**Summary**

**Text Classification in a Hierarchical Mixture Model Small Training Set**

* Topic hierarchies can be utilized to overcome the sparseness problem in text categorization with a large number of categories.
* This paper presents a **hierarchical mixture model** which extends the standard **naïve Bayes classifier**.
* Statistical methods developed to handle sparse data:
  + **Shrinkage**
  + **Deleted Interpolation**
  + These methods use term distribution estimated for more general, coarser text classes to provide better, smoothed estimates of .
* Hierarchical Shrinkage Model
  + Assume a predefined hierarchy of text categories.
  + Model parameters are learned through EM method
  + New documents are classified following Baye’s Rule
* This paper compares Naïve Bayes, the Hierarchical Shrinkage Model, Probabilistic Latent Semantic Analysis, KNN and SVMs.
* Data: a subset of 20 News-group dataset (each document has a single class) and Reuters-21578 (multi-label classification)
* **Hierarchical Mixture Model VS Other**
  + Hierarchy of topics
    - Provide better estimates for
    - Obtain a differentiation of words
  + Each word in a document is generated from some node on the path from the document class node to the root nodes
  + Term probability at inner nodes are effectively shared among multiple terminal nodes
  + Compared to CAM, this model has labelled data to train and predefined hierarchy.
* **Hierarchical Mixture Model**
  + Document
  + The document is assumed to be generated from repeated independent random trials. Number of trials = document length.
  + Probability of generating word in document that belongs to :
  + Model Parameters: . Estimated using tempered EM
  + Expectation:
  + Maximization
  + More general word become more probable at higher nodes level of tree.
* Preprocessing
  + Include subject header
  + Tokenized and lowercased
  + No stemming
  + Remove Stop words
  + Remove words that occurred less than 3 times

**The Cluster-Abstraction Model: Unsupervised Learning of Topic Hierarchies from Text Data**

* The proposed technique is purely data-driven and does not make use of domain-dependent background information, and it does not rely on predefined document categories or give list of topics
* means number of occurrence for word w in document d
* **Standard Latent Class (non-hierarchical)**
  + each document d belongs to one cluster, where number of clusters is assumed to be fixed.
  + Expectation Step:
  + Maximization Step:
* **Cluster-Abstraction Model**
  + Assume each word occurrence has an associated abstraction node
  + Additional latent variable vector is introduced to assign words in d to exactly one of the nodes in the hierarchy
* **Hold-out Method**
  + Simplest version of cross validation
  + Just split into training set and test set
  + To tune the parameter of model
* **EM Algorithms**
  + Expectation:
  + Maximization:
* Annealed EM Algo
  + Temperature parameter T
  + T is utilized as a control parameter which is initialized at a high value and lowered until the performance on the held-out data start to increase.
* Preprocessing
  + Stemming
  + Standard stop word list.
  + Very rarely occurring words have also been eliminated.

**Distributional Clustering of English Words**

* Deterministic Annealing
* For large enough corpora, the number of possible joint events is much larger than the number of event occurrences in the corpus, so many events are seen rarely or never, making their frequency counts unreliable estimate of their probabilities
* Dealing with the sparseness problem by estimating the likelihood of unseen events from “similar” events that have been seen. This requires a reasonable definition of verb similarity. Words are similar if they tend to participate in the same events.
* Word association tendency = association of words to hidden **sense class (cluster)** + association between sense class
* **Problem Setting**
  + Two major word class: Verb and Noun
  + Single Relationship
  + Distance between two clusters: , use it for two words to measure how likely they are in the same cluster
* **Theoretical Basis**
  + Objects are nouns, contexts are verbs and nouns as direct object
  + Basic Clustering model:
  + This model only has noun clusters with cluster memberships determined by P(n|c) and centroid distribution determined by P(v|c)
  + Average cluster distortion (cost):