## Quaranteam

Christie McDaniel – 800 635 720

Tyler Norton – 800 679 931

Ruslan Cahnovsky – 800 617 436

Dylan Sholtes – 800 623 380

# Improving Upon REGMAPR Model for NLP

## Background and Motivation

Comprehending the meaning of sentences and their relationships to each other can come quickly to humans, even instantly. Contradictions between claims can be simple to spot, and methods like process of elimination can be used when they aren't as obvious. However, Natural Language Processing (NLP) is still growing and has much further to progress before it completely catches up to human capability. While machine learning can already tackle question answering and sentence generation among other tasks, it is still lacking in ability in some of the more intricate nuances of language and comprehension.

Time permitting, our goal is to create a novel approach to NLP and more specifically Natural Language Inference (NLI) in order to solve this problem. However, as this is our first foray into examining the implications of NLP, we stand instead on the shoulders of giants such as Stanford University's Computer Science department. The Stanford Natural Language Inference (SNLI) Corpus is a dataset of over 570,000 human-labelled sets of sentences, hypotheses, and labels for which to train NLI algorithms and test their accuracy.

After viewing research on the published models built using the SNLI dataset, we found the REGMAPR model lacking in information pertaining to the parameters and the training accuracy of the model. We plan to improve on the REGMAPR model since it has so little information published. We also have motivation to improve upon the recorded test accuracy of the model, which is currently 85.9%.

## Method

REGMAPR is a model that does not use bidirectional LSTMs like most other neural models. It first starts with a Siamese architecture and then is augmented with additional features. It will receive a pair of sentences, with each word mapped to its corresponding distributed representation or word embedding.

## Intended Experiments

The dataset we will use includes three text files: a dev file containing 10,000 sentence pairs, a training file containing 550,152 pairs, and a test file containing 10,000 pairs. The data has eight different fields, including the sentence pairs and parses produced by the Stanford Parser. Two different parses are included, one in Penn Treebank format, and another in binary format for use in tree-structured neural networks. Other fields include annotator labels and a gold label used for evaluating classification accuracy, as well as a captionID and a pairID as unique identifiers to each sentence pair. We plan to implement a REGMAPR model, since the Stanford publications lack any information aside from the testing accuracy of this model. This will also allow us to evaluate the performance of our model against the performance of a similar model to improve upon the test accuracy. Additionally, this will allow us to find a training accuracy for the REGMAPR model, as there is no recorded training accuracy in the Stanford publications. In the Stanford publications, the REGMAPR model used to achieve the 85.9% accuracy only includes BASE and REG from REGMAPR. In our REGMAPR model, we could improve upon the accuracy by including MA and PR.