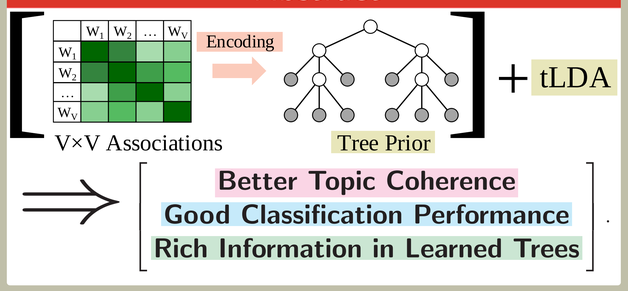
Adapting Topic Models using Lexical Associations with Tree Priors

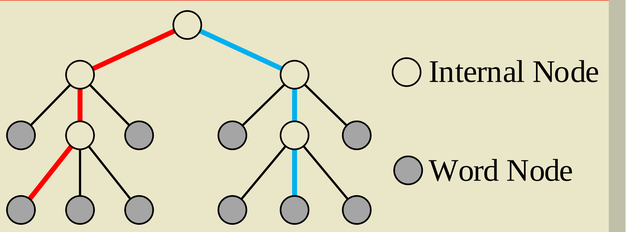
Weiwei Yang[1,2], Jordan Boyd-Graber[1,2,3,4], and Philip Resnik[1,2,3,5]

[1]Computer Science, [2]UMIACS, [3]Language Science Center, [4]iSchool, [5]Linguistics, University of Maryland, College Park

Abstract

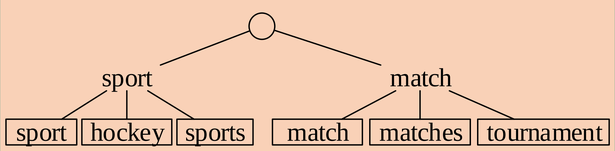


Tree Prior and tLDA



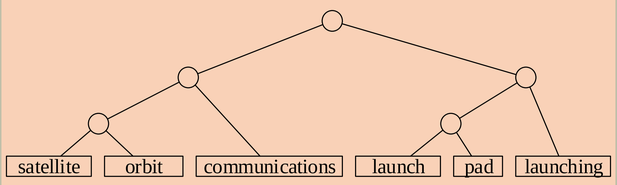
- Tree prior encodes word associations in its structure.
- Words are encoded in leaves and can be reached from the root via paths (red and blue lines).
- tLDA topics are multinomial distributions over the paths from the root to leaves.

Two-Level (2LV)



- Every word matches an internal node.
- Its child nodes are itself and the top N words with the highest association scores with the word.

Hierarchical Agglomerative Clustering (HAC)



- Every word is initially assigned to a cluster.
- Then repeatedly merges the two clusters with the highest average pairwise association score.

