Have You Seen That Number? Investigating Extrapolation in Question Answering Models

Jeonghwan Kim, Giwon Hong, Kyung-min Kim, Junmo Kang , Sung-Hyon Myaeng

IR&NLP Lab

Overview

Machine reading comprehension (MRC) models in DROP unable to do **numerical extrapolation** in textual reasoning.

Proposing a novel **E-digit** surface form to alleviate **extrapolation** issue in MRC models.

Motivation

"Kasay hitting a 45-yard field goal … with Kasay again hitting a 49-yard field goal…" → **Model** → "Kasay hitting a 4500-yard field goal … with Kasay again hitting a 4900-yard field goal…"

Training Data | Test Data

Change in the number range causes significant performance drop.

Related Work

- Extrapolation addressed in Arithmetic Word Problems (AWP) setting *(Trask et al., 2018 ; Kim et al., 2021)*.

  "What is 24 + 5?"

- Digit-position information to improve arithmetic reasoning capability of Transformer models *(Nogueira et al., 2021)*.

Probing Models for Extrapolation Capability

1. **Stanza NER** to extract:
   QUANTITY , CARDINAL , MONEY type numbers
2. **Data Perturbation**
   ADD(10), ADD(100), FACTOR(10), FACTOR(100)

"How many people, households, and families reside in the county according to the 2000 census?"

As of the census of **2000**, there were **49,927** people, **18,009** households, and **12,192** families residing in the county. The population density was **48** people per square mile (**19**/km²).

Answer: **80,128**

FACTOR(100)

As of the census of **2000**, there were **4,992,700** people, **1,800,900** households, and **1,219,200** families residing in the county. The population density was **4,800** people per square mile (**19**/km²).

Answer: **8,012,800**

Perturbated DROP data evaluated on models on the DROP leaderboard led to performance drop:

NAQANet → -23.17 in Exact Match (EM)
NumNet → -37.37 in EM
NumNet+ (RoBERTa) → -26.03 in EM
GenBERT → -26.02 in EM

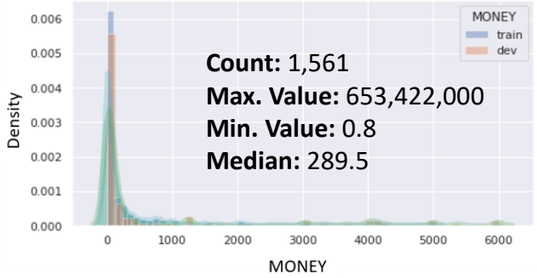
(for FACTOR(100) perturbation)

Reasons behind performance degradation

1. **Sub-word Representation**
   Different tokenization for similar numbers

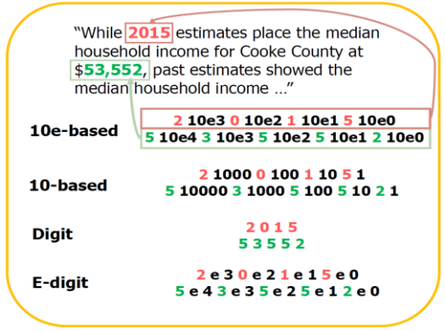
   21260 → '212', '##60'
   21262 → '212', '##6', '##2'

2. **Limited Number Distribution**
   Number occurrence is cluttered and sparse



Leads to lack of **Inductive Bias** for numbers

Method



Digit-position information is important *(Nogueira et al., 2021).*

E-digit decouples the digit-position from position-dependent embeddings

Result

Model – GenBERT (Geva et al., 2020)
Dataset – Perturbated DROP – FACTOR(100)

Model	DROP – FACTOR(100)	
	EM	F1
Original	42.78	45.10
10e-based	43.02	49.94
10based	49.24	56.47
Digit	44.97	51.76
E-digit	**57.91**	**63.98**

Takeaways

Proposing an evaluation benchmark for more challenging, but necessary number reasoning.

A simple yet readily applicable **E-digit** method.

Further investigation on unidentified issues causing the extrapolation issue