

# The Gap of Semantic Parsing: A Survey on Automatic Math Word Problem Solvers

Dongxiang Zhang, Lei Wang, Nuo Xu, Bing Tian Dai and Heng Tao Shen

**Abstract**—Automatically solving mathematical word problems (MWP) is challenging, primarily due to the semantic gap between human-readable words and machine-understandable logics. Despite a long history dated back to the 1960s, MWPs has regained intensive attention in the past few years with the advancement of Artificial Intelligence (AI). To solve MWPs successfully is considered as a milestone towards general AI. Many systems have claimed promising results in self-crafted and small-scale datasets. However, when applied on large and diverse datasets, none of the proposed methods in the literatures achieves a high precision, revealing that current MWPs solvers are still far from intelligent. This motivated us to present a comprehensive survey to deliver a clear and complete picture of automatic math problem solvers. In this survey, we emphasize on algebraic word problems, summarize their extracted features and proposed techniques to bridge the semantic gap, and compare their performance in the publicly accessible datasets. We will also cover automatic solvers for other types of math problems such as geometric problems that require the understanding of diagrams. Finally, we will identify several emerging research directions for the readers with interests in MWPs.

**Index Terms**—math word problem, semantic parser, reasoning, survey, natural language processing, machine learning

## 1 INTRODUCTION

Designing an automatic solver for mathematical word problems (MWPs) has a long history dated back to the 1960s [1], [2], [3], and continues to attract intensive research attention. In the past three years, more than 40 publications on this topic have emerged in the premier venues of artificial intelligence. The problem is particularly challenging because there remains a wide semantic gap to parse the human-readable words into machine-understandable logics so as to facilitate quantitative reasoning. Hence, MWPs solvers are broadly considered as good test beds to evaluate the intelligence level of agents in terms of natural language understanding [4], [5] and the successful solving of MWPs would constitute a milestone towards general AI.

We categorize the evolving of MWP solvers into three major stages according to the technologies behind these solvers, as depicted in Figure 1. In the first pioneering stage, roughly from the year 1960 to 2010, systems such as STUDENT [1], DEDUCOM [6], WORDPRO [7] and ROBUST [8], manually craft rules and schemas for pattern matchings. Thereupon, these solvers heavily rely on human interventions and can only resolve a limited number of scenarios that are defined in advance. Those early efforts for automatic understanding of natural language mathematical problems have been thoroughly reviewed in [9]. We exclude them from the scope of this survey paper and focus on the recent technology developments that have not been covered in the previous survey [9].

In the second stage, MWP solvers made use of semantic parsing [10], [11], with the objective of mapping the sentences from problem statements into structured logic representations so as to facilitate quantitative reasoning. It has regained considerable

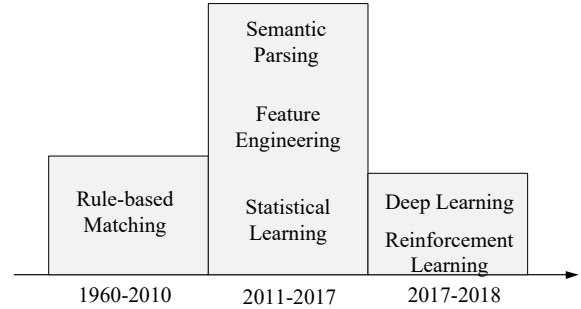


Fig. 1. Technology evolving trend in solving MWPs.

interests from the academic, and a booming number of methods have been proposed in the past years. These methods leveraged various strategies of feature engineering and statistical learning for performance boosting. The authors of these methods also claimed promising results in their public or manually harvested datasets. In this paper, one of our tasks is to present a comprehensive review on the proposed methods in this stage. The methods will be first organized according to the sub-tasks of MWPs which they were designed to solve, such as arithmetic word problem (in Section 2), equation set word problem (in Section 3) and geometric word problem (in Section 5). We then examine the proposed techniques in each sub-task with a clear technical organization and accountable experimental evaluations.

MWP solvers in the third stage originated from an empirical work [12] deserve some special attention. Its experimental results on a large-scale and diversified dataset showed that the status of MWP solvers was not as optimistic as they claimed to be. In fact, the accuracies of many approaches dropped sharply and there is a great room for improvement in this research area. To design more accurate and robust solutions, the subsequent publications are forked into two directions. One is to continue refining the technology of semantic parsing. For instance, Huang et al. proposed a new type of semantic representation to conduct fine-grained inference [13]. The other direction attempts to exploit the

- D. Zhang, L. Wang, N. Xu and H. T. Shen are with the Center for Future Media and School of Computer Science and Engineering, University of Electronic Science and Technology of China. Email: zhangdo@uestc.edu.cn; demolei@outlook.com; nuon-uoxu1128@gmail.com; shenhengtao@hotmail.com
- B. T. Dai is with School of Information Systems, Singapore Management University. Email: btdai@smu.edu.sg

advantages of deep learning models, with the availability of large-scale training datasets. This is an emerging research direction for MWP solvers and we observed two instances, including Deep Neural Solver [14] using recurrent neural network to generate the equation template, and MathDQN [15] relying on the deep reinforcement learning framework. It is expected to witness more and more deep learning based methods to solve MWPs in the future.

To sum up, we present a comprehensive survey to review the MWP solvers proposed in recent years. Researchers in the community can benefit from this survey in the following ways:

- 1) We provide a wide coverage on the math word problems, including arithmetic word problem, equation set problem, geometry word problem and miscellaneous sub-tasks related to automatic math solvers. The practitioners can easily identify all relevant approaches for performance evaluations. We observed that the unawareness of relevant competitors is not uncommon in the past literature. We are positive that the availability of our survey can help avoid such unawareness.
- 2) The solvers designed for arithmetic word problems (AWP) with only one unknown variable and equation set problems (ESP) with multiple unknown variables are often not differentiated by previous works. In fact, the methods proposed for ESP are more general and can be used to solve AWP. In this survey, we clearly identify the difference and organize them in separate sections. We also generalize their methods in terms of the number of operands and the type of operators they are capable to support.
- 3) Feature engineering plays a vital role to bridge the gap of semantic parsing. Almost all MWP solvers state their strategies of crafting effective features, resulting in a very diversified group of features, and there is a lack of clear organization among these features. In this survey, we will be the first to summarize all these proposed features in Table 5.
- 4) As for the fairness on performance evaluations, ideally, there should be a benchmark dataset well accepted and widely adopted by the MWP research community, just like ImageNet [16] for visual object recognition and VQA [17], [18] for visual question answering. Unfortunately, we observed that many approaches tend to compile their own datasets to verify their superiorities, which result in missing of relevant competitors as mentioned above. Tables 2 and 3 integrate the results of existing methods on all the public datasets. After collecting the accuracies that have been reported in the past literature, we observed many empty cells in the table. Each empty cell refers to a missing experiment on a particular algorithm and dataset. In this survey, we make our best efforts to fill the missing results by conducting a considerable number of additional experiments, and guarantee that this survey provides comprehensive comparison and delivers explicit experimental analysis.

The remainder of the paper is organized as follows. We first review the arithmetic word problem solvers in Section 2, followed by equation set problem solvers in Section 3. Since feature engineering deserves special attention, we summarize the extracted features as well as the associated pre-processing techniques in Section 4. The geometric word problem solvers are reviewed in Section 5. We also cover miscellaneous automatic solvers related to math problems in Section 6. We conclude the paper and point out several future directions in MWPs that are

worth examination in the final section.

## 2 ARITHMETIC WORD PROBLEM SOLVER

The arithmetic word problems are targeted for elementary school students. The input is the text description for the math problem, represented in the form of a sequence of  $k$  words  $\langle w_0, w_1, \dots, w_k \rangle$ . There are  $n$  quantities  $q_1, q_2, \dots, q_n$  mentioned in the text and an unknown variable  $x$  whose value is to be resolved. Our goal is to extract the relevant quantities and map this problem into an arithmetic expression  $E$  whose evaluation value provides the solution to the problem. There are only four types of fundamental operators  $O = \{+, -, \times, \div\}$  involved in the expression  $E$ .

An example of arithmetic word problem is illustrated in Figure 2. The relevant quantities to be extracted from the text include 17, 7 and 80. The number of hours spent on the bike is the unknown variable  $x$ . To solve the problem, we need to identify the correct operators between the quantities and their operation order such that we can obtain the final equation  $17 + 7x = 80$  or expression  $x = (80 - 17) \div 7$  and return 9 as the solution to this problem.

Word Problem
Oceanside Bike Rental Shop charges 17 dollars plus 7 dollars an hour for renting a bike. Tom paid 80 dollars to rent a bike. How many hours did he pay to have the bike checked out?
Equation
$17 + (7 * x) = 80$
Solution
$x = 9$

Fig. 2. An example of arithmetic word problem.

In this section, we consider feature extraction as a black box and focus on the high-level algorithms and models. The details of feature extraction will be comprehensively present in Section 4. We classify existing algebra word problem solvers into three categories: rule-based, statistic-based and tree-based methods.

### 2.1 Rule-based Methods

The early approaches to math word problems are rule-based systems based on hand engineering. Published in 1985, WORDPRO [7] solves one-step arithmetic problems. It predefines four types of schemas, including *change-in*, *change-out*, *combine* and *compare*. The problem text is transformed into a set of propositions and the answer is derived with simple reasoning based on the propositions. Another system ROBUST, developed by Bakman [8], could understand free-format multi-step arithmetic word problems. It further expands the *change* schema of WORDPRO [7] into six distinct categories. The problem text is split into sentences and each sentence is mapped to a proposition. Yun et al. also proposed to use schema for multi-step math problem solving [19]. However, the implementation details are not explicitly revealed in [19]. Since these systems have been out of date, we only provide such a brief overview to cover the

representative ones. Readers can refer to [9] for a comprehensive survey of early rule-driven systems for automatic understanding of natural language math problems.

## 2.2 Statistic-based Methods

The statistic-based methods leverage traditional machine learning models to identify the entities, quantities and operators from the problem text and yield the numeric answer with simple logic inference procedure. The scheme of quantity entailment proposed in [20] can be used to solve arithmetic problems with only one operator. It involves three types of classifiers to detect different properties of the word problem. The *quantity pair classifier* is trained to determine which pair of quantities would be used to derive the answer. The *operator classifier* picks the operator  $op \in \{+, -, \times, \div\}$  with the highest probability. The *order classifier* is relevant only for problems involving subtraction or division because the order of operands matters for these two types of operators. With the inferred expression, it is straightforward to calculate the numeric answer for the simple math problem.

To solve math problems with multi-step arithmetic expression, the statistic-based methods require more advanced logic templates. This usually incurs additional overhead to annotate the text problems and associate them with the introduced template. As an early attempt, ARIS [21] defines a logic template named *state* that consists of a set of entities, their containers, attributes, quantities and relations. For example, “*Liz has 9 black kittens*” initializes the number of *kitten* (referring to an entity) with *black* color (referring to an attribute) and belonging to *Liz* (referring to a container). The solution splits the problem text into fragments and tracks the update of the states by verb categorization. More specifically, the verbs are classified into seven categories: *observation*, *positive*, *negative*, *positive transfer*, *negative transfer*, *construct* and *destroy*. To train such a classifier, we need to annotate each split fragment in the training dataset with the associated verb category. Another drawback of ARIS is that it only supports addition and subtraction. [22] follows a similar processing logic to ARIS. It predefines a corpus of logic representation named *schema*, inspired by [8]. The sentences in the text problem are examined sequentially until the sentence matches a schema, triggering an update operation to modify the number associated with the entities.

In [23], Mitra et al. proposed a new logic template named *formula*. Three types of formulas are defined, including *part whole*, *change* and *comparison*, to solve problems with addition and subtraction operators. For example, the text problem “*Dan grew 42 turnips and 38 cantelopes. Jessica grew 47 turnips. How many turnips did they grow in total?*” is annotated with the part-whole template:  $\langle \text{whole} : x, \text{parts} : \{42, 47\} \rangle$ . To solve a math problem, the first step connects the assertions to the formulas. In the second step, the most probable formula is identified using the log-linear model with learned parameters and converted into an algebraic equation.

