**Yarowsky's log-Likelihood decision ist model**

**Task at hand:**

The task at hand is a classification system based on a decision list that defines best log likelihood valued rules with preference to highest rule. The algorithm begins by listing a simple bag of words which helps in determining the ambiguities in the classes. The next step here is to collect the training context which is creating a labeled data with the class presence removed(Removal of accents) and having them in separate labeled set of data, after that we measure the collocational distribution by extracting features from both the class data available to us, the features taken in to considerations are linguistic features for that for the model are listed below-

1] Word to right

2] Word to left

3] Pair of words at offset

4] Word found in +-K window

After careful, feature extraction we compile them to get the collocational distribution, also the paper encourages use of more linguistic based features if possible to extract like Parts of Speech, lemmatization etc. The next step is very important the log-likelihood for all features is extracted and a rank-order list called decision rule list is created with best log-likelihood having the highest spot. Now, the testing is done on this basis and the decision rule list is used upon the extracted features of the testing set to determine the class for that test data. A further optional pruning step is also defined. The main step is that of training the decision lists for general classes of ambiguity which determines which feature rules are to be used for evaluation.

The primary assumption being made here is that the model will work quite well even on test data with no matching decision rule and other assumption is that of a binary classification system which might be very narrowed for classification in general.

The decision list approach can be used for many different spectrums of classification even incorporated into non-binary classification, a key application of such a list can be in clustering of text of images where the decision list approach will be very important in determining which cluster to place the object in from the domain. The decision list with modification can also be used for prediction based tasks like regression or decision trees.

**My model and conceptual decisions:**

Conceptually, the task was to implement a binary classification that will align itself to a class based on features partial towards the class. This is achieved first by extracting the features which is a bag -of – words form of values for many linguistics based data, then this data is used upon the training context developed and finally a value is achieved which determines which features are more likely to align with which class for that train data. That is different training data will have different decision lists, the decision list is used to match the same features for a test data and if the same features are detected then based on the rules a class is assigned to that test data instance. The features, used by my model are listed below-

**1-POS at -k**

**2-POS at +k**

**3-last word in context**

**4- Word at +1**

**5- Word at -1**

A major decision that I took was related to the baseline, when the feature extracted seem sparse that is many instances of test data do not match to that on the decision list, the model after considering all features on the list assign the baseline majority class to that data. This although makes the data more like a baseline model, improves on it in some instances.

**Top 10 decision rules and anylisis:**

**A table highlighting the top 10 decision list for both BASE and SAKE is given below,**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **TOP 10** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
| **SAKE** | **Last word**  **‘the’** | **Word**  **at -1**  **‘a’** | **POS at**  **-2**  **‘PRP$’** | **POS at**  **2**  **‘NN’** | **POS at**  **2**  **‘OT’** | **Last word**  **‘and’** | **Word**  **at -1**  **‘the’** | **POS at -2**  **‘IN’** | **Word**  **at 1**  **‘’for’** | **Word**  **at 1**  **‘for’** |
| **BASS** | **Word**  **at -1**  **‘the’** | **Last word**  **‘the’** | **POS at**  **2**  **‘NN’** | **POS at**  **2**  **‘IN’** | **POS at**  **-2**  **‘NN’** | **Word**  **at 1**  **‘a’** | **POS at**  **-2**  **‘DT’** | **Last word**  **‘and’** | **Word**  **at 1**  **‘the’** | **Word**  **at -1**  **‘and’** |

Here, we can see that the influence of last word and POS features is heavy in both the data sets, although the top decision for BASS is the word at -1 but further decisions down the list are full of POS based features. Also, it is observed that words before and after bass are generally nouns while in case of Sake words before it are generally pronouns with words after it being personal pronouns. In both cases the last word that is most prominent is ‘the’, this might be an indication of the cut-off contexts given in the data instead of entire sentences. In case of Bass the word ‘the’ defines a lot of features defining it’s implication when it comes to the sense of it while in case of sake it is found that word ‘for’ is influential which shows that the data in fact had very less number of \*sake words and sake is usually associated with ‘for’ around it. With removal of stop words one can further improve the model thus drastically increasing the chances of having more linguistically rich words showing up in the decision lists. This will help Sake but might harm bass as the stoop words help to an extent when it comes to determining bass’s sense.

