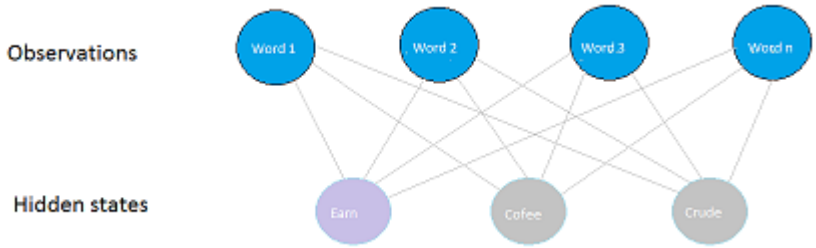

Figure 1: Text Categorization states in HMM

Related Works

HMM has been a useful statistical model for many applications in natural language processing for example part of speech tagging, speech recognition etc. In the field of information retrieval the model has been used very effectively - Miller et al 1999 and elke and Shauble 1994 (Information retrieval) Conroy and O'leary 2001 text summarization. Frasconi used HMM to classify multipage documents. The classification of the documents was based on the structure of the documents. Kwan yi and Jamshed Behesti used HMM to classify documents based on LCC(library of congress classification) scheme. Ludovic denoyer discussed the passage model for classification using HMM.

HMM based Text Classification

Document Preprocessing

In case of supervised learning document processing involves the following steps

- Removing the punctuation and numbers from the text.
- Tokenization and stemming, i.e. separating words by spaces and taking the root
- Removing the stop words like a, the ,and etc

After the documents are processed the documents are used feature selection. Since after removing the stop words from the dataset the number of words in the bag of words representation in the document set are huge. Feature selection involves reducing the number of features or words representing a particular category. Various methods have been used for feature reduction like chi-square, gini index, TFIDF etc. In this paper we have used the information gain approach of reducing the number of words for a given category.

$IG(t)=-\sum_{i=1...m} p(Ci)\log p(Ci) + P(t) \sum_{i=1...m} p(Ci/t) \log p(Ci/t) + P(t`) \sum_{i=1...m} P(Ci/t`)\log p(Ci/t`)$

- P(Ci) is the probability of category Ci
- P(t) is the parobability that t occurs in collection
- P(Ci/t) is the probability that a category is Ci given the term t appears
- P(Ci/t`) is the probability that the category is Ci given the term t does not appear

Table 1: Weighted Information Gain for Subset of Reuters Corpus

FACTORS	0.025
FALL	0.026
FARM	0.076
FARMER	0.025
FBC	0.029
FEBRUARY	0.091
FEE	0.029
FIGURES	0.025
FILES	0.029
FINANCE	0.025
FIRB	0.058

3 Building the Classifier

The HMM is trained using Reuters dataset on several categories. The goal of this phase is to build an HMM using preprocessed training data as input. The output of this learning process is the five parameters of the HMM, (S, V, π, A, B). We estimate the model parameters using the maximum likelihood estimate. Three sets of probabilities calculated using MLE is:

State transition probability distribution A

A={aij }: aij stores the probability of state j following state i.

N(si , sj): Number of times we move from state si to state sj
N(si): Number of transitions from state si.
V: entire vocabulary (all output symbols).
aij = P(qt = sj / qt-1 = si) = (N(si , sj) + 1) / (N(si) + N), i ≥1 and j ≥ N.

Transition probability gives the probability of changing from one particular category to another.

Observation symbol probability distribution:

B = {bj(k)} is the output symbol array that stores the probability of an
observation vk being produced from the state j, independent of time t.
B = {bj(k)} , bi(k) = P(xt = vk/qt = Sj) 1≤ j ≤ N and 1≤ k ≤ M.
N(k,i): Number of times state i has seen output symbol k.
N(s): Number of occurrences of state i.
V: entire vocabulary (all output symbols)
bi(k) = P(k|i) = (N(k,i) + 1) / (N(i) + |V|)
Initial state distribution: Π

Π= {Πi} is the initial probability array that stores the probability of the system starting at state si in an observation.
π = {Πi}, Πi = P (q1 = Si), 1≤ i ≤ N
Πi , the probability of being in state si at time t=1.
N(si): Number of times we start from state si.
N: Number of input sequences.
Πi = N(si) / N

Initial state probability matrix in our HMM model gives the probability for every particular catagory to be the first catagory in a sequence.

Categories for a block of text are determined by applying the Viterbi algorithm determined from the text. Viterbi algorithm involves multiplying many probabilities together. Since each of these numbers is less than one, we can end up working with numbers that are tiny enough to be indistinguishable from zero. To avoid this we worked with log of probabilities. The most likely path through the HMM is calculated and the categories are determined to be the unique collection of all categories from the path.

