EQUATE : A Benchmark Evaluation Framework for Quantitative Reasoning in Natural Language Inference

Abhilasha Ravichander*, **Aakanksha Naik***,
Carolyn Rose, **Eduard Hovy**
Language Technologies Institute, Carnegie Mellon University
{aravicha, anaik, cprose, hovy}@cs.cmu.edu

Abstract

Quantitative reasoning is a higher-order reasoning skill that any intelligent natural language understanding system can reasonably be expected to handle. We present EQUATE[1] (Evaluating Quantitative Understanding Aptitude in Textual Entailment), a new framework for quantitative reasoning in textual entailment. We benchmark the performance of 9 published NLI models on EQUATE, and find that on average, state-of-the-art methods do not achieve an absolute improvement over a majority-class baseline, suggesting that they do not implicitly learn to reason with quantities. We establish a new baseline Q-REAS that manipulates quantities symbolically. In comparison to the best performing NLI model, it achieves success on numerical reasoning tests (+24.2%), but has limited verbal reasoning capabilities (-8.1%). We hope our evaluation framework will support the development of models of quantitative reasoning in language understanding.

1 Introduction

Numbers play a vital role in our lives. We reason with numbers in day-to-day tasks ranging from handling currency to reading news articles to understanding sports results, elections and stock markets. As numbers are used to communicate information accurately, reasoning with them is an essential core competence in understanding natural language (Levinson, 2001; Frank et al., 2008; Dehaene, 2011). A benchmark task in natural language understanding is natural language inference (NLI)(or recognizing textual entailment (RTE)) (Cooper et al., 1996; Condoravdi et al., 2003; Bos and Markert, 2005; Dagan et al., 2006), wherein a model determines if a natural language hypothesis can be justifiably inferred from a given premise[2]. Making such inferences often necessitates reasoning about quantities.

RTE-QUANT
P: After the deal closes, Teva will generate sales of about $ 7 billion a year, the company said.
H: Teva earns $ 7 billion a year.
AWP-NLI
P: Each of farmer Cunningham's 6048 lambs is either black or white and there are 193 white ones.
H: 5855 of Farmer Cunningham's lambs are black.
NEWSNLI
P: Emmanuel Miller, 16, and Zachary Watson, 17, are charged as adults, police said.
H: Two teen suspects charged as adults.
REDDITNLI
P: Oxfam says richest one percent to own more than rest by 2016.
H: Richest 1% To Own More Than Half Worlds Wealth By 2016 Oxfam.

Table 1: Examples from evaluation sets in EQUATE

Consider the following example from Table 1,

P: With 99.6% of precincts counted , Dewhurst held 48% of the vote to 30% for Cruz .
H: Lt. Gov. David Dewhurst fails to get 50% of primary vote.

To conclude the hypothesis is inferable, a model must reason that since 99.6% precincts are counted, even if all remaining precincts vote for Dewhurst, he would fail to get 50% of the primary vote. Scant attention has been paid to building datasets to evaluate this reasoning ability. To address this gap, we present EQUATE (Evaluating Quantity Understanding Aptitude in Textual Entailment) (§3), which consists of five evaluation sets, each

*The first two authors contributed equally to this work.
[1]Code and data available at https://github.com/AbhilashaRavichander/EQUATE.
[2]Often, this is posed as a three-way decision where the hypothesis can be inferred to be true (entailment), false (contradiction) or cannot be determined.

Proceedings of the 23rd Conference on Computational Natural Language Learning, pages 349–361
Hong Kong, China, November 3-4, 2019.

featuring different facets of quantitative reasoning in textual entailment (Table 1) (including verbal reasoning with quantities, basic arithmetic computation, dealing with approximations and range comparisons.).

We evaluate the ability of existing state-of-the-art NLI models to perform quantitative reasoning (§4.1), by benchmarking 9 published models on EQUATE. Our results show that most models are incapable of quantitative reasoning, instead relying on lexical cues for prediction. Additionally, we build Q-REAS , a shallow semantic reasoning baseline for quantitative reasoning in NLI (§4.2). Q-REAS is effective on synthetic test sets which contain more quantity-based inference, but shows limited success on natural test sets which require deeper linguistic reasoning. However, the hardest cases require a complex interplay between linguistic and numerical reasoning. The EQUATE evaluation framework makes it clear where this new challenge area for textual entailment stands.

2 Related Work

NLI has attracted community-wide interest as a stringent test for natural language understanding (Cooper et al., 1996; Fyodorov; Glickman et al., 2005; Haghighi et al., 2005; Harabagiu and Hickl, 2006; Romano et al., 2006; Dagan et al., 2006; Giampiccolo et al., 2007; Zanzotto et al., 2006; Malakasiotis and Androutsopoulos, 2007; MacCartney, 2009; de2; Dagan et al., 2010; Angeli and Manning, 2014; Marelli et al., 2014). Recently, the creation of large-scale datasets (Bowman et al., 2015; wil; Khot et al., 2018) spurred the development of many neural models (Parikh et al., 2016; Nie and Bansal, 2017; Conneau et al., 2017; Balazs et al., 2017; Chen et al., 2017a; Radford et al., 2018; Devlin et al., 2018).

However, state-of-the-art models for NLI treat the task like a matching problem, which appears to work in many cases, but breaks down in others. As the field moves past current models of the matching variety to ones that embody more of the reasoning we know is part of the task, we need benchmarks that will enable us to mark progress in the field. Prior work on challenge tasks has already made headway in defining tasks for subproblems such as lexical inference with hypernymy, co-hyponymy, antonymy (Glockner et al., 2018; Naik et al., 2018). In this work, we specifically probe into quantitative reasoning.