Another type of annotation is introduced in [24], [25] to facilitate solving a math problem. A group of *logic forms* are predefined and the problem text is converted into the logic form representation by certain mapping rules. For instance, the sentence “*Fred picks 36 limes*” will be transformed into  $\text{verb}(v_1, \text{pick}) \ \& \ \text{nsubj}(v_1, \text{Fred}) \ \& \ \text{dobj}(v_1, n_1) \ \& \ \text{head}(n_1, \text{limes}) \ \& \ \text{nummod}(n_1, 36)$ . Finally, logic inference is performed on the derived logic statements to obtain the answer.

To sum up, these statistical-based methods have two drawbacks that limit their usability. First, it requires additional annotation overhead that prevents them from handling large-scale datasets. Second, these methods are essentially based on a set of pre-defined templates, which are brittle and rigid. It will take great efforts to extend the templates to support other operators like multiplication and division. It is also not robust to diversified datasets. In the following, we will introduce the tree-based solutions, which are widely adopted and become the mainstreaming solutions to arithmetic word problems.

## 2.3 Tree-Based Methods

The arithmetic expression can be naturally represented as a binary tree structure such that the operators with higher priority are placed in the lower level and the root of the tree contains the operator with the lowest priority. The idea of tree-based approaches [26], [27], [28], [15] is to transform the derivation of the arithmetic expression to constructing an equivalent tree structure step by step in a bottom-up manner. One of the advantages is that there is no need for additional annotations such as equation template, tags or logic forms. Figure 3 shows two tree examples derived from the math word problem in Figure 2. One is called *expression tree* that is used in [26], [28], [15] and the other is called *equation tree* in [27]. These two types of trees are essentially equivalent and result in the same solution, except that equation tree contains a node for the unknown variable  $x$ .

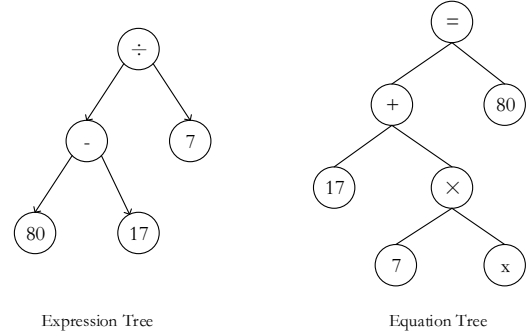


Fig. 3. Examples of expression tree and equation tree for Figure 2.

The overall algorithmic framework among the tree-based approaches consists of two processing stages. In the first stage, the quantities are extracted from the text and form the bottom level of the tree. The candidate trees that are syntactically valid, but with different structures and internal nodes, are enumerated. In the second stage, a scoring function is defined to pick the best matching candidate tree, which will be used to derive the final solution. A common strategy among these algorithms is to build a local classifier to determine the likelihood of an operator being selected as the internal node. Such local likelihood is taken into account in the global scoring function to determine the likelihood of the entire tree.

Roy et al. [26] proposed the first algorithmic approach that leverages the concept of expression tree to solve arithmetic word problems. Its first strategy to reduce the search space is training a binary classifier to determine whether an extracted quantity is relevant or not. Only the relevant ones are used for tree construction and placed in the bottom level. The irrelevant quantities are discarded. The tree construction procedure is mapped to a collection of simple prediction problems, each

determining the lowest common ancestor operation between a pair of quantities mentioned in the problem. The global scoring function for an enumerated tree takes into account two terms. The first one, denoted by  $\phi(q)$ , is the likelihood of quantity  $q$  being irrelevant, i.e.,  $q$  is not used in creating the expression tree. In the ideal case, all the irrelevant quantities are correctly predicted with high confidence, resulting in a large value for the sum of  $\phi(q)$ . The other term, denoted by  $\phi(op)$ , is the likelihood of selecting  $op$  as the operator for an internal tree node. With these two factors,  $\text{Score}(E)$  is formally defined as

$$\text{Score}(E) = w_1 \sum_{q \in I(E)} \phi(q) + \sum_{op \in \mathcal{N}} \phi(op) \quad (1)$$

where  $I(E)$  is the group of irrelevant quantities that are not included in expression  $E$ , and  $\mathcal{N}$  refers to the set of internal tree nodes. To further reduce the tree enumeration space, beam search is applied in [26]. To generate the next state  $T'$  from the current partial tree, the algorithm avoids choosing all the possible pairs of terms and determining their operator. Instead, only top- $k$  candidates with the highest partial scores are retained. Experimental results with  $k = 200$  show that the strategy achieves a good balance between accuracy and running time. In [29], the authors publish the service as a web tool and it can respond promptly to a math word problem.

The solution in [27], named ALGES, differs from [26] in two major ways. First, it adopts a more brutal-force manner to exploit all the possible equation trees. More specifically, ALGES does not discard irrelevant quantities, but enumerates all the syntactically valid trees. Integer Linear Programming (ILP) is applied as it can help enforce the constraints such as syntactic validity, type consistence and domain specific simplicity considerations. Consequently, its computation cost is dozens of times higher than that in [27], according to an efficiency evaluation in [15]. Second, its scoring function is different from Equation 1. There is no need for the term  $\phi(q)$  because ALGES does not build a classifier to check the quantity relevance. Besides the monotonic aggregation of the likelihood from local operator classifiers, the scoring function incorporates a new term  $\phi(P) = \theta^T f_P$  to assign a coherence score for the tree instance. Here,  $f_P$  is the global feature extracted from a problem text  $P$ , and  $\theta$  refers to the parameter vector.

The goal of [30] is also to build an equation tree by parsing the problem text. It makes two assumptions that can simplify the tree construction, but also limit its applicability. First, the final output equation form is restricted to have at most two variables. Second, each quantity mentioned in the sentence can be used at most once in the final equation. The tree construction procedure consists of a pipeline of predictors that identify irrelevant quantities, recognize grounded variables, and generate the final equation tree. With customized feature selection and SVM based classifier, the relevant quantities and variables are extracted and used as the leaf nodes of the equation tree. The tree is built in a bottom-up manner. It is worth noting that to reduce the search space and simplify the tree construction, only adjacent nodes are combined to generate their parent node.

UnitDep [28] can be viewed as an extension work of [26] by the same authors. An important concept, named Unit Dependency Graph (UDG), is proposed to enhance the scoring function. The vertices in UDG consist of the extracted quantities. If the quantity correspond to a rate (e.g., 8 dollars per hour), the vertex is marked as RATE. There are six types of edge relations to be considered,

such as whether two quantities are associated with the same unit. Building the UDG requires additional annotation overhead as we need to train two classifiers for the nodes and edges. The node classifier determines whether a node is associated with a rate. The edge classifier predicts the type of relationship between any pair of quantity nodes. Given a valid unit dependency graph  $G$  generated by the classifiers, its likelihood is defined as

$$\phi(G) = \sum_{v \in G \wedge \text{LABEL}(v) = \text{rate}} P(v) + \lambda \sum_{e \in G} P(e) \quad (2)$$

In other words, we sum up the prediction probability for the RATE nodes and all the edges. The new scoring function for an expression tree extends Equation 1 by incorporating  $\phi(G)$ . Rules are defined to enforce the rate consistence between an expression tree  $T$  and a candidate graph  $G$ . For example, if  $v_i$  is the only node in the tree that is labeled RATE and it appears in the question, there should not exist a path from the leaf node to the root which only contains operators of addition and subtraction. Finally, the candidate graph  $G$  with the highest likelihood and rate-consistent with  $T$  is used to calculate the total score of  $T$ .

In [15], Wang et al. made the first attempt of applying deep reinforcement learning to solve arithmetic word problems. The motivation is that deep Q-network has witnessed success in solving various problems with big search space such as playing text-based games [31], information extraction [32], text generation [33] and object detection in images [34]. To fit the math problem scenario, they formulate the expression tree construction as a Markov Decision Process and propose the MathDQN that is customized from the general deep reinforcement learning framework. Technically, they tailor the definitions of states, actions, and reward functions which are key components in the reinforcement learning framework. By using a two-layer feed-forward neural network as the deep Q-network to approximate the Q-value function, the framework learns model parameters from the reward feedback of the environment. Compared to the aforementioned approaches, MathDQN iteratively picks the best operator for two selected quantities. This procedure can be viewed as beam search with  $k = 1$  when exploiting candidate expression trees. Its deep Q-network acts as the operator classifier and guides the model to select the most promising operator for tree construction.

## 2.4 Dataset Repository and Performance Analysis

The accuracy of arithmetic word problems is evaluated on the datasets that are manually harvested and annotated from online websites. These datasets are small-scale and contain hundreds of math problems. In this subsection, we make a summary on the datasets that have been used in the aforementioned datasets. Moreover, we organize the performance results on these datasets into one unified table. We also make our best efforts to conduct additional experiments. The new results are highlighted in blue color. In this way, readers can easily identify the best performers in each dataset.

### 2.4.1 Datasets

There have been a number of datasets collected for the arithmetic word problems. We present their descriptions in the following and summarize the statistics of the datasets in Table 1.

- 1) **A12** [21]. There are 395 single-step or multi-step arithmetic word problems for the third, fourth, and fifth graders. It

involves problems that can be solved with only addition and subtraction. The dataset is harvested from two websites: math-aids.com and ixl.com and comprises three subsets: MA1 (from math-aids.com), IXL (from ixl.com) and MA2 (from math-aids.com). Among them, IXL and MA2 are more challenging than MA1 because IXL contains more information gaps and MA2 includes more irrelevant information in its math problems.

- 2) **IL** [26]. The problems are collected from websites k5learning.com and dadsworksheets.com. The problems that require background knowledge (e.g., “apple is fruit” and “a week comprises 7 days”) are pruned. To improve the diversity, the problems are clustered by textual similarity. For each cluster, at most 5 problems are retained. Finally, the dataset contains 562 single-step word problems with only one operator, including addition, subtraction, multiplication, and division.
- 3) **CC** [26]. The dataset is designed for multi-step math problems. It contains 600 multi-step problems without irrelevant quantities, harvested from commoncoresheets.com. The dataset involves various combinations of four basic operators, including (a) addition followed by subtraction; (b) subtraction followed by addition; (c) addition and multiplication; (d) addition and division; (e) subtraction and multiplication; and (f) subtraction and division. It is worth noting that this dataset does not incorporate irrelevant quantities in the problem text. Hence, there is no need to apply the quantity relevance classifier for the algorithms containing this component.
- 4) **SingleEQ** [27]. The dataset contains both single-step and multi-step arithmetic problems and is a mixture of problems from a number of sources, including math-aids.com, k5learning.com, ixl.com and a subset of the data from **AI2**. Each problem involves operators of multiplication, division, subtraction, and addition over non-negative rational numbers.
- 5) **AllArith** [28]. The dataset is a mixture of the data from **AI2**, **IL**, **CC** and **SingleEQ**. All mentions of quantities are normalized into digit representation. To capture how well the automatic solvers can distinguish between different problem types, near-duplicate problems (with over 80% match of unigrams and bigrams) are removed. Finally, there remain 831 math problems.
- 6) **Dolphin-S**. This is a subset of Dolphin18K [12] which originally contains 18,460 problems and 5,871 templates with one or multiple equations. The problems whose template is associated with only one problem are extracted as the dataset of **Dolphin-S**. It contains 115 problems with single operator and 6,955 problems with multiple operators.
- 7) **Math23K** [14]. The dataset contains Chinese math word problems for elementary school students and is crawled from multiple online education websites. Initially, 60,000 problems with only one unknown variable are collected. The equation templates are extracted in a rule-based manner. To ensure high precision, a large number of problems that do not fit the rules are discarded. Finally, 23,161 math problems with 2,187 templates are remained.

#### 2.4.2 Performance Analysis

Given the aforementioned datasets, we merge the experimental results reported from previous works into one table. Such a unified organization can facilitate readers in identifying the methods with

TABLE 1  
Statistics of arithmetic word problem datasets.

