Fall 2020 Introduction to Natural Language Processing
Project Proposal

Project Topic: Machine Translation

Project Team members

1. Sam Perlmutter
2. Ethan Mendel

Prior work (5 points): Each group member must contribute at least 1 source (citations, resources, links) based on prior work in this topic area.

Team member 1: I found multiple datasets of English-Hebrew sentence pairs to use a training set – http://casmacat.eu/corpus/global-voices.html | http://opus.nlpl.eu/Wikipedia.php

Team member 2: I read through googletrans API to help build a hebrew corpus as a starting point and a checker for our English/Hebrew machine translation - https://pypi.org/project/googletrans/

Next: Indicate which of the ideas was selected to continue for the project.

We are going to develop an attention-based neural machine translation model to translate between English and Hebrew.

Data sources (4 points): Each group member must propose at least 1 data source for the chosen topic and document it with at least 1 source (see note below) in the proposal.

Team member 1: http://casmacat.eu/corpus/global-voices.html | http://opus.nlpl.eu/Wikipedia.php

Team member 2: http://www.manythings.org/bilingual/heb/

Next: Indicate which of the data sources was selected to continue for the project.

We are going to aggregate all sentence pairs from all the datasets to train and test using as large a dataset as possible.

Fall 2020 Introduction to Natural Language Processing
Project Proposal

Project responsibilities (1 point)

Aggregate data together, build and test model

Team member 1: Aggregate data from sources provided, test model

Team member 2: Aggregate data from source provided, build model