Fine-tuning Language models to solve Natural Language Understanding tasks

- Dhruv Dhamani (ddhamani@uncc.edu)
- Saloni Gupta (sgupta38@uncc.edu)
- Himanshu Sunil Dhawale (hdhawale@uncc.edu)
- Bhavya Chawla (bchawla@uncc.edu)

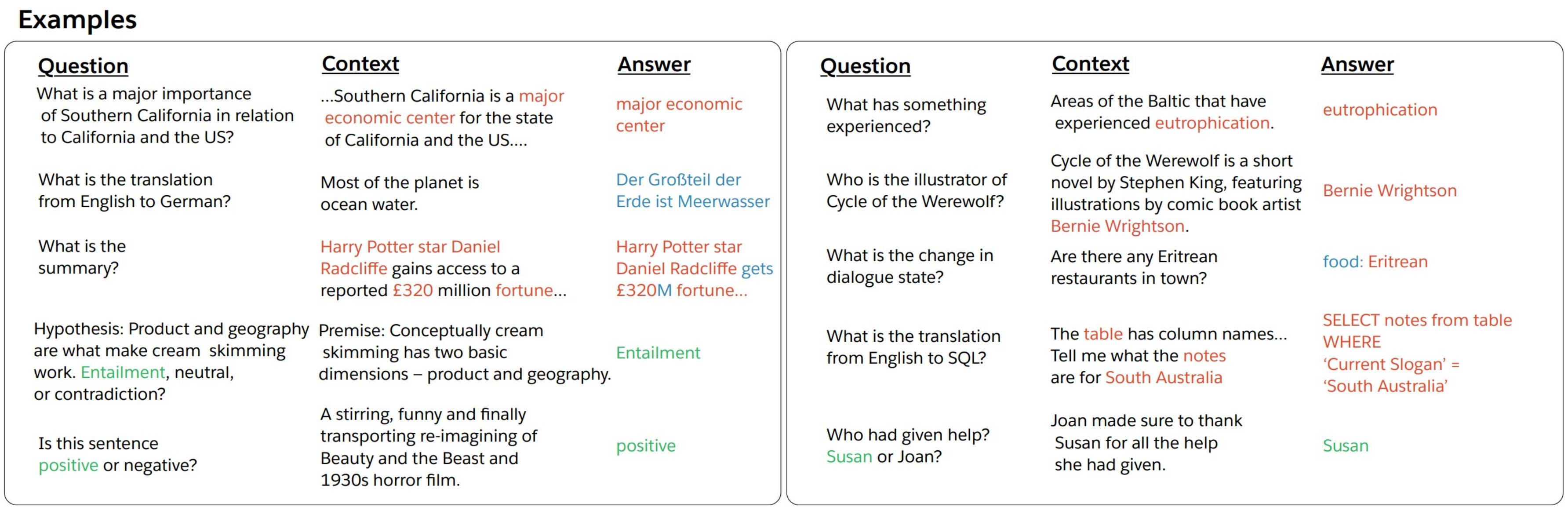
Table of contents

Introduction

Omni-Supervised learning was defined as *a special regime of semi-supervised learning in which the learner exploits all available labeled data plus internet-scale sources of unlabeled data* in a paper by Facebook AI Research (FAIR) in the 2017 paper Data Distillation: Towards Omni-Supervised Learning .

A fancy name to give to something researchers in NLP have been doing for years. Word embeddings have been sourced from *internet-scale data*, and then applied to several tasks achieving state-of-the-art results.

A paper The Natural Language Decathlon: Multitask Learning as Question Answering has demonstrated how labelled data can be used to train a language model to perform multiple tasks by casting all tasks as question-answers over a context, while another recent paper Language Models are Unsupervised Multitask Learners by researchers at OpenAI has shown how better quality data, and a more complex transformer based architecture results in a model that can achieve state-of-the-art results without any finetuning whatsoever.



