Injecting Numerical Reasoning Skills into Language Models

Mor Geva*
Tel Aviv University,
Allen Institute for AI
morgeva@mail.tau.ac.il

Ankit Gupta*
Tel Aviv University
ankitgupta.iitkanpur@gmail.com

Jonathan Berant
Tel Aviv University,
Allen Institute for AI
joberant@cs.tau.ac.il

- high-level reasoning skills, such as numerical reasoning, are difficult to learn from a language-modeling objective only.

- In this work, we show that numerical reasoning is amenable to automatic data generation, and thus one can inject this skill into pre-trained LMs, by generating large amounts of data, and training in a multi-task setup.

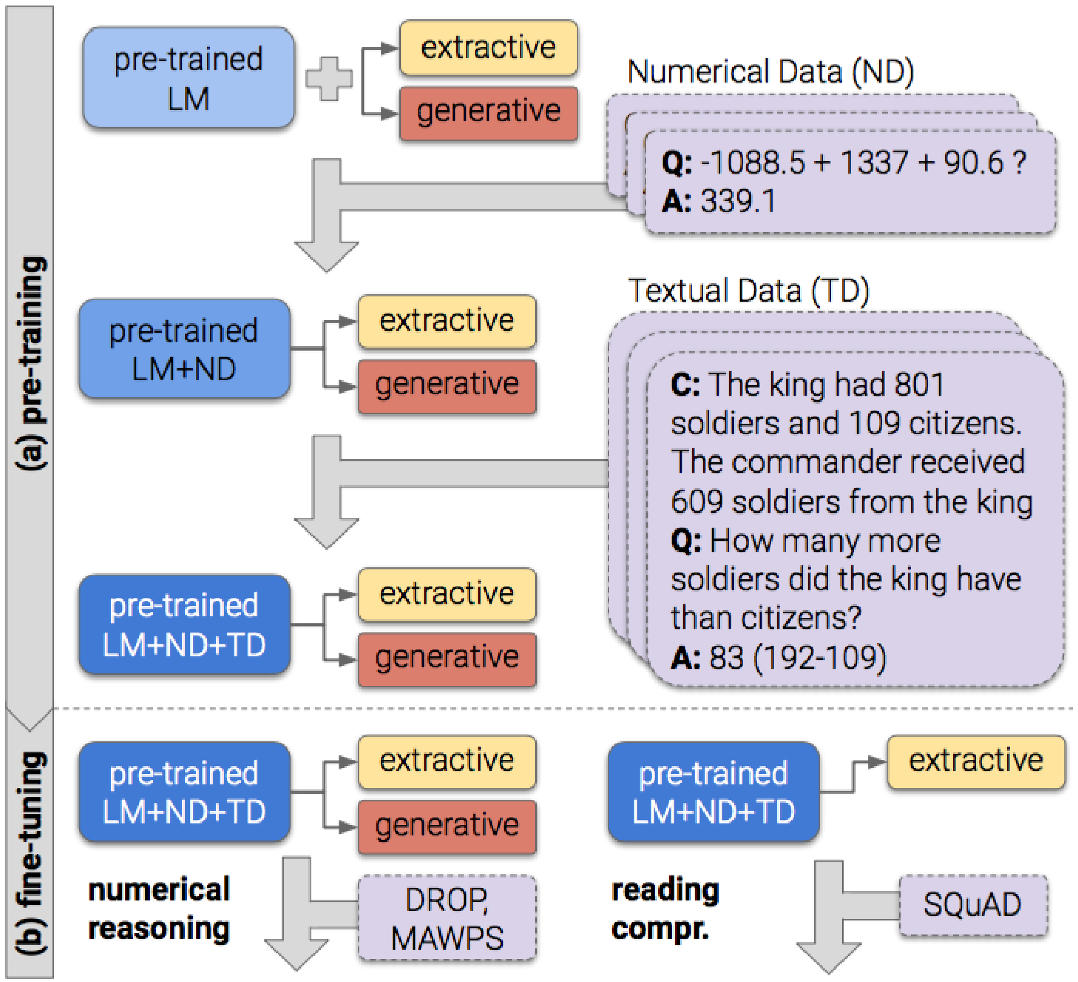

Figure 1: An overview of our approach for injecting numerical skills into a pre-trained LM. (a) We add two pre-training steps over large amounts of synthetic numerical data (ND) and textual data (TD); (b) we further fine-tune the model over either numerical reasoning datasets (DROP, MAWPS) or reading comprehension datasets (SQuAD).

