CSCI - B 659: Assignment #1

Due on Sunday, February 26, 2017

Prof. Cavar

poosingh/rraavi/vpatani

Contents

Text Corpus Selection

We selected our corpuse of text to be **Driver.**

From the below given meanings (**NLTK**):

1. The operator of a motor vehicle
2. Someone who drives animals that pull a vehicle
3. A golfer who hits the golf ball with a driver
4. (Computer science) a program that determines how a computer will communicate with a peripheral device
5. A golf club (a wood) with a near vertical face that is used for hitting long shots from the tee

We are only interested in the **first**, **second** and **fourth** meaning of the words for our disambiguation.
We have followed 3 approaches to disambiguate the given word:

- Bag Of Words Approach
- Naive Bayes
- Naive Bayes with TF - IDF

Approach 1 - Bag of Words

Model Selection

Our Bag Of Words approach comes from the family of Lexical Sample Task. This contains a small set of pre selected words on which we train our model.
We first build a vocabulary of words at the begining, find synonyms and antonyms for it, and put them in a vocabulary collection. This is our feature set. Each vocabulary word (or its related meaning) represents a position in the feature vector.

Optimise Feature Extraction

While preparing the vocabulary, we use stemming (sometimes stemming adversely affects our output but we selected to go ahead with it anyway) and lemmatisation.
We pass each word of the vocabulary through Stemming and Lemmatisation, followed by looking for similar (synonyms and antonyms) words. This will also contain the word to be disambiguated.

Making the Training/Testing Model

We read each line of the training set and check for similar words from the vocabulary and if it does exist then we mark that location of feature vector as 1, which in the hindisght means that this sentence represents some relation (maybe it is negatively co-related, but it is) and it is good to learn.
We prepare the test set in a similar fashion.

Results

Once TimBL trains and test data is given, we need to interpret the result.
If on given test data, TimBL classifies the data as 0, then it has correctly classified that data point and 1 means it has incorrectly classified that data point. We had fairly variable results.
As a whole test data set, it has given us an accuracy within in a range of 0% to 60%, due to unpredictive nature of IB1 algorithm.
On an average we get 5/10 results correctly classified with the correct definition of that word.

Approach 2 - Naive Bayes

Approach 3 - Naive Bayes - TF-IDF

Comparison

References

1. Stanford Slides

2. PYWSD