Do NLP Models Know Numbers?
Probing Numeracy in Embeddings

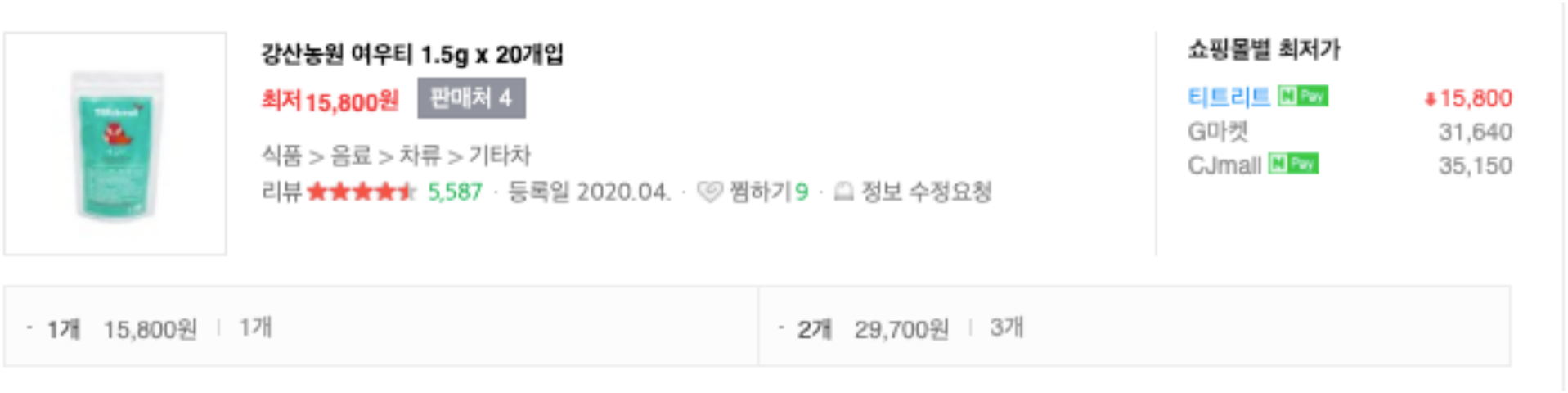
Empirical Methods in Natural Language Processing(EMNLP), 2019

Eric Wallace, Yizhong Wang, Sujian Li
Sameer Singh, Matt Gardner

Soojung Kim
April 20, 2020

Before start

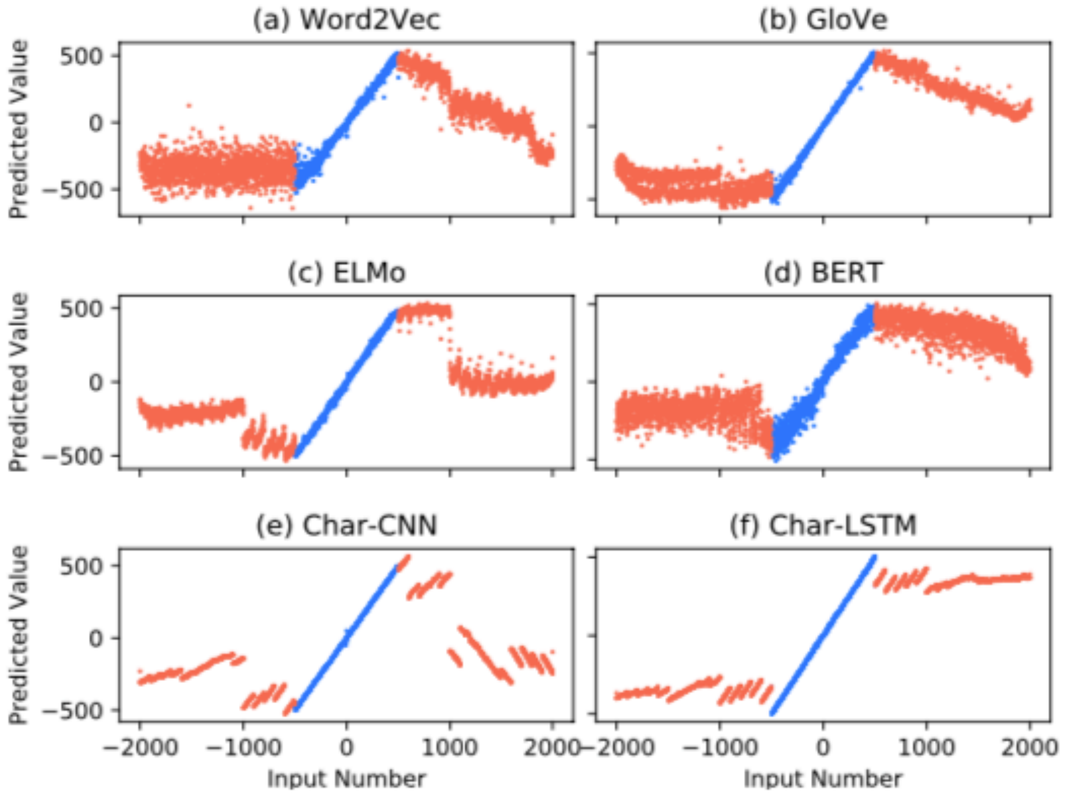
왜 이 논문을 선택하게 됐는지



Abstract

NLP model treat numbers in the same way as other tokens. e.g., distributed vector