Dataset	# problems	# single-op	# multi-op	operators $O$
<b>MA1</b>	134	112	22	{+, −}
<b>IXL</b>	140	119	21	{+, −}
<b>MA2</b>	121	96	25	{+, −}
<b>AI2</b>	395	327	68	{+, −}
<b>IL</b>	562	562	0	{+, −, ×, ÷}
<b>CC</b>	600	0	600	{+, −, ×, ÷}
<b>SingleEQ</b>	508	390	118	{+, −, ×, ÷}
<b>AllArith</b>	831	634	197	{+, −, ×, ÷}
<b>Dolphin-S</b>	7,070	115	6,955	{+, −, ×, ÷}
<b>Math23K</b>	23,129	3,139	19,990	{+, −, ×, ÷}

superior performance in each dataset. As shown in Table 2, the rows refer to the corpus of datasets and the columns are the statistic-based and tree-based methods. The cells are filled with the accuracies of these algorithms when solving math word problems in different datasets. We conduct additional experiments to cover all the cells by the tree-based solutions. These new experiment results are highlighted in blue color. Those with missing value are indicated by “-” and it means that there was no experiment conducted for the algorithm in the particular dataset. The main reason is that they require particular efforts on logic templates and annotations, which are very trivial and cumbersome for experiment reproduction. There is no algorithm comparison for the dataset **Math23K** because the problem text is in Chinese and the feature extraction technologies proposed in the statistic-based and tree-based approaches are not applicable. From the results in Table 2, we derive the following observations worth noting and provide reasonings to explain the results.

First, the statistic-based methods with advanced logic representation, such as Schema [22], Formula [23] and LogicForm [24], [25], achieve dominating performance in the **AI2** dataset. Their superiority is primarily owned to the additional efforts on annotating the text problem with more advanced logic representation. These annotations allow them to conduct fine-grained reasoning. In contrast, ARIS [21] does not work as good because it focuses on “change” schema of quantities and does not fully exploit other schemas like “compare” [22]. Since there are only hundreds of math problems in the datasets, it is feasible to make an exhaustive scan on the math problems and manually define the templates to fit these datasets. For instance, all quantities and the main-goal are first identified by rules in LogicForm [24], [25] and explicitly associated with their role-tags. Thus, with sufficient human intervention, the accuracy of statistic-based methods in **AI2** can boost to 88.64%, much higher than that of tree-based methods. Nevertheless, these statistic-based methods are considered as brittle and rigid [12] and not scalable to handle large and diversified datasets, primarily due to the heavy annotation cost to train an accurate mapping between the text and the logic representation.

Second, the results of tree-based methods in **AI2**, **IL** and **CC** are collected from [15] where the same experimental setting of 3-fold cross validation is applied. It is interesting to observe that ALGES [27], ExpressionTree [26] and UNITDEP [28] cannot perform equally well on the three datasets. ALGES works poorly in **AI2** because irrelevant quantities exist in its math problems and ALGES is not trained with a classifier to get rid of them. However, it outperforms ExpressionTree and UNITDEP by a wide

TABLE 2  
Accuracy of statistic-based and tree-based methods in solving arithmetic problems.

Methods		Publish Year	AI2	IL	CC	SingleEQ	AllArith	Dolphin-S
Statistic-based	ARIS [21]	2014	77.7	-	-	48	-	-
	Schema [22]	2015	<b>88.64</b>	-	-	-	-	-
	Formula [23]	2016	86.07	-	-	-	-	-
	LogicForm [24], [25]	2016	84.8	<b>80.1</b>	53.5	-	-	-
Tree-based	ALGES [27]	2015	52.4	72.9	65	72	60.4	-
	ExpressionTree [26]	2015	72	73.9	45.2	<b>66.38</b>	79.4	<b>26.11</b>
	UNITDEP [28]	2017	56.2	71.0	53.5	<b>72.25</b>	<b>81.7</b>	<b>28.78</b>
	MathDQN [15]	2018	78.5	73.3	<b>75.5</b>	<b>52.96</b>	<b>72.68</b>	<b>30.06</b>

margin in the **CC** dataset because **CC** does not involve irrelevant quantities. In addition, this dataset only contains multi-step math problems. ALGES exploits the whole search space to enumerate all the possible trees, whereas ExpressionTree and ALGES use beam search for efficiency concern. UNITDEP does not work well in **AI2** because this dataset only involves operators  $+$  and  $-$  and the unit dependency graph does not take effect. Moreover, its proposed context feature poses a negative impact on the **AI2** dataset. After removing this feature from the input vector fed to the classifier, the accuracy in **AI2** increases from 56.2% to 74.7%, but the result in **CC** drops from the current accuracy of 53.5% to 47.3%. Such an observation implies the limitation of hand-crafted features in UNITDEP. Among the three datasets **AI2**, **IL** and **CC**, MathDQN [15] achieves leading or comparable performance. In the **CC** dataset, which contains only multi-step problems and is considered as the most challenging one, MathDQN yields remarkable improvement and boosts the accuracy from 65% (derived by ALGES) to 75.5%. This is because MathDQN models the tree construction as Markov Decision Process and leverage the strengths of deep Q-network (DQN). By using a two-layer feed-forward neural network as the deep Q-network to approximate the Q-value function, the framework learns model parameters from the reward feedback of the environment. Consequently, the RL framework demonstrates higher generality and robustness than the other tree-based methods when handling complicated scenarios. In the **IL** dataset, its performance is not superior to ExpressionTree as **IL** only contains one-step math problems. There is no need for hierarchical tree construction and cannot expose the strength of Markov Decision Process in MathDQN or the exhaustive enumeration strategy in ALGES.

As to the datasets of SingleEQ and AllArith, UNITDEP is a winner in both datasets, owing to the effectiveness of the proposed unit dependency graph (UDG). In the math problems with operators  $\{\times, \div\}$ , the unit and rate are important clues to determine the correct quantities and operators in the math expression. The UDG poses constraints on unit compatibility to filter the false candidates in the expression tree construction. It can alleviate the brittleness of the unit extraction system, even though it requires additional annotation overhead in order to induce UDGs.

Last but not the least, these experiments were conducted on small-scale datasets and their performances on larger and more diversified datasets remain unclear. Recently, Huang et al. have noticed the gap and released Dolphin18K [12] which contains 18,460 problems and 5,871 templates with one or multiple equations. The findings in [12] are astonishing. The accuracies of existing approaches for equation set problems, which will be introduced in the next section, degrade sharply to less than

25%. These methods cannot even perform better than a simple baseline that first uses text similarity to find the most similar text problem in the training dataset and then fills the number slots in its associated equation template. To discover the performance of existing arithmetic problem solvers in large-scale datasets, we conduct similar experiments to examine the performance of tree-based approaches in **Dolphin-S**. The results are shown in the last two columns of Table 2, leading to two conclusions. First, the overall performances of existing tree-based arithmetic problem solvers are not promising in large-scale and diversified datasets. There is still great room for improvement in the research area of arithmetic math word problem solver. Second, MathDQN exhibits superior performance over its two tree-based competitors, verifying its robustness by using the deep reinforcement learning framework.

### 3 EQUATION SET SOLVER

The equation set problems are much more challenging because they involve multiple unknown variables to resolve and require to formulate a set of equations to obtain the final solution. The aforementioned arithmetic math problem can be viewed as a simplified variant of equation set problem with only one unknown variable. Hence, the methods introduced in this section can also be applied to solve the problems in Section 2.

Figure 4 shows an example of equation set problem. There are two unknown variables, including the acres of corn and the acres of wheat, to be inferred from the text description. A standard solution to this problem is to use variables  $x$  and  $y$  to represent the number of corn and wheat, respectively. From the text understanding, we can formulate two equations  $42x + 30y = 18600$  and  $x + y = 500$ . Finally, the values of  $x$  and  $y$  can be inferred.

Compared to the arithmetic problem, the equation set problem contains more numbers of unknown variables and numbers in the text, resulting in a much larger search space to enumerate valid candidate equations. Hence, the methods designed for arithmetic problems can be hardly applied to solve equation set problems. For instance, the tree-based methods assume that the objective is to construct a single tree to maximize a scoring function. They require substantial revision to adjust the objective to building multiple trees which has exponentially higher search space. This would be likely to degrade the performance. In the following, we will review the existing methods, categorize them into four groups from the technical perspective, and examine how they overcome the challenge.

Word Problem
The Johnson Farm has 500 acres of land allotted for cultivating corn and wheat. The cost of cultivating corn and wheat is 42 dollars for corn and 30 dollars for wheat. Mr. Johnson has 18,600 dollars available for cultivating these crops. If he used all the land and entire budget, how many acres of corn and how many acres of wheat should he plant?
Equations
$2x + 30y = 18600$ $x + y = 500$
Solution
$x = 300 \quad y = 200$

Fig. 4. An example of equation set problem.

### 3.1 Parsing-Based Methods

The work of [35] can be viewed as an extension of tree-based approaches to solve a math problem with multiple equations. Since the objective is no longer to build an equation tree, a meaning representation language called *DOL* is designed as the structural semantic representation of natural language text. The core component is a semantic parser that transforms the textual sentences into DOL trees. The parsing algorithm is based on context-free grammar (CFG) [36], [37], a popular mathematical system for modeling constituent structure in natural languages. For every DOL node type, the lexicon and grammar rules are constructed in a semi-supervised manner. The association between math-related concepts and their grammar rules is manually constructed. Finally, the CFG parser is built on top of 9,600 grammar rules. During the parsing, a score is calculated for each DOL node and the derivation of the DOL trees with the highest score is selected to obtain the answer via a reasoning module.

### 3.2 Similarity-Based Methods

The work of [12] plays an important role in the research line of automatic math problem solvers because it rectifies the understanding of technology development in this area. It, for the first time, examines the performance of previous approaches in a large and diversified dataset and derives astonishing experimental findings. The methods that claimed to achieve an accuracy higher than 70% in a small-scale and self-collected dataset exhibit very poor performance in the new dataset. In other words, none of the methods proposed before [12] is really general and robust. Hence, the authors reach a conclusion that the math word problems are still far from being satisfactorily solved.

A new baseline method based on text similarity, named SIM, is proposed in [12]. In the first step, the problem text is converted into a word vector whose values are the associated TF-IDF scores [38]. The similarity between two problems (one is the query problem to solve and the other is a candidate in the training dataset with known solutions) is calculated by the Jaccard similarity between their converted vectors. The problem with the highest similarity score is identified and its equation template is used to help solve

the query math problem. In the second step, the unknown slots in the template are filled. With the availability of a large training dataset, the number filling is conducted in a simple and effective way. It finds an instance in the training dataset associated with the same target template and the minimum edit-distance to the query problem, and aligns the numbers in these two problems with ordered and one-to-one mapping. It is considered a failure if these two problems do not contain the same number of quantities.

### 3.3 Template Based Methods

There are two methods [39], [40] that pre-define a collection of equation set templates. Each template contains a set of *number slots* and *unknown slots*. The *number slots* are filled by the numbers extracted from the text and the *unknown slots* are aligned to the nouns. An example of template may look like

$$\begin{aligned}
 u_1 + u_2 - n_1 &= 0 \\
 n_2 \times u_1 + n_3 \times u_2 - n_4 &= 0
 \end{aligned}$$

where  $n_i$  is a number slot and  $u_i$  represents an unknown variable.

To solve an equation set problem, these approaches first identify a candidate template from the corpus of pre-defined templates. The next task is to fill the *number slots* and *unknown slots* with the information extracted from the text. Finally, the instance with the highest probability is returned. This step often involves a scoring function or a rank-aware classifier such as RankSVM [41]. A widely-adopted practice is to define the probability of each instance of derivation  $y$  based on the feature representation  $x$  for a text problem and a parameter vector  $\theta$ , as in [30], [39], [40]:

$$p(y|x;\theta) = \frac{e^{\theta \cdot \Phi(x,y)}}{\sum_{y' \in \mathcal{Y}} e^{\theta \cdot \Phi(x,y')}}$$

With the optimal derivation instance  $y^{opt}$ , we can obtain the final solution.

In [39], the objective is to maximize the total probabilities of  $y$  that leads to the correct answer. The latent variables  $\theta$  are learned by directly optimizing the marginal data log-likelihood. More specifically, L-BFGS [42] is used to optimize the parameters. The search space is exponential to the number of slots because each number in the text can be mapped to any *number slot* and the nouns are also candidates for the *unknown slots*. In practice, the search space is too huge to find the optimal  $\theta$  and beam search inference procedure is adopted to prevent enumerating all the possible  $y$  leading to the correct answer. For the completion of each template, the next slot to be considered is selected according to a pre-defined canonicalized ordering and only top-k partial derivations are maintained.

