CS224N Project Proposal

Jeffrey Zhang, Ann He SUNet ID(s): jz5003, annhe

February 9, 2017

Mentor

Our mentor is Danqi Chen.

Problem Description

Given two questions, can we successfully predict whether or not they are duplicates? This is an interesting question to Quora because they actively merge questions so that relevant information can be found in the same place. Furthermore, the dataset is newly released, so not many people have worked on it yet.

Data

First Quora Dataset Release: Question Pairs–a binary-labeled dataset of questions pairs indicating whether they are semantically identical or not. Dataset found here:

https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs

Methodology/Algorithm and Related Work

We will first use a naive approach looking at the cosine-similarity of question vectors created from the average of word vectors in a question.

We can calculate the conditional probability of the second question being the translation of the first question. This is good because there have been lots of papers on learning this conditional probability. Usually it is formulated as follows:

Let x be the vector of words in the first sentence (question in our case) and y be the vector of words for the second. The conditional probability is given by:

$$p(y \mid x) = \prod_{t=1}^{T} p(y_t \mid y_1, y_2, \cdots, y_{t-1}, x).$$

The paper Bahdanau, Cho, Bengio 2015 computes these individual probabilities with a bidirectional RNN. Other related papers cited in the paper use other methods to compute this probability. We will implement some of these methods.

Also, we can try an LSTM network to see if improvements can be made from the naive approach (summing word embeddings). The paper Mueller Thyagarajan 2016 describes such an approach for predicting sentence semantic similarity with LSTMs.

Finally, we will use the attention mechanism approach to see if further improvements can be made to the LSTM approach.

Further extensions can include pre-processing the input questions into dependency-parsed representations. We believe that this may be a promising approach since questions are oftentimes more logically structured than sentences in general. In fact, Tai, Socher, Manning 2015 describes Dependency Tree-LSTM's and other Tree-LSTM approaches.

Evaluation Plan

We are ultimately performing binary classification. For the purposes of the duplicate question problem, false positives are especially undesirable, so we will use precision and recall to analyze our results, producing plots for the precision/recall curve and also just raw accuracy for each of our approaches and hyperparameter tunings.

Links to Relevant work

1. Bahdanau, Cho, Bengio 2015:

   `https://arxiv.org/pdf/1409.0473.pdf`

2. Mueller, Thyagarajan 2016:

   `www.mit.edu/~jonasm/info/MuellerThyagarajan_AAAI16.pdf`

3. Tai, Socher, Manning 2015:

   `http://www-nlp.stanford.edu/pubs/tai-socher-manning-acl2015.pdf`