**Report and analysis:**

**1] Comaprison against a majority class baseline:**

**A table representing the discrepancies between baseline and my model is given below-**

|  |  |  |  |
| --- | --- | --- | --- |
| **Comparison** | **Baseline Accuracy** | **My Accuracy** | **Baseline Label** |
| **BASE** | **56%** | **54%** | **BASE** |
| **SAKE** | **94%** | **90%** | **SAKE** |

The table above suggests that the baseline accuracy enables it to assign the non-astrix sense to both bass and sake with bass having a near 50% accuracy and sake mirroring the data set with having 94% accuracy. The silly baseline helps shine light on the fact that the data set for sake is not balanced and heavily partial towards one class while it also shows that the data for bass is too evenly distributed for it to extract good aligned features thus the low probability.

**2] Improvement over baseline:**

No improvement was observed for both sake and bass, but the normal model’s accuracy is very close to the baseline one with difference just in range of 3-4%. The major reason for this is the sparsity in features selected and the data on which it is trained. Specially the Sake data set is heavily aligned towards non-astrix sake which shows in its baseline accuracy and reflected in mine.

**3] Confusion matrix:**

**The confusion matrix for both bass and sake are given below-**

For Bass,

|  |  |  |
| --- | --- | --- |
| **Confusion Matrix (BASS)** | **BASE** | **\*BASE** |
| **BASE** | **30** | **26** |
| **\*BASE** | **20** | **24** |

For Sake,

|  |  |  |
| --- | --- | --- |
| **Confusion Matrix (SAKE)** | **SAKE** | **\*SAKE** |
| **SAKE** | **87** | **7** |
| **\*SAKE** | **3** | **3** |

This observation sums up what we already have deduced the low numbers of \*Sake is reflected with 50% of it being false negative. While the near perfect distribution for bass is also evident with equal numbers of bass and \*bass in the set.

**4] Examples from each test set:**

**Bass**

**Correct:**

1-presentation of a whole black bass slit and sprinkled with

2-the rocks or wreck with bass fishing gear on saturday the

3-the frantic drums and pulsing bass unleashed by roni size and

**Wrong:**

1-friends earshot in a worldrenowned bass lake like castaic pros dismiss

2-timbres into a single muddy bass line the treble fared no

3-laurent movingly portrayed by the bass robert lloyd secretly marries them

Sake

**Correct:**

1-had been done for the sake of knowledge and he had

2-had tasted the generic hot sake served at many of the

3-been the hallmarks of the sake industry in the 20th century

**Wrong:**

1-a type of fine traditional sake for which the rice is

2-are doing so for the sake of peace absent from the

3-bury the hatchet for the sake of their children and the

**5] Metrics (Marco-averaged precision with equal weighed classes):**

The metric is marco – averaged precision (recall) with both classes weighed equally,

For BASS,

Recall=True positive/(True positive+ False Negative)

=30/50

=0.6

For SAKE,

Recall = True positive/(True positive+ False Negative)

=87/90

=0.966

Absolute change in accuracy Is also given as

Absolute change= abs(Baseline accuracy – My Accuracy)/100

For BASS,

Absolute change= abs(Baseline accuracy – My Accuracy)/100

=abs(56-54)/100

=2%

For SAKE,

Absolute change= abs(Baseline accuracy – My Accuracy)/100

=abs(94-90)/100

=4%

|  |  |  |
| --- | --- | --- |
| **Metric:Case** | **BASS** | **SAKE** |
| **Recall** | **0.6** | **0.966** |
| **Absolute Change** | **2%** | **4%** |

**Reflection on results:**

The results achieved are self-proving that is the metric is an indication for values found in confusion matrix, the confusion matrix shows that the low numbers of \*Sake is reflected with 50% of it being false negative. While the near perfect distribution for bass is also evident with equal numbers of bass and \*bass in the set. Also, the metrics show that the sake has almost near perfect recall which suggest that the model trained well but what is not observed is the sparsity of one class compared to other class in sake database. The absolute change suggests the model accuracy is not far from baseline which suggest average performance which should be improved by interpolation or pruning or selection of better features. The primary concern is for the distribution of data in the data set, which was sparse in one while very well distributed in other. The sake testing although seem right is inherently very inaccurate while the bass feels wrong but is considerably accurate.