Topic Model Stability for Hierarchical Summarization

John E. Miller and **Kathleen F. McCoy**
Computer & Information Sciences
University of Delaware, Newark, DE 19711
millerje@udel.edu, mccoy@udel.edu

Abstract

We envisioned responsive generic hierarchical text summarization with summaries organized by topic and paragraph based on hierarchical structure topic models. But we had to be sure that topic models were stable for the sampled corpora. To that end we developed a methodology for aligning multiple hierarchical structure topic models run over the same corpus under similar conditions, calculating a representative centroid model, and reporting stability of the centroid model. We ran stability experiments for standard corpora and a development corpus of Global Warming articles. We found *flat* and *hierarchical* structures of two levels plus the root offer stable centroid models, but *hierarchical* structures of three levels plus the root didn't seem stable enough for use in hierarchical summarization.

1 Introduction

We envisioned a responsive generic hierarchical text summarization process for complex subjects and multiple page documents with resulting text summaries organized by topic and paragraph. Information extraction and summary construction would be based on hierarchical structure topic models learned in the analysis phase.[1] The *hierarchical* topic structure would provide the organization as well as the information quantity budget and extraction criteria for sections and paragraphs in hierarchical summarization. Initial attempts along this path offered promise for a more coherent and organized summary for a small corpus of Global Warming articles from (Live Science, 2015) versus that obtained by *flat* topic structures.

However, multiple analyses of the same Global Warming corpus and various standard corpora under similar conditions rendered seemingly different hierarchical topic models. Model differences remained even after transforming and reducing models based on required summary size and other extrinsic summary requirements. So we decided to examine topic model stability with the goal of assuring that stable, representative, and credible topic models would be produced in our analysis phase. This paper documents our effort at assuring hierarchical topic model stability for hierarchical summarization.

It is inherent in Bayesian probabilistic topic modeling and similar methods that repeat analyses of the same corpus under the same conditions give different results. But we must have substantially similar results to do credible hierarchical summarization (or other application). We require topic model stability, i.e., similar topic models for analyses performed under similar conditions. Without stable results, we do not know which analyses to believe, if any, and we mistrust the methodology itself. Furthermore, any application of the resulting topic model is not credible.

Organization of Paper Bayesian probabilisitic topic analysis (§2.1) expresses a corpus as the matrix product of topic compositions of words with document mixtures of topics. In *flat* topic analysis, the matrix of topic-word compositions is organized as a flat vector of individual topics. With *hierarchical* structure topic analysis, the topics take on a hierarchical tree structure.

Topic model quality (§2.2) is typically assessed by predictive likelihood of words for a test corpus or by assessment of topic coherence. Our stability assessment methodology seems largely com-

[1] Phases are the somewhat standard: corpus preparation, analysis, information extraction, summary construction.

Proceedings of the Workshop on New Frontiers in Summarization, pages 64–73
Copenhagen, Denmark, September 7, 2017.

plementary to quality assessment.

The Hungarian assignment algorithm (Kuhn, 1955) has been used for aligning *flat* topic model pairs (§2.3), based on a cost matrix of pairwise topic alignments. We will use a pairwise topic similarity measure for populating the Hungarian algorithm's cost matrix.

Topic models, including hierarchical models, are being used to construct text summaries (§2.4), including hierarchical text summaries. This provides sufficient reason to want to assure the stability of *flat* and *hierarchical* structure topic models.

We introduce the particular *flat* and *hierarchical* structure topic models (§3.1) used for this paper.

In a simple yet significant innovation, we extend topic alignment (§3.2) to hierarchical structure topic model pairs via a recursive application of the Hungarian assignment algorithm starting with root topics of the model pair. Surprisingly, we find time complexity of the *hierarchical* topic structure improves versus *flat* structure with increasing level of the hierarchy. [2]

We measure stability (§3.3) as alignment (proportion of aligned topics), similarity (weighted cosine similarity over topic compositions), and divergence (Jensen-Shannon divergence over topic distributions). Measures are defined for *flat* and then extended to *hierarchical* structure topic models.

