Expanding Numerical Reasoning Capabilities with ModernBERT and Flan-T5

Aman Kumar and
Deric Liang

01

Introduction

Overview and contributions

SemEval 2024 Task 7 - NumEval; Numerical Reasoning with Headlines

Overview

- Using numeral-heavy news articles to develop understanding of numeric semantics
- Predict the missing number from the corresponding news headline
- "Stealing $10" versus "Stealing $100,000"

Objectives

- Evaluate the performance of low-resource approaches with fine-tuning relative to the existing literature
- Expand beyond the accuracy evaluation of the Numerical Reasoning task and demonstrate how additional metrics can provide additional insight

Contributions

- Showcase how models released after SemEval 2024 perform
- Demonstrate performance of low-resource approaches compared to previous literature
- Demonstrate power of relatively little fine-tuning on model performance

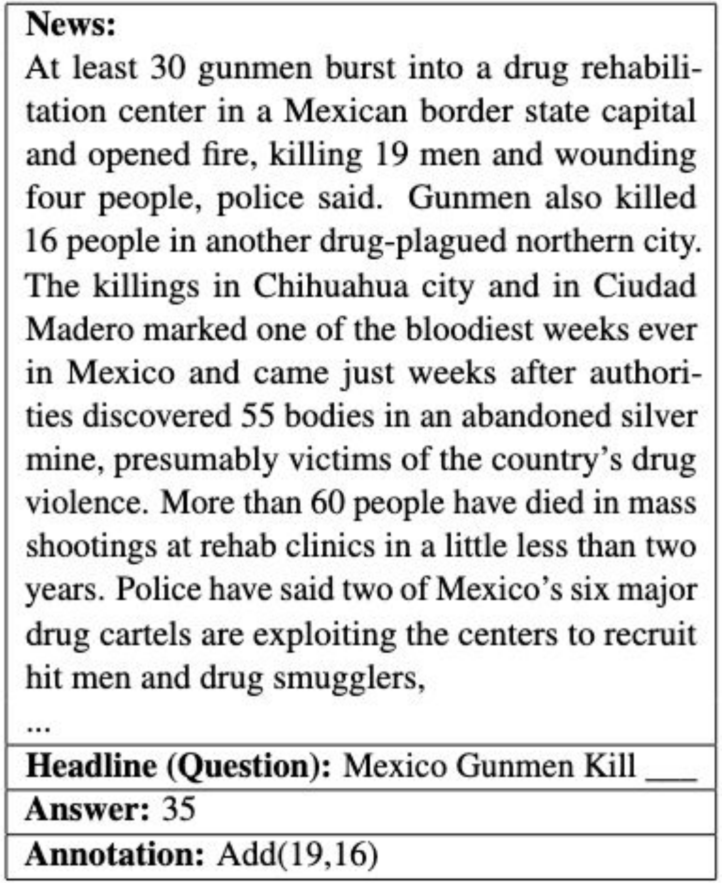
Models

- ModernBERT
- Flan-T5

02

Methods

Data



News:

At least 30 gunmen burst into a drug rehabilitation center in a Mexican border state capital and opened fire, killing 19 men and wounding four people, police said. Gunmen also killed 16 people in another drug-plagued northern city. The killings in Chihuahua city and in Ciudad Madero marked one of the bloodiest weeks ever in Mexico and came just weeks after authorities discovered 55 bodies in an abandoned silver mine, presumably victims of the country's drug violence. More than 60 people have died in mass shootings at rehab clinics in a little less than two years. Police have said two of Mexico's six major drug cartels are exploiting the centers to recruit hit men and drug smugglers,

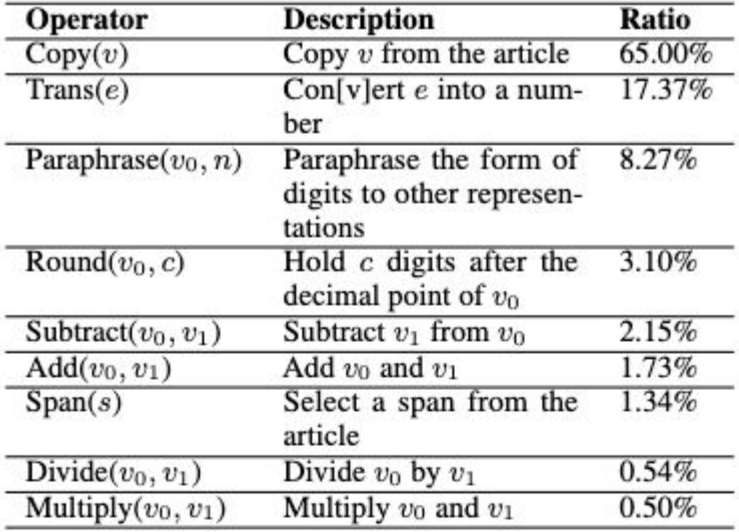
...