De Marneffe et al. (2008) find that in a corpus of real-life contradiction pairs collected from Wikipedia and Google News, *29%* contradictions arise from numeric discrepancies, and in the RTE-3 (Recognizing Textual Entailment) development set, numeric contradictions make up 8.8% of contradictory pairs. Naik et al. (2018) find that model inability to do numerical reasoning causes 4% of errors made by state-of-the-art models. Sammons et al. (2010); Clark (2018) argue for a systematic knowledge-oriented approach in NLI by evaluating specific semantic analysis tasks, identifying quantitative reasoning in particular as a focus area. Bentivogli et al. (2010) propose creating specialized datasets, but feature only 6 examples with quantitative reasoning. Our work bridges this gap by providing a more comprehensive examination of quantitative reasoning in NLI.

While to the best of our knowledge, prior work has not studied quantitative reasoning in NLI, Roy (2017) propose a model for a related subtask called *quantity entailment*, which aims to determine if a given quantity can be inferred from a sentence. In contrast, our work is concerned with general-purpose textual entailment which considers if a given *sentence* can be inferred from another. Our work also relates to solving arithmetic word problems (Hosseini et al., 2014; Mitra and Baral, 2016; Zhou et al., 2015; Upadhyay et al., 2016; Huang et al., 2017; Kushman et al., 2014a; Koncel-Kedziorski et al., 2015; roy; Roy, 2017; Ling et al., 2017a). A key difference is that word problems focus on arithmetic reasoning, while the requirement for linguistic reasoning and world knowledge is limited as the text is concise, straightforward, and self-contained (Hosseini et al., 2014; Kushman et al., 2014b). Our work provides a testbed that evaluates basic arithmetic reasoning while incorporating the complexity of natural language.

Recently, Dua et al. (2019) also recognize the importance of quantitative reasoning for text understanding. They propose DROP, a reading comprehension dataset focused on a limited set of discrete operations such as counting, comparison, sorting and arithmetic. In contrast, EQUATE features diverse phenomena that occur naturally in text, including reasoning with approximation, ordinals, implicit quantities and quantifiers, requiring NLI models to reason comprehensively about the interplay between quantities and language. Ad-

ditionally, through EQUATE we suggest the inclusion of controlled synthetic tests in evaluation benchmarks. Controlled tests act as basic validation of model behaviour, isolating model ability to reason about a property of interest.

3 Quantitative Reasoning in NLI

Our interpretation of "quantitative reasoning" draws from cognitive testing and education (Stafford, 1972; Ekstrom et al., 1976), which considers it "verbal problem-solving ability". While inextricably linked to mathematics, it is an inclusive skill involving everyday language rather than a specialized lexicon. To excel at quantitative reasoning, one must interpret quantities expressed in language, perform basic calculations, judge their accuracy, and justify quantitative claims using verbal and numeric reasoning. These requirements show a reciprocity: NLI lends itself as a test bed for quantitative reasoning, which conversely, is important for NLI (Sammons et al., 2010; Clark, 2018). Motivated by this, we present the EQUATE (Evaluating Quantity Understanding Aptitude in Textual Entailment) framework.

3.1 The EQUATE Dataset

EQUATE consists of five NLI test sets featuring quantities. Three of these tests for quantitative reasoning feature language from real-world sources such as news articles and social media (§3.2; §3.3; §3.4). We focus on sentences containing quantities with numerical values, and consider an entailment pair to feature quantitative reasoning if it is at least one component of the reasoning required to determine the entailment label (but not necessarily the only reasoning component). Quantitative reasoning features quantity matching, quantity comparison, quantity conversion, arithmetic, qualitative processes, ordinality and quantifiers, quantity noun and adverb resolution[3] as well as verbal reasoning with the quantity's textual context[4]. Appendix B gives some examples for these quantitative phenomena. We further filter sentence pairs which require only temporal reasoning, since specialized knowledge is needed to reason about time. These three test sets contain pairs which conflate multiple lexical and quantitative reasoning phenomena. In order to study aspects of quantitative reasoning in isolation, EQUATE further features two controlled synthetic tests (§3.5; §3.6).

3.2 RTE-Quant

This test set is constructed from the RTE subcorpus for quantity entailment (Roy, 2017), originally drawn from the RTE2-RTE4 datasets (Dagan et al., 2006). The original sub-corpus conflates temporal and quantitative reasoning. We discarded pairs requiring temporal reasoning, obtaining a set of 166 entailment pairs.

3.3 NewsNLI

This test set is created from the CNN corpus (Hermann et al., 2015) of news articles with abstractive summaries. We identify summary points with quantities, filtering out temporal expressions. For a summary point, the two most similar sentences[5] from the article are chosen, flipping pairs where the premise begins with a first-person pronoun (eg:⟨"He had nine pears", "Bob had nine pears"⟩ becomes ⟨"Bob had nine pears", "He had nine pears"⟩). The top 50% of similar pairs are retained to avoid lexical overlap bias. We crowdsource annotations for a subset of this data from Amazon Mechanical Turk. Crowdworkers[6] are shown two sentences and asked to determine whether the second sentence is definitely true, definitely false, or not inferable given the first. We collect 5 annotations per pair, and consider pairs with lowest token overlap between premise and hypothesis and least difference in premise-hypothesis lengths when stratified by entailment label. Top 1000 samples meeting these criteria form our final set. To validate crowdsourced labels, experts are asked to annotate 100 pairs. Crowdsourced gold labels match expert gold labels in 85% cases, while individual crowdworker labels match expert gold labels in 75.8%. Disagreements are manually resolved by experts and examples not featuring quantitiative reasoning are filtered, leaving a set of 968 samples.

3.4 RedditNLI

This test set is sourced from the popular social forum \reddit[7]. Since reasoning about quanti-

[3] Such as the quantities represented in *dozen*, *twice*, *teenagers*.

[4] For example, ⟨Obama cuts tax rate to 28%, Obama wants to cut tax rate to 28% as part of overhaul⟩.

[5] According to Jaccard similarity.

[6] We require crowdworkers to have an approval rate of 95% on at least 100 tasks and pass a qualification test.

[7] According to the Reddit User Agreement, users grant Reddit the right to make their content available to other organizations or individuals.