In [40], Zhou et al. proposed an enhanced algorithm for the template-based learning framework. First, they only consider assigning the number slots with numbers extracted from the text. The underlying logic is that when the number slots have been processed, it would be an easy task to fill the unknown slots. In this way, the hypothesis space can be significantly reduced. Second, the authors argue that the beam search used in [39] does not exploit all the training samples, and its resulting model may be sub-optimal. To resolve the issue, the max-margin objective [43] is used to train the log-linear model. The training process is turned into a QP problem that can be efficiently solved with the constraint generation algorithm [44].

Since the annotation of equation templates is expensive, a key challenge to KAZB and ZDC is the lack of sufficient annotated data. To resolve the issue, in [45], Upadhyay et al. attempted to exploit the large number of algebra word problems that have been posted and discussed in online forums. These data are not explicitly annotated with equation templates but their numeric answers are extracted with little or no manual effort. The goal of [45] is to improve a strong solver trained by fully annotated data with a large number of math problems with noisy and implicit supervision signals. The proposed *MixedSP* algorithm makes use of both explicit and implicit supervised examples mixed at the training stage and learns the parameters jointly. With the learned model to formulate the mapping between an algebra word problem and an equation template, the math problem solving strategy is similar to KAZB and ZDC. All the templates in the training set have to be exploited to find the best alignment strategy.

The aforementioned template-based methods suffer from two drawbacks [13]. First, the math concept is expressed as an entire template and may fail to work well when the training instances are sparse. Second, the learning process relies on lexical and syntactic features such as the dependency path between two slots in a template. Such a huge and sparse feature space may play a negative impact on effective feature learning. Based on these two arguments, FG-Expression [13] parses an equation template into fine-grained units, called *template fragment*. Each template is represented in a tree structure as in Figure 3 and each fragment represents a sub-tree rooted at an internal node. The main objective and challenge in [13] are learning an accurate mapping between textual information and template fragments. For instance, a text piece “20% discount” can be mapped to a template fragment  $1 - n_1$  with  $n_1 = 0.2$ . Such mappings are extracted in a semi-supervised way from training datasets and stored as part of the sketch for templates. The proposed solution to a math problem consists of two stages. First, RankSVM model [41] is trained to select top-k templates. The features used for the training incorporate textual features, quantity features and solution features. It is worth noting that the proposed template fragment is applied in the feature selection for the classifier. The textual features preserve one dimension to indicate whether the problem text contains textual expressions in each template fragment. In the second stage, the alignment is conducted for the k templates and the one with the highest probability is used to solve the problem. The features and rank-based classifier used to select the best alignment are similar to those used in the first stage. Compared to the previous template-based methods, FG-Expression also significantly reduces the search space because only top-k templates are examined whereas previous methods align numbers for all the templates in the training dataset.

### 3.4 DL-Based Methods

In recent years, deep learning (DL) has witnessed great success in a wide spectrum of “smart” applications, such as video captioning [46], video event recognition [47], human action recognition [48], visual question answering [49], and question answering with knowledge base [50]. The main advantage is that with sufficient amount of training data, DL is able to learn an effective feature representation in a data-driven manner without human intervention. It is not surprising to notice that several efforts have been attempted to apply DL for math word problem solving. Deep Neural Solver (DNS) [14] is the first deep learning

based algorithm that does not rely on hand-crafted features. This is a milestone contribution because all the previous methods (including MathDQN) require human intelligence to help extract features that are effective. The deep model used in DNS is a typical sequence to sequence (seq2seq) model [51], [52], [53]. The words in the problem are vectorized into features through word embedding techniques [54], [55]. In the encoding layer, GRU [56] is used as the Recurrent Neural Network (RNN) to capture word dependency because compared to LSTM [57], GRU has fewer parameters in the model and is less likely to be over-fitting in small datasets. This seq2seq model translates math word problems to equation templates, followed by a number mapping step to fill the slots in the equation with the quantities extracted from the text. To ensure that the output equations by the model are syntactically correct, five rules are pre-defined as validity constraints. For example, if the  $i^{\text{th}}$  character in the output sequence is an operator in  $\{+, -, \times, \div\}$ , then the model cannot result in  $c \in \{+, -, \times, \div, \cdot, =\}$  for the  $(i+1)^{\text{th}}$  character.

To further improve the accuracy, DNS enhances the model in two ways. First, it builds a LSTM-based binary classification model to determine whether a number is relevant. This is similar to the relevance model trained in ExpressionTree [26] and UNITDEP [28]. The difference is that DNS uses LSTM as the classifier with unsupervised word-embedding features whereas ExpressionTree and UNITDEP use SVM with hand-crafted features. Second, the seq2seq model is integrated with a similarity-based method [12] introduced in Section 3.2. Given a pre-defined threshold, the similarity-based retrieval strategy is selected as the solver if the maximal similarity score is higher than the threshold. Otherwise, the seq2seq model is used to solve the problem. Another follow-up of DNS was proposed recently in [58]. Instead of using GRU and LSTM, the math solver examines the performance of other seq2seq models when applied in mapping the problem text to equation templates. In particular, two models including BiLSTM [57] and structured self-attention [59], were examined respectively for the equation template classification task. Results show that both models achieve comparable performance.

### 3.5 Dataset Repository and Performance Analysis

Similar to the organization of Section 2, we summarize the dataset repository and performance analysis for the equation set solvers.

#### 3.5.1 Benchmark Datasets

There have been four datasets specifically collected for the equation set problems that involve multiple unknown variables. We present their descriptions in the following and summarize the statistics of the datasets in Table 3. We use  $\frac{\# \text{ of problems}}{\# \text{ of templates}}$  to report the average number of problems associated with each template. We noticed that in each dataset, a small fraction of problems are associated with one unknown variable in the template. Thus, we also report the number of single-equation problems in each dataset.

- 1) **ALG514** [39]. The dataset is crawled from Algebra.com, a crowd-sourced tutoring website. The problems are posted by students. The problems with information gap or require explicit background knowledge are discarded. Consequently, a set of 1024 questions is collected and cleaned by crowd-workers in Amazon Mechanical Turk. These problems are further filtered as the authors require each equation template



to appear for at least 6 times. Finally, 514 problems are left in the dataset.

- 2) **Dolphin1878** [35]. Its math problems are crawled from two websites: `algebra.com` and `answers.yahoo.com`. For math problems in `answers.yahoo.com`, the math equations and answers are manually added by human annotators. Finally, the dataset combined from the two sources contains 1,878 math problems with 1183 equation templates.
- 3) **DRAW1K** [60]. The authors of **DRAW1K** argued that **Dolphin1878** has limited textual variations and lacks narrative. This motivated them to construct a new dataset that is diversified in both vocabularies and equation systems. With these two objectives, they constructed **DRAW1K** with exactly 1,000 linear equation problems that are crawled and filtered from `algebra.com`.
- 4) **Dolphin18K** [12]. The dataset is collected and rectified mainly from the math category of Yahoo! Answers<sup>1</sup>. The problems, equation system annotations, and answers are extracted semi-automatically, with great intervention of human efforts. The procedure consists of four stages: removing irrelevant problems, cleaning problem text, extracting gold answers and constructing equation system annotations. The harvested dataset is so far the largest one, with 18,460 problems and 5,871 equation templates.

### 3.5.2 Performance Analysis

The performances of the equation set solvers on the existing datasets are reported in Table 4. From the experimental results, we derive the following observations and discussions.

First, there are many empty cells in the table. In the ideal case, the algorithms should be conducted on all the available benchmark datasets and compared with all the previous competitors. The reasons are multi-fold, such as limitation of the implementation (e.g., as Upadhyay stated in [45], they could not run ZDC on **DRAW1K** because ZDC can only handle limited types of equation systems), delayed release of the dataset (e.g., the **Dolphin18K** dataset has not been released when the work of DNS is published) or unfitness in certain scenarios (e.g., the experiments of FG-Expression [13] were only conducted on **Dolphin18K** because the authors considered that the previous datasets are not suitable due to their limitation in scalability and diversity). It is noticeable that such an incomplete picture brings difficulty to judge the performance and may miss certain insightful findings.

Second, **ALG514** is the smallest dataset and also the most widely adopted dataset for performance evaluation. Among the template-based methods, MixedSP outperforms KAZB and ZDC because it benefits from the mined implicit supervision from an external source of additional 2,000 samples. As reported in [45], if only the explicit dataset (i.e., the problems in **ALG514**) is used, its performance is slightly inferior to ZDC. A possible reason to explain this is that ZDC uses a richer set of features based on POS tags, coreference and dependency parses. In contrast, MixedSP only uses features based on POS tags. It is also interesting to see that SIM and DNS obtain the same accuracy on **ALG514**. This is the dataset is too small to train an effective deep learning model. The reported accuracy of seq2seq model is only 16.1% in **ALG514**. DNS is a hybrid approach that combines a seq2seq model and similarity retrieval model. It means the deep learning

model does not take any effect when handling problems in **ALG514**.

Finally, the datasets of **Dolphin1878** and **DRAW1K** are released for the approaches of DOL and MixedSP, respectively. In the experimental settings, simple baselines such as SIM based on textual similarity or KAZB which is the earliest template-based method, are selected. It is not surprising to see that **Dolphin1878** and **DRAW1K** outperform their competitors by a large margin. Nevertheless, the research for equation set solvers has shifted to proposing methods that can work well in a large and diversified dataset such as **Dolphin18K**. We implemented our own version of DNS and evaluated its performance on the large dataset. Unfortunately, we did not observe a higher accuracy derived from DNS. The reasons could be two-fold. First, our implementation may not be optimized and the model parameters may not be well tuned. Second, there are thousands of templates in the datasets, which may bring challenges for the classification task. Nevertheless, with the availability of large-scale datasets, applying deep learning models for MWPs is still an interesting research direction that deserves intensive attention.

## 4 FEATURE EXTRACTION

Feature extraction has been a vital component in the machine learning (ML) workflow. Effective feature construction, in an either supervised or semi-supervised manner, can significantly boost the accuracy. For instance, SIFT [61], [62] and other local descriptors [63], [64], [65] have been intensively used in the domain of object recognition and image retrieval for decades as they are invariant to scaling, orientation and illumination changes. Consequently, a large amount of research efforts have been devoted to design effective features to facilitate ML tasks. Such a discipline was partially changed by the emergence of deep learning. In the past years, deep learning has transformed the world of artificial intelligence, due to the increasing processing power afforded by graphical processing units (GPUs), the enormous amount of available data, and the development of more advanced algorithms. With well-annotated sufficient training data, the methodology can automatically learn the effective latent feature representation for the classification or prediction tasks. Hence, it can help replace the manual feature engineering process, which is non-trivial and labor-intensive.

In the area of automatic math problem solver, as reviewed in the previous sections, the only DL-based approach without feature engineering is DSN [14]. It applies word embedding to vectorize the text information and encodes these vectors by GRU network for automatic feature extraction. The limitation is that a large amount of labeled training data is required to make the model effective. Before the appearance of DSN, most of the math problem solvers were designed with the availability of small-scale datasets. Thus, feature engineering plays an important role in these works to help achieve a high accuracy. In this section, we provide a comprehensive review on the features engineering process in the literature and show how they help bridge the gap between textual/visual information and the semantic/logical representation.

### 4.1 Preprocessing

Before we review the feature space defined in various MWP solvers, we first present the preliminary background on the preprocessing steps that have been commonly adopted to facilitate the subsequent feature extraction.

1. <https://answers.yahoo.com/>

TABLE 3  
Statistics of datasets for equation set problems.

Datasets	Proposed in	# of problems	# of templates	$\frac{\# \text{ of problems}}{\# \text{ of templates}}$	# of single-eq problems	# of words	# of sentences
<b>ALG514</b>	KAZB [39]	514	28	18.36	91	1.62k	19.3k
<b>Dolphin1878</b>	DOL [35]	1,878	1,183	1.59	712	3.30k	41.4k
<b>DRAWIK</b>	MixedSP [45]	1,000	232	4.31	255	1.38k	13.8k
<b>Dolphin18K</b>	SIM [12]	18,460	5,871	3.14	8,333	49.9k	604k

TABLE 4  
Accuracies of equation set problem solvers on existing datasets.

