해결하는 능력이 과장되어 평가되었다고 생각
- Body에 존재하는 간단한 heuristics에 기대 예측한다고 판단

03 Existing Datasets - Problems

Constrained Model

Model	MAWPS	ASDiv-A
FFN + LSTM Decoder (S)	75.1	46.3
FFN + LSTM Decoder (R)	**77.9**	**51.2**

Table 5: 5-fold cross-validation accuracies (↑) of the constrained model on the datasets. (R) denotes that the model is provided with non-contextual RoBERTa pre-trained embeddings while (S) denotes that the model is trained from scratch.

- Seq2Seq 기반의 간단한 model
- LSTM Encoder를 제거하고 Feed-Forward Network를 넣어 input embeddings가 hidden representation에 mapping되도록 함
- LSTM Decoder가 hidden representations의 평균 초기 hidden states를 받음
- Decoding 할 때 attention mechanism을 사용하여 각 input tokens의 hidden representations에 가중치를 부여함

- Model이 단어 순서에 대한 정보가 없더라도 datasets의 대다수 문제를 해결할 수 있음
- 문제에 등장하는 특정 단어들만 보고서 그에 맞는 수식을 도출할 수 있음

03 Existing Datasets - Problems

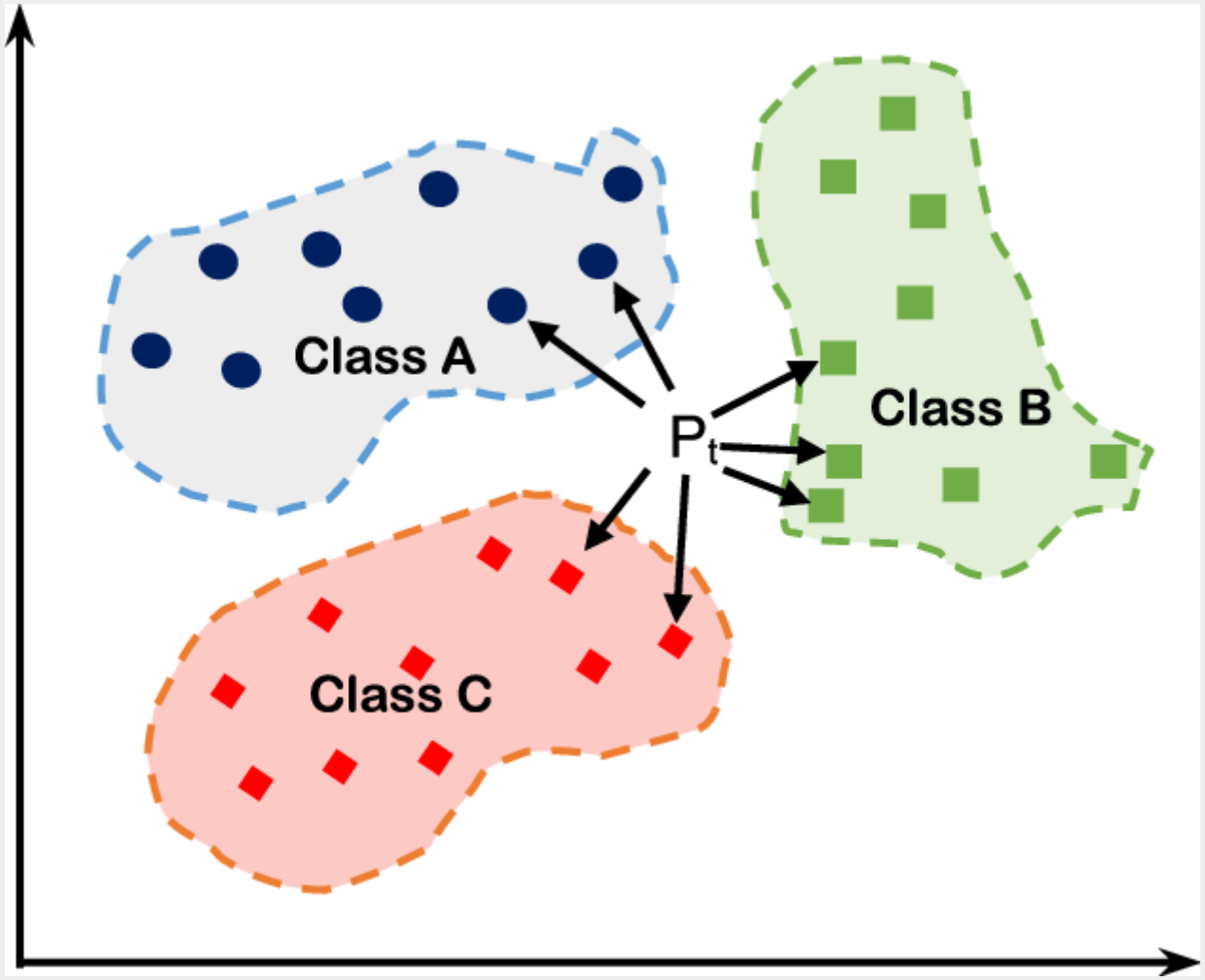
Attention Weights

Input Problem	Predicted Equation	Answer