Source	Test Set	Size	Classes	Data Source	Annotation Source	Quantitative Phenomena
Natural	RTE-Quant	166	2	RTE2-RTE4	Experts	Arithmetic, Ranges, Quantifiers
	NewsNLI	968	2	CNN	Crowdworkers	Ordinals, Quantifiers, Arithmetic, Approximation, Magnitude, Ratios
	RedditNLI	250	3	Reddit	Experts	Range, Arithmetic, Approximation, Verbal
Synthetic	Stress Test	7500	3	AQuA-RAT	Automatic	Quantifiers
	AwpNLI	722	2	Arithmetic Word Problems	Automatic	Arithmetic

Table 2: An overview of test sets included in EQUATE. RedditNLI and Stress Test are framed as 3-class (entailment, neutral, contradiction) while RTE-Quant, NewsNLI and AwpNLI are 2-class (entails=yes/no). RTE 2-4 formulate entailment as a 2-way decision. We find that few news article headlines are contradictory, thus NewsNLI is similarly framed as a 2-way decision. For algebra word problems, substituting the wrong answer in the hypothesis necessarily creates a contradiction under the event coreference assumption (De Marneffe et al., 2008), thus it is framed as a 2-way decision as well.

ties is important in domains like finance or economics, we scrape all headlines from the posts on \r\economics, considering titles that contain quantities and do not have meta-forum information. Titles appearing within three days of each other are clustered by Jaccard similarity, and the top 300 pairs are extracted. After filtering out nonsensical titles, such as concatenated stock prices, we are left with 250 sentence pairs. Similar to RTE, two expert annotators label these pairs, achieving a Cohen's kappa of 0.82. Disagreements are discussed to resolve final labels.

3.5 Stress Test

We include the numerical reasoning stress test from (Naik et al., 2018) as a synthetic sanity check. The stress test consists of 7500 entailment pairs constructed from sentences in algebra word problems (Ling et al., 2017b). Focusing on quantifiers, it requires models to compare entities from hypothesis to the premise while incorporating quantifiers, but does not require them to perform the computation from the original algebra word problem (eg: ⟨"NHAI employs 100 men to build a highway of 2 km in 50 days working 8 hours a day","NHAI employs less than 700 men to build a highway of 2 km in 50 days working 8 hours a day"⟩).

3.6 AwpNLI

To evaluate arithmetic ability of NLI models, we repurpose data from arithmetic word problems (roy). They have the following characteristic structure. First, they establish a world and optionally update its state. Then, a question is posed about the world. This structure forms the basis of our pair creation procedure. World building and update statements form the premise. A hypothesis template is generated by identifying modal/auxiliary verbs in the question, and subsequent verbs, which we call secondary verbs. We identify the agent and conjugate the secondary verb in present tense followed by the identified unit to form the final template (for example, the algebra word problem 'Gary had 73.0 dollars. He spent 55.0 dollars on a pet snake. How many dollars did Gary have left?' would generate the hypothesis template 'Agent(Gary) Verb(Has) Answer(18.0) Unit(dollars) left'). For every template, the correct guess is used to create an entailed hypothesis. Contradictory hypotheses are created by randomly sampling a wrong guess ($x \in \mathbb{Z}^+$ if correct guess is an integer, and $x \in \mathbb{R}^+$ if it is a real number) [8]. We check for grammaticality, finding only 2% ungrammatical hypotheses, which are manually corrected leaving a set of 722 pairs.

4 Models

We describe the 9 NLI models[9] used in this study, as well as our new baseline. The interested reader is invited to refer to the corresponding publications for further details.

4.1 NLI Models

1) **Majority Class (MAJ):** Simple baseline that always predicts the majority class in test set.

[8] From a uniform distribution over an interval of 10 around the correct guess (or 5 for numbers less than 5), to identify plausible wrong guesses.

[9] Accuracy of all models on MultiNLI closely matches original publications (numbers in appendix A).

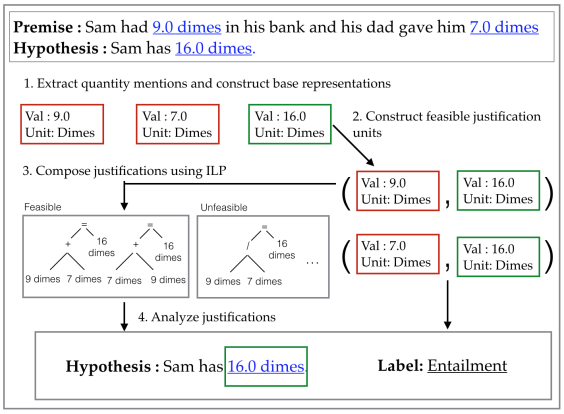

Figure 1: Overview of Q-REAS baseline.

2) **Hypothesis-Only (HYP):** FastText classifier (Mikolov et al., 2018) trained on only hypotheses (Gururangan et al., 2018).
3) **ALIGN:** A bag-of-words alignment model inspired by MacCartney (2009).[10]
4) **CBOW:** A simple bag-of-embeddings sentence representation model (wil).
5) **BiLSTM:** The simple BiLSTM model described by wil.
6) **Chen (CH):** Stacked BiLSTM-RNNs with shortcut connections and character word embeddings (Chen et al., 2017b).
7) **InferSent:** A single-layer BiLSTM-RNN model with max-pooling (Conneau et al., 2017).
8) **SSEN:** Stacked BiLSTM-RNNs with shortcut connections (Nie and Bansal, 2017).
9) **ESIM:** Sequential inference model proposed by Chen et al. (2017a) which uses BiLSTMs with an attention mechanism.
10) **OpenAI GPT:** Transformer-based language model (Vaswani et al., 2017), with finetuning on NLI (Radford et al., 2018).
11) **BERT:** Transformer-based language model (Vaswani et al., 2017), with a cloze-style and next-sentence prediction objective, and finetuning on NLI (Devlin et al., 2018).