Methods		Publish Year	ALG514	Dolphin1878	DRAWIK	Dolphin18K
Statistic-based	DOL [35]	2015		60.2		
Similarity-based	SIM [12]	2016	70.1	29	25.5	18.4
Template-based	KAZB [39]	2014	68.7		43.2	
	ZDC [40]	2015	79.7			17.9
	MixedSP [45]	2016	83.0		59.5	
	FG-Expression [13]	2017				28.4
DL-based	DNS [14]	2017	70.1	28.29	31	21.6

#### 4.1.1 Syntactic Parsing

Syntactic parsing focuses on organizing the lexical units and their semantic dependency in a tree structure, which serves as a useful resource for effective feature selection. Sorts of parsers have been developed, among which the Stanford parser works as the most comprehensive and widely-adopted one. It is a package consisting of different probabilistic natural language parsers. To be more specific, its neural-network parser [66] is a transition-based dependency parser that uses high-order features to achieve high speed and good accuracy; the Compositional Vector Grammar parser [67] can be seen as factoring discrete and continuous parsing in one model; and the (English) Stanford Dependencies representation [68] is an automatic system to extract typed dependency parses from phrase structure parses, where a dependency parse represents dependencies between individual words and a phrase structure parse represents nesting of multi-word constituents. Besides Stanford parser, there exist other effective dependency parsers with their own traits. For example, [69] presents an easy-fist parsing algorithm that iteratively selects the best pair of neighbors in the tree structure to connect at each parsing step.

Those parsers account in WMP solvers. For instance, the neural-network parser [66] is adopted in [70] for coreference resolution, which is another pre-processing step for MWP solvers. UnitDep [28] automatically generates features from a given math problem by analyzing its derived parser tree using the Compositional Vector Grammar parser [67]. Additionally, the Stanford Dependencies representation [68] has been applied in multiple solvers. We observed its occurrence in Formula [23] and ARIS [21] to extract attributes of entities (the subject, verb, object, preposition and temporal information), in KAZB [39] to generate part-of-speech tags, lematizations, and dependency parses to compute features, and in ALGES [27] to obtain syntactic information used for grounding and feature computation. ExpressionTree [26] is an exceptional case without using Stanford Parser. Instead, it uses the easy-fist parsing algorithm [69] to detect the verb associated with each quantity.

#### 4.1.2 Coreference Resolution

Co-reference resolution involves the identification and clustering of noun phrases mentions that refer to the same real-world entity. The MWP solvers use it as a pre-processing step to ensure the correct arithmetic operations or value update on the same entity. [71] is an early deterministic approach which is driven entirely by the syntactic and semantic compatibility learned from a large, unlabeled corpus. It allows proper and nominal mentions to only corefer with antecedents that have the same head, but pronominal mentions to corefer with any antecedent. On top of [71], Raghunathan et al. [72] proposed an architecture based on tiers of deterministic coreference models. The tiers are processed from the highest to the lowest precision and the entity output of a tier is forwarded to the next tier for further processing. [73] is another model that integrates a collection of deterministic coreference resolution models. Targeting at exploring rich feature space, [74] proposed a simple classification model for coreference resolution with a well-designed set of features. NECo is proposed in [75] and capable of solving both named entity linking and co-reference resolution jointly.

As to applying coreference resolvers in MWP solvers, the Illinois Coreference Resolver [74] [76] is used in [20] to identify pronoun referents and facilitate semantic labeling. In [70], a rule function  $\text{Coref}(A, B)$ , which is true when A and B represent the same entity, is derived as a component of the declarative rules to determine the math operators. Given a pair of sentences, each containing a quantity, ZDC [40] takes into account the existence of coreference relationship between these two sentences for feature exploitation. Meanwhile, ARIS [21] adopts the [72] for coreference resolution and uses the predicted coreference relationships to replace pronouns with their coreferent links.

#### 4.2 Common Features

There have been various types of features proposed in the past literature. We separate them into *common features* and *unique features*, according to the number of solvers that have adopted a particular type of feature. The unique features were proposed once and not reused in another work, implying that their effect could be limited. The *common features* are considered to be more general and effective, and they are the focus of this survey.

In Table 5, we categorize the *common features* according to their syntactic sources for feature extraction, such as quantities, questions, verbs, etc. For each type of proposed feature, we identify its related MWP solvers, and provide necessary examples to explain features that are not straightforward to figure out.

#### 4.2.1 Quantity-related Features

The basic units in an arithmetic expression or an equation set consist of quantities, unknown variables and operators. Hence, a natural idea is to extract quantity-related features to help identify the relevant operands and their associated operators. As shown in Table 5, a binary indicator to determine whether a quantity refers to a rate is adopted in many solvers [26] [28] [15] [40] [45]. It signals a strong connection between the quantity and operators of  $\{\times, \div\}$ . The value of the quantity is also useful for operator classifier or quantity relevance classifier. For instance, a quantity whose value is a real number between  $[0, 1]$  is likely to be associated with multiplication or division operators [40], [45]. It is also observed that quantities in the text format of “one” or “two” are unlikely to be relevant with the solution [39] [40], [45], [13]. Examples include “if *one* airplane averages 400 miles per hour,...” and “the difference between *two* numbers is 36”.

#### 4.2.2 Context-related Features

The information embedded in the text window centered at a particular quantity can also provide important clues for solving math word problems. To differentiate two quantities both in the numeric format, we can leverage the word lemmas, part of speech (POS) tags and dependence types within the window as the features. In this manner, quantities associated with the same operators would likely to share similar context information. A trivial trick used in [26] [28] [15] is to examine whether there exists comparative adverbs. For example, terms “more”, “less” and “than” indicate operators of  $\{+, -\}$ .

#### 4.2.3 Quantity-pair Features

The relationship between two quantities is helpful to determine their associated operator. A straightforward example is that if two quantities are associated with the same unit, they can be applied with addition and subtraction [26] [28] [15] [40]. If one quantity is related to a rate and the other is associated with a unit that is part of the rate, their operator is likely to be multiplication or division [26] [27] [28] [15].

Numeric relation and context similarity are two types of quantity-pair features proposed in [40] [45]. The former obtains two sets of nouns located within the same sentence as the two quantities and sorts them by the distance in the dependency tree. Then, a scoring function is defined to measure the similarity between these two sorted noun lists. Higher similarity implies that the two quantities are more likely to be connected by addition or subtraction operators. The latter extracts features for equation template classifier. It is observed that the contextual information between two numbers is similar, they are likely to be located within in a template with symmetric number slots. For example, given a template  $n_1 \times u_1 + n_2 \times u_2$ , “ $n_1$ ” and “ $n_2$ ” are symmetric. The context similarity is measured by the Jaccard similarity on two sets of words among the context windows. Given a problem text “A plum costs 2 dollars and a peach costs 1 dollars”, “2” and “1” are two quantities with similar context.

Two types of quantity-pair features were both adopted in the template-based solutions to equation set problems [39] [40]. The

first type is the dependency path between a pair of quantities. Their similarity may be helpful to determine the corresponding positions (or number slots) in the equation template. For example, given a sentence “2 footballs and 3 soccer balls cost 220 dollars”, the dependency paths between two quantity pairs (2, 220) and (3, 220) are identical, implying that 2 and 3 refer to similar types of number slots in the template. The other feature is whether two quantities appear in the same sentence. If so, they are likely to appear in the same equation of the template. Finally, a popular quantity-pair feature used in [26] [28] [15] [39] [40] [45] examines whether the value of one quantity is greater than the other, which is helpful to determine the correct operands for subtraction operator.

#### 4.2.4 Question-related Features

Distinguishing features can also be derived from questions. It is straightforward to figure out that the unknown variable can be inferred from the question and if a quantity whose unit appears in the question, this quantity is likely to be relevant. The remain question-related features presented in Table 5 were proposed by Roy et al. [26], [28] and followed by MathDQN [15]. Their feature design leverages the number of matching tokens between the related noun phrase of a quantity and the question text. The quantities with the highest number of matching tokens are considered as useful clues. They also check whether the question contains rate indicators such as “each” and “per”, or comparison indicators such as “more” or “less”. The former is related to  $\{\times, \div\}$  and the latter is related to  $\{+, -\}$ . Moreover, if the question text contains “how many”, it implies that the solution is a positive number.

#### 4.2.5 Verb-related Features

Verbs are important indicators for correct operator determination. For example, “lose” is a verb indicating quantity loss for an entity and related to the subtraction operator. Given a quantity, we call the verb closest to it in the dependency tree as its *dependent verb*. [26] [27] [28] [15] directly use dependent verb as one of the features. Another widely-adopted verb-related feature is a vector capturing the distance between the dependent verb and a small pre-defined collection of verbs that are found to be useful in categorizing arithmetic operations. Again, the remaining features come from the works [26], [28], [15]. The features indicate whether two quantities have the same dependent verbs or whether their dependent verbs refer to the same verb mention. As we can see from the examples in Table 5, the difference between these two types of features is the occurrence number of the dependent verb in the sentence.

#### 4.2.6 Global Features

There are certain types of global features in the document-level proposed by existing solvers. [26], [28], [15] use the number of quantities in the problem text as part of feature space. Unigrams and bigrams are also applied in [20] [39]. They may play certain effect in determining the quantities and their order. Note that the unigrams and bigrams are defined in the word level rather than the character level.

## 5 GEOMETRIC WORD PROBLEM (GWP) SOLVER

Geometry solvers have been studied for a long history. Visual diagram understanding is a sub-domain that has attracted significant attention. As an early work for understanding line drawings, [77]

TABLE 5  
Common Features.

Feature Type	Description	Used In	Remark
Quantity-related Features	Whether the quantity refers to a rate	[26] [28] [15] [40] [45]	For “each ride cost 5 tickets”, the quantity “5” is a rate
	Is between 0 and 1	[40] [45]	
	Is equal to one or two	[39] [40] [45] [13]	
Context-related Features	Word lemma	[39] [40] [45]	For “Connie has 41.0 red markers.”, the word lemmas around the quantity “41.0” are {Connie, have, red, marker}.
	POS tags	[28] [39] [40] [45] [20]	For “A chef needs to cook 16.0 potatoes.”, the POS tags within a window of size 2 centered at the quantity “16.0” are {TO, VB, NNS}.
	Dependence type	[39] [40] [45]	For “Ned bought 14.0 boxes of chocolates candy”, we can detect multiple dependencies within the window of size 2 around the “14.0”: (boxes, 14.0) → (num), (boxes, of) → (prep), (bought, Ned) → (nsubj). The dependence root is “bought”.
	Comparative adverbs	[26] [28] [15]	For “If she drank 25 of them and then bought 30 more.”, “more” is a comparative term in the window of quantity “30”.
Quantity-pair Features	Whether both quantities have the same unit	[26] [28] [15] [40]	For “Student tickets cost 4 dollars and general admission tickets cost 6 dollars”, quantities “4” and “6” have the same unit.
	If one quantity is related to a rate and the other is associated with a unit that is part of the rate	[26] [27] [28] [15]	For “each box has 9 pieces” and “Paul bought 6 boxes of chocolate candy”, “9” is related to a rate ( i.e., pieces/box) and “6” is associated to the unit “box”.
	Numeric relation of two quantities	[40] [45]	For each quantity, the nouns around it are extracted and sorted by the distance in the dependency tree. Then, a scoring function is defined on the two sorted lists to measure the numeric relation.
	Context similarity between two quantities	[40] [45]	The context is represented by the set of words around the quantity.
	Dependency path between two quantities.	[39] [40]	For “2 footballs and 3 soccer balls cost 220 dollars”, the dependency path for the quantity pair (2,3) is $num(footballs,2) - conj(footballs, balls) - num(balls, 3)$ .
	Whether both quantities appear in the same sentence	[39] [40]	
	Whether the value of the first quantity is greater than the other	[26] [28] [15] [39] [40] [45]	
Question-related Features	Whether the unit or related noun phrase of a quantity appears in the question	[20] [26] [27] [28] [15] [39]	
	Whether the unit or related noun phrase of a quantity has the highest number of match tokens with the question text	[26] [28] [15]	For the question “How many apples are left in the box?” and a quantity 77 that appears in “77 apples in a box”, there are two matching tokens (“apples” and “box”).
	Number of quantities which happen to have the maximum number of matching tokens with the question	[26] [28] [15]	For “Rose have 9 apples and 12 erasers. ... 3 friends. How many apples dose each friend get?”, the number of matching tokens for quantities 9, 12 and 3 is 1, 0 and 1. Hence, there are two quantities with the maximum matching token number.
	Whether any component of the rate is present in the question	[26] [28] [15]	Given a question “How many blocks does George have?” and a quantity 6 associated with rate “blocks/box”, the feature indicator is set to 1 since block appears in the question.
	Whether the question contains terms like “each” or “per”	[26] [28] [15]	
	Whether the question contains comparison-related terms like “more” or “less”	[26] [28] [15]	
	Whether the question contains terms like “how many”	[39] [40] [45] [13]	It implies that the solution is positive.
Verb-related Features	Dependent verb of a quantity	[26] [27] [28] [15] [21] [24] [27]	the verb closest to the quantity in the dependency tree
	Distance vector between the dependent verb and a small collection of pre-defined verbs that are useful for arithmetic operator classification		
	Whether two quantities have the same dependent verbs	[26] [28] [15]	For “In the first round she scored 40 points and in the second round she scored 50 points”, the quantities “40” and “50” both have the same verb “scored”. Note that “scored” appeared twice in the sentence.
Global Features	Whether both dependent verbs refer to the same verb mention	[26] [28] [15]	For “She baked 4 cupcakes and 29 cookies.”, the quantities “4” and “29” both shared the verb “baked”. Note that “baked” appeared only once in the sentence.
	Number of quantities mentioned in text	[26] [28] [15]	
	Unigrams and bigrams of sentences in the problem text	[20] [39]	