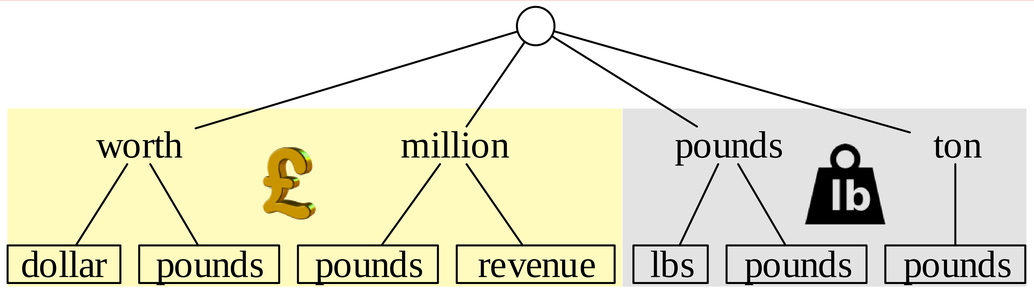
Pipeline

1. Prepare Word Associations

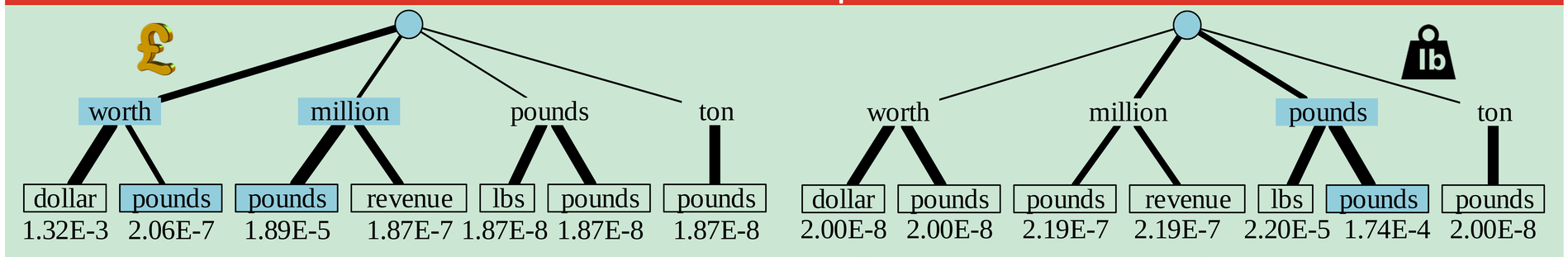
- Take "pounds" as an example.
  - British currency vs. weight unit.

worth	dollar	.5529
worth	**pounds**	.5095
million	**pounds**	.5161
million	revenue	.5628
lbs	**pounds**	.6070
ton	**pounds**	.4767

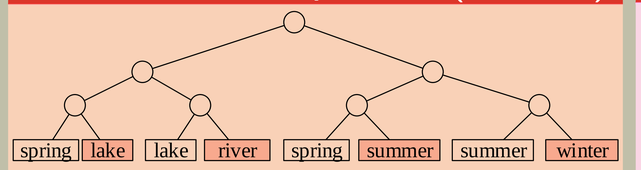
2. Construct Tree Priors



3. Learn Topics



HAC with Leaf Duplication (HAC-LD)

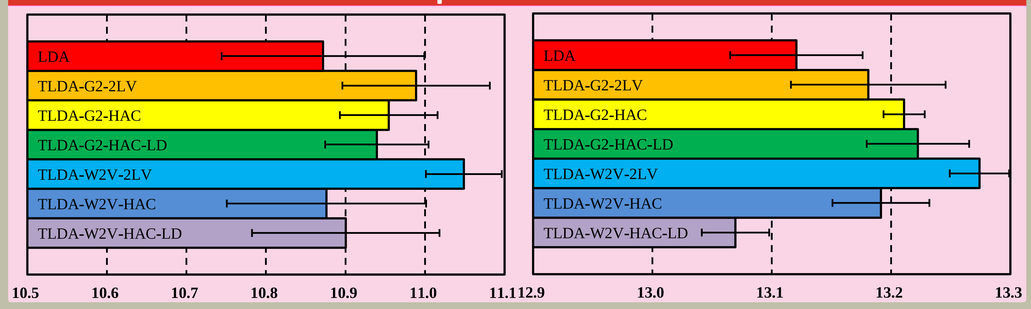


- Initialize every cluster with a word and its most associated word. Then applies HAC.
- A word with multiple senses can be assigned to multiple clusters close to its senses.

Topic Coherence

- **Datasets**: 20NewsGroups (20NG, left) and Amazon product reviews (right).
- **Associations:** word2vec (W2V) and Dunning likelihood (G2).
- **Metric**: Average pairwise **PMI** value of models' topics' top 10 words, on Wikipedia reference corpus. Higher is better.
- **Baselines**: LDA and latent concept topic model (LCTM).
- **Summary**: tLDA generally yields more coherent topics quantitatively. LCTM performs too poorly to be included.
- For topic words and qualitative analysis, see the paper. ↓

Topic Coherence



Paper

Classification

Model	Tree	Path	20NG	Amazon
BOW	–	–	86.64	86.73
BOW+VEC	–	–	86.59	**87.30**
LDA	–	–	86.67	86.99
LCTM	–	–	86.52	86.83
tLDA (W2V)	2LV	N	86.75	87.07
		Y	86.73	87.13
	HAC	–	86.79	87.19
	HAC-LD	N	86.73	87.02
		Y	86.94	86.88
tLDA (G2)	2LV	N	86.82	87.15
		Y	**86.96**	87.05
	HAC	–	86.63	87.11
	HAC-LD	N	86.73	87.07
		Y	86.91	86.94

- **20NG**: Multi-class SVM classification. Documents' groups are their labels.
- **Amazon**: Binary SVM classification. 4-5 stars are positive and 1-2 stars are negative.

wwyang@cs.umd.edu, jbg@umiacs.umd.edu, resnik@umd.edu