4.2 Q-REAS Baseline System

Figure 1 describes the Q-REAS baseline for quantitative reasoning in NLI. The model manipulates quantity representations symbolically to make entailment decisions, and is intended to serve as a strong heuristic baseline for numerical reasoning on the EQUATE benchmark. This model has four stages: Quantity mentions are extracted and parsed into semantic representations called NUMSETS (§4.2.1, §4.2.2); compatible NUMSETS are extracted (§4.2.3) and composed (§4.2.4) to form *justifications*; Justifications are analyzed to determine entailment labels (§4.2.5).

[10]Model accuracy on RTE-3 test is 61.12%, comparable to the reported average model performance in the RTE competition of 62.4% .

INPUT	
P_c	Set of "compatible" single-valued premise quantities
P_r	Set of "compatible" range-valued premise quantities
H	Hypothesis quantity
O	Operator set $\{+, -, *, /, =, \cap, \cup, \setminus, \subseteq\}$
L	Length of equation to be generated
SL	Symbol list ($P_c \cup P_r \cup H \cup O$)
TL	Type list (set of types from P_c, P_r, H)
N	Length of symbol list
K	Index of first range quantity in symbol list
M	Index of first operator in symbol list
OUTPUT	
e_i	Index of symbol assigned to i^{th} position in postfix equation
VARIABLES	
x_i	Main ILP variable for position i
c_i	Indicator variable: is e_i a single value?
r_i	Indicator variable: is e_i a range?
o_i	Indicator variable: is e_i an operator?
d_i	Stack depth of e_i
t_i	Type index for e_i

Table 3: Input, output and variable definitions for the Integer Linear Programming (ILP) framework used for quantity composition

4.2.1 Quantity Segmenter

We follow Barwise and Cooper (1981) in defining quantities as having a number, unit, and an optional approximator. Quantity mentions are identified as least ancestor noun phrases from the constituency parse of the sentence containing cardinal numbers.

4.2.2 Quantity Parser

The quantity parser constructs a grounded representation for each quantity mention in the premise or hypothesis, henceforth known as a NUMSET [11]. A NUMSET is a tuple (val, unit, ent, adj, loc, verb, freq, flux)[12] with:

1. val $\in [\mathbb{R}, \mathbb{R}]$: quantity value represented as a range
2. unit $\in S$: unit noun associated with quantity
3. ent $\in S^{\phi}$: entity noun associated with unit (e.g., *donations* worth 100$)

[11]A NUMSET may be a composition of other NUMSETS .

[12]As in (Koncel-Kedziorski et al., 2015) S denotes all possible spans in the sentence, ϕ represents the empty span, and S^{ϕ}=$S \cup \phi$

Definitional Constraints	
Range restriction	$x_i < K$ or $x_i = M-1$ for $i \in [0, L-1]$ if $c_i = 1$
	$x_i \geq K$ and $x_i < M$ for $i \in [0, L-1]$ if $r_i = 1$
	$x_i \geq M$ for $i \in [0, L-1]$ if $o_i = 1$
Uniqueness	$c_i + r_i + o_i = 1$ for $i \in [0, L-1]$
Stack definition	$d_0 = 0$ (Stack depth initialization)
	$d_i = d_{i-1} - 2o_i + 1$ for $i \in [0, L-1]$ (Stack depth update)
Syntactic Constraints	
First two operands	$c_0 + r_0 = 1$ and $c_1 + r_1 = 1$
Last operator	$x_{L-1} \geq N-1$ (Last operator should be one of $\{=, \subseteq\}$)
Last operand	$x_{L-2} = M-1$ (Last operand should be hypothesis quantity)
Other operators	$x_i \leq N-2$ for $i \in [0, L-3]$ if $o_i = 1$
Other operands	$x_i < K$ for $i \in [0, L-3]$ if $c_i = 1$
	$x_i < M$ for $i \in [0, L-3]$ if $r_i = 1$
Empty stack	$d_{L-1} = 0$ (Non-empty stack indicates invalid postfix expression)
Premise usage	$x_i \neq x_j$ for $i, j \in [0, L-1]$ if $o_i \neq 1, o_j \neq 1$
Operand Access	
Right operand	$op2(x_i) = x_{i-1}$ for $i \in [0, L-1]$ such that $o_i = 1$
Left operand	$op1(x_i) = x_l$ for $i, l \in [0, L-1]$ where $o_i = 1$ and l is the largest index such that $l \leq (i-2)$ and $d_l = d_i$

Table 4: Mathematical validity constraint definitions for the ILP framework. Functions $op1()$ and $op2()$ return the left and right operands for an operator respectively. Variables defined in table 3.

Type Consistency Constraints	
Type assignment	$t_i = TL[k]$ for $i \in [0, L-1]$ if $c_i + r_i = 1$ and $type(SL_i) = k$
Two type match	$t_i = t_a = t_b$ for $i \in [0, L-1]$ such that $o_i = 1, x_i \in \{+, -, *, /, =, \cap, \cup, \setminus, \subseteq\}, a = op1(x_i), b = op2(x_i)$
One type match	$t_i \in \{t_a, t_b\}, t_a \neq t_b$ for $i \in [0, L-1]$ such that $o_i = 1, x_i = *, a = op1(x_i), b = op2(x_i)$
	$t_i = t_a \neq t_b$ for $i \in [0, L-1]$ such that $o_i = 1, x_i = /, a = op1(x_i), b = op2(x_i)$
Operator Consistency Constraints	
Arithmetic operators	$c_a = c_b = 1$ for $i \in [0, L-1]$ such that $o_i = 1, x_i \in \{+, -, *, /, =\}, a = op1(x_i), b = op2(x_i)$
Range operators	$r_a = r_b = 1$ for $i \in [0, L-1]$ such that $o_i = 1, x_i \in \{\cap, \cup, \setminus\}, a = op1(x_i), b = op2(x_i)$
	$r_b = 1$ for $i \in [0, L-1]$ such that $o_i = 1, x_i = \subseteq, b = op2(x_i)$

Table 5: Linguistic consistency constraint definitions for the ILP framework. Functions $op1()$ and $op2()$ return the left and right operands for an operator respectively. Variables defined in table 3.