Q. Is this a good representation for numerical tasks?
A: Yes! Pre-trained vectors(Glove, ELMo, BERT) know numbers



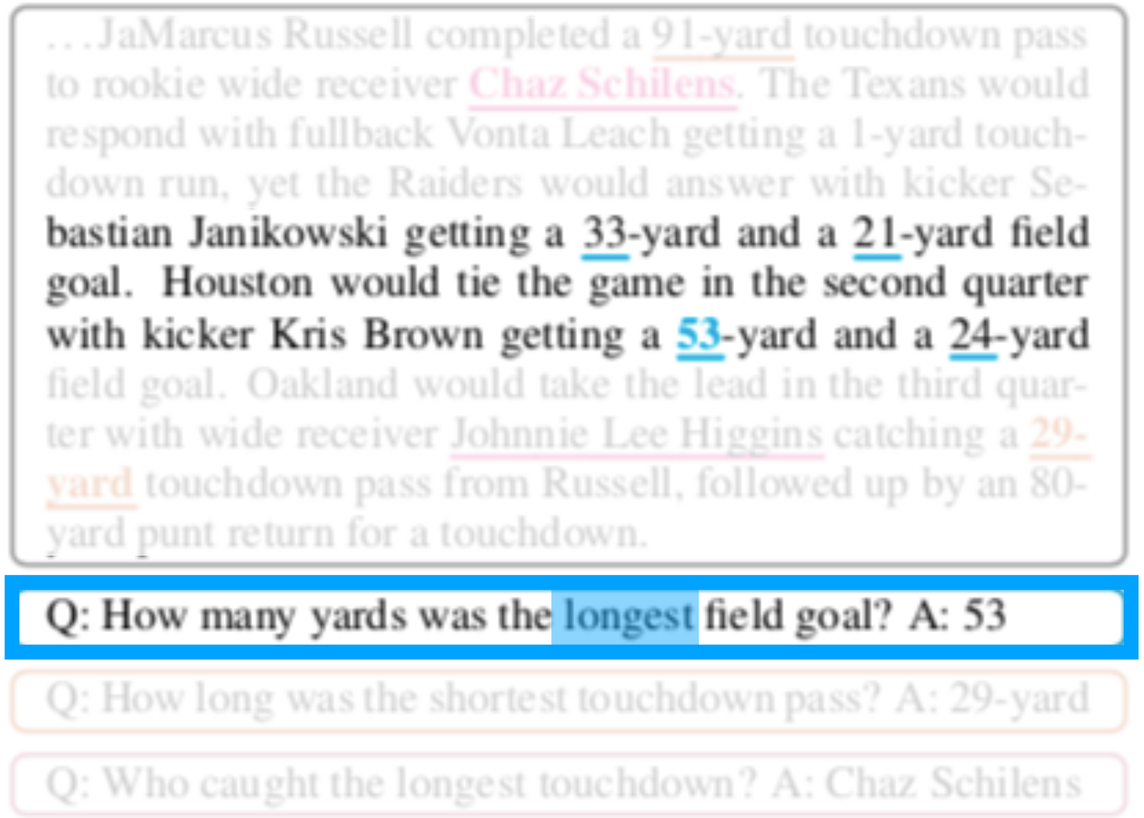
Introduction

Numerical reasoning ability:
the ability to understand and work with numbers in digit or word form

String "23" is bigger than "twenty-two".

Introduction

Numerical reasoning ability:
Extracting the list of field goals and computing that list's maximum



if and how NLP models learn to reason numerically
over paragraphs with only question-answer supervision