presented an efficient characteristic pattern detection method by scanning the distribution of black pixels and generating feature points graph. In [78], a structure mapping engine named *GeoRep* was proposed to generate qualitative spatial descriptions from line diagrams. After that, the visual elements can be formulated through a two-level representation architecture. This work was also applied to the repetition and symmetry detection model in MAGI [79]. Inspired by human cognitive process of reading juxtaposition diagrams, MAGI detects repetition by aligning visual and conceptual relational structure to analyze repetition-based diagrams.

The problem of rectangle and parallelogram detection in diagram understanding has also received a considerable amount of interest. The proposed techniques fall into two main categories, either based on primitive or Hough transform [80]. The primitive-based methods combine line segments or curves to form possible edges of a quadrangle. For examples, Lin and Nevatia [81] proposed the approach of parallelogram detection from a single aerial image by linear feature extraction and formation of hypothesis following certain geometric constraints. Similarly, Lagunovsky and Ablameyko [82] studied the problem of rectangular detection based on line primitives. As to the Hough transform based techniques, [83] presented an approach for automatic rectangular particle detection in cryo-electron microscopy through Hough transform, but this method can only work well when all rectangles have the same dimensions and the dimensions must be aware in advance. Jung et.al. [84] proposed a window Hough transform algorithm to tackle the problem of rectangle detection with varying dimensions and orientations.

Geometry theorem proving (GTP) [85], [86] was initially viewed as an artificial intelligence problem that was expected to be easily tackled by machines. The difficulties of solving GTP problems lie in the visual reasoning in geometry and the generation of elegant and concise geometry proofs. Moreover, completing the proof requires the ingenuity and insights to the problem. The first automated GTP was developed by Gelernter [85], which used the diagram as pruning heuristic. The system rejects geometry goals that fail to satisfy the diagram. Whereas the limitation of the method is the true sub-goal may be pruned erroneously due to the insufficient precise arithmetic applied to the diagram. Fleuriot et. al. [86] studied Newton's geometric reasoning procedures in his work *Principia* and presented theorem prover Isabelle to formalize and generate the style of reasoning performed by Newton. By combining the existing geometry theorem proving techniques and the concepts of Nonstandard Analysis, the prover Isabelle can produce proofs of lemmas and theorem in *Principia*. Readers can refer to the book chapter [86] for the survey of early development of GTP.

In this survey, we are more interested to examine the math problems that are required to consider visual diagram and textual mentions simultaneously. As illustrated in Figure 5, a typical geometry word problem contains text descriptions or attribute values of geometric objects. The visual diagram may contain essential information that are absent from the text. For instance, points O, B and C are located on the same line segment and there is a circle passing points A, B, C and D. To well solve geometry word problems, three main challenges need to be tackled: 1) diagram parsing requires the detection of visual mentions, geometric characteristics, the spatial information and the co-reference with text; 2) deriving visual semantics which refer to the textual information related to visual analogue involves

the semantic and syntactic interpretation to the text; and 3) the inherent ambiguities lie in the task of mapping visual mentions in the diagram to the concepts in real world.

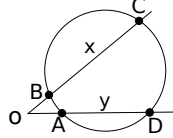
Geometric Problem
In the figure below triangle OAB has an area of 72 and triangle ODC has an area of 288. Find x and y.

Equation
<p>area of OAB = <math>72 = (1/2) \sin(\angle AOB) * OA * OB</math></p> <p>solve the above for <math>\sin(\angle AOB)</math> to find <math>\sin(\angle AOB) = 1/2</math></p> <p>area of ODC = <math>288 = (1/2) \sin(\angle DOC) * OD * OD</math></p> <p>Note that <math>\sin(\angle DOC) = \sin(\angle AOB) = 1/2</math>, <math>OD = 18 + y</math> and <math>OC = 16 + x</math> and substitute in the above to obtain the first equation in x and y</p> <p><math>1152 = (18 + y)(16 + x)</math></p> <p>We now use the theorem of the intersecting lines outside a circle to write a second equation in x and y</p> <p><math>16 * (16 + x) = 14 * (14 + y)</math></p>
Solution
$x=20, y=14$

Fig. 5. An example of geometric problem.

## 5.1 Text-Aligned Diagram Understanding

The very early computer program, BEATRIX [87], [88], parses the English text and diagram components of the elementary physics problems together by establishing the coreference between the text and diagram. Watanabe et al. proposed a framework to combine layout information and natural language to analyze the pictorial book of flora diagrams [89]. An overview of the research on integration of visual and linguistic information was provided in the survey paper by Srihar [90]. However, these early approaches rely on written rules or manual regulations, i.e., the visual elements needed to be recognized with human intervention and their performances were usually dependent on specified diagrams.

G-ALINGER [91] is an algorithmic work that addresses the geometry understanding and text understanding simultaneously. To detect primitives from a geometric diagram, Hough transform [92] is first applied to initialize lines and circles segments. An objective function that incorporates pixel coverage, visual coherence and textual-visual alignment. The function is sub-modular and a greedy algorithm is designed to pick the primitive with the maximum gain in each iteration. The algorithm stops when no positive gain can be obtained according to the objective function. In [93], the problem of visual understanding is addressed in the context of science diagrams. The objective is to identify the graphic representation for the visual entities and their relations such as temporal transitions, phase transformations and inter object dependencies. An LSTM-based network is proposed for syntactic parsing of diagrams and learns the graphic structure.

## 5.2 GWP Solvers

GEOS [94] can be considered as the first work to tackle a complete geometric word problem as shown in Figure 5. The

method consists of two main steps: 1) parsing text and diagram respectively by generating a piece of logical expression to represent the key information of the text and diagram as well as the confidence scores, and 2) addressing the optimization problem by aligning the satisfiability of the derived logical expression in a numerical method that requires manually defining indicator function for each predicate. It is noticeable that G-ALINGER is applied in GEOS [91] for primitive detection. Despite the superiority of automated solving process, the performance of the system would be undermined if the answer choices are unavailable in a geometry problem and the deductive reasoning based on geometric axiom is not used in this method. A subsequent improver of GEOS is presented in [95]. It harvests an axiomatic knowledge from 20 publicly available math textbooks and builds a more powerful reasoning engine that leverages the structured axiomatic knowledge for logical inference.

GeoShader [96], as the first tool to automatically handle geometry problem with shaded area, presents an interesting reasoning technique based on analysis hypergraph. The nodes in the graph represent intermediate facts extracted from the diagram and the directed edges indicate the relationship of deductibility between two facts. The calculation of the shaded area is represented as the target node in the graph and the problem is formulated as finding a path in the hypergraph that can reach the target node.

## 6 MISCELLANEOUS MATH TASKS

### 6.1 Other Variants of Math Problems

Apart from geometric problems, there are also assorted variants of math problems that AI system focuses on. Aristo [97] is able to solve non-diagram multiple-choice questions through five parallel solvers, one for pure text, two for statistic and two for inference. Finally, the combiner of Aristo outputs a comprehensive score of each option based on scores from the five solvers. A similar work on multiple-choice questions is [98], which takes Wikipedia as a knowledge base. After ranking and filtering relevant pages retrieved from Wikipedia, it presents a new scoring function to pick the best answer from the candidates. Another variant is targeted at solving IQ test and a noticeable number of computer models have been proposed in [99], [100]. Taking [100] for example, it proposed a framework for solving verbal IQ questions, which classifies questions into several categories and each group of questions are solved by a specific solver respectively. Furthermore, logic puzzles are addressed in [101] by transforming robust natural language to precise semantics. For other forms of math problems, [102] solves probability problems automatically by a two-step approach, namely first formulating questions in a declarative language and then computing the answer through a solver implemented in ProbLog [103]. And algebraic word problems are solved by generating answer rationales written in natural language in [104] through a sequence-to-sequence model.

### 6.2 Math Problem Solver in Other Languages

Solving math word problems in other languages also attracts research attention. Yu et al. addressed the equation set problem solver in Chinese [105], [106]. A pool of rule-based semantic models are crafted to map the sentences in Chinese text into equation templates. The experiments were conducted on a very

small-scale dataset with 104 problems. Recently, there has been the first attempt to solve Arabic arithmetic word problems [107]. Its test dataset was collected by translating the **A12** dataset [21] from English to Arabic. The proposed techniques also rely on the verb categorization, similar to those proposed in [21], except that customization for the Arabic language needs to be made for the tasks of syntactic parser and named entity recognition. To conclude, the math word problem solvers in other languages than English are still at a very early stage. The datasets used are neither large-scale nor challenging and the proposed techniques are obsolete. This research area has great room for improvement and calls for more efforts to be involved.

### 6.3 Math Problem Generator

We also review automatic math word problem generators that can efficiently produce a large, diverse and configurable corpus of question-answer database. The topics covered in this survey include algebra word problems with basic operators  $\{+, -, \times, \div\}$  and geometry problems.

In [108], Wang et al. leveraged the concept of *expression tree* to generate a math word problem. The tree structure can provide the skeleton of the story, and meanwhile allow the story to be constructed recursively from the sub-stories. Each sub-story can be seen as a text template with value slots to be filled. These sub-stories will be concatenated into an entire narrative. Different from [108], the work of [109] rewrites a given math word problem to fit a particular theme such as *Star War*. In this way, students may stay more engaged with their homework assignments. The candidate are scored with the coherence in multiple factors (e.g., syntactic, semantic and thematic). [110] generates math word problems that match the personal interest of students. The generator uses Answer Set Programming [111], in which programs are composed of facts and rules in a first-order logic representation, to satisfy a collection of pedagogical and narrative requirements. Its objective is to produce coherent and personalized story problems that meet pedagogical requirements.

In the branch of geometry problem generator, GeoTutor [112], [113] is designed to generate geometry proof problems for high school students. The input contains a figure and a set of geometry axioms. The output is a pair  $(I, G)$ , where  $I$  refers to the assumptions for the figure and goals in  $G$  are sets of explicit facts to be inferred. Singhal et al. also tackled the automated generation of geometry questions for high school students [114], [115]. Its input interface allows users to select geometric objects, concepts and theorems. Compared with [112], [113], its geometric figure is generated by the algorithm rather than specified by the user. Based on the figure, the next step of generating facts and solutions is similar to that in [112], [113]. It requires pre-knowledge on axioms and theorems and results in the formation capturing the relationships between its objects.

