Cue and Scope detection

Lot of work is going on Natural language processing in today's world. Sentiment analysis is one among the subfields of NLP where research is going on from decades. Several methods were invented and their drawbacks were identified slowly. The thought of making the prediction 100% correct led the data scientists to search for new methods which improve the accuracy. This methods Bag of Words, Using Long short-term memory, Recursive tree etc.,

Identification of Cue and scope in a sentence is the subfield of sentiment analysis. Cue is nothing but the word that alters the polarity of the sentence, Scope is the word that is affected by the cue. The given data is The Conan Doyle corpus with different stories. Training is done using the sentences in the book The Hound of Baskerville. The training process is feature based where the input to the model features.

There are mainly 2 kinds of -ve words

1. Words that are completely negative
   Eg: Not, None, Nothing etc.
2. Words that contain affix or prefix word that makes the entire word -ve
   Eg: Infrequent, Invisible etc

In the given data there are only 6 possible affixes or prefixes

“In”, ”I'm”, “less”, “Ir”, “dis”, “un”

Pure negative words (type 1) are detected using the bag of words but for type 2 we need a model and at the same time, there are cases where the word contain the affixes or prefixes but are not negative Eg: underlying. In order to make our model detect them accurately, 5 features are used.

1. After removing the affixed word in the original word will it make sense
2. Will the split word and original word have same parts of speech(90% cases true)
3. Will the list of opposite words of the split word contains the original word
4. Type of split
5. Length of an affix or prefix word

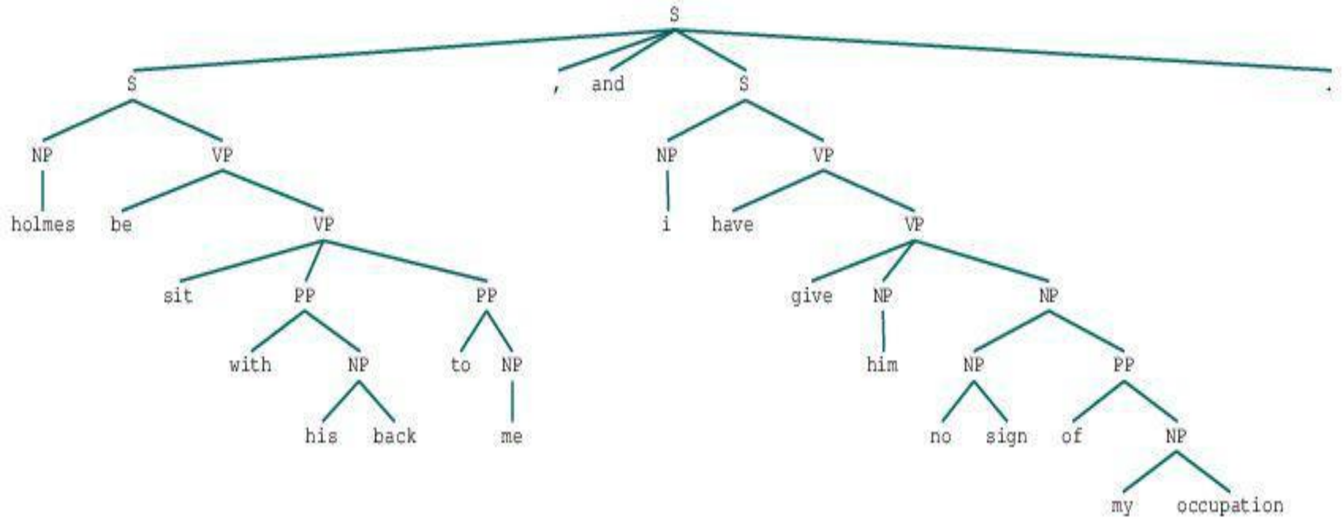
Based on all the file feature the model is trained. Support vector classifier with the non-linear model is used. The idea to choose these features is based on the study of various research papers. The selection of features is quite different in each research paper.

After cue detection, the sentences which contain at least 1 cue is stored. Now each sentence is taken individually and scope words in the sentence are again identified. There are mainly 4 feature in this identification process

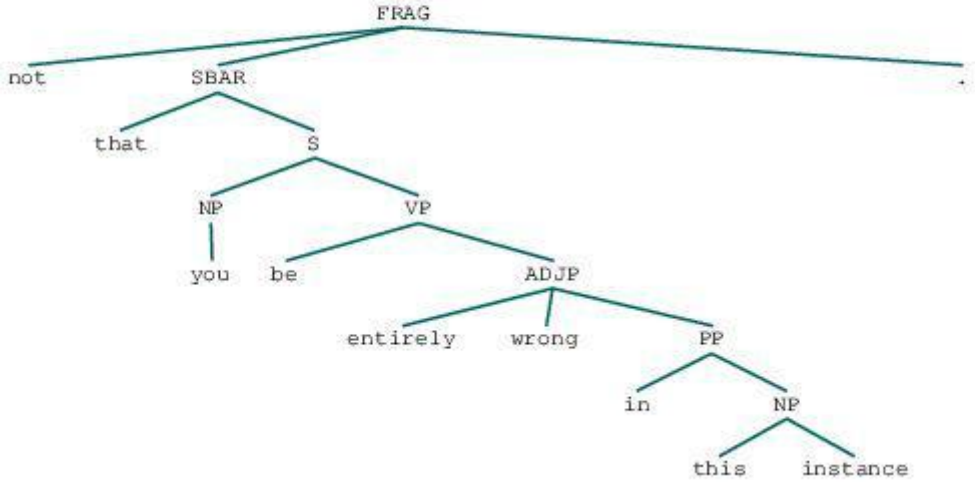
1. Left side distance of the current word to the cue in the sentence
2. Right side distance of the current word to the cue in the sentence
3. The shortest path from the current word to cue in the parse tree
4. Number of commas in the path from current word to cue in the sentence

The parse tree is something that is constructed using geometrical rules. Here are some examples of the parse trees.

For the sentence “holmes be sit with his back to me , and i have give him no sign of my occupation .” the parse tree is



For the sentence “not that you be entirely wrong in this instance .”



The shortest path is identified using this tree.

As all the four features are obtained a new SVC non-linear classifier is used for scope detection. These classifiers are used in test case prediction but the entire test case has to go along all these preprocessing steps before applying the predictor to a word or a sentence.

The challenges faced during this process are

1. The given data is skewed data: To avoid this proper features and model were selected. Selection of many features or much stronger classifier like artificial neural network leads to the decrease in the f1 score of prediction
2. Data-preprocessing: Using logical approach this problem was solved.

This model is not highly accurate but it gives a good output.

Thank you
Nikhil Konijeti