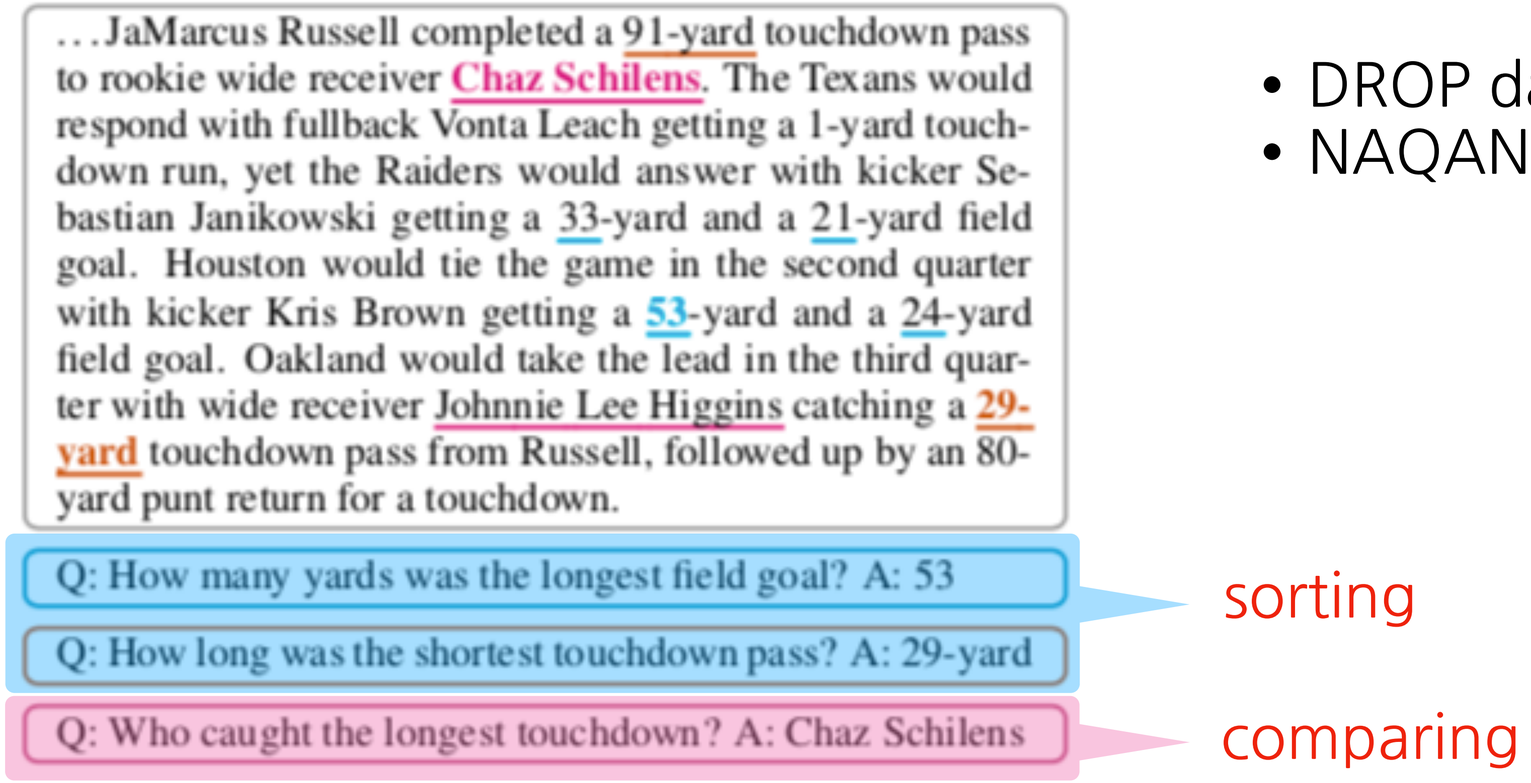
1. Analyzing the sota NAQANet model

2. Probe token embedding methods

https://demo.allennlp.org/reading-comprehension/MTczNDY1Mw==

Analyzing the sota NAQANet model

Testing QA models on questions that evaluate numerical reasoning taken DROP dataset (e.g., sorting, comparing, or summing numbers.)



- DROP dataset
- NAQANet (sota)

DROP dataset: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs

Reading comprehension is the task of answering questions about a passage of text to show that the system understands the passage.