## 7 CONCLUSIONS AND FUTURE DIRECTIONS

In this paper, we present a comprehensive survey to review the development of math word problem solvers in recent years. The topics discussed in this survey cover arithmetic word problems, equation set problems, geometry word problems and miscellaneous others related to math. We compared the techniques proposed for each math task, provided a rational categorization, and conducted accountable experimental analysis. Moreover, we took a close examination on the subject of feature engineering

proposed for MWP solvers and summarized the diversified proposal of syntactic features.

Overall speaking, the current status of MWP solvers is far from promising and has great room for improvement. There is no doubt that the topic would continue to attract more and more research attention in the next few years, especially after the public release of large-scale datasets such as Dolphin18K and Math23K. In the following, we present a number of future directions that may be of interest to the community and worth further exploration.

Firstly, DNS [14] was the first attempt that used deep learning models in MWP solvers so as to avoid non-trivial feature engineering. This work shed light on the feasibility of designing end-to-end models to enhance the accuracy and reduce human intervention. Considering that DNS only uses a basic seq-to-seq network structure, with LSTM and GRU as the encoding and decoding units, we expect more advanced networks to be developed. Moreover, as a common practice, these models can be integrated with attention mechanism [116] for performance advancement.

Secondly, aligning visual understanding with text mention is an emerging direction that is particularly important to solve geometry word problems. However, this challenging problem has only been evaluated in self-collected and small-scale datasets, similar to those early efforts on evaluating the accuracy of solving algebra word problem. There is a chance that these proposed aligning methods fail to work well in a large and diversified dataset. Hence, it calls for a new round of evaluation for generality and robustness with a better benchmark dataset for geometry problems.

Thirdly, interpretability plays a key role in measuring the usability of MWP solvers in the application of online tutoring, but may pose new challenges for the deep learning based solvers [100]. For instance, AlphaGo [117] and AlphaZero [118] have achieved astonishing superiority over human players, but their near-optimal actions could be difficult for human to interpret. Similar issues may occur in the domain of automatic math problem solver and they deserve an early examination.

Last but not the least, solving math word problems in English plays a dominating role in the literature. We only observed a very rare number of math solvers proposed to cope with other languages. This research topic may grow into a direction with significant impact. To our knowledge, many companies in China have harvested an enormous number of word problems in K12 education. As reported in 2015<sup>2</sup>, Zuoyebang, a spin off from Baidu, has collected 950 million questions and solutions in its database. When coupled with deep learning models, this is an area with immense imagination and exciting achievements can be expected.

## REFERENCES

- [1] D. Bobrow, "Natural language input for a computer problem solving system," in *Semantic information processing*, M. Minsky, Ed. MIT Press, 1964, pp. 146–226.
- [2] E. A. Feigenbaum and J. Feldman, *Computers and Thought*. New York, NY, USA: McGraw-Hill, Inc., 1963.
- [3] E. Charniak, "Computer solution of calculus word problems," in *Proceedings of the 1st International Joint Conference on Artificial Intelligence*, ser. IJCAI'69, 1969, pp. 303–316.
- [4] P. Clark, "Elementary school science and math tests as a driver for AI: take the aristo challenge!" in *AAAI*, 2015, pp. 4019–4021.
- [5] P. Clark and O. Etzioni, "My computer is an honor student - but how intelligent is it? standardized tests as a measure of AI," *AI Magazine*, vol. 37, no. 1, pp. 5–12, 2016.
- [6] J. R. Slagle, "Experiments with a deductive question-answering program," *Commun. ACM*, vol. 8, no. 12, pp. 792–798, Dec. 1965.
- [7] C. R. Fletcher, "Understanding and solving arithmetic word problems: A computer simulation," *Behavior Research Methods, Instruments, & Computers*, vol. 17, no. 5, pp. 565–571, Sep 1985.
- [8] Y. Bakman, "Robust Understanding of Word Problems with Extraneous Information," *ArXiv Mathematics e-prints*, Jan. 2007.
- [9] A. Mukherjee and U. Garain, "A review of methods for automatic understanding of natural language mathematical problems," *Artif. Intell. Rev.*, vol. 29, no. 2, pp. 93–122, Apr. 2008.
- [10] D. Goldwasser and D. Roth, "Learning from natural instructions," in *IJCAI*, 2011, pp. 1794–1800.
- [11] T. Kwiatkowski, E. Choi, Y. Artzi, and L. S. Zettlemoyer, "Scaling semantic parsers with on-the-fly ontology matching," in *EMNLP*, 2013, pp. 1545–1556.
- [12] D. Huang, S. Shi, C. Lin, J. Yin, and W. Ma, "How well do computers solve math word problems? large-scale dataset construction and evaluation," in *ACL*, 2016.
- [13] D. Huang, S. Shi, J. Yin, and C.-Y. Lin, "Learning fine-grained expressions to solve math word problems," in *EMNLP*, 2017, pp. 805–814.
- [14] Y. Wang, X. Liu, and S. Shi, "Deep neural solver for math word problems," in *EMNLP*, 2017, pp. 845–854.
- [15] L. Wang, D. Zhang, L. Gao, J. Song, L. Guo, and H. T. Shen, "Mathdqn: Solving arithmetic word problems via deep reinforcement learning," in *AAAI*. AAAI Press, 2018.
- [16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.
- [17] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec 2015, pp. 2425–2433.
- [18] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the V in VQA matter: Elevating the role of image understanding in visual question answering," in *CVPR*, 2017, pp. 6325–6334.
- [19] R. Yun, M. Yuhui, C. Guangzuo, H. Ronghuai, and Z. Ying, "Frame-based calculus of solving arithmetic multi-step addition and subtraction word problems," in *Education Technology and Computer Science, International Workshop on (ETCS)*, vol. 02, 03 2010, pp. 476–479. [Online]. Available: doi.ieeecomputersociety.org/10.1109/ETCS.2010.316
- [20] S. Roy, T. Vieira, and D. Roth, "Reasoning about quantities in natural language," *TACL*, vol. 3, pp. 1–13, 2015.
- [21] M. J. Hosseini, H. Hajishirzi, O. Etzioni, and N. Kushman, "Learning to solve arithmetic word problems with verb categorization," in *EMNLP*, 2014, pp. 523–533.
- [22] S. S. Sundaram and D. Khemani, "Natural language processing for solving simple word problems," in *Proceedings of the 12th International Conference on Natural Language Processing*. Trivandrum, India: NLP Association of India, December 2015, pp. 394–402.
- [23] A. Mitra and C. Baral, "Learning to use formulas to solve simple arithmetic problems," in *ACL (1)*. The Association for Computer Linguistics, 2016.
- [24] C.-C. Liang, K.-Y. Hsu, C.-T. Huang, C.-M. Li, S.-Y. Miao, and K.-Y. Su, "A tag-based english math word problem solver with understanding, reasoning and explanation," in *NAACL*, 2016.
- [25] C. Liang, K. Hsu, C. Huang, C. Li, S. Miao, and K. Su, "A tag-based statistical english math word problem solver with understanding, reasoning and explanation," in *IJCAI*, 2016, pp. 4254–4255.
- [26] S. Roy and D. Roth, "Solving general arithmetic word problems," in *EMNLP*, 2015, pp. 1743–1752.
- [27] R. Koncel-Kedziorski, H. Hajishirzi, A. Sabharwal, O. Etzioni, and S. D. Ang, "Parsing algebraic word problems into equations," *TACL*, vol. 3, pp. 585–597, 2015.
- [28] S. Roy and D. Roth, "Unit dependency graph and its application to arithmetic word problem solving," in *AAAI*. AAAI Press, 2017, pp. 3082–3088.
- [29] —, "Illinois math solver: Math reasoning on the web," in *NAACL*, 2016.
- [30] S. Roy, S. Upadhyay, and D. Roth, "Equation parsing : Mapping sentences to grounded equations," in *EMNLP*. The Association for Computational Linguistics, 2016, pp. 1088–1097.
- [31] K. Narasimhan, T. D. Kulkarni, and R. Barzilay, "Language understanding for text-based games using deep reinforcement learning," in *EMNLP*, 2015, pp. 1–11.

2. <http://www.marketing-interactive.com/baidus-zuoyebang-attracts-outside-investors/>