The more topic models in the study, the more credible the stability analysis, since we are aligning more models and measuring stability based on more analyses. For complex problems, however, more models also makes it more likely we would encounter alternative topic models, just as human topic modelers might. We perform agglomerative clustering on topic model similarity (§3.4) to test whether models form a single or multiple stable topic model groups, or are unstable.

For each cluster, we align models and calculate topic frequency weighted centroids (§3.5) of topic-word compositions for aligned topics. Then we assess stability versus the centroid model (§3.6) similarly to that done previously for model pairs.

We demonstrate the methodology (§4) over *flat* and *hierarchical* structure models in an 18 run factorial experiment on three corpora, and in a separate *ad hoc* 16 run experiment on a larger corpus.

We return to our work on hierarchical summarization (§5) now armed with stable hierarchical topic models and examine our next steps as well as options for further research.

[2] Software engineering already knows this – that hierarchical structure is less time complex than monolithic.

2 Previous Work

We use Bayesian probabilistic topic modeling in the analysis phase of our hierarchical summarization process. Here we briefly review topic modeling, topic model quality, topic model stability, and use of topic models in hierarchical summarization.

2.1 Topic Models

The Latent Dirichlet analysis (LDA) Bayesian probabilistic topic model, introduced and popularized by Blei et al. (2003); Griffiths and Steyvers (2004), factors a corpus of document-word occurrences as the matrix product of topic compositions of words and document mixtures of topics (figure 1). The topic structure is *flat* and the number of topics, K, and vocabulary size, V, are fixed. In the generative probabilistic model, topic-word compositions are distributed symmetric Dirichlet with parameter η, and document-topic mixtures are distributed Dirichlet with concentration parameter α.

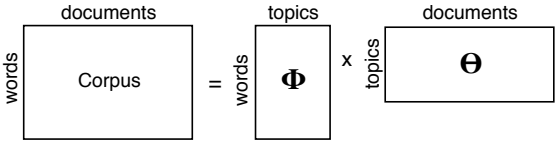

Figure 1: Topic Model Factorization of Corpus

Teh et al. (2005, 2006) generalized the LDA model in two important ways: (1) the number of topics, K, is made open ended by treating the topic model as a Dirichlet process (DP) with growth parameter γ for sampling a new topic, and (2) documents are sampled from Dirichlet processes (DPs) which are themselves sampled from corpus DPs thus forming hierarchical Dirichlet processes, HDPs, even while the topic structure remains *flat*.

Blei et al. (2010) developed *hierarchical* topic analysis where the generative model of the corpus consists of a hierarchy of nested Dirichlet processes (DPs) and each document is generated as a single non-branching path down the corpus hierarchical structure. *Stay-or-go* stochastic switches are used at each document node to determine whether to *stay* on the current topic or *go* to a topic further down the tree.

Paisley et al. (2015) extended the non-branching document paths to a nested hierarchical structure

Dirichlet process model with branching in both the document and global models. In figure 2, the grey represents the corpus tree and the black overlaid trees the individual document trees. Each document parent node is a DP sampled from its corresponding corpus node DP. Analysis infers the corpus topic structure and compositions, and document topic mixtures and *stay-or-go* switches.

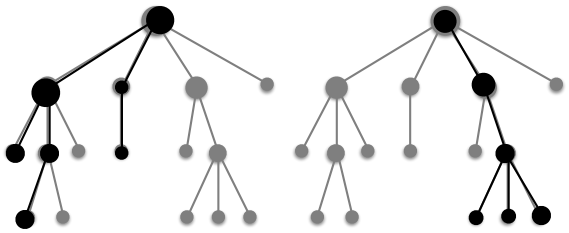

Figure 2: Hierarchical Corpus Structure

2.2 Quality

Predictive log likelihood for words, test LL($\mathbf{x}$), is a popular measure of topic analysis quality. Test LL($\mathbf{x}$) shows the predictability of words on test data given the model fit to training data (corpus topics and compositions). While not a stability measure, test LL($\mathbf{x}$) does give an objective indication of predictability. Teh et al. (2007) provides formulas for calculating test LL($\mathbf{x}$) for the *flat* topic structure in both Gibbs sampler and variational inference analysis methods.

