QUESTION OR NOT
TEXT CLASSIFICATION OF ENGLISH SENTENCES

Anfal Alatawi, Bayader Alsahafi, Najwa Noorwali

Dr. Areej Alhothali | CPCS433

Introduction

Can we make machines determine whether a sentence is a question or not? To answer this question, we attempted to train and test two machine learning models that are widely used in natural language processing: Convolutional neural network (CNN) and Support Vector Machin (SVM).

Models

SVM:

Finds the margin that maximally separates question and non-question sentences.

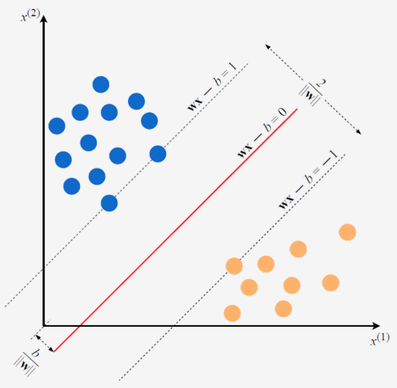

Figure 1. Optimal Hyperplane using the SVM algorithm

CNN:

CNN uses a variation of multilayer perceptron designed to require minimal preprocessing. It's generally used in computer vision, however they've recently been applied to various NLP tasks and the results were promising

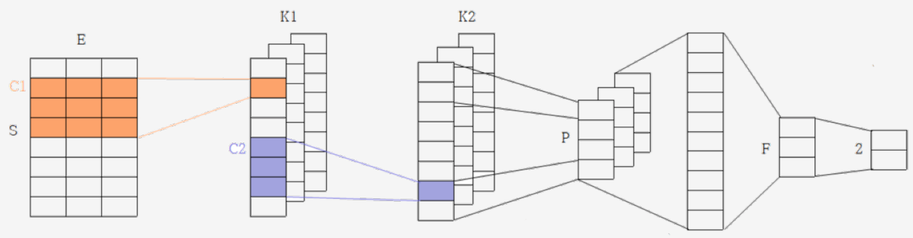

Figure 2. Representation of CNN layers

Results

	CNN	SVM
Training Accuracy	94.3%	89.7%
Testing Accuracy	93.9%	80.4%

Generating Dataset

Dataset:

- 8351 questions from Quora website
- 8347 sentences from OPUS

Labeling:

- 0 for sentences
- 1 for questions

Data preprocessing:

- Convert text data to lower case
- Remove punctuations
- Tokenize text data
- Pad each sentence to the maximum sentence length

Feature Extraction

We performed feature extraction on our text data using Glove pretrained word vectors. The Glove has multiple embedding vector sizes, including 50, 100, 200 and 300 dimensions. We chose the 100-dimensional version

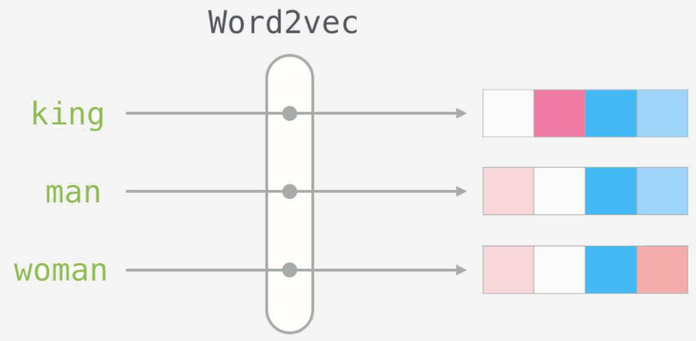

Figure 3. Words are represented as a d-dimensional vector

Analysis and Future Work

Both models reported good results on training and testing. As we can see, CNN achieved a higher accuracy than SVM. However, more investigation is needed to check the reliability of the obtained results.

Data Improvement:

- Remove stop words that doesn't convey any notion of meaning
- Collect more data

Model Accuracy:

- Try different values for the CNN parameters
- Explore the effect of text augmentation