Reasoning	**Passage** (some parts shortened)	**Question**	**Answer**	**BiDAF**
Subtraction (28.8%)	That year, his **Untitled (1981)**, a painting of a haloed, black-headed man with a bright red skeletal body, depicted amid the artists signature scrawls, was **sold by Robert Lehrman for $16.3 million, well above its $12 million high estimate**.	How many more dollars was the Untitled (1981) painting sold for than the 12 million dollar estimation?	4300000	$16.3 million
Comparison (18.2%)	In **1517, the seventeen-year-old King sailed to Castile**. There, his Flemish court .... **In May 1518, Charles traveled to Barcelona in Aragon**.	Where did Charles travel to first, Castile or Barcelona?	Castile	Aragon
Selection (19.4%)	In 1970, to commemorate the 100th anniversary of the founding of Baldwin City, **Baker University professor and playwright Don Mueller and Phyllis E. Braun, Business Manager, produced a musical play entitled The Ballad Of Black Jack** to tell the story of the events that led up to the battle.	Who was the University professor that helped produce The Ballad Of Black Jack, Ivan Boyd or Don Mueller?	Don Mueller	Baker
Addition (11.7%)	Before the UNPROFOR fully deployed, the HV clashed with an armed force of the RSK in the village of Nos Kalik, located in a pink zone near Šibenik, and captured the village at 4:45 p.m. on **2 March 1992**. The JNA formed a battlegroup to counterattack the **next day**.	What date did the JNA form a battlegroup to counterattack after the village of Nos Kalik was captured?	3 March 1992	2 March 1992
Count (16.5%) and Sort (11.7%)	Denver would retake the lead with kicker **Matt Prater nailing a 43-yard field goal**, yet Carolina answered as kicker **John Kasay ties the game with a 39-yard field goal**. ... Carolina closed out the half with **Kasay nailing a 44-yard field goal**. ... In the fourth quarter, Carolina sealed the win with **Kasay's 42-yard field goal**.	Which kicker kicked the most field goals?	John Kasay	Matt Prater
Coreference Resolution (3.7%)	**James Douglas** was the second son of Sir George Douglas of Pittendreich, and Elizabeth Douglas, daughter David Douglas of Pittendreich. Before **1543 he married Elizabeth**, daughter of James Douglas, 3rd Earl of Morton. **In 1553 James Douglas succeeded to the title and estates of his father-in-law**.	How many years after he married Elizabeth did James Douglas succeed to the title and estates of his father-in-law?	10	1553
Other Arithmetic (3.2%)	Although the movement initially gathered some **60,000 adherents**, the subsequent establishment of the Bulgarian Exarchate **reduced their number by some 75%**.	How many adherents were left after the establishment of the Bulgarian Exarchate?	15000	60,000
Set of spans (6.0%)	According to some sources 363 civilians were killed in **Kavadarci**, 230 in **Negotino** and 40 in **Vatasha**.	What were the 3 villages that people were killed in?	Kavadarci, Negotino, Vatasha	Negotino and 40 in Vatasha
Other (6.8%)	This **Annual Financial Report** is our principal financial statement of accountability. The **AFR gives a comprehensive view** of the Department's financial activities ...	What does AFR stand for?	Annual Financial Report	one of the Big Four audit firms

https://arxiv.org/pdf/1903.00161.pdf

DROP dataset: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs

focus on Comparative and Superlative questions

Comparative questions

- quantities or events "larger", "smaller", "longer", ...
- "either-or" relations

Question Type	Example	Reasoning Required
Comparative (Binary)	Which country is a bigger exporter, Brazil or Uruguay?	Binary Comparison
Comparative (Non-binary)	Which player had a touchdown longer than 20 yards?	Greater Than

Superlative questions

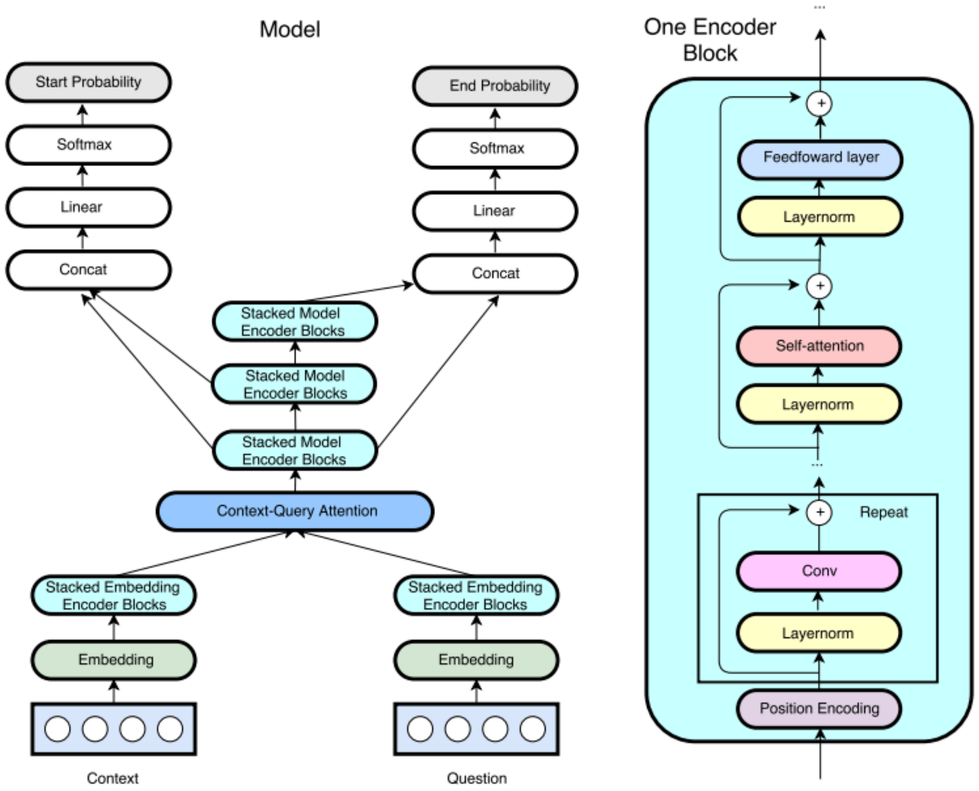
- "shortest", "longest", "biggest"
- finding the maximum or minimum of a list
- argmax operation

