PARSING ALGEBRAIC WORD PROBLEMS INTO EQUATIONS

Rik Koncel-Kedzioriski
Hannaneh Hajishirzi
Ashish Sabharwal
Oren Etzioni
Sienna Dumas Ang
Problem definition

- Map math word problem to an equation (any equation) that allows for deducing the correct numerical solution.

On Monday, 375 students went on a trip to the zoo. All 7 buses were filled and 4 students had to travel in cars. How many students were in each bus?
Why Math Word problems?

Math Textbooks

Math section of standardized tests
Why Math Word problems?

- Math Word problems are short narratives: lead to general narrative understanding
- Evaluable: if we can answer the question, we’ve understood it
- Challenging!
Challenges

- Open Vocabulary describing arbitrary entities and events
Challenges

- Discourse/Narrative Understanding
  1. Co-reference
  2. Ellipsis
  3. Predicate effects
  4. Quantifier effects
  5. Non-linear storylines
Challenges

Solving Math Word Problems requires a precise understanding of the text.
Previous Approaches: Verb Categorization

Hosseini et al. 2014 “Learning to Solve Arithmetic Word Problems with Verb Categorization”

■ Pros
- Uses a semantics of Entities and Containers
- Tracks container state across sentences
- Learns effect of verb on container states

■ Cons
- Semantics is domain specific and impoverished
- Sentences often have more than nouns and verbs
- Seed-verb lexicon
- Only handles + or - operations
Previous Approaches: Templates

- Kushman et al 2015 “Learning to Automatically Solve Algebra Word Problems”

<table>
<thead>
<tr>
<th>Derivation 1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Word problem</strong></td>
</tr>
</tbody>
</table>
| **Aligned template** | $u_1 + u_2 - n_1 = 0$  
$n_2 \times u_1^2 + n_3 \times u_2^2 - n_4 = 0$ |
| **Instantiated equations** | $x + y - 278 = 0$  
$1.5x + 4y - 792 = 0$ |
| **Answer** | $x = 128$  
$y = 150$ |
Previous Approaches: Templates

■ Pros
- Can handle all arithmetic operations
- Introduces a difficult dataset
- ...successfully solves systems of equations

■ Cons
- Brittle: Must have seen template >20x to ensure answer accuracy >55%
- Brittle: Memorizes spurious lexical clues
- Non-semantic: System has no internal representation of the narrative beyond the template
- Semantics hurt: Zhou et al. 2015 beats Kushman by removing non-numeric variables
Our Approach: Semantically augmented equation trees

- Equation trees where each node is associated with a type label
- Operations are type-constrained
  - Addition and subtraction and equality prefer similar types
  - Multiplication and division prefer dissimilar types
- Types derived automatically from each problem
On Monday, 375 students went on a trip to the zoo. All 7 buses were filled and 4 students had to travel in cars. How many students were in each bus?
On Monday, 375 students went on a trip to the zoo. All 7 buses were filled and 4 students had to travel in cars. How many students were in each bus?
Semantically augmented equation trees

375 students
4 students
x students

7 buses
Semantically augmented equation trees

375 students + 7 buses * x students = 4 students
Semantically augmented equation trees

\[
\begin{align*}
375 \text{ students} & \quad 7 \text{ buses} & \quad x \text{ students} & \quad 4 \text{ students} \\
\quad \quad = & \quad + & \quad + & \quad = \\
\text{Local Scores} & \quad \text{students} & \quad \text{students} & \\
\text{Global Score} & 
\end{align*}
\]
Optimization

- Combine Global and Local tree scores

\[ p(t|w) \propto \left( \prod_{t_j \in t} \mathcal{L}_{local}(t_j|w) \right) \times \mathcal{G}_{global}(t|w) \]
On Monday, 375 students went on a trip to the zoo. All 7 buses were filled and 4 students had to travel in cars. How many students were in each bus?

1. Ground text $w$ into base Qsets

- Qnt: 375
  - Ent: Student
- Qnt: $x$
  - Ent: Student
- Qnt: 7
  - Ent: Bus
- Qnt: 4
  - Ent: Student

2. Use ILP to generate $M$ equation trees $T(w)$

- $T_i(w)$: subset of $T(w)$ yielding correct solution
- $T(w) \setminus T_i(w)$

3. Train local model

$T_{local}$: operator nodes in $T_i(w)$

<table>
<thead>
<tr>
<th>Training example</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(7_b, x_s)$</td>
<td>*</td>
</tr>
<tr>
<td>$(375_s, \text{combine}(7_b, x_s))$</td>
<td>−</td>
</tr>
<tr>
<td>$(7_b, x_s)$</td>
<td>*</td>
</tr>
<tr>
<td>$(\text{combine}(7_b, x_s), 4_s)$</td>
<td>+</td>
</tr>
</tbody>
</table>