Examples

Question	Context	Answer
What is a major importance of Southern California in relation to California and the US?	...Southern California is a major economic center for the state of California and the US....	major economic center
What is the translation from English to German?	Most of the planet is ocean water.	Der Großteil der Erde ist Meerwasser
What is the summary?	Harry Potter star Daniel Radcliffe gains access to a reported £320 million fortune...	Harry Potter star Daniel Radcliffe gets £320M fortune...
Hypothesis: Product and geography are what make cream skimming work. Entailment, neutral, or contradiction?	Premise: Conceptually cream skimming has two basic dimensions – product and geography.	Entailment
Is this sentence positive or negative?	A stirring, funny and finally transporting re-imagining of Beauty and the Beast and 1930s horror film.	positive

Question	Context	Answer
What has something experienced?	Areas of the Baltic that have experienced eutrophication.	eutrophication
Who is the illustrator of Cycle of the Werewolf?	Cycle of the Werewolf is a short novel by Stephen King, featuring illustrations by comic book artist Bernie Wrightson.	Bernie Wrightson
What is the change in dialogue state?	Are there any Eritrean restaurants in town?	food: Eritrean
What is the translation from English to SQL?	The table has column names... Tell me what the notes are for South Australia	SELECT notes from table WHERE 'Current Slogan' = 'South Australia'
Who had given help? Susan or Joan?	Joan made sure to thank Susan for all the help she had given.	Susan

Figure 1: Overview of the decaNLP dataset with one example from each decaNLP task in the order presented in Section 2. They show how the datasets were pre-processed to become question answering problems. Answer words in red are generated by pointing to the context, in green from the question, and in blue if they are generated from a classifier over the output vocabulary.

While the researchers at OpenAI made no attempts at finetuning the GPT2 on various tasks - the whole point of the paper was that language models trained with quality data can achieve competitive results on various tasks *without any finetuning*. However, we couldn't help but be *very excited* about finding out how such a model would perform with finetuning, considering that, to the best of our knowledge, there has never been any language model trained with data of this quality, and scale without the data being bastardized by any harsh pre-processing.

And so that is what we plan on doing. Utilizing the smallest pretrained GPT2 model released by OpenAI, we would be finetuning the model and evaluating it's performance on either the open, crowd-sourced NLU benchmark by Snips.ai, or the NLU Evaluation Corpora (Braun et al.), whichever proves to be easier to work with.

We'll also be experimenting with the use of data augmentation in the question-context-answer format proposed by McCann et al., by paraphrasing the questions and answers, which we hypothesize will result in a model that generalizes better.

The Problem

The problem which we refer to as Natural Language Understanding is succinctly described in this blog post, to quote:

> ### Natural Language Understanding - It's all about filling slots.
>
> The trickiest part in Natural Language Understanding is extracting the attributes of the query the user is making. This problem is called slot-filling. Let's take the following example:
>
> > *"Is it gonna be sunny on **Sunday after lunch**, so we can go to **Coney Island?**"*
>
> A first model will first identify that the user is asking for the weather conditions. The slot filler will then be looking for typical attributes of such queries: a location, a date, etc. In short, the NLU component receives the sentence in natural language: "Is it gonna be sunny on Sunday after lunch, so we can go to Coney Island?", and returns a structured object that a classic algorithm can act upon:
>
> ```
> {
>  "intent": "GetWeather",
>  "slots":[
>    "datetime": "2017-06-04T14:00:00-05:00",
>    "location": "Coney Island"
>  ]
> }
> ```