Passage: *Taunton has four art galleries... **Hughes/ Donahue Gallery** founded in 2007, a local community gallery serving local Taunton artists... **Art Euphoric** founded in 2008 has both visual and craft exhibits...*
Q1: How many years **after** founding of **Hughes/ Donahue** was **Art Euphoric** founded? **A1**: 1 (number)
Q2: Which gallery was founded **later**, **Hughes/ Donahue** or **Art Euphoric**? **A2**: Art Euphoric (span)

Table 1: Example passage from DROP, and two questions with different answer types.

Numerical Reasoning Over Text
hybrid approach

- *Context span head*: computes a distribution over all spans in the *context* using a feed-forward network (FFN) $\mathbf{FF_c}(\mathbf{L})$.
- *Question span head*: computes a distribution over spans in the *question* using a FFN $\mathbf{FF_q}(\mathbf{L})$.
- *Count head*: computes a distribution over the numbers $\{0, \ldots, 9\}$ using a FFN $\mathbf{FF_{cnt}}(\mathbf{L})$.
- *Arithmetic head*: computes a distribution over all signed combinations of numbers in the context using a FFN $\mathbf{FF_{cmb}}(\mathbf{L})$ (the numbers in the context are identified in a pre-processing step).

Finally, for deciding which answer head to use for a given input, a *type* head $\mathbf{FF_{typ}}(\mathbf{L})$ outputs a probability distribution $p_{\text{head}}(h \mid \mathbf{q}, \mathbf{c})$ (using a FFN). Thus the model probability for an answer is

$$p(a \mid \mathbf{q}, \mathbf{c}) = \sum_{h \in \text{heads}} p_{\text{head}}(\text{h} \mid \mathbf{c}, \mathbf{q}) \cdot p(a \mid \mathbf{c}, \mathbf{q}, h).$$