Superlative (Number)	How many yards was the shortest field goal?	List Minimum
Superlative (Span)	Who kicked the longest field goal?	Argmax

https://arxiv.org/pdf/1903.00161.pdf

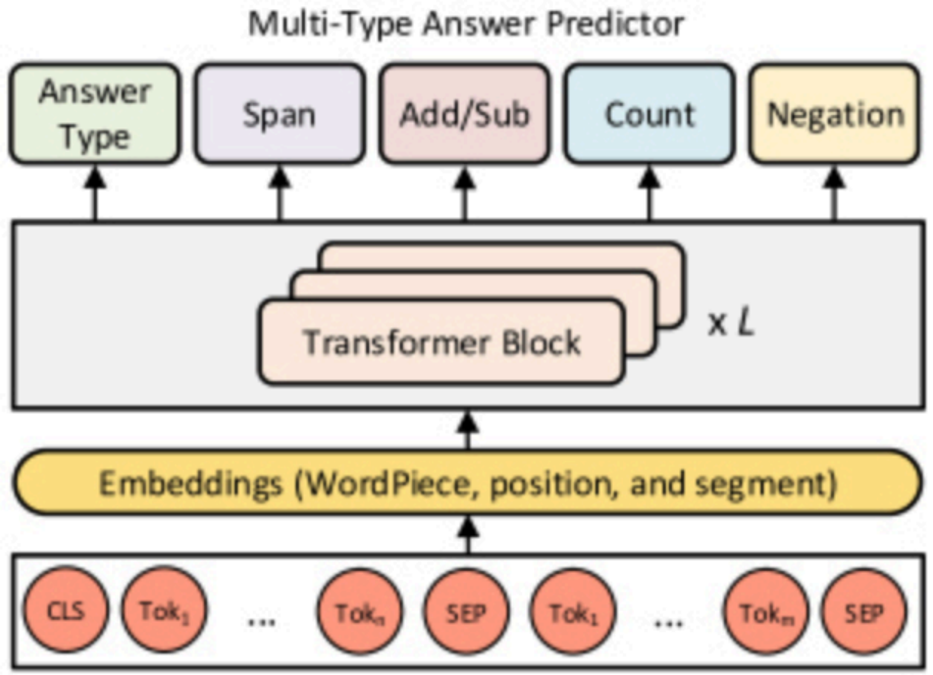
NAQANet model

QANet + four output branches(passage span, question span, count, add/sub)



QANet

https://arxiv.org/pdf/1804.09541.pdf



Multi type answer predictor

https://deepai.org/publication/a-multi-type-multi-span-network-for-reading-comprehension-that-requires-discrete-reasoning

Analyzing the sota NAQANet model

NAQANet achieving only 49 F1 on the entire validation set
It scores 89 F1 on numerical comparsion questions.

Question Type	Count	EM	F1
Human (Test Set)	9622	92.4	96.0
Full Validation	9536	46.2	49.2
Number Answers	5842	44.3	44.4
Comparative	704	73.6	76.4
Binary (either-or)	477	86.0	89.0
Non-binary	227	47.6	49.8
Superlative Questions	861	64.6	67.7
Number Answers	475	68.8	69.2
Span Answers	380	59.7	66.3

Stress Testing NAQANet's Numeracy

NAQANet has a strong understanding of numeracy for numbers in the training range,
but the model can fail to extrapolate to other values

training range [0, 100]

Stress Test Dataset	All Questions F1	All Questions Δ	Superlative F1	Superlative Δ
Original Validation Set	49.2	-	67.7	-
Add [1, 20]	47.7	-1.5	64.1	-3.6
Add [21, 100]	41.4	-7.8	40.4	-27.3
Multiply [2, 10]	41.1	-8.1	39.3	-28.4
Multiply [11, 100]	38.8	-10.4	32.0	-35.7
Digits to Words [0, 20]	45.5	-3.7	63.8	-3.9
Digits to Words [21, 100]	41.9	-7.3	46.1	-21.6

* Digits to Words: "75" -> "seventy-five"

1. Analyzing the sota NAQANet model

2. Probe token embedding methods

https://demo.allennlp.org/reading-comprehension/MTczNDY1Mw==

Probe token embedding methods(e.g., BERT, GloVe)

To understand how models can capture numerical reasoning,
we use three synthetic tasks to probe the numeracy of token embeddings.