John delivered 3 letters at every house. If he delivered for 8 houses, how many letters did John deliver?	3 * 8	24 ✓
John delivered 3 letters at every house. He delivered 24 letters in all. How many houses did John visit to deliver letters?	3 * 24	72 ✗
Sam made 8 dollars mowing lawns over the Summer. He charged 2 bucks for each lawn. How many lawns did he mow?	8 / 2	4 ✓
Sam mowed 4 lawns over the Summer. If he charged 2 bucks for each lawn, how much did he earn?	4 / 2	2 ✗
10 apples were in the box. 6 are red and the rest are green. how many green apples are in the box?	10 - 6	4 ✓
10 apples were in the box. Each apple is either red or green. 6 apples are red. how many green apples are in the box?	10 / 6	1.67 ✗

Table 6: Attention paid to specific words by the constrained model.

- 학습한 constrained model의 attention weights를 분석함
- Context를 보지 않고 하나의 단어만으로 예측하는 모습을 보임
- 문제에 변형을 가하면 도출되는 수식이 달라져야 하지만 계속해서 같은 단어만 보아 같은 수식(틀린 수식)이 도출됨
- Datasets에 학습된 모든 모델이 이렇다는 것은 아님
- 그러나 constrained model과 같은 모델도 좋은 성적을 낼 수 있음을 보임
- Dataset이 model의 성능을 확실하게 측정하기에는 모자람