- [32] K. Narasimhan, A. Yala, and R. Barzilay, “Improving information extraction by acquiring external evidence with reinforcement learning,” in *EMNLP*, 2016, pp. 2355–2365.
- [33] H. Guo, “Generating Text with Deep Reinforcement Learning,” *ArXiv e-prints*, Oct. 2015.
- [34] J. C. Caicedo and S. Lazebnik, “Active object localization with deep reinforcement learning,” in *ICCV*, 2015, pp. 2488–2496.
- [35] S. Shi, Y. Wang, C. Lin, X. Liu, and Y. Rui, “Automatically solving number word problems by semantic parsing and reasoning,” in *EMNLP*, 2015, pp. 1132–1142.
- [36] D. E. Knuth, “Semantics of context-free languages,” *Mathematical Systems Theory*, vol. 2, no. 2, pp. 127–145, 1968.
- [37] J. Earley, “An efficient context-free parsing algorithm,” *Commun. ACM*, vol. 13, no. 2, pp. 94–102, 1970. [Online]. Available: <http://doi.acm.org/10.1145/362007.362035>
- [38] D. Zhang, L. Nie, H. Luan, K. Tan, T. Chua, and H. T. Shen, “Compact indexing and judicious searching for billion-scale microblog retrieval,” *ACM Trans. Inf. Syst.*, vol. 35, no. 3, pp. 27:1–27:24, 2017.
- [39] N. Kushman, L. Zettlemoyer, R. Barzilay, and Y. Artzi, “Learning to automatically solve algebra word problems,” in *ACL*, 2014, pp. 271–281.
- [40] L. Zhou, S. Dai, and L. Chen, “Learn to solve algebra word problems using quadratic programming,” in *EMNLP*, 2015, pp. 817–822.
- [41] R. Herbrich, T. Graepel, and K. Obermayer, “Large margin rank boundaries for ordinal regression,” in *Advances in Large Margin Classifiers*, P. J. Bartlett, B. Schölkopf, D. Schuurmans, and A. J. Smola, Eds. MIT Press, 2000, pp. 115–132.
- [42] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed. New York: Springer, 2006.
- [43] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer New York, 2013. [Online]. Available: <https://books.google.com/books?id=EoDSBwAAQBAJ>
- [44] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, 2009.
- [45] S. Upadhyay, M. Chang, K. Chang, and W. Yih, “Learning from explicit and implicit supervision jointly for algebra word problems,” in *EMNLP*. The Association for Computational Linguistics, 2016, pp. 297–306.
- [46] L. Gao, Z. Guo, H. Zhang, X. Xu, and H. T. Shen, “Video captioning with attention-based LSTM and semantic consistency,” *IEEE Trans. Multimedia*, vol. 19, no. 9, pp. 2045–2055, 2017.
- [47] L. Yu, Y. Yang, Z. Huang, P. Wang, J. Song, and H. T. Shen, “Web video event recognition by semantic analysis from ubiquitous documents,” *IEEE Trans. Image Processing*, vol. 25, no. 12, pp. 5689–5701, 2016.
- [48] X. Wang, L. Gao, P. Wang, X. Sun, and X. Liu, “Two-stream 3-d convnet fusion for action recognition in videos with arbitrary size and length,” *IEEE Transactions on Multimedia*, vol. 20, no. 3, pp. 634–644, March 2018.
- [49] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, “VQA: visual question answering,” in *ICCV*, 2015, pp. 2425–2433.
- [50] L. Dong, F. Wei, M. Zhou, and K. Xu, “Question answering over freebase with multi-column convolutional neural networks,” in *ACL*, 2015, pp. 260–269.
- [51] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” in *NIPS*, 2014, pp. 3104–3112.
- [52] M. Luong, Q. V. Le, I. Sutskever, O. Vinyals, and L. Kaiser, “Multi-task sequence to sequence learning,” *CoRR*, vol. abs/1511.06114, 2015.
- [53] S. Wiseman and A. M. Rush, “Sequence-to-sequence learning as beam-search optimization,” in *EMNLP*, 2016, pp. 1296–1306.
- [54] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *NIPS*, 2013, pp. 3111–3119.
- [55] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in *EMNLP*, 2014, pp. 1532–1543.
- [56] K. Cho, B. van Merriënboer, Ç. Gülçehre, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” *CoRR*, vol. abs/1406.1078, 2014. [Online]. Available: <http://arxiv.org/abs/1406.1078>
- [57] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [58] B. Robaidek, R. Koncel-Kedziorski, and H. Hajishirzi, “Data-Driven Methods for Solving Algebra Word Problems,” *ArXiv e-prints*, Apr. 2018.
- [59] Z. Lin, M. Feng, C. Nogueira dos Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, “A Structured Self-attentive Sentence Embedding,” *ArXiv e-prints*, Mar. 2017.
- [60] S. Upadhyay and M. Chang, “Annotating derivations: A new evaluation strategy and dataset for algebra word problems,” in *EACL*, 2017, pp. 494–504.
- [61] D. G. Lowe, “Object recognition from local scale-invariant features,” in *ICCV*, 1999, pp. 1150–1157.
- [62] —, “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
- [63] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *CVPR*, 2005, pp. 886–893.
- [64] H. Bay, A. Ess, T. Tuytelaars, and L. J. V. Gool, “Speeded-up robust features (SURF),” *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.
- [65] G. Mori, S. J. Belongie, and J. Malik, “Efficient shape matching using shape contexts,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 11, pp. 1832–1837, 2005.
- [66] D. Chen and C. D. Manning, “A fast and accurate dependency parser using neural networks,” in *ACL*, 2014, pp. 740–750.
- [67] R. Socher, J. Bauer, C. D. Manning, and A. Y. Ng, “Parsing with compositional vector grammars,” in *ACL*, 2013, pp. 455–465.
- [68] M. de Marneffe, B. MacCartney, and C. D. Manning, “Generating typed dependency parses from phrase structure parses,” in *LREC*, 2006, pp. 449–454.
- [69] Y. Goldberg and M. Elhadad, “An efficient algorithm for easy-first non-directional dependency parsing,” in *Human Language Technologies: the 2010 Conference of the North American Chapter of the Association for Computational Linguistics*, 2010, pp. 742–750.
- [70] S. Roy and D. Roth, “Mapping to Declarative Knowledge for Word Problem Solving,” *ArXiv e-prints*, dec 2017.
- [71] A. Haghighi and D. Klein, “Simple coreference resolution with rich syntactic and semantic features,” in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*. Association for Computational Linguistics, 2009, pp. 1152–1161.
- [72] K. Raghunathan, H. Lee, S. Rangarajan, N. Chambers, M. Surdeanu, D. Jurafsky, and C. D. Manning, “A multi-pass sieve for coreference resolution,” in *EMNLP*, 2010, pp. 492–501.
- [73] H. Lee, Y. Peirsman, A. X. Chang, N. Chambers, M. Surdeanu, and D. Jurafsky, “Stanford’s multi-pass sieve coreference resolution system at the conll-2011 shared task,” in *CoNLL*, 2011, pp. 28–34.
- [74] E. Bengtson and D. Roth, “Understanding the value of features for coreference resolution,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP ’08, 2008, pp. 294–303.
- [75] H. Hajishirzi, L. Zilles, D. S. Weld, and L. S. Zettlemoyer, “Joint coreference resolution and named-entity linking with multi-pass sieves,” in *EMNLP*, 2013, pp. 289–299.
- [76] K.-W. Chang, R. Samdani, and D. Roth, “A Constrained Latent Variable Model for Coreference Resolution,” in *EMNLP*, 2013.
- [77] X. Lin, S. Shimotsuji, M. Minoh, and T. Sakai, “Efficient diagram understanding with characteristic pattern detection,” *Computer Vision, Graphics, and Image Processing*, vol. 30, no. 1, pp. 84–106, 1985. [Online]. Available: [https://doi.org/10.1016/0734-189X\(85\)90020-9](https://doi.org/10.1016/0734-189X(85)90020-9)
- [78] R. W. Ferguson and K. D. Forbus, “Georep: A flexible tool for spatial representation of line drawings,” in *AAAI*, 2000, pp. 510–516.
- [79] —, “Telling juxtapositions: Using repetition and alignable difference in diagram understanding,” 1998.
- [80] R. O. Duda and P. E. Hart, “Use of the hough transformation to detect lines and curves in pictures,” *Commun. ACM*, vol. 15, no. 1, pp. 11–15, 1972. [Online]. Available: <http://doi.acm.org/10.1145/361237.361242>
- [81] C. Lin and R. Nevatia, “Building detection and description from a single intensity image,” *Computer Vision and Image Understanding*, vol. 72, no. 2, pp. 101–121, 1998. [Online]. Available: <https://doi.org/10.1006/cviu.1998.0724>
- [82] D. Lagunovsky and S. Ablameyko, “Straight-line-based primitive extraction in grey-scale object recognition,” *Pattern Recognition Letters*, vol. 20, no. 10, pp. 1005–1014, 1999. [Online]. Available: [https://doi.org/10.1016/S0167-8655\(99\)00067-7](https://doi.org/10.1016/S0167-8655(99)00067-7)
- [83] Y. Zhu, B. Carragher, F. Mouche, and C. S. Potter, “Automatic particle detection through efficient hough transforms,” *IEEE Trans. Med. Imaging*, vol. 22, no. 9, pp. 1053–1062, 2003.
- [84] C. R. Jung and R. Schramm, “Rectangle detection based on a windowed hough transform,” in *SIBGRAPI*, 2004, pp. 113–120.
- [85] H. Gelernter, “Computers & thought,” E. A. Feigenbaum and J. Feldman, Eds. MIT Press, 1995, ch. Realization of a Geometry-theorem Proving Machine, pp. 134–152.
- [86] J. Fleuriot, *Geometry Theorem Proving*. London: Springer London, 2001, pp. 11–30.



- [87] W. C. Bulko, "Understanding text with an accompanying diagram," in *IEA/AIE (Vol. 2)*, 1988, pp. 894–898.
- [88] G. S. Novak and W. C. Bulko, "Understanding natural language with diagrams," in *Proceedings of the 8th National Conference on Artificial Intelligence, Boston, Massachusetts, July 29 - August 3, 1990, 2 Volumes.*, 1990, pp. 465–470.
- [89] Y. Watanabe and M. Nagao, "Diagram understanding using integration of layout information and textual information," in *COLING-ACL*, 1998, pp. 1374–1380.
- [90] R. K. Srihari, "Computational models for integrating linguistic and visual information: A survey," *Artif. Intell. Rev.*, vol. 8, no. 5-6, pp. 349–369, 1994.
- [91] M. J. Seo, H. Hajishirzi, A. Farhadi, and O. Etzioni, "Diagram understanding in geometry questions," in *AAAI*, 2014, pp. 2831–2838.
- [92] L. G. Shapiro and G. C. Stockman, *Computer Vision*. Prentice Hall, 2001.
- [93] A. Kembhavi, M. Salvato, E. Kolve, M. J. Seo, H. Hajishirzi, and A. Farhadi, "A diagram is worth a dozen images," in *ECCV*, 2016, pp. 235–251.
- [94] M. J. Seo, H. Hajishirzi, A. Farhadi, O. Etzioni, and C. Malcolm, "Solving geometry problems: Combining text and diagram interpretation," in *EMNLP*, 2015, pp. 1466–1476.
- [95] M. Sachan, A. Dubey, and E. P. Xing, "From textbooks to knowledge: A case study in harvesting axiomatic knowledge from textbooks to solve geometry problems," in *EMNLP*, 2017, pp. 784–795.
- [96] C. Alvin, S. Gulwani, R. Majumdar, and S. Mukhopadhyay, "Synthesis of problems for shaded area geometry reasoning," in *Artificial Intelligence in Education - 18th International Conference, AIED 2017, Wuhan, China, June 28 - July 1, 2017, Proceedings*, 2017, pp. 455–458.
- [97] P. Clark, O. Etzioni, T. Khot, A. Sabharwal, O. Tafjord, P. D. Turney, and D. Khashabi, "Combining retrieval, statistics, and inference to answer elementary science questions," in *AAAI*, 2016, pp. 2580–2586.
- [98] G. Cheng, W. Zhu, Z. Wang, J. Chen, and Y. Qu, "Taking up the gaokao challenge: An information retrieval approach," in *IJCAI*, 2016, pp. 2479–2485.
- [99] J. Hernández-Orallo, F. Martínez-Plumed, U. Schmid, M. Siebers, and D. L. Lowe, "Computer models solving intelligence test problems: Progress and implications (extended abstract)," in *IJCAI*, 2017, pp. 5005–5009.
- [100] H. Wang, F. Tian, B. Gao, C. Zhu, J. Bian, and T. Liu, "Solving verbal questions in IQ test by knowledge-powered word embedding," in *EMNLP*, 2016, pp. 541–550.
- [101] I. Lev, B. MacCartney, C. D. Manning, and R. Levy, "Solving logic puzzles: From robust processing to precise semantics," in *Proceedings of the 2Nd Workshop on Text Meaning and Interpretation*, ser. TextMean '04. Stroudsburg, PA, USA: Association for Computational Linguistics, 2004, pp. 9–16.
- [102] A. Dries, A. Kimmig, J. Davis, V. Belle, and L. D. Raedt, "Solving probability problems in natural language," in *IJCAI*, 2017, pp. 3981–3987.
- [103] L. De Raedt, A. Kimmig, and H. Toivonen, "Problog: A probabilistic prolog and its application in link discovery," in *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, ser. IJCAI'07, 2007, pp. 2468–2473.
- [104] W. Ling, D. Yogatama, C. Dyer, and P. Blunsom, "Program induction by rationale generation: Learning to solve and explain algebraic word problems," in *ACL*, 2017, pp. 158–167.
- [105] X. Yu, M. Wang, Z. Zeng, and J. Fan, "Solving directly-stated arithmetic word problems in chinese," in *2015 International Conference of Educational Innovation through Technology (EITT)*, Oct 2015, pp. 51–55.
- [106] X. Yu, P. Jian, M. Wang, and S. Wu, "Extraction of implicit quantity relations for arithmetic word problems in chinese," in *2016 International Conference on Educational Innovation through Technology (EITT)*, Sept 2016, pp. 242–245.
- [107] B. Siyam, A. A. Saa, O. Alqaryouti, and K. Shaalan, "Arabic arithmetic word problems solver," in *ACLING*, 2017, pp. 153–160.
- [108] K. Wang and Z. Su, "Dimensionally guided synthesis of mathematical word problems," in *IJCAI*, 2016, pp. 2661–2668.
- [109] R. Koncel-Kedziorski, I. Konstas, L. Zettlemoyer, and H. Hajishirzi, "A theme-rewriting approach for generating algebra word problems," in *EMNLP*, 2016, pp. 1617–1628.
- [110] O. Polozov, E. O'Rourke, A. M. Smith, L. Zettlemoyer, S. Gulwani, and Z. Popovic, "Personalized mathematical word problem generation," in *IJCAI*, 2015, pp. 381–388.
- [111] M. Gebser, R. Kaminski, B. Kaufmann, and T. Schaub, *Answer Set Solving in Practice*, ser. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2012.
- [112] C. Alvin, S. Gulwani, R. Majumdar, and S. Mukhopadhyay, "Synthesis of geometry proof problems," in *AAAI*, 2014, pp. 245–252.
- [113] —, "Automatic synthesis of geometry problems for an intelligent tutoring system," *CoRR*, vol. abs/1510.08525, 2015.
- [114] R. Singhal, M. Henz, and K. McGee, "Automated generation of high school geometric questions involving implicit construction," in *CSEDU*, 2014, pp. 467–472.
- [115] —, "Automated generation of geometry questions for high school mathematics," in *CSEDU*, 2014, pp. 14–25.
- [116] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *CoRR*, vol. abs/1409.0473, 2014.
- [117] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, pp. 484–503, 2016. [Online]. Available: <http://www.nature.com/nature/journal/v529/n7587/full/nature16961.html>
- [118] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, and A. Bolton, "Mastering the game of go without human knowledge," *Nature*, vol. 550, no. 7676, p. 354, 2017.