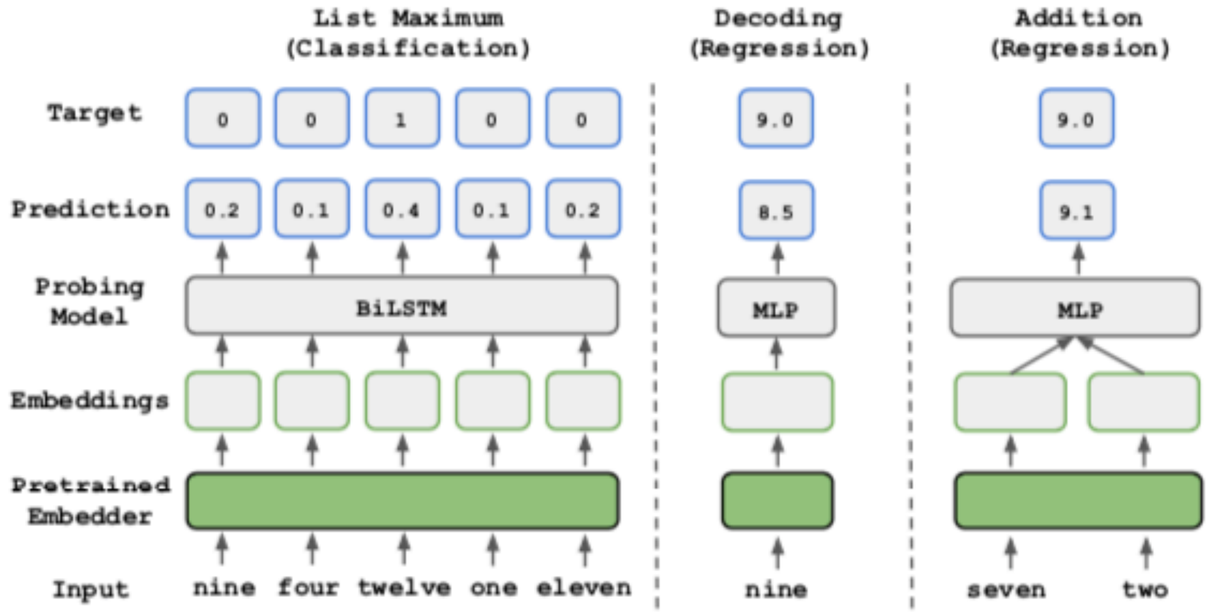


Figure 3: Our probing setup. We pass numbers through a pre-trained embedder (e.g., BERT, GloVe) and train a probing model to solve numerical tasks such as finding a list's maximum, decoding a number, or adding two numbers. If the probing model generalizes to held-out numbers, the pre-trained embeddings must contain numerical information. We provide numbers as either words (shown here), digits ("9"), floats ("9.1"), or negatives ("-9").

Results: Embeddings capture numeracy

word2vec and GloVe:
continuous bag of words objective can teach fine-grained number magnitude.

Interpolation	**List Maximum (5-classes)**			**Decoding (RMSE)**			**Addition (RMSE)**		
Integer Range	[0,99]	[0,999]	[0,9999]	[0,99]	[0,999]	[0,9999]	[0,99]	[0,999]	[0,9999]
Random Vectors	0.16	0.23	0.21	29.86	292.88	2882.62	42.03	410.33	4389.39
Untrained CNN	0.97	0.87	0.84	2.64	9.67	44.40	1.41	14.43	69.14
Untrained LSTM	0.70	0.66	0.55	7.61	46.5	210.34	5.11	45.69	510.19
Value Embedding	**0.99**	0.88	0.68	**1.20**	11.23	275.50	**0.30**	15.98	654.33
Pre-trained									
Word2Vec	0.90	0.78	0.71	2.34	18.77	333.47	0.75	21.23	210.07
GloVe	0.90	0.78	0.72	2.23	13.77	174.21	0.80	16.51	180.31
ELMo	0.98	0.88	0.76	2.35	13.48	62.20	0.94	15.50	45.71
BERT	0.95	0.62	0.52	3.21	29.00	431.78	4.56	67.81	454.78
Learned									
Char-CNN	0.97	**0.93**	**0.88**	2.50	**4.92**	**11.57**	1.19	**7.75**	**15.09**
Char-LSTM	0.98	0.92	0.76	2.55	8.65	18.33	1.21	15.11	25.37
DROP-trained									
NAQANet	0.91	0.81	0.72	2.99	14.19	62.17	1.11	11.33	90.01
- GloVe	0.88	0.90	0.82	2.87	5.34	35.39	1.45	9.91	60.70

Results: Embeddings capture numeracy

Character-level methods

strength of the char-level convolutions seems to lie in the architectural prior

Interpolation	**List Maximum (5-classes)**			**Decoding (RMSE)**			**Addition (RMSE)**		
Integer Range	[0,99]	[0,999]	[0,9999]	[0,99]	[0,999]	[0,9999]	[0,99]	[0,999]	[0,9999]
Random Vectors	0.16	0.23	0.21	29.86	292.88	2882.62	42.03	410.33	4389.39
Untrained CNN	0.97	0.87	0.84	2.64	9.67	44.40	1.41	14.43	69.14
Untrained LSTM	0.70	0.66	0.55	7.61	46.5	210.34	5.11	45.69	510.19
Value Embedding	**0.99**	0.88	0.68	**1.20**	11.23	275.50	**0.30**	15.98	654.33
Pre-trained									
Word2Vec	0.90	0.78	0.71	2.34	18.77	333.47	0.75	21.23	210.07
GloVe	0.90	0.78	0.72	2.23	13.77	174.21	0.80	16.51	180.31
ELMo	0.98	0.88	0.76	2.35	13.48	62.20	0.94	15.50	45.71
BERT	0.95	0.62	0.52	3.21	29.00	431.78	4.56	67.81	454.78
Learned									
Char-CNN	0.97	**0.93**	**0.88**	2.50	**4.92**	**11.57**	1.19	**7.75**	**15.09**
Char-LSTM	0.98	0.92	0.76	2.55	8.65	18.33	1.21	15.11	25.37
DROP-trained									
NAQANet	0.91	0.81	0.72	2.99	14.19	62.17	1.11	11.33	90.01
- GloVe	0.88	0.90	0.82	2.87	5.34	35.39	1.45	9.91	60.70