4. adj $\in S^{\phi}$: adjective associated with unit if any[13],
5. loc $\subseteq S^{\phi}$: location of unit (e.g.,'in the bag')[14]
6. verb $\in S^{\phi}$: action verb associated with quantity[15].
7. freq $\subseteq S^{\phi}$: if quantity recurs[16] (e.g, 'per hour'),
8. flux $\in \{$increase to, increase from, decrease to, decrease from$\}^{\phi}$: if quantity is in a state of flux[17].

To extract **values** for a quantity, we extract cardinal numbers, recording contiguity. We normalize the number[18]. We also handle simple ratios such as quarter, half etc, and extract bounds (eg: *fewer than 10 apples* is parsed to $[-\infty, 10]$ apples.)

To extract **units**, we examine tokens adjacent to cardinal numbers in the quantity mention and identify known units. If no known units are found, we assign the token in a *numerical modifier* relationship with the cardinal number, else we assign the nearest noun to the cardinal number as the unit. A quantity is determined to be **approximate** if the word in an *adverbial modifier* relation with the cardinal number appears in a gazetteer[19]. If approximate, range is extended to (+/-)2% of the

[13]Extracted as governing verb linked to entity by an *amod* relation.

[14]Extracted as prepositional phrase attached to the quantity and containing noun phrase.

[15]Extracted as governing verb linked to entity by *dobj* or *nsubj* relation.

[16]extracted using keywords *per* and *every*

[17]using gazetteer: *increasing*, *rising*, *rose*, *decreasing*, *falling*, *fell*, *drop*

[18](remove ","s, convert written numbers to float, decide the numerical value, for example hundred fifty eight thousand is 158000, two fifty eight is 258, 374m is 3740000 etc.). If cardinal numbers are non-adjacent, we look for an explicitly mentioned range such as 'to' and 'between'.

[19]roughly, approximately, about, nearly, roundabout, around, circa, almost, approaching, pushing, more or less, in the neighborhood of, in the region of, on the order of,something like, give or take (a few), near to, close to, in the ballpark of

M \ D	RTE-Q	Δ	NewsNLI	Δ	RedditNLI	Δ	NR ST	Δ	AWPNLI	Δ	Nat. Avg. Δ	Synth. Avg. Δ	All Avg. Δ
MAJ	57.8	0.0	50.7	0.0	**58.4**	**0.0**	33.3	0.0	50.0	0.0	+0.0	+0.0	+0.0
HYP	49.4	-8.4	52.5	+1.8	40.8	-17.6	31.2	-2.1	50.1	+0.1	-8.1	-1.0	-5.2
ALIGN	62.1	+4.3	56.0	+5.3	34.8	-23.6	22.6	-10.7	47.2	-2.8	-4.7	-6.8	-5.5
CBOW	47.0	-10.8	61.8	+11.1	42.4	-16.0	30.2	-3.1	50.7	+0.7	-5.2	-1.2	-3.6
BiLSTM	51.2	-6.6	63.3	+12.6	50.8	-7.6	31.2	-2.1	50.7	+0.7	-0.5	-0.7	-0.6
CH	54.2	-3.6	64.0	+13.3	55.2	-3.2	30.3	-3.0	50.7	+0.7	+2.2	-1.2	+0.9
InferSent	66.3	+8.5	65.3	+14.6	29.6	-28.8	28.8	-4.5	50.7	+0.7	-1.9	-1.9	-1.9
SSEN	58.4	+0.6	65.1	+14.4	49.2	-9.2	28.4	-4.9	50.7	+0.7	+1.9	-2.1	+0.3
ESIM	54.8	-3.0	62.0	+11.3	45.6	-12.8	21.8	-11.5	50.1	+0.1	-1.5	-5.7	-3.2
GPT	**68.1**	**+10.3**	72.2	+21.5	52.4	-6.0	36.4	+3.1	50.0	+0.0	+8.6	+1.6	+5.8
BERT	57.2	-0.6	**72.8**	**+22.1**	49.6	-8.8	36.9	+3.6	42.2	-7.8	+4.2	-2.1	+1.7
Q-REAS	56.6	-1.2	61.1	+10.4	50.8	-7.6	**63.3**	**+30**	**71.5**	**+21.5**	+0.5	+25.8	**+10.6**

Table 6: Accuracies(%) of 9 NLI Models on five tests for quantitiative reasoning in entailment. M and D represent *models* and *datasets* respectively. Δ captures improvement over majority-class baseline for a dataset. Column Nat.Avg. reports the average accuracy(%) of each model across 3 evaluation sets constructed from natural sources (RTE-Quant, NewsNLI, RedditNLI), whereas Synth.Avg. reports the average accuracy(%) on 2 synthetic evaluation sets (Stress Test, AwpNLI). Column Avg. represents the average accuracy(%) of each model across all 5 evaluation sets in EQUATE.

current value.

4.2.3 Quantity Pruner

The pruner constructs "compatible" premise-hypothesis NUMSET pairs. Consider the pair "Insurgents killed *7 U.S. soldiers*, set off a car bomb that killed *four Iraqi policemen*" and "*7 US soldiers* were killed, and *at least 10 Iraqis* died". Our parser extracts NUMSETS corresponding to *"four Iraqi policemen"* and *"7 US soldiers"* from premise and hypothesis respectively. But these NUMSETS should not be compared as they involve different units. The pruner discards such incompatible pairs. Heuristics to identify unit-compatible NUMSET pairs include three cases- 1) direct string match, 2) synonymy/hypernymy relations from WordNet, 3) one unit is a nationality/ job[20] and the other unit is synonymous with person (Roy, 2017).