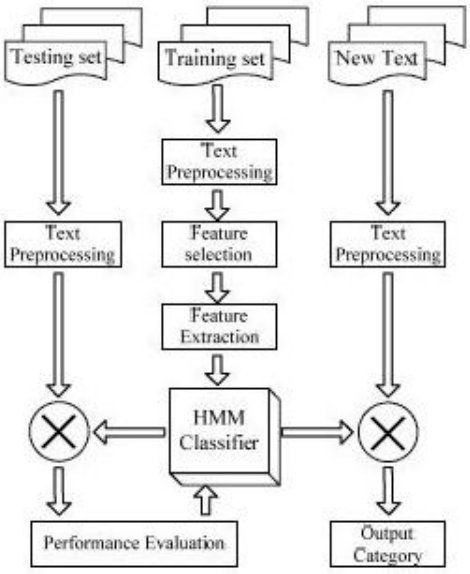

Figure 2: Text Categorization Process

Testing the Model

The model created is tested against the new document to be classified. The goal of this phase is to compute the most probable sequence of states for a given sequence of outputs using the HMM that we built from the training data. After the document is preprocessed the text set consists of all the tokens (output symbols) from the test set and the category set consists of C={c1,c2…cn}for which the model was created. After setting the HMM for each category and calculating P(wi/ λ) for each wi and using the viterbi algorithm for comparing the probability C_{max}=argmax P(wi/ λ) the final category of test set can be computed.

Table 2: Logarithmic Computations using viterbi algorithm - final category in bold

Time	*Token*	*Acq*	*Crude*	*Coffee*
0	Factors	-0.08	-0.48	-0.17
1	Agriculture	-.0045	-.0056	-0.038
2	crude	-9.84	-2.823	-3.677
3	File	-16.11	-12.15	-19.40
4	Finance	-19.86	-19.15	-25.19
5	Eugene	-25.91	-23.63	-27.68
6	Analyst	-30.26	-25.91	-34.18
7	Expert	-32.84	-26.69	-34.29

Performance Evaluation

The performance of text categorization model built is evaluated based on standard precision, recall, and F1 values. Precision, recall, and F1 values are calculated on the test data from the collection. Let TP be the number of true positives, i.e., the number of documents that both experts and the model agreed as belonging to the same category. Let FP be the number of false positives, i.e., the number of documents that are wrongly categorized by the model as belonging to that category.
Precision is defined as:

$$precision = \frac{TP}{TP + FP}$$

Let FN be the number of false negatives, that is, the number of documents that are not labeled as belonging to the category but should have been.

Recall is defined as:

$$recall = \frac{TP}{TP + FN}$$

The harmonic mean of precision and recall is called the F1 measure, and is defined as

$$F1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}}$$

4 Conclusion and Future works

We have presented a text classification model based on the hidden morkov model. The model was tested on the Reuters corpus. Only three categories from the Reuters corpus viz Acq, Crude, Coffee with 175 training documents and 60 test documents was used for training and testing the classifier respectively. The model uses only unigram tokens for training and testing. Since limited number of documents and terms were used, the performance of the model was satisfactory when compared to other machine learning text classification methods. However the performance of the model can be improved by using a large corpus of labeled training dataset. The performance can also be improved by considering two or more tokens together having some sort of semantic relation between them. By taking two or more tokens together the accuracy of the HMM based classifier can be improved. Also the structure of the documents under consideration can also be used to enhance the performance of HMM based text classification.

References

[1] Kwan yi and jamshid Behishti “Text categorization Model based on Hidden Morkov Model” CAIS/ACSI 2003R. Caves, Multinational Enterprise and Economic Analysis, Cambridge University Press, Cambridge, 1982. (book style)
[2] Ludovic Denoyer and Hugo Zaragoza (2001)“HMM based Passage Models for Document Classification and Ranking. 23rd BCS European Annual Colloquium on Information Retrieval, 2001
[3] Phil Blussom “A tutorial on Hidden Morkov Model” - 2004.
[4] P. Frasconi(2002) “ Hidden Morkov Model for Text Catagorization for Multipage Documents
[5] Honglak Lee and Andrew Y. Ng “ Spam Deobfuscation using Hidden Morkov Model”
[6] Flavia A. Barros, Eduardo F. A. “Hidden Markov Models and Text Classifiers for Information Extraction on Semi-Structured Texts”
[7] Max Bramer “Principles of Data mining” p-135- 152 2nd edition Springer.
[8] Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys, vol. 34, no 1, p. 1-47.
[9] Grezegorz Szymanski and Zygmunt Ciota “Hidden Markov Models Suitable for Text Generation”
[10] J. Jiang and C. Zhai, \Extraction of coherent relevant passages using hidden markov models," ACM Trans. Inf. Syst., vol. 24, no. 3, pp. 295{319, 2006.
[11] D. R. H. Miller, T. Leek, and R. M. Schwartz, \Bbn at trec7: Using hidden markov models for information retrieval," in In Proceedings of TREC-7, 1999, pp. 133{142.
[12] Yiming Yang, Jan O. Pederson “A comparative study on feature selection in text categorization”
[13] Yang Jian, Wang Hai-hang. "Text Classification Algorithm based on Hidden Markov Model"."Journal of Computer Applications", pages 2348-2350, 2361, 2010