Results: Embeddings capture numeracy

Sub-word models:

poor method to encode digits: similar two numbers can have very different sub-word divisions

Interpolation	**List Maximum (5-classes)**			**Decoding (RMSE)**			**Addition (RMSE)**		
Integer Range	[0,99]	[0,999]	[0,9999]	[0,99]	[0,999]	[0,9999]	[0,99]	[0,999]	[0,9999]
Random Vectors	0.16	0.23	0.21	29.86	292.88	2882.62	42.03	410.33	4389.39
Untrained CNN	0.97	0.87	0.84	2.64	9.67	44.40	1.41	14.43	69.14
Untrained LSTM	0.70	0.66	0.55	7.61	46.5	210.34	5.11	45.69	510.19
Value Embedding	**0.99**	0.88	0.68	**1.20**	11.23	275.50	**0.30**	15.98	654.33
Pre-trained									
Word2Vec	0.90	0.78	0.71	2.34	18.77	333.47	0.75	21.23	210.07
GloVe	0.90	0.78	0.72	2.23	13.77	174.21	0.80	16.51	180.31
ELMo	0.98	0.88	0.76	2.35	13.48	62.20	0.94	15.50	45.71
BERT	0.95	0.62	0.52	3.21	29.00	431.78	4.56	67.81	454.78
Learned									
Char-CNN	0.97	**0.93**	**0.88**	2.50	**4.92**	**11.57**	1.19	**7.75**	**15.09**
Char-LSTM	0.98	0.92	0.76	2.55	8.65	18.33	1.21	15.11	25.37
DROP-trained									
NAQANet	0.91	0.81	0.72	2.99	14.19	62.17	1.11	11.33	90.01
- GloVe	0.88	0.90	0.82	2.87	5.34	35.39	1.45	9.91	60.70

Results: Embeddings capture numeracy

A linear subspace exists:
for small ranges on the decoding task, this capture number magnitude.

Interpolation	**Decoding (RMSE)**		
Integer Range	[0,50]	[-50,50]	[0,999]
Random	13.86	29.46	275.41
Word2Vec	4.15	8.93	29.04
GloVe	3.21	5.76	23.27
ELMo	1.20	2.89	21.53
BERT	3.23	7.86	64.42

Table 8: *Number Decoding interpolation accuracy with linear regression.* Linear regression is competitive to the fully connected probe for smaller numbers.

Results: Embeddings capture numeracy

Extrapolation on list maximum.
trained on the integer range [0, 150] and evaluate on integers from the Test Range

Extrapolation	**List Maximum (5-classes)**		
Test Range	[151,160]	[151,180]	[151,200]
Rand. Vectors	0.17	0.22	0.15
Untrained CNN	0.80	0.47	0.41
Pre-trained			
Word2Vec	0.14	0.16	0.11
GloVe	0.19	0.17	0.21
ELMo	0.65	0.57	0.38
BERT	0.35	0.11	0.14
Learned			
Char-CNN	0.81	0.75	0.73
Char-LSTM	**0.88**	**0.84**	**0.82**
DROP			
NAQANet	0.31	0.29	0.25
- GloVe	0.58	0.53	0.48

Results: Embeddings capture numeracy

Interpolation	**List Maximum (5-classes)**	
Float Range	[0.0,99.9]	[0.0,999.9]
Rand. Vectors	0.18 ± 0.03	0.21 ± 0.04
ELMo	0.91 ± 0.03	0.59 ± 0.01
BERT	0.82 ± 0.05	0.51 ± 0.04
Char-CNN	0.87 ± 0.04	0.75 ± 0.03
Char-LSTM	0.81 ± 0.05	0.69 ± 0.02

Table 5: *Interpolation with floats (e.g., "18.1") for list maximum.* Pre-trained embeddings capture numeracy for float values. The probing model is trained on a randomly shuffled 80% of the *Float Range* and tested on the remaining 20%. See the text for details on selecting decimal values. We show the mean alongside the standard deviation over 5 different random shuffles.

