Things:

### Dataset:

* Yorai: Finish up the dataset uploads.
* Nikud: Get rid of it (optionally). To aid with having less distinct words.
  + re.sub(r'[\u0591-\u05BD\u05BF-\u05C2\u05C4-\u05C7]', '', token)
* Debug: single letter words + weird wrong words.
* Any characters that are not letters (e.g. dashes) to spaces.
  + Apostrophes to empty-strings.

### Models:

* **CopyNet** - the one Jacob sent us
* **Transformer from PyTorch.** Mostly based on **"**Attention is all you need"
* Baseline: HMM from pset.
* Maybe **GPT-2** (OpenAI)

Doron Hazan, Yorai Shaoul, Dana Rosenfarb, Zachary Metzman

A 1–2 page summary of project progress so far. What experiments have youalready run? How long is model training taking? What information have youlearned that you didn’t anticipate in your initial project proposal? What stillneeds to happen for the final report submission?

Intro and Project Description

[Doron] I can write

Data Collection and cleaning [Doron]

* Scraping and parsing
* Removing netkedot
* Remove weird characters
* God’s name
* Say how many sentence and words
* Reversing
* Source

Models

No attention ?

**Dana Model [Dana]**

Transformers with attention

**CopyNet [Zach]**

Since modern hebrew and biblical hebrew share many of the same words, it was suggested that we explore a copy mechanism. Therefore, we have been experimenting with CopyNet described in the following paper <https://arxiv.org/pdf/1603.06393.pdf>. We found an older implementation of the model, and we are modifying it to suit our needs. So far, we there have been some issues with training. First, there is a vocabulary size parameter that should be in the tens of thousands, but with a vocab size of that magnitude it is difficult to find a setting that does not cause a CUDA memory error. With a smaller vocab size, say 2000, the validation BLEU scores are in the thousands place. In addition, while the BLEU on the validation set may be increasing slightly, we find that the validation loss is increasing too. Given the time for this project and the resources available, a CopyNet implementation with experiments may not be viable. Instead, experiments involving transfer learning of pretrained seq2seq models from PyTorch seems far more doable. It may be interesting to see if certain pretrained translation models allow for faster convergence with the two types of hebrew. If so, it may suggest that there are abstract language concepts that the model captures that are related to hebrew.

* Not working well. Need to figure out what is wrong. Loss does not seem to improve at all. I’ve tried many epochs. There are some other parameters that can be changed but it is likely that it is an issue with decoding (look for different implementations)

Maybe transfer learning of other models.

GPT3 https://github.com/openai/gpt-3

**Baseline seq2seq [Yorai]**

[Yorai] I can run some experiments with that. [Doron] I can write about the model and add too :)

Evaluations + Experiments:

* Nikud, no nikud (not many words so no convergence -- will try on Dana's model).
* Baseline scores from seq2seq. How do we want to measure success? BLEU?
* Try to translate modern-hebrew inputs and hope for it to stay the same-ish
* Warm starting Dana Model?
* **BLEU Scores for initial model.[Yorai]**
* **Initial Ablation Studies [Yorai]**