04 SVAMP Simple Variations on Arithmetic Math word Problems



04 SVAMP

Question Sensitivity	Same Object, Different Structure	**Original:** Allan brought two balloons and Jake brought four balloons to the park. How many balloons did Allan and Jake have in the park? **Variation:** Allan brought two balloons and Jake brought four balloons to the park. How many more balloons did Jake have than Allan in the park?
	Different Object, Same Structure	**Original:** In a school, there are 542 girls and 387 boys. 290 more boys joined the school. How many pupils are in the school? **Variation:** In a school, there are 542 girls and 387 boys. 290 more boys joined the school. How many boys are in the school?
	Different Object, Different Structure	**Original:** He then went to see the oranges being harvested. He found out that they harvest 83 sacks per day and that each sack contains 12 oranges. How many sacks of oranges will they have after 6 days of harvest? **Variation:** He then went to see the oranges being harvested. He found out that they harvest 83 sacks per day and that each sack contains 12 oranges. How many oranges do they harvest per day?
Reasoning Ability	Add relevant information	**Original:** Every day, Ryan spends 4 hours on learning English and 3 hours on learning Chinese. How many hours does he spend on learning English and Chinese in all? **Variation:** Every day, Ryan spends 4 hours on learning English and 3 hours on learning Chinese. If he learns for 3 days, how many hours does he spend on learning English and Chinese in all?
	Change Information	**Original:** Jack had 142 pencils. Jack gave 31 pencils to Dorothy. How many pencils does Jack have now? **Variation:** Dorothy had 142 pencils. Jack gave 31 pencils to Dorothy. How many pencils does Dorothy have now?
	Invert Operation	**Original:** He also made some juice from fresh oranges. If he used 2 oranges per glass of juice and he made 6 glasses of juice, how many oranges did he use? **Variation:** He also made some juice from fresh oranges. If he used 2 oranges per glass of juice and he used up 12 oranges, how many glasses of juice did he make?
Structural Invariance	Change order of objects	**Original:** John has 8 marbles and 3 stones. How many more marbles than stones does he have? **Variation:** John has 3 stones and 8 marbles. How many more marbles than stones does he have?
	Change order of phrases	**Original:** Matthew had 27 crackers. If Matthew gave equal numbers of crackers to his 9 friends, how many crackers did each person eat? **Variation:** Matthew gave equal numbers of crackers to his 9 friends. If Matthew had a total of 27 crackers initially, how many crackers did each person eat?
	Add irrelevant information	**Original:** Jack had 142 pencils. Jack gave 31 pencils to Dorothy. How many pencils does Jack have now? **Variation:** Jack had 142 pencils. Dorothy had 50 pencils. Jack gave 31 pencils to Dorothy. How many pencils does Jack have now?

Dataset	# Problems	# Equation Templates	# Avg Ops	CLD
MAWPS	2373	39	1.78	0.26
ASDiv-A	1218	19	1.23	0.50
SVAMP	1000	26	1.24	0.22

Table 9: Statistics of our dataset compared with MAWPS and ASDiv-A.

- Evaluation templates와 avg ops 개수를 봤을 때 타 datasets와 크게 다른 수준을 지니지 않음
- Corpus Lexicon Diversity (CLD)
  - lexical diversity를 측정하는 척도
  - 보통 더 높을수록 dataset의 질이 높음을 뜻함
  - SVAMP는 CLD가 낮음에도 불구하고 타 datasets보다 더 평가하기 좋음을 보임
  - CLD는 datasets의 질을 평가하는 좋은 척도가 아니라고 생각

04 SVAMP

	Seq2Seq		GTS		Graph2Tree	
	S	*R*	*S*	*R*	*S*	*R*
Full Set	24.2	40.3	30.8	41.0	36.5	**43.8**
One-Op	25.4	42.6	31.7	44.6	42.9	**51.9**
Two-Op	20.3	**33.1**	27.9	29.7	16.1	17.8
ADD	28.5	**41.9**	35.8	36.3	24.9	36.8
SUB	22.3	35.1	26.7	36.9	41.3	**41.3**
MUL	17.9	**38.7**	29.2	**38.7**	27.4	35.8
DIV	29.3	56.3	39.5	61.1	40.7	**65.3**

Table 10: Results of models on the SVAMP challenge set. *S* indicates that the model is trained from scratch. *R* indicates that the model was trained with RoBERTa embeddings. The first row shows the results for the full dataset. The next two rows show the results for subsets of SVAMP composed of examples that have equations with one operator and two operators respectively. The last four rows show the results for subsets of SVAMP composed of examples of type Addition, Subtraction, Multiplication and Division respectively.

- SVAMP는 사람에게 있어 다른 dataset와 다르지 않은 단계를 다루고 있음 (Grade 4)
- Model을 학습한 결과 타 datasets보다 더 challenging 함을 보여줌

Model	SVAMP w/o ques	ASDiv-A w/o ques
Seq2Seq	29.2	58.7
GTS	28.6	60.7
Graph2Tree	**30.8**	**64.4**

Table 11: Accuracies (↑) of models on SVAMP without questions. The 5-fold CV accuracy scores for ASDiv-A without questions are restated for easier comparison.