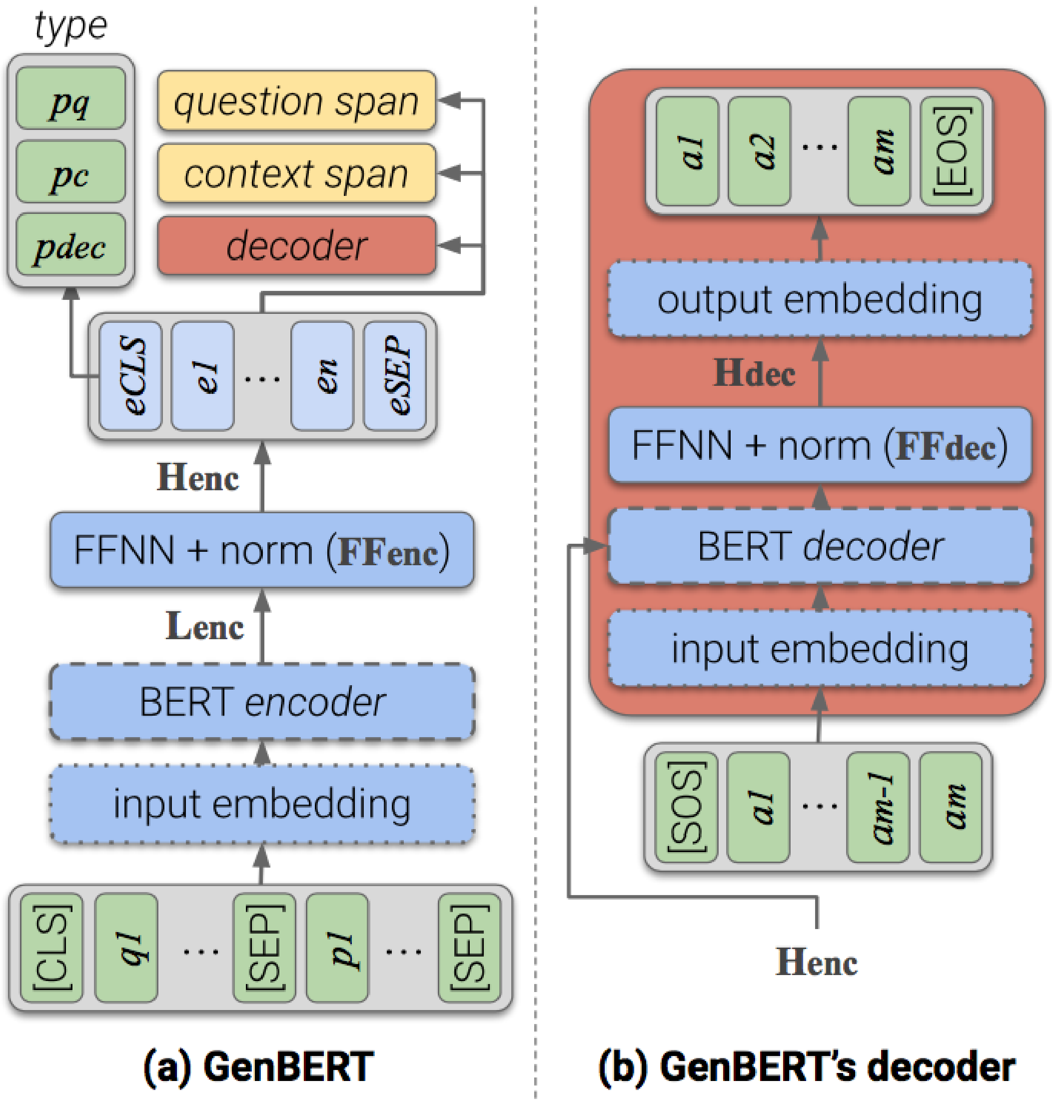

Figure 2: GENBERT's network architecture: (a) a high-level overview of the network, including a generative head (red), two span-extraction heads (yellow), and an answer type head. (b) a closer overview of GENBERT's generative head.

- Digit Tokenization
    - Hence, we tokenize numbers digit-by-digit.

- Random Shift (RS)
    - the model can potentially over-fit and learn to perform numerical reasoning only when numbers are at the beginning of an input
    - when the input length $n1 + n2 + 3 < 512$, we shift all position IDs by a random integer in $(0, 1, \ldots, 512 - (n1 + n2 + 3))$

Generating Numerical Data (ND)

Operation	Template	Example instantiation
signed float combination	$s_1\ f_1\ s_2\ f_2\ s_3\ f_3\ s_4\ f_4$	517.4 - 17484 - 10071.75 + 1013.21
min/max/avg	$o(f_1,\ f_2,\ f_3,\ f_4)$	largest(13.42, 115.5, 72.76)
arg max, arg min	$arg(w_1\ f_1,\ w_2\ f_2,\ w_3\ f_3,\ w_4\ f_4)$	arg min(highish 137.1, sightliness 43.2)
date min/max	$dsup(d_1,\ d_2,\ d_3,\ d_4)$	oldest(June 04, 959; 01 May 959)
date difference	diff in $prd(d_1, d_2)$	diff in days(05 April 112; June 01, 112)
percentage	$pcent\ w\ ::\ w_1\ p_1\%,\ w_2\ p_2\%,\ w_3\ p_3\%,\ w_4\ p_4\%$	percent not sunbird :: sunbird 33.2%, defector 60.77%, molehill 6.03%

Table 2: Templates for generating synthetic numerical examples and the numerical operations required to answer them. **Domains** (defined in App. A.1): $s_i \in \{-,+\}$, $f_i \in \mathbb{R}^+$, $o \in \mathcal{O}$: superlative words like *"longest"*, $arg \in \{\arg\min, \arg\max\}$, $w_i \in \mathcal{W}$: words from NTLK Words Corpus, $d_i \in \mathcal{D}$: dates until Sep 2019, $dsup \in \mathcal{DSUP}$: superlative words like *"latest"*, $prd \in$ {*"days"*, *"months"*, *"years"*}, $p_i \in (0, 100)$, $pcent \in$ {*"percent"*, *"percent not"*}.

Generating Textual Data (TD)

- Passage generation
  - A framework a *world state* consists of *entities*, which are objects that are being counted, and *containers*, which are objects that own entities. Sentences use *verb categories* to describe how the number of entities in a container changes, and thus a world state can be updated given a sentence

- numbers (NUM), entities (ENT), containers (CONT) and attributes (ATTR)

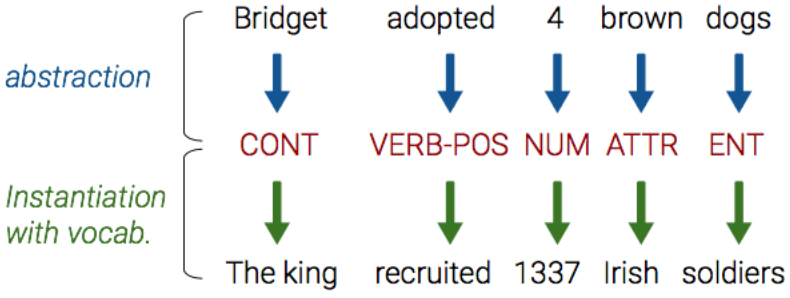

Figure 3: Template extraction and instantiation. A template (in red) is extracted from a MWP sentence, using categories for containers, entities, verbs, attributes and numbers, according to Hosseini et al. (2014). For generation, the categories are instantiated with a domain-specific vocabulary.