4.2.4 Quantity Composition

The composition module detects whether a hypothesis NUMSET is justified by composing "compatible" premise NUMSETS . For example, consider the pair "I had *3 apples* but gave *one* to my brother" and "I have *two apples*". Here, the premise NUMSETS P_1 ("*3 apples*") and P_2 ("*one apple*") must be composed to deduce that the hypothesis NUMSET H_1 ("*2 apples*") is justified. Our framework accomplishes this by generating postfix arithmetic equations[21] from premise NUMSETS , that justify the hypothesis NUMSET [22]. In this example, the expression $< P_1, P_2, -, H_1, =>$ will be generated.

The set of possible equations is exponential in number of NUMSETS , making exhaustive generation intractable. But a large number of equations are invalid as they violate constraints such as unit consistency. Thus, our framework uses integer linear programming (ILP) to constrain the equation space. It is inspired by prior work on algebra word problems (Koncel-Kedziorski et al., 2015), with some key differences:

1. **Arithmetic equations:** We focus on arithmetic equations instead of algebraic ones.
2. **Range arithmetic:** Quantitative reasoning involves ranges, which are handled by representing then as endpoint-inclusive intervals and adding the four operators ($\cup, \cap, \setminus, \subseteq$)
3. **Hypothesis quantity-driven:** We optimize an ILP model for each hypothesis NUMSET because a sentence pair is marked "entailment" iff every hypothesis quantity is justified.

Table 3 describes ILP variables. We impose the following types of constraints:
1. Definitional Constraints: Ensure that ILP variables take on valid values by constraining initialization, range, and update.
2. Syntactic Constraints: Assure syntactic validity of generated postfix expressions by limiting

[20]Lists of jobs, nationalities scraped from Wikipedia.

[21]Note that arithmetic equations differ from algebraic equations in that they do *not* contain unknown variables

[22]Direct comparisons are incorporated by adding "=" as an operator.

operator-operand ordering.

3. Operand Access: Simulate stack-based evaluation correctly by choosing correct operator-operand assignments.

4. Type Consistency: Ensure that all operations are type-compatible.

5. Operator Consistency: Force range operators to have range operands and mathematical operators to have single-valued operands.

Definitional, syntactic, and operand access constraints ensure mathematical validity while type and operator consistency constraints add linguistic consistency. Constraint formulations are provided in Tables 4 and 5. We limit tree depth to 3 and retrieve a maximum of 50 solutions per hypothesis NUMSET , then solve to determine whether the equation is mathematically correct. We discard equations that use invalid operations (division by 0) or add unnecessary complexity (multiplication/ division by 1). The remaining equations are considered plausible justifications .

4.2.5 Global Reasoner

The global reasoner predicts the final entailment label as shown in Algorithm 1[23], on the assumption that every NUMSET in the hypothesis *has* to be justified [24] for entailment.

Algorithm 1 PredictEntailmentLabel(P, H, C, E)

Input: Premise quantities P, Hypothesis quantities H, Compatible pairs C, Equations E
Output: Entailment label $l \in$ { e, c, n }

if $C = \emptyset$ **then return** n
$J \leftarrow \emptyset$
$L \leftarrow []$
for $q_h \in H$ **do**
    $J_h \leftarrow \{q_p \mid q_p \in P, (q_p, q_h) \in C\}$
    $J \leftarrow J \cup \{(q_h, J_h)\}$
    $L \leftarrow L + [false]$
for $(q_h, J_h) \in J$ **do**
    if $J_h = \emptyset$ **then return** n
    for $q_p \in J_h$ **do**
        $s \leftarrow$ MaxSimilarityClass(q_p, q_h)
        if $s = e$ **then**
            if ValueMatch(q_p, q_h) **then**
                $L[q_h] = true$
            if !ValueMatch(q_p, q_h) **then**
                $L[q_h] = false$
        if $s = c$ **then**
            if ValueMatch(q_p, q_h) **then**
                $L[q_h] = c$
for $q_h \in H$ **do**
    $E_q \leftarrow \{e_i \in E \mid hyp(e_i) = q_h\}$
    if $E_q \neq \emptyset$ **then**
        $L[q_h] = true$
if $c \in L$ **then return** c
if count($L, true$) = len(L) **then return** e
return n

5 Results and Discussion

Table 6 presents results on EQUATE. All models, except Q-REAS are trained on MultiNLI. Q-REAS utilizes WordNet and lists from Wikipedia. We observe that neural models, particularly OpenAI GPT excel at verbal aspects of quantitative reasoning (RTE-Quant, NewsNLI), whereas Q-REAS excels at numerical aspects (Stress Test, AwpNLI).

5.1 Neural Models on NewsNLI:

To tease apart contributory effects of numerical and verbal reasoning in natural data, we experiment with NewsNLI. We extract all entailed pairs where a quantity appears in both premise and hypothesis, and perturb the quantity in the hypothesis generating contradictory pairs. For example, the pair ⟨'In addition to 79 fatalities , some 170 passengers were injured.'⟩ 'The crash took the lives of 79 people and injured some 170', 'entailment' is changed to ⟨'In addition to 79 fatalities , some 170 passengers were injured.', 'The crash took the lives of 80 people and injured some 170', 'contradiction'⟩, assuming scalar implicature and event coreference. Our perturbed test set contains 218 pairs. On this set, GPT[25] achieves an accuracy of 51.18%, as compared to 72.04% on the unperturbed set, suggesting the model relies on verbal cues rather than numerical reasoning. In comparison, Q-REAS achieves an accuracy of 98.1% on the perturbed set, compared to 75.36% on the unperturbed set, highlighting

[23] MaxSimilarityClass() takes two quantities and returns a probability distribution over entailment labels based on unit match. Similarly, ValueMatch() detects whether two quantities match in value (this function can also handle ranges).

[24] This is a necessary but not sufficient condition for entailment. Consider the example, ⟨'Sam believed Joan had 5 apples', 'Joan had 5 apples'⟩. The hypothesis quantities of 5 apples is justified but is not a sufficient condition for entailment.

[25] the best-performing neural model on EQUATE.

reliance on quantities rather than verbal information. Closer examination reveals that OpenAI switches to predicting the 'neutral' category for perturbed samples instead of entailment, accounting for 42.7% of its errors, possibly symptomatic of lexical bias issues (Naik et al., 2018).