4. Train global model

$T_{global}$: problem-tree pairs

- Positive examples (from $T_i(w)$)
  - $375 - (7 \times x) = 4$
  - $375 = (7 \times x) + 4$
  - $375 = (x \times 7) + 4$
- Negative examples (from $T(w) \setminus T_i(w)$)
  - $375 + (7 \times x) = 4$
  - $375 = (7 / x) + 4$
  - $375 - (x + 7) = 4$
Qsets
Semantic representation of a quantity including:
- Quantity
- Entity
- Container
- Adjectives
- Verbs
- Locations

Extracted from Stanford Dependency Parse according to deterministic rules

Recursive: Combination of 2 Qsets is a Qset
ILP

Constrains space of possible equations according to:

- Syntactic Validity
- Definitional Constraints
- Type Consistency
- Qset Ordering
- Domain Considerations
Local Model

- Learns best operation for combining two Qsets
- Multi-class SVM ( + - * / )
- Probabilistic
- Features Include:
  - Entity, verb, adjective, location distances
  - Distances from verbs to Hosseini verb class exemplars
  - Textual Distances, intervening discourse markers

1. Ground text \( w \) into base Qsets
   - Qnt: 375
   - Ent: Student
   - Qnt: 7
   - Ent: Bus
   - Qnt: 4
   - Ent: Student

2. Use ILP to generate \( M \) equation trees \( T(w) \)

3. Train local model
   \( T_{\text{local}} \): operator nodes in \( T_i(w) \)

4. Train global model
   \( T_{\text{global}} \): problem-tree pairs
Global Model

- Combines local scores
- Reranks based on global textual features and root node
- Incorporates ILP constraints into final decision boundary
Experiments

- Accuracy on 508 single equation questions and subsets
- Template overlap is reuse of equation templates in a dataset
- Lexical overlap is reuse of lexemes

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALGES</td>
<td>0.72</td>
</tr>
<tr>
<td>Kushman et al 2014</td>
<td>0.67</td>
</tr>
<tr>
<td>Hosseini et al 2014</td>
<td>0.48</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Template Overlap</th>
<th>Lexical Overlap</th>
<th>ALGES</th>
<th>Kushman et al 2014</th>
<th>Hosseini et al 2014</th>
<th>Error Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.4</td>
<td>4.3</td>
<td>0.72</td>
<td>0.67</td>
<td>0.48</td>
<td>15%</td>
</tr>
<tr>
<td>7.7</td>
<td>3.3</td>
<td>0.66</td>
<td>0.60</td>
<td>0.46</td>
<td>15%</td>
</tr>
<tr>
<td>6.3</td>
<td>2.6</td>
<td>0.66</td>
<td>0.46</td>
<td>0.26</td>
<td>33%</td>
</tr>
<tr>
<td>2.1</td>
<td>2.5</td>
<td>0.63</td>
<td>0.26</td>
<td>0.26</td>
<td>50%</td>
</tr>
</tbody>
</table>

Accuracy on Full SingleEQ dataset

Accuracy on datasets with reduced Lexical/Template overlap
Qualitative Results

Luke had 20 stickers. He bought 12 stickers from a store in the mall and got 20 stickers for his birthday. Then Luke gave 5 of the stickers to his sister and used 8 to decorate a greeting card. How many stickers does Luke have left?

\[
(20 + ((12 + 20) - 8)) - 5 = x
\]

Maggie bought 4 packs of red bouncy balls, 8 packs of yellow bouncy balls, and 4 packs of green bouncy balls. There were 10 bouncy balls in each package. How many bouncy balls did Maggie buy in all?

\[
x = ((((4 + 8) + 4) * 10)
\]
Sara, Keith, Benny, and Alyssa each have 96 baseball cards. How many dozen baseball cards do they have in all?

A restaurant sold 63 hamburgers last week. How many hamburgers on average were sold each day?

The sum of three consecutive even numbers is 162. What is the smallest of these numbers?
Conclusion

- Precision Narrative Understanding is an important problem
- We have understood narratives when we can prove we’ve understood them
- Solving the math word problem proves we’ve understood the narrative
Thanks!

kedzior@uw.edu
Extrapolative Models for Text Generation

Rik Koncel-Kedziorski
Extrapolation
Extrapolation

- Intelligent Behavior
- Creativity
Story Telling

- Requires Extrapolation
- Basically the most “human” thing that humans do
  - Human behavior is intelligent
    - Not counting the president’s
Difficulties

- **Coherence**
  - Jamie was painting her toenails. She was wearing them when she got a flat tire. She went to the store and bought a new pair.