Literature Review

- The Natural Language Decathlon: Multitask Learning as Question Answering

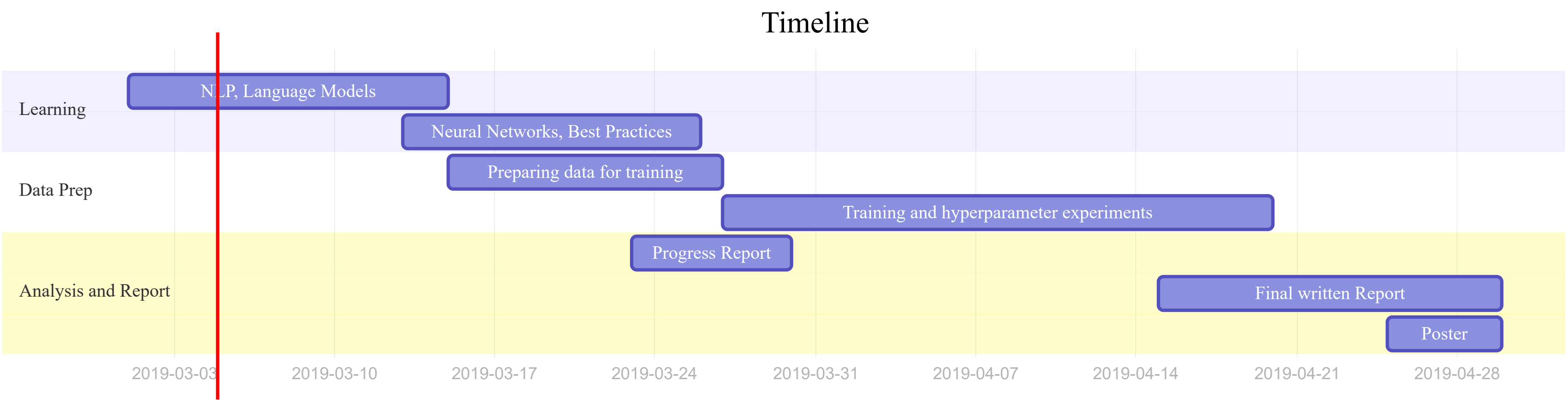
  > Deep learning has improved performance on many natural language processing (NLP) tasks individually. However, general NLP models cannot emerge within a paradigm that focuses on the particularities of a single metric, dataset, and task. We introduce the Natural Language Decathlon (decaNLP), a challenge that spans ten tasks: question answering, machine translation, summarization, natural language inference, sentiment analysis, semantic role labeling, zero-shot relation extraction, goal-oriented dialogue, semantic parsing, and commonsense pronoun resolution. We cast all tasks as question answering over a context. Furthermore, we present a new Multitask Question Answering Network (MQAN) jointly learns all tasks in decaNLP without any task-specific modules or parameters in the multitask setting. MQAN shows improvements in transfer learning for machine translation and named entity recognition, domain adaptation for sentiment analysis and natural language inference, and zero-shot capabilities for text classification. We demonstrate that the MQAN's multi-pointer-generator decoder is key to this success and performance further improves with an anti-curriculum training strategy. Though designed for decaNLP, MQAN also achieves state of the art results on the WikiSQL semantic parsing task in the single-task setting. We also release code for procuring and processing data, training and evaluating models, and reproducing all experiments for decaNLP.

- Language Models are Unsupervised Multitask Learners

  > Natural language processing tasks, such as question answering, machine translation, reading comprehension, and summarization, are typically approached with supervised learning on task specific datasets. We demonstrate that language models begin to learn these tasks without any explicit supervision when trained on a new dataset of millions of webpages called WebText. When conditioned on a document plus questions, the answers generated by the language model reach 55 F1 on the CoQA dataset - matching or exceeding the performance of 3 out of 4 baseline systems without using the 127,000+ training examples. The capacity of the language model is essential to the success of zero-shot task transfer and increasing it improves performance in a log-linear fashion across tasks. Our largest model, GPT-2, is a 1.5B parameter Transformer that achieves state of the art results on 7 out of 8 tested language modeling datasets in a zero-shot setting but still underfits WebText. Samples from the model reflect these improvements and contain coherent paragraphs of text. These findings suggest a promising path towards building language processing systems which learn to perform tasks from their naturally occurring demonstrations.

Logistics

Timeline



Division of work

Unlike a traditional software engineering project, there is no easy way to divide labour when it comes to training a model. As such we would shy away from attempting to do such a thing, and work as a single unit until we are in the training phase, at which time all of us will run independent experiments for training the network; the best model will be used.

Miscellaneous Information

Questions that we want to answer during the project

- How much better does a fine-tuned GPT2 perform vs vanilla GPT2,
- How does the augmentation of the questions and answers posed affect the generalising power of the model.

Expectation of what you will be able to learn from the project

- Working with neural network frameworks,
- Best practices for training a neural network,
- Answers to the questions posed above.

Is our idea novel?

In a way - yes, in a way - no. While we couldn't find evidence of what we propose being done before, it being such a natural extension of earlier works it wouldn't be very ethical to call the work novel.