5.2 What Quantitative Phenomena Are Hard?

We sample 100 errors made by Q-REAS on each test in EQUATE, to identify phenomena not addressed by simple quantity comparison. Our analysis of causes for error suggest avenues for future research:

1. **Multi-step numerical-verbal reasoning:** Models do not perform well on examples requiring interleaved verbal and quantitative reasoning, especially multi-step deduction. Consider the pair ⟨"Two people were injured in the attack", "Two people perpetrated the attack"⟩. Quantities "two people" and "two people" are unit-compatible, but must not be compared. Another example is the NewsNLI entailment pair in Table 1. This pair requires us to identify that 16 and 17 refer to Emmanuel and Zachary's ages (quantitative), deduce that this implies they are teenagers (verbal) and finally count them (quantitative) to get the hypothesis quantity "two teens". Numbers and language are intricately interleaved and developing a reasoner capable of handling such complex interplay is challenging.

2. **Lexical inference:** Lack of real world knowledge causes errors in identifying quantities and valid comparisons. Errors include mapping abbreviations to correct units ("m" to "meters"), detecting part-whole coreference ("seats" can be used to refer to "buses"), and resolving hypernymy/hyponymy ("young men" to "boys").

3. **Inferring underspecified quantities:** Quantity attributes can be implicitly specified, requiring inference to generate a complete representation. Consider "A mortar attack killed four people and injured 80". A system must infer that the quantity "80" refers to people. On RTE-Quant, 20% of such cases stem from zero anaphora, a hard problem in coreference resolution.

4. **Arithmetic comparison limitations:** These examples require composition between incompatible quantities. For example, consider ⟨"There were 3 birds and 6 nests", "There were 3 more nests than birds"⟩. To correctly label this pair "3 birds" and "6 nests" must be composed.

6 Conclusion

In this work, we present EQUATE, an evaluation framework to estimate the ability of models to reason quantitatively in textual entailment. We observe that existing neural approaches rely heavily on the lexical matching aspect of the task to succeed rather than reasoning about quantities. We implement a strong symbolic baseline Q-REAS that achieves success at numerical reasoning, but lacks sophisticated verbal reasoning capabilities. The EQUATE resource presents an opportunity for the community to develop powerful hybrid neuro-symbolic architectures, combining the strengths of neural models with specialized reasoners such as Q-REAS . We hope our insights lead to the development of models that can more precisely reason about the important, frequent, but understudied, phenomena of quantities in natural language.

Acknowledgments

This research was supported in part by grants from the National Science Foundation Secure and Trustworthy Computing program (CNS-1330596, CNS-15-13957, CNS-1801316, CNS-1914486) and a DARPA Brandeis grant (FA8750-15-2-0277). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the NSF, DARPA, or the US Government. The author Naik was supported by a fellowship from the Center of Machine Learning and Health at Carnegie Mellon University. The authors would like to thank Graham Neubig, Mohit Bansal and Dongyeop Kang for helpful discussion regarding this work, and Shruti Rijhwani and Siddharth Dalmia for reviews while drafting this paper. The authors are also grateful to Lisa Carey Lohmueller and Xinru Yan for volunteering their time for pilot studies.

References

Gabor Angeli and Christopher D Manning. 2014. Naturalli: Natural logic inference for common sense reasoning. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 534–545.

Jorge Balazs, Edison Marrese-Taylor, Pablo Loyola, and Yutaka Matsuo. 2017. Refining raw sentence representations for textual entailment recognition via attention. In *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*, pages 51–55, Copenhagen, Denmark. Association for Computational Linguistics.

Jon Barwise and Robin Cooper. 1981. Generalized quantifiers and natural language. In *Philosophy, Language, and Artificial Intelligence*, pages 241–301. Springer.

Luisa Bentivogli, Elena Cabrio, Ido Dagan, Danilo Giampiccolo, Medea Lo Leggio, and Bernardo Magnini. 2010. Building textual entailment specialized data sets: a methodology for isolating linguistic phenomena relevant to inference. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Languages Resources Association (ELRA).

Johan Bos and Katja Markert. 2005. Recognising textual entailment with logical inference. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 628–635. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017a. Enhanced lstm for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668, Vancouver, Canada. Association for Computational Linguistics.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017b. Recurrent neural network-based sentence encoder with gated attention for natural language inference. In *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*, pages 36–40, Copenhagen, Denmark. Association for Computational Linguistics.

Peter Clark. 2018. What knowledge is needed to solve the rte5 textual entailment challenge? *arXiv preprint arXiv:1806.03561*.

Cleo Condoravdi, Dick Crouch, Valeria De Paiva, Reinhard Stolle, and Daniel G Bobrow. 2003. Entailment, intensionality and text understanding. In *Proceedings of the HLT-NAACL 2003 workshop on Text meaning-Volume 9*, pages 38–45. Association for Computational Linguistics.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Robin Cooper, Dick Crouch, Jan Van Eijck, Chris Fox, Johan Van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, et al. 1996. Using the framework. Technical report.

Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. 2010. The fourth pascal recognizing textual entailment challenge. *Journal of Natural Language Engineering*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*, pages 177–190. Springer.

Marie-Catherine De Marneffe, Anna N. Rafferty, and Christopher D. Manning. 2008. Finding contradictions in text. In *Proceedings of ACL-08: HLT*, pages 1039–1047, Columbus, Ohio. Association for Computational Linguistics.

Stanislas Dehaene. 2011. *The number sense: How the mind creates mathematics*. OUP USA.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *CoRR*, abs/1903.00161.

Ruth B Ekstrom, Diran Dermen, and Harry Horace Harman. 1976. *Manual for kit of factor-referenced cognitive tests*, volume 102. Educational Testing Service Princeton, NJ.

Michael C Frank, Daniel L Everett, Evelina Fedorenko, and Edward Gibson. 2008. Number as a cognitive technology: Evidence from pirahã language and cognition. *Cognition*, 108(3):819–824.

