A Bayesian Learning Approach to Concept-Based Document Classification

Speaker: Georgiana Ifrim
Supervisors: Prof. Gerhard Weikum
Martin Theobald

Outline

- Problem Statement
- Motivation
- Our Approach
- Experimental Results
- Conclusion & Future Work

Problem Statement

- Automatic text categorization
  - Assignment of natural language text to one or more predefined categories, based on their content
- Classical solutions
  - Learn prediction rules based on frequency statistics from a training collection
  - Apply the learned rules to classify new test documents
- Problem with existing solutions
  - No semantics of natural language involved

Motivation

- Improve classification accuracy by combining existing techniques from
  - Statistical Learning
    - EM algorithm, Bayesian classifier
  - Natural Language Processing
    - Stemming, PoS tagging, Word Sense Disambiguation
- Use existing knowledge resources
  - WordNet ontology
- Achieve robustness to language variation
  - Elimination of synonymy, polysemy

Classical Solution

- Naïve Bayes Classifier
  - Documents are generated from a parametric distribution P[d|t]
  - Naïve assumption: Given the topic label, features observed in a document are independent

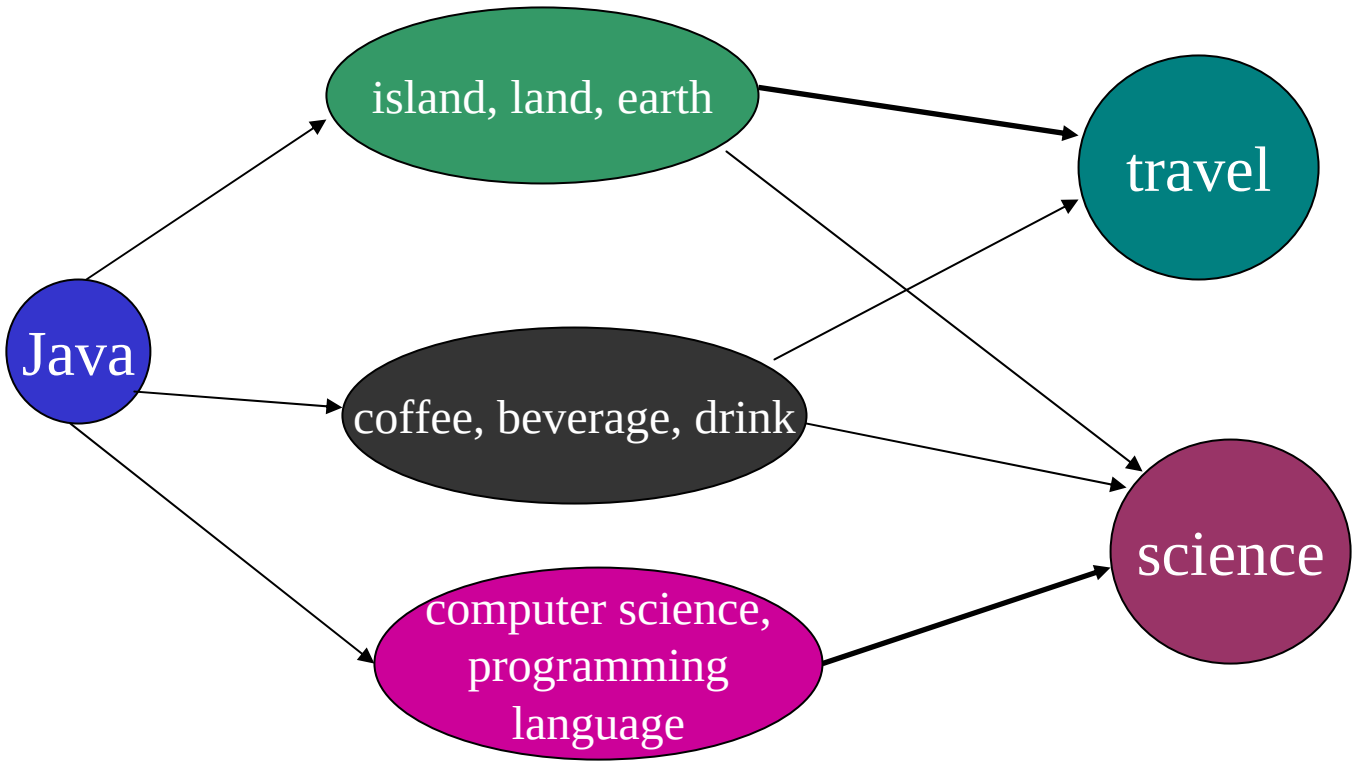
$$P[d \mid t] = \prod_{f \in d} P[f \mid t]$$