- Drop dataset   from wiki crowdsource
  - Train 5565
  - Dev 582 test 588

- MAWPS
  - Math Word Problem Repository
  - Templet
  - 3000

P: The commander recruited 1949 Polish families in Spain. The householder recruited 1996 Japanese families in Spain. There were 10913 white rebels and 77 Chinese families in Spain. 6641 British soldiers, 476 asian rebels, and 338 Germans families were recruited in Russia.
Q: How many Japanese families were in Spain? **A**: 1996 **Q**: How many more Japanese families were in Spain than Polish families? **A**: 47 (1996-1949) **Q**: How many families of Spain were not Polish families? **A**: 2073 (4022-1949)

Table 3: An example synthetic passage (P) and questions. Questions (Q) were generated from templates and answers (A) were calculated based on the world state.

Question generation

- To create questions, we craft 13 question templates that are instantiated with objects from the world state
- Vocab

Reasoning	**Templates**
Selection	How many ATTR−1 ENT−1 were in CONT−1−ENV? How many ATTR−1 ENT−1 did CONT−1−AGT VERB−POS?
Intra-entity difference	How many more ATTR−1 ENT−1 were in CONT−1−ENV than ATTR−2 ENT−2 ? How many more ATTR−1 ENT−1 did CONT−1−AGT have than ATTR−2 ENT−2 ?
Intra-entity subset	How many ENT−1 of CONT−1 were ATTR−1 ENT−1 ? How many ENT−1 of CONT−1 were not ATTR−1 ENT−1 ?
Inter-entity comparison	Were there {more | less} ATTR−1 ENT−1 in CONT−1−ENV or in CONT−2−ENV ? Who had {more | less} ATTR−1 ENT−1, CONT−1−AGT or CONT−2−AGT ?
Inter-entity superlative	Who had the {highest | lowest} number of ATTR−1 ENT−1 in total ?
Intra-entity superlative	What was the {highest | lowest} number of ATTR−1 ENT−1 VERB−POS in CONT−1−ENV ? What is the {highest | lowest} number of ATTR−1 ENT−1 CONT−1−AGT VERB−POS ?
Inter-entity sum	How many ATTR−1 ENT−1 were in CONT−1−ENV (, CONT−⋆−ENV) and CONT−2−ENV {in total | combined} ? How many ATTR−1 ENT−1 did CONT−1−ENV (, CONT−⋆−ENV) and CONT−2−ENV have {in total | combined} ?

Table 9: Templates for questions about generated synthetic passages, testing for numerical reasoning. The template placeholders are filled-in with values from the world state obtained after generating the synthetic passage.

Training

- To ensure that the model does not lose its language understanding abilities, we employ a multi-task setup, and include a standard *masked LM* objective from BERT

- Data from wiki

	Development		Test	
	EM	F_1	EM	F_1
GenBERT	46.1	49.3	-	-
GenBERT$_{\text{+ND-LM-RS}}$	61.5	65.4	-	-
GenBERT$_{\text{+ND-LM}}$	63.8	67.2	-	-
GenBERT$_{\text{+ND}}$	64.7	68.2	-	-
GenBERT$_{\text{+TD}}$	64.4	67.8	-	-
GenBERT$_{\text{+ND+TD}}$	**68.8**	72.3	**68.6**	**72.4**
NABERT+	63.0	66.0	61.6	65.1
MTMSN$_{\text{BASE}}$	68.2	**72.8**	-	-

Table 4: Performance of GenBERT and comparable models on the development and test sets of DROP.

	AddSub	SOp	SEq
GenBERT	2	1.2	1.3
$\text{GenBERT}_{\text{+ND}}$	22.8	26.5	23
$\text{GenBERT}_{\text{+TD}}$	10.4	21.5	12.1
$\text{GenBERT}_{\text{+ND+TD}}$	22.8	**28.3**	22.3
NABERT+	19.2	19.6	17.4
$\text{MTMSN}_{\text{BASE}}$	**32.2**	28	**32.5**

Table 6: EM on MWP datasets.

Sop = singleop
Seq = singleq

	EM	F_1
BERT	**81.1**	**88.6**
GenBERT$_{+ND-LM}$	78.1	85.8
GenBERT$_{+ND}$	80.7	88.1
GenBERT$_{+TD}$	80.7	88.2
GenBERT$_{+ND+TD}$	**81.3**	**88.6**

Table 7: Performance on SQuAD v1 development set. Scores for BERT are using wordpiece tokenization.