Yaroslav Fyodorov. A natural logic inference system. Citeseer.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*, pages 1–9. Association for Computational Linguistics.

Oren Glickman, Ido Dagan, and Moshe Koppel. 2005. Web based probabilistic textual entailment.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking nli systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*.

Aria Haghighi, Andrew Ng, and Christopher Manning. 2005. Robust textual inference via graph matching. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*.

Sanda Harabagiu and Andrew Hickl. 2006. Methods for using textual entailment in open-domain question answering. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 905–912. Association for Computational Linguistics.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701.

Mohammad Javad Hosseini, Hannaneh Hajishirzi, Oren Etzioni, and Nate Kushman. 2014. Learning to solve arithmetic word problems with verb categorization. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 523–533.

Danqing Huang, Shuming Shi, Chin-Yew Lin, and Jian Yin. 2017. Learning fine-grained expressions to solve math word problems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 805–814.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. Scitail: A textual entailment dataset from science question answering.

Rik Koncel-Kedziorski, Hannaneh Hajishirzi, Ashish Sabharwal, Oren Etzioni, and Siena Dumas Ang. 2015. Parsing algebraic word problems into equations. *Transactions of the Association for Computational Linguistics*, 3:585–597.

Nate Kushman, Yoav Artzi, Luke Zettlemoyer, and Regina Barzilay. 2014a. Learning to automatically solve algebra word problems. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 271–281.

Nate Kushman, Luke S. Zettlemoyer, Regina Barzilay, and Yoav Artzi. 2014b. Learning to automatically solve algebra word problems. In *ACL*.

Stephen C Levinson. 2001. Pragmatics. In *International Encyclopedia of Social and Behavioral Sciences: Vol. 17*, pages 11948–11954. Pergamon.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017a. Program induction by rationale generation: Learning to solve and explain algebraic word problems. *arXiv preprint arXiv:1705.04146*.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017b. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 158–167, Vancouver, Canada. Association for Computational Linguistics.

Bill MacCartney. 2009. *Natural language inference*. Stanford University.

Prodromos Malakasiotis and Ion Androutsopoulos. 2007. Learning textual entailment using svms and string similarity measures. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 42–47. Association for Computational Linguistics.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, Roberto Zamparelli, et al. 2014. A sick cure for the evaluation of compositional distributional semantic models.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Arindam Mitra and Chitta Baral. 2016. Learning to use formulas to solve simple arithmetic problems. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2144–2153.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference.

In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yixin Nie and Mohit Bansal. 2017. Shortcut-stacked sentence encoders for multi-domain inference. In *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*, pages 41–45, Copenhagen, Denmark. Association for Computational Linguistics.

Ankur P Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. *arXiv preprint arXiv:1606.01933*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Lorenza Romano, Milen Kouylekov, Idan Szpektor, Ido Dagan, and Alberto Lavelli. 2006. Investigating a generic paraphrase-based approach for relation extraction.

Subhro Roy. 2017. *Reasoning about quantities in natural language*. Ph.D. thesis, University of Illinois at Urbana-Champaign.

Mark Sammons, VG Vydiswaran, and Dan Roth. 2010. Ask not what textual entailment can do for you... In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1199–1208. Association for Computational Linguistics.

Richard E Stafford. 1972. Hereditary and environmental components of quantitative reasoning. *Review of Educational Research*, 42(2):183–201.

Shyam Upadhyay, Ming-Wei Chang, Kai-Wei Chang, and Wen-tau Yih. 2016. Learning from explicit and implicit supervision jointly for algebra word problems. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 297–306.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

F Zanzotto, Alessandro Moschitti, Marco Pennacchiotti, and M Pazienza. 2006. Learning textual entailment from examples. In *Second PASCAL recognizing textual entailment challenge*, page 50. PASCAL.

Lipu Zhou, Shuaixiang Dai, and Liwei Chen. 2015. Learn to solve algebra word problems using quadratic programming. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 817–822.

Appendix

A Baseline performance on MultiNLI-Dev Matched

Model	MultiNLI Dev
Hyp Only	53.18%
ALIGN	45.0%
CBOW	63.5%
BiLSTM	70.2%
Chen	73.7%
NB	74.2%
InferSent	70.3%
ESIM	76.2%
OpenAI Transformer	81.35%
BERT	83.8%

Table 7: Performance of all baseline models used in the paper on the matched subset of MultiNLI-Dev

Table 7 presents classification accuracies of all baseline models used in this work on the matched subset of MultiNLI-Dev. These scores are very close to the numbers reported by the original publications, affirming the correctness of our baseline setup.

B Examples of quantitative phenomena present in EQUATE

Table 8 presents some examples from EQUATE which demonstrate interesting quantitative phenomena that must be understood to label the pair correctly.

Phenomenon	Example
Arithmetic	**P:** Sharper faces charges in Arizona and California **H:** Sharper has been charged in two states
Ranges	**P:** Between 20 and 30 people were trapped in the casino **H:** Upto 30 people thought trapped in casino
Quantifiers	**P:** Poll: Obama over 50% in Florida **H:** New poll shows Obama ahead in Florida
Ordinals	**P:** Second-placed Nancy celebrated their 40th anniversary with a win **H:** Nancy stay second with a win
Approximation	**P:** Rwanda has dispatched 1917 soldiers **H:** Rwanda has dispatched some 1900 soldiers
Ratios	**P:** Londoners had the highest incidence of E. Coli bacteria (25%) **H:** 1 in 4 Londoners have E. Coli bacteria
Comparison	**P:** Treacherous currents took four lives on the Alabama Gulf coast **H:** Rip currents kill four in Alabama
Conversion	**P:** If the abuser has access to a gun, it increases chances of death by 500% **H:** Victim five times more likely to die if abuser is armed
Numeration	**P:** Eight suspects were arrested **H:** 8 suspects have been arrested
Implicit Quantities	**P:** The boat capsized two more times **H:** His sailboat capsized three times

Table 8: Examples of quantitative phenomena present in EQUATE