Headline (Question): Mexico Gunmen Kill ___

Answer: 35

Annotation: Add(19,16)

Table 2: An example from NumHG.



Operator	Description	Ratio
Copy(v)	Copy v from the article	65.00%
Trans(e)	Con[v]ert e into a number	17.37%
Paraphrase(v_0, n)	Paraphrase the form of digits to other representations	8.27%
Round(v_0, c)	Hold c digits after the decimal point of v_0	3.10%
Subtract(v_0, v_1)	Subtract v_1 from v_0	2.15%
Add(v_0, v_1)	Add v_0 and v_1	1.73%
Span(s)	Select a span from the article	1.34%
Divide(v_0, v_1)	Divide v_0 by v_1	0.54%
Multiply(v_0, v_1)	Multiply v_0 and v_1	0.50%

Table 3: Overview of predefined operators. v, v_0, and v_1 denote the selected numerals, and e denotes the English word. s and c denote a span from the article and a constant, respectively.

Modeling

- Train on multiple dimensions:
  - Mixed Training Set
  - Reasoning Training Set
- Base ModernBERT (150M) and Flan-T5 (248M) models
- Versions:
  - Baseline
  - Mixed Reasoning
  - Reasoning-only
- Mix of PEFT (LoRA) and non-PEFT

Metrics

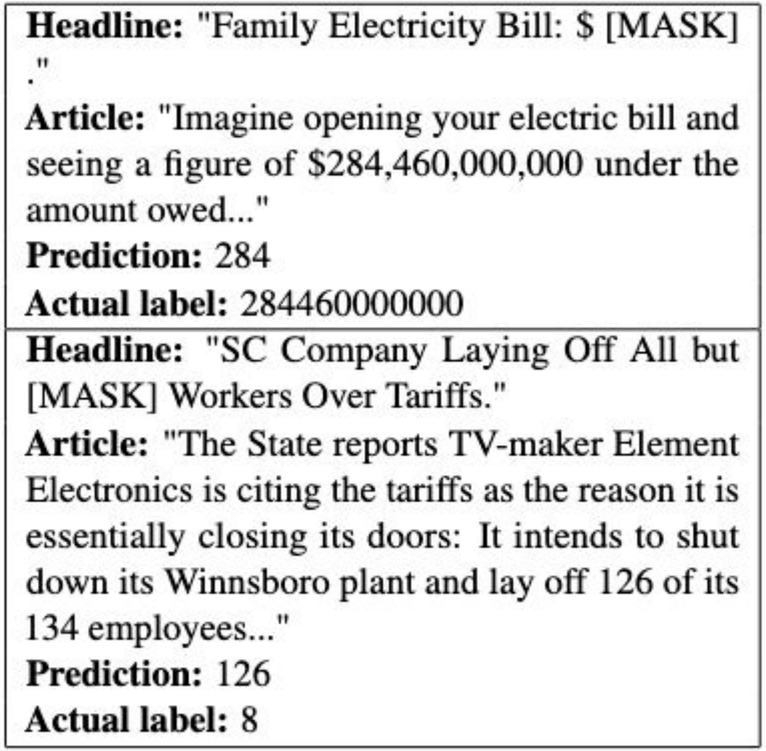
- Accuracy
- Mean Absolute Percentage Error (MAPE)
- Symmetrical Mean Absolute Percentage Error (SMAPE)

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i - \hat{Y}_i}{Y_i}\right|$$

$$SMAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{(\hat{y}_i + y_i)/2}\right|$$

Note on metrics

- Example 1:
  - MAPE = 100%
  - SMAPE = 200%
- Example 2:
  - MAPE = 1475%
  - SMAPE = 176%



Headline: "Family Electricity Bill: $ [MASK] ."

Article: "Imagine opening your electric bill and seeing a figure of $284,460,000,000 under the amount owed..."

Prediction: 284

Actual label: 284460000000

Headline: "SC Company Laying Off All but [MASK] Workers Over Tariffs."

Article: "The State reports TV-maker Element Electronics is citing the tariffs as the reason it is essentially closing its doors: It intends to shut down its Winnsboro plant and lay off 126 of its 134 employees..."

