Examining the Ability of LMs to Reason and Interact with Numbers

Karan Raj Bagri & Dhamale Vivek Shrikrishna & Pawar Rushikesh Gajanansa
Indian Institute of Science
Bengaluru, KA, India
{karanraj,viveksd,rushikeshp}@iisc.ac.in

Abstract

Language Models have achieved impressive feats in Natural Language Processing, excelling at tasks like text generation, machine translation, Question Answers, text summarization. However NLP systems rarely give special considerations to numbers. They are treated just like any other text tokens, but there is fundamental difference between words/letters and numbers. Also, during preprocessing most of numbers get mapped to <UNK> token because of not being in vocabulary. Which results in poor number play despite great wordplay that these models can perform. In this project we intend to come up with set of properties that LMs should know about numbers, and we will try to build tests to check how much of numeracy is captured by current models.

1 Introduction

NLP systems neglects numbers. They are treated just like any other word/character tokens. Which affects the ability of model to handle and interact with numbers on numerical tasks such as comparing, sorting, identifying patterns in sequences. Numbers are semantically different from texts, and their operations are also different. So, it is necessary to identify/define properties of numbers that should be learnt by LMs in order to truly understand and perform numerical tasks, along with tests that can check whether models are learning these properties.

First we will do extensive literature survey on this topic. To motivate this study We will test numerical ability of current LMs like LLama 2 and Bloom and try to find out peculiar numerical tasks (e.g; sorting, finding max-min) in which model fails. Then we will find the properties required for such problem solving which model perhaps fails to learn.

2 Related Work:

Wallace et al. (2019)Some previously conducted studies have investigated the numerical reasoning capabilities of the then state-of-the-art models on datasets like the DROP dataset. One such model was the NAQANet. They found that although the performance was mediocre (49 F1) on the entire validation dataset, the model scored well (89 F1) on numerical comparison questions. It was also found that the model fails to extrapolate well to numbers outside its training range. Upon making some changes in the validation paragraphs, like generating a random number and multiplying the numbers in each paragraph, the performance drop was significant (35.7 F1). Also, upon converting small numbers to word-form, the performance dip was marginal (3.9 F1) when compared to the drop experienced upon converting larger numbers to word-form (21.6).

The source of numeracy are the token embeddings themselves. The reason behind the success or failure of a model on numerical tasks lies in the information encoded within these embeddings. Past approaches have tried to probe various token embedding methods, to see how they compare with each other on tasks such as finding the maximum element in a list, finding the value of a number given its embedding as well as addition of two embeddings and

then decoding the resultant sum. It was found that the finer, character-level models had a clear advantage over the sub-word models like BERT. The extrapolation problem persisted for tasks like decoding and addition, while it had minimal effect on the list-maximum finding problem. We plan to evaluate whether these problems remain as significant for the current state-of-the-art models

Thawani et al. (2021)In this survey, authors have arranged recent NLP work on numeracy into a comprehensive taxonomy of tasks and methods and broken down the subjective notion of numeracy into 7 subtasks, arranged along two dimensions: granularity (exact vs approximate) and units (abstract vs grounded) and analyzed the myriad representational choices made by 18 previously published number encoders and decoders. We basically want to build up upon this work and come up with a holistic evaluation for checking number representations.

Geva et al. (2020)This work proposes a general method to "inject" specific skills into LLMs using generated(synthetic) data. By applying this method to numerical reasoning with text, they achieve results comparable to state-of-the-art models, demonstrating the effectiveness of our approach for enhancing LM capabilities.

Contributions

Literature review:

Thawani et al. (2021) : Rushikesh Pawar

Wallace et al. (2019): Karan Raj Bagri

Geva et al. (2020) : Vivek Dhamale

References

Mor Geva, Ankit Gupta, and Jonathan Berant. Injecting numerical reasoning skills into language models, 2020.

Avijit Thawani, Jay Pujara, Pedro A. Szekely, and Filip Ilievski. Representing numbers in nlp: a survey and a vision, 2021.

Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. Do nlp models know numbers? probing numeracy in embeddings, 2019.