Project Proposal

Souvik Bagchi, Ikhlas Attarwala

What task will you address, including evaluation metrics, and why is it relevant? You want to make this convincing, so spend about two paragraphs on this.

In this project we will be focusing on the following task -

- Named entity recognition
- Question Answering tasks (if time permits)

Named entity recognition (NER) has been sought after in Natural Language Processing (NLP) since decades and is an important part of natural language understanding. Most NLP algorithms today based off of neural language models (NLM) or statistical language models (such as latent semantic analysis, matrix factorization etc.) give us pretty accurate answers but are usually unable to answer as to why the model is working. There have been many interpretations in the past as to why the models work. For example, many claim that SLMs work due to inherent capture of statistical knowledge of words in a corpora, while, in NLMs the more recent use of Long short term memory (LSTMs), claims to "*preserve"* the essence of the document in the weight matrix which makes them perform well.

We believe that using NER along with NLMs will give us a better understanding of why the the NLMs perform the way they do and hence consider the task of NER important in the space of NLP.

One of the holy grails of NLP has been the task reading comprehension, viz. given a text can a computer answer a question whose answer can be found in the text. In our project given that we have enough time we will try implement the question answering task in our project.

Metric for evaluation for both tasks-

Precision -

- For NER it is simple to calculate the precision

- For the question answer task - Given you have predicted answer vector **x** and corresponding correct answer vector **a**,we treat **a** as a bag of words and then calculate the precision

Recall - similar to calculating the precision
F1 score - harmonic mean of precision and recall

How will you acquire your data? Most projects should be completed with off-the- shelf corpora, but if you are inventing a new data set, be as specific as possible here.

NER task -

For the NER task we have a cleaned dataset with annotation at https://www.kaggle.com/abhinavwalia95/entity-annotated-corpus

Question Answer Task -

For the question answer task we will be using the Squad 2.0 data set which can be found here - https://rajpurkar.github.io/SQuAD-explorer/

What model will anchor you project and what modifications (if any) to the model will be explored?

For our model we will be using the Bidirectional Encoder Representations from transformers (BERT) model.

For the NER task (which will be our primary focus) we will use BERT by feeding the output vector of each token into a classification layer that predicts NER. We plan to use a softmax for this last layer but would love some guidance as to what else we can do to improve our model.

For the question answer the model can be trained by learning two extra vectors that mark the beginning and the end of the answer.

Which features/attributes will you use to perform your task?

NER task –

We will be using annotations from the data set which will be fed into the model and expect the model to learn the annotations from the backprop.

Question Answer task -

For this task we will be feeding the text and the expected output answer to the model. The transformer encoder should be able to learn this model once provided with this information.