**Hybrid Distributional and Definitional Word Vectors**

Team members: Harold Mei, Ranjani Venkatesh Iyer

Mentor: Andrey Kurenkov

* Introduction

Word vectors are typically arrived at through distributional statistics (such as co-occurence), but it is surprising that the most logical source of words' meanings - dictionaries - are not leverages in the process. We want to investigate the ability to use word definitions in the process of creating word vectors.

The project is a continuation of Andrey Kurenkov and Tony Duan’s previous work on Def2Vec, in there the authors quantitatively and qualitatively demonstrate that, by leveraging definitions alone, Def2Vec is able to embed words into a semantically meaningful space comparable to that of pretrained GloVe embeddings; they also demonstrated the utility of Def2Vec in improving the performance of a Neural Machine Translation model when the pre-trained vectors vocabulary is limited.

However, Andrey’s team realized that by using merely definitional vectors is unable to perform as well as distributional vectors in the NMT system; This motivates the introduction of a combined distributional and definitional word vectors - Hybrid Distributional and Definitional Word Vectors. Including both types of representation captures different aspects of a given word’s meaning and so the combination of them may outperform either one alone.

* How we plan to do it

As a continuation of a previous project, our plan is to follow the previous work by applying the similar methods that Andrey has done already. We will first build a word definition encoder by utilizing a Seq2Seq auto-encoder with attention, and a recurrent LSTM model to maximize the bag-of-words unigram probability for each word in the definition (? Harold). After that the word definition vectors will be applied on an open-source OpenNMT system by concatenating its GloVe vector to train and evaluate the model (this part should be one of the unfinished by Andrey?).

We will also apply an intrinsic evaluation and an extrinsic evaluation for HybridVec. In Intrinsic evaluation the goal is to assess the quality of the definition encoder model by measuring the mean distances for Def2Vec and GloVe vectors for words in Stanford Rare Words Dataset, and to verify its accuracy in reconstructing the corresponding ground truth GloVe vectors through t-SNE. In the extrinsic evaluation the job is to very how useful HybridVec is in downstream tasks by tools of Perplexities and BLEU metrics.