CS 224N Project Proposal: Contextualization of Reddit Post Titles

Tyler Chase
tchase56@stanford.edu

Rolland He
rhe@stanford.edu

William Qiu
willqiu@stanford.edu

February 9, 2017

Mentor: Danqi Chen

Problem Description

For our project, we will be looking at Reddit posts and specifically trying to identify patterns of post titles for various subreddits. To this effect, the goal of our project will be twofold. First, we will analyze post titles from different subreddits and attempt to predict the subreddit using only the title. Additionally, we aim to implement a bot that is able to generate post titles for particular subreddits.

Data

The project will utilize the full Reddit Submission Corpus, (obtained from http://files.pushshift.io/), which contains all reddit submissions (both posts and comments) categorized by subreddits from January 01, 2008 thru August 31, 2015 (with partial data for years 2006 and 2007). Because the total number of subbreddits on Reddit exceeds that of thousands, we will only consider the most popular subreddits by subscribers accessible from http://redditlist.com/. We will filter through posts recent enough that they occurred after reddit had accomplished their large following, but not recent enough that we encounter posts where vote statistics haven't stabilized. We will filter through this post data to take a large number of post titles (x) and labellings of the titles based on their subreddit (y).

Methodology/Algorithm

For subreddit prediction, we plan on using a Bag of Words/Word2vec model. For the post generation portion of our project, we simply utilize a recurrent neural network structure (RNN) with a LSTM layer. In both of the above tasks, we will use TensorFlow to implement the models. We do not know of existing implementations for our exact problems, though we will likely reuse some code from the homework assignments when implementing the different models.

Related Work

Since our problem relies heavily on using LSTM's, we will be reading various resources and papers on some state of the art LSTM techniques.

In addition, the NIPS paper *Sequence to Sequence Learning with Neural Networks* [**sutskever2014sequence**] details an effective method to train an LSTM on datasets where the output is a sequence of words instead of a fixed representation (expressed as some vector). Since we are building a post generator and cannot evaluate the performance of it via traditional methods where there is a fixed target output, it is necessary to use a method that maps sequences to sequences.

Finally, we plan to read through *A Neural Conversational Model* [**vinyals2015neural**] and *A Persona-Based Neural Conversation Model* [**li2016persona**] to better understand how to build conversational models, which are commonly used in sequence to sequence frameworks, since they can be trained end to end.

Evaluation Plan

As mentioned in the related work section, we cannot perform evaluation in the traditional sense, where we have a fixed target output. Instead, we will use two main methods of evaluation that is currently used in the NLP community. First, we will simply hand score a batch of generated posts (~ 100) generated for each given subreddit. We will also utilize the BLEU method [**papineni2002bleu**] frequently used in machine translation as a more automatic approach. The drawback of this approach, however, is simply that since our model does not conform to conventional task of having labeled outputs to input data, the BLEU method can only be an approximate assessment of our model. We will use the combination of these two methods when assessing the power of our models when making adjustments to, for example, the hyper-parameters. Given the ambiguous nature of our generated posts however, even with these two asessment methods, there is no guarantee that we are able to determine the true optimal set of hyper-parameters for our model.