- SVAMP에서 question을 제거한 후 test
- 정확도가 반으로 줄어드는 모습을 보임
- Question의 정보도 이용하도록 잘 구성됨을 증명

04 SVAMP

Model	SVAMP
FFN + LSTM Decoder (S)	17.5
FFN + LSTM Decoder (R)	**18.3**
Majority Template Baseline	11.7

Table 12: Accuracies (↑) of the constrained model on SVAMP. (R) denotes that the model is provided with non-contextual RoBERTa pretrained embeddings while (S) denotes that the model is trained from scratch.

- SVAMP에 constrained model을 학습시켜 정확도 확인
- Baseline보다 약간 더 높아진 성능을 보임
- 간단한 pattern으로 구성된 model도 풀 수 있도록 vulnerable 하지 않음
- Datasets의 문제를 해결하기 위해서는 contextual information이 필요함

Dataset	2 nums	3 nums	4 nums
ASDiv-A	93.3	59.0	47.5
SVAMP	78.3	25.4	25.4

Table 15: Accuracy break-up according to the number of numbers in the input problem. **2 nums** refers to the subset of problems which have only 2 numbers in the problem text. Similarly, **3 nums** and **4 nums** are subsets that contain 3 and 4 different numbers in the problem text respectively.

- 가장 좋은 성적을 낸 Graph2Tree로 input text에 나오는 숫자의 개수 별로 성능을 확인함
- 2개 이상의 숫자가 나오면 model의 성능이 매우 떨어짐을 보임
- 현재 방법들은 context와 숫자를 연관 짓는 능력이 부족함을 보여줌

04 SVAMP

Removed Category	# Removed Examples	Change in Accuracy (Δ)
Question Sensitivity	462	+13.7
Reasoning Ability	649	-3.3
Structural Invariance	467	+4.5

Table 13: Change in accuracies when categories are removed. The Change in Accuracy $\Delta = Acc(Full - Cat) - Acc(Full)$, where $Acc(Full)$ is the accuracy on the full set and $Acc(Full - Cat)$ is the accuracy on the set of examples left after removing all examples which were created using Category Cat either by itself, or in use with other categories.

Removed Variation	# Removed Examples	Change in Accuracy (Δ)
Same Obj, Diff Struct	325	+7.3
Diff Obj, Same Struct	69	+1.5
Diff Obj, Diff Struct	74	+1.3
Add Rel Info	264	+5.5
Change Info	149	+3.2
Invert Operation	255	-10.2
Change order of Obj	107	+2.3
Change order of Phrases	152	-3.3
Add Irrel Info	281	+6.9

Table 14: Change in accuracies when variations are removed. The Change in Accuracy $\Delta = Acc(Full - Var) - Acc(Full)$, where $Acc(Full)$ is the accuracy on the full set and $Acc(Full - Var)$ is the accuracy on the set of examples left after removing all examples which were created using Variation Var either by itself, or in use with other variations.

- Variation의 각 category 마다 정확도에 어느정도 영향을 끼치는지 확인
- Question sensitivity와 structural invariance가 SVAMP를 더 challenging하게 만들어 줌
- Reasoning ability의 경우 invert operation 때문에 정확도가 감소했다고 판단
  - ASDiv-A와 가장 비슷한 형태

05 Conclusion

정말 현재의 NLP methods는 Math Word Problems를 해결할 수 있을까?

NO!

- 현존하는 대다수의 datasets는 word-order information이나 question text도 필요로 하지 않은 simple heuristics로도 해결됨
- SVAMP가 MWPs에 대한 성능을 좀 더 정확히 측정하는 척도가 되길 바람
- 아직까지 현존하는 model은 one-unknown arithmetic word problems도 충분히 해결하지 못한다고 판단됨
- 더 어려운 문제로 넘어가기에는 아직 이름

Q & A