Interpolation	**List Maximum (5-classes)**
Integer Range	[-50,50]
Rand. Vectors	0.23 ± 0.12
Word2Vec	0.89 ± 0.02
GloVe	0.89 ± 0.03
ELMo	0.96 ± 0.01
BERT	0.94 ± 0.02
Char-CNN	0.95 ± 0.07
Char-LSTM	0.97 ± 0.02

Table 6: *Interpolation with negatives (e.g., "-18") on list maximum.* Pre-trained embeddings capture numeracy for negative values.

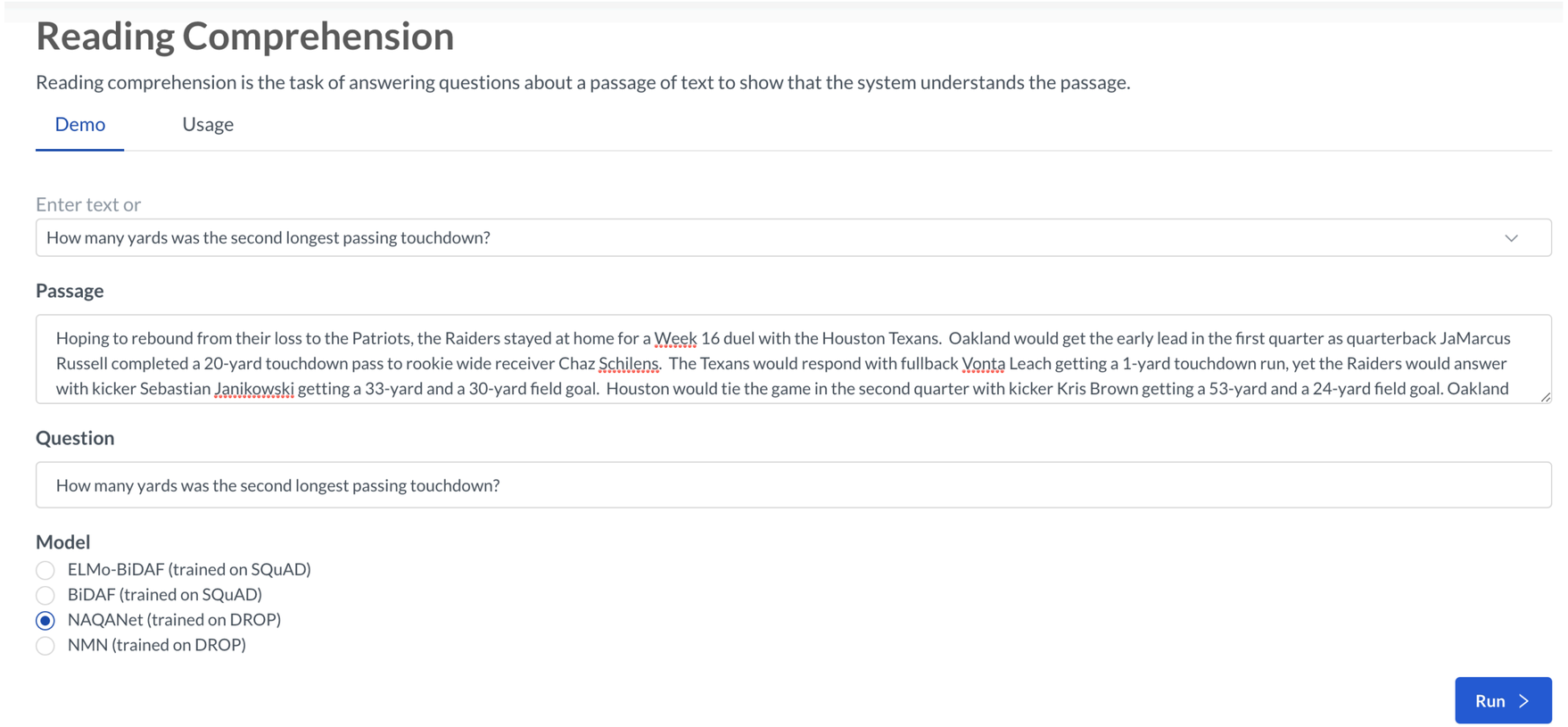
Augmenting Data to aid Extrapolation

	Superlative		Comparative		All Validation	
	Original	Bigger	Original	Bigger	Original	Bigger
NAQANet	64.5 / 67.7	30.0 / 32.2	73.6 / 76.4	70.3 / 73.0	**46.2** / 49.2	38.7 / 41.4
+ Data Augmentation	**67.6 / 70.9**	**59.2 / 62.4**	**76.0 / 77.7**	**75.0 / 76.8**	46.1 / **49.3**	**42.8 / 45.8**

Table 11: Data augmentation improves NAQANet's interpolation and extrapolation results. We created the *Bigger* version of DROP by multiplying numbers in the passage by a random integer from [11, 20] and then adding a random integer from [21, 40]. Scores are shown in EM / F1 format.

Conclusion

pre-trained token representations naturally encode numeracy.
But, model fails to extrapolate to numbers outside the training range



https://demo.allennlp.org/reading-comprehension/MTczNDk0NQ==

 Thank you