- **Event Ordering**
  - Jamie had been married for six years. He had been engaged for a long time. He was finally ready to propose to his fiance.

- **Blandness**
  - Jamie went to the fair with his friends. They had a lot of fun. They had a lot of fun. They had a great time. They had a great time.
Applications

- Anticipating Likely Outcomes (Planning)
- Dynamic Video Game Plot Lines
- Educational Technologies
- Other ideas that I can’t tell you
Neural Text Generation
seq2seq and cetera
Neural Network
Neural Network
Training

Diagram:

- Input: training data
- Learning system
- Actual output
- Error signal
- Output desired

Σ
A Neural Network is a **Function**
Maps a set of **weights** to a **loss**
**Backpropagation**
- Derivative of $\text{NN}(W_t)$
- Picks $W_{t+1}$ so that $\text{NN}(W_{t+1}) < \text{NN}(W_t)$
**Recompute and Repeat**
Sequence Prediction

- Recurrent Neural Networks
Long Short Term Memory

Diagram showing a Long Short Term Memory (LSTM) network with input $x_t$, output $h_t$, and gates $i_t$, $o_t$, and $f_t$.
Gated Recurrent Unit
seq2seq

"le chat est noir" <EOS>
[ 02 85 03 12 99 ]

Encoder

Context

Decoder

"the cat is black"
[SOS] [ 00 42 82 16 04 ]

[ 42 82 16 04 99 ]

"the cat is black" <EOS>
Words are Numbers now

SOS EOS the a is and or
00 01 02 03 04 05 06 ...

and = < 0 0 0 0 0 1 0 ... >
Encoding
Decoding
Attention
Attention
Viola
Extrapolation

Slightly different
Extrapolation

- Input is short:
  - Nothing to attend to
- Make up some additional inputs:
  - Focus our learning on key parts of the output
  - Controllable text generation
Jamie was in the Caribbean on vacation. He decided to go snorkeling on his second day...

- Narrative Chains (Chambers and Jurafsky 2008)
  - Jamie was; He decided; He go; He snorkeling...

- Subject Verb Object Modifier (Pichotta and Mooney 2016, Martin et al 2018)
  - <Jamie, was, Caribbean, vacation>, <He, decided, -, ->, <he, go, -, ->, <He, snorkeling, -, day>

- Events and Entities only
  - Jamie was Caribbean vacation; he decided go snorkeling day

- Novel information
  - Jamie vacation; decided snorkeling
Hierarchical Model

Tasha was cold. … she liked rain. … she liked rain Tasha …
the storm

Tasha liked rain
Experiments

Etc…
Data

- ROCStories (Mostafazadeh et al 2016)
  - ~100k crowdsourced stories with titles
  - 5 sentences (50 words)
  - “Have beginning, middle, and end”

- Jamie decided she wanted to go to the beach. She invited all her friends to go along. They had a great time, but covered in a lot of sticky sand. They searched for a shower for what felt like ages. Finally they found one and decided it was the best trip ever.
Details

- **Encoders**
  - 500 dimension bidirectional RNN with GRU

- **Surface and Intermediate State decoders**
  - Luong Attention (Luong et al 2015)
  - Input feeding
  - 500 dimension
  - Dropout

- **Intermediate representation**
  - Sequences model
  - Novel Event and Entity Lemmas

- **Optimization**
  - SGD with learning rate decay
  - Early Stopping based on validation accuracy
## Evaluation

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vanilla seq2seq</td>
<td>0.31</td>
<td>0.5</td>
</tr>
<tr>
<td>Sequences Model</td>
<td>0.22</td>
<td>0.5</td>
</tr>
</tbody>
</table>
- **getting an id**
  - **Vanilla seq2seq**: Jamie was going to the bar. She needed to get her ID. She went to the store. She found a ID. She bought it.
  - **Sequences Model**: Jamie was at the bar. He was very drunk. He didn't know what to do. He decided to go anyway. He got drunk.

- **Jamie’s popcorn**
  - **Vanilla seq2seq**: Jamie wanted to make popcorn. He went to the store and bought all the ingredients. He went to the store and bought all the ingredients. He put the kernels in the microwave. He put the kernels in the microwave.
  - **Sequences Model**: Jamie loved popcorn. One day she decided to make popcorn. She bought popcorn and popcorn. It was delicious. Jamie loved popcorn.
Analysis

- Sequences model tends toward shorter texts (negatively impacting BLEU)
- Sometimes forgets the title and goes off topic based on intermediate representation
Solutions

- Make that stuff not happen anymore somehow...
Future Directions

- Self-attention
- Editing Prototypes (Guu and Hashimoto et al. 2017)
- Focus on just “main event”
- From first and last sentence, generate an interesting middle