  - Estimate model's parameters P[f|t] from a training collection (maximum likelihood)
  - Use Bayes' rule to reverse the generative model and predict which topic generated a certain document

$$P[t \mid d] = \frac{P[d \mid t] \bullet P[t]}{P[d]} = \frac{P[d \mid t] \bullet P[t]}{\sum_{t} P[d \mid t] \bullet P[t]}$$

Our Approach

- Relate features to topics through latent concepts

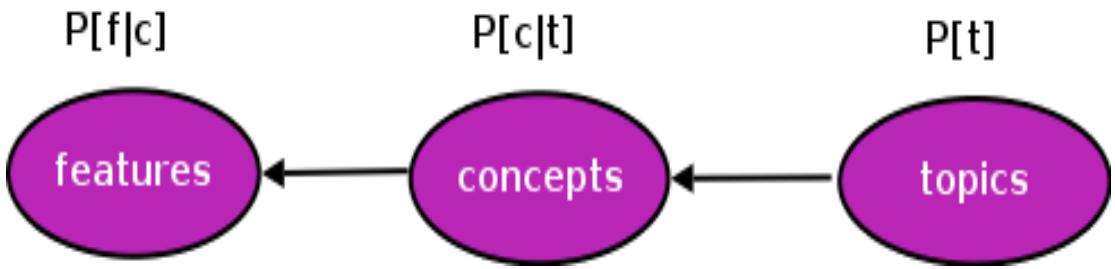


Our Approach

- **Given**
  - A data collection (Reuters-21578, Amazon)
    - A set of training and test documents with known topic labels and observed features, but latent concepts
  - An ontology DAG of concepts (WordNet)
    - Each concept has a set of synonyms, a short textual description and is linked by hierarchical relations
- **Goal**
  - For a given document, predict its topic label

Latent Generative Model

- **Generation process**



- Select a topic t with probability P[t]
- Pick a latent variable c with probability P[c|t]
  (prob that concept c describes topic t)
- Generate a feature f with probability P[f|c]
  (prob that word f means concept c)

Latent Generative Model

- Estimate model's parameters: EM algorithm
- Problems with EM
  - Large number of model parameters sparsely represented in the observed training data
  - Possibility of converging to a local maximum of the likelihood function
- Solutions proposed
  - Prune the parameter space to reflect only meaningful combinations
  - Pre-initialize model parameters to get closer to the global maximum

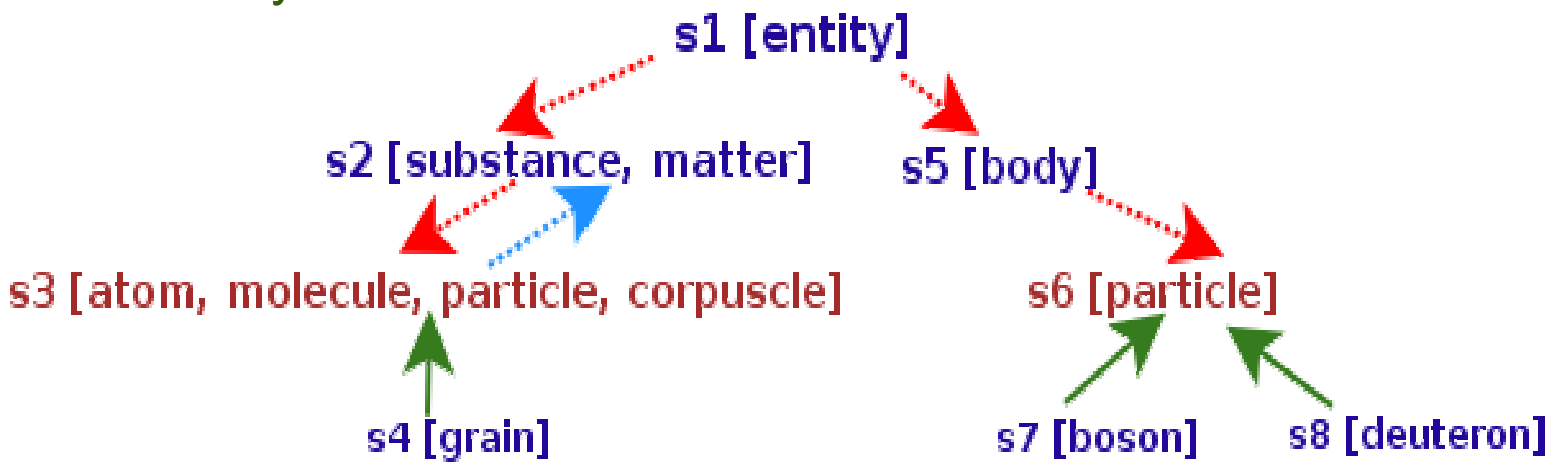
Latent Generative Model

- Pruning the parameter space
  1. Feature selection (Mutual Information)
  2. Concept selection (from the ontology) that reflects the semantics of the training collection well
  - ➢ Effect:
    - ➢ Reduce computational complexity, Increase model robustness