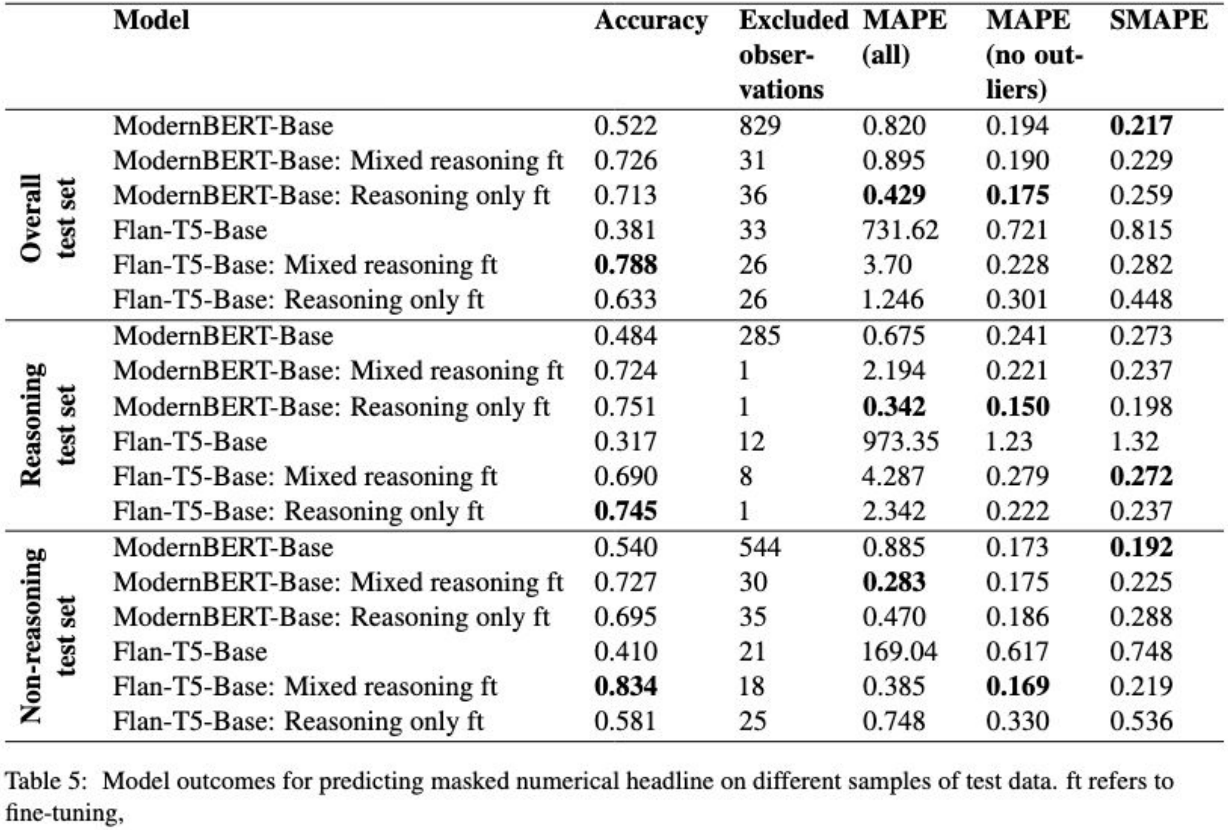
Prediction: 126

Actual label: 8

Table 4: Examples of extreme under-prediction and over-prediction from ModernBERT. Example 1 also demonstrates a mis-predicted label because the actual label is not in the vocabulary.

03

Results



	Model	Accuracy	Excluded observations	MAPE (all)	MAPE (no outliers)	SMAPE
Overall test set	ModernBERT-Base	0.522	829	0.820	0.194	**0.217**
	ModernBERT-Base: Mixed reasoning ft	0.726	31	0.895	0.190	0.229
	ModernBERT-Base: Reasoning only ft	0.713	36	**0.429**	**0.175**	0.259
	Flan-T5-Base	0.381	33	731.62	0.721	0.815
	Flan-T5-Base: Mixed reasoning ft	**0.788**	26	3.70	0.228	0.282
	Flan-T5-Base: Reasoning only ft	0.633	26	1.246	0.301	0.448
Reasoning test set	ModernBERT-Base	0.484	285	0.675	0.241	0.273
	ModernBERT-Base: Mixed reasoning ft	0.724	1	2.194	0.221	0.237
	ModernBERT-Base: Reasoning only ft	0.751	1	**0.342**	**0.150**	0.198
	Flan-T5-Base	0.317	12	973.35	1.23	1.32
	Flan-T5-Base: Mixed reasoning ft	0.690	8	4.287	0.279	**0.272**
	Flan-T5-Base: Reasoning only ft	**0.745**	1	2.342	0.222	0.237
Non-reasoning test set	ModernBERT-Base	0.540	544	0.885	0.173	**0.192**
	ModernBERT-Base: Mixed reasoning ft	0.727	30	**0.283**	0.175	0.225
	ModernBERT-Base: Reasoning only ft	0.695	35	0.470	0.186	0.288
	Flan-T5-Base	0.410	21	169.04	0.617	0.748
	Flan-T5-Base: Mixed reasoning ft	**0.834**	18	0.385	**0.169**	0.219
	Flan-T5-Base: Reasoning only ft	0.581	25	0.748	0.330	0.536

Table 5: Model outcomes for predicting masked numerical headline on different samples of test data. ft refers to fine-tuning,

Limitations & Future Work

- Small models (yet outperforms zero-shot approaches with large models)
- Out-of-vocabulary labels
  - ModernBERT - only one token can be predicted
- Future work
  - Multiple tokenization representations of labels
  - Could incorporate data augmentation or secondary data approaches

Questions?

THANK YOU