- Pre-initialize model parameters
  1. P[f|c] = sim(context(f), context(c))
  2. P[c|t] = sim(context(c), context(t))
  - ➢ Effect:
    - ➢ Get closer to the global maximum of the likelihood function

Ontology

- WordNet
  - Senses of particle
    - Hypernym
    - Hyponym
    - Meronym



Experimental Results

- Evaluation measures
  - Per topic:
    - Precision = $\frac{\#\ correct\ positive\ predictions}{\#\ positive\ predicitons}$
    - Recall = $\frac{\#\ correct\ positive\ predictions}{\#\ positive\ examples}$
    - F1–measure: harmonic mean of Precision and Recall
  - All topics
    - Microaveraged: Precision, Recall, F1
    - Macroaveraged: Precision, Recall, F1

Experimental Results – Reuters-21578

- News collection
- Select 5 most populated topics: earn, acq, crude, trade, money-fx
- Training: 5,000 documents
- Test: 2,000 documents

- Small text example:
  - " Crude oil prices rallied today, moving over 17.00 dlrs a barrel because of Saudi Arabia's determined effort to support prices, analysts said."

  - "USAir offered of buy 50 pct of that airline's stock for 71 dlrs cash per share and the balance for 73 dlrs per share in USAir stock."

Experimental Results – Reuters-21578

- Sensitivity to training set size:
  - Number of features: 300.
  - Vary number of training documents. Average over 3 random selections of training sets.

Training per topic	Concepts LatentM	Concepts LatentMPoS	Microavg F1 NBayes	Microavg F1 LatentM	Microavg F1 LatentMPoS	Microavg F1 SVM
10	2669	1560	88.9%	88.7%	87.8%	90.0%
20	2426	1395	89.6%	92.2%	90.7%	92.1%
30	2412	1321	92.7%	94.0%	92.2%	93.6%
40	2364	1447	92.1%	93.0%	91.2%	94.5%
50	2411	1317	93.8%	95.0%	93.8%	93.8%
100	2475	1372	95.3%	94.9%	93.8%	95.5%
150	2477	1385	96.0%	95.0%	94.4%	95.4%
200	2480	1387	95.9%	95.8%	94.5%	95.9%

Experimental Results - Amazon

- Collection of natural language text extracted from www.amazon.com (books' editorial reviews)
- Select 3 topics: Biological Sciences, Mathematics, Physics
- Total number of documents: 6,000
- Split into:
  - Training: 1,500 documents (500 per topic)
  - Test: 4,500 documents
- Small text example:
  - **"You will learn about classical control theory and its application to physiological systems, and contemporary topics and methodologies shaping bioengineering research today. Discussions on the latest developments in system identification, optimal control, and nonlinear dynamical analysis will keep you up-to-date with recent bioengineering advances."**

Experimental Results - Amazon

- Sensitivity to the number of features
  - Training size: 500 docs per topic
  - Vary features: 100 - 1,000

Number of features	Concepts LatentM	Concepts LatentMPoS	Microavg F1 NBayes	NBayes PoS	Microavg F1 LatentM	Microavg F1 LatentMPoS	Microavg F1 SVM	SVM PoS
100	1099	509	75.9%	77.5%	78.3%	79.0%	77.3%	78.1%
200	1957	936	77.0%	78.8%	79.5%	81.3%	73.1%	72.1%
300	2886	1390	78.3%	79.4%	81.0%	81.9%	65.4%	66.2%
400	3677	1922	78.6%	79.9%	81.3%	82.0%	68.4%	66.4%
500	4623	2232	78.7%	80.3%	81.8%	82.5%	69.6%	67.9%
600	5354	2547	79.0%	80.2%	82.6%	82.7%	69.9%	69.7%
700	5973	2867	78.8%	80.5%	82.8%	83.1%	71.4%	70.8%
800	6551	3231	78.9%	80.3%	82.9%	83.1%	73.0%	72.2%
900	7230	3677	78.4%	79.9%	83.0%	83.2%	73.3%	71.9%
1,000	7877	3959	78.4%	79.9%	83.2%	83.5%	72.8%	73.0%

Conclusions & Future Work

- Generative model approach to text categorization
- Combines SL and NLP techniques to achieve robustness to variations in word usage
  - Latent variable model
  - Semantical knowledge resources
  - NLP techniques
- Increases classification accuracy by exploiting semantics of natural language text

- More experiments to further assess the proposed model
  - Different setups
  - Different collections

Thank you!