Assessing quality of individual topics can be as simple as noting topics below a minimum frequency or comparing divergence of topics from any of uniform, corpus, or power distributions of word frequencies. More powerful methods assess individual and aggregate topic coherence. The current standard is to measure coherence by normalized pairwise mutual information (NPMI) (Aletras and Stevenson, 2013; Lau et al., 2014; Röder et al., 2015) versus pairwise probabilities calculated from some very large pertinent corpus.

We view test likelihood and topic coherence as largely complementary to topic model stability.

2.3 Topic Alignment and Stability

Topic models must be aligned on topics before assessing stability. de Wall and Barnard (2008) calculates similarity weights between topics from different models over documents, constructs a cost matrix from negative similarity weights, and applies the Hungarian assignment algorithm (Kuhn, 1955) to determine the optimal pairwise topic model alignment. Stability is defined as the correlation between aligned topics over documents.

Greene et al. (2014) calculates the average of Jaccard scores on sets of popular word ranks between topic combinations of a topic model pair, and determines the model agreement (i.e., stability) as the average over topics of Jaccard scores resulting from the optimal topic alignment by the Hungarian assignment algorithm.

Chuang et al. (2015) notes that model alignment is "ill-defined and computationally intractable" with multiple-to-multiple mappings between topics, and adopts the solution of mapping topics *up-to-one* topic.[3]

Yang et al. (2016) aligns topics for *flat* topic structures also using the Hungarian assignment algorithm and *up-to-one* topic correspondence. Stability is measured as agreement between token topic assignments over aligned topic models.

We use the Hungarian algorithm and the *up-to-one* topic correspondence. We choose to emphasize topic correspondence based on topic word compositions, as in the generative model, and so base our cost matrix on similarity of topic word compositions between models.

2.4 Topic Model Based Summarization

Haghighi and Vanderwende (2009) examined several hybrid topic models using LDA as a building block and demonstrated the superior efficacy of their hybrid model (general topic, general content topic, detail content topics, and document specific topics) in constructing short summaries for Document Understanding Conferences (U.S. Department of Commerce: National Institute of Standards and Technology, 2015). Delort and Alfonseca (2011); Mason and Charniak (2011) used similar models in short summaries for the Text Analysis Conferences (of Commerce: National Institute of Standards and Technology, 2010, 2011). Celikyilmaz and Hakkani-Tur (2010, 2011) used a more general *hierarchical* LDA topic model structure, doing *hierarchical* summarization for longer summaries. Christensen et al. (2014) developed "hierarchical summarization" using temporal hierarchical clustering and budgeting summary component size by cluster.

We use a more general hierarchical structured Bayesian topic model similar to Paisley et al.

[3]Indeed, the issue of mapping 1 topic to 2+ topics would be an interesting and useful problem to solve.

(2015). Essential for any of these related hierarchical topic model or cluster based methods is the stability of the model used to drive summarization.

3 Methodology

We present a process for aligning topic models and measuring topic model stability for both *flat* and *hierarchical* structure cases. The resulting stable hierarchical structure topic centroid model would be further transformed to take into account extrinsic summarization requirements.

Stability – Measurement Process

1. Infer multiple topic models for the same corpus run under similar conditions.
2. Determine pairwise topic model alignments.
3. Calculate stability over pairs.
4. Cluster topic models using agglomerative clustering over pairwise stability.
5. For each cluster:
   (a) Align member topic models and calculate topic model centroids.
   (b) Align member topic models with topic centroid model.
   (c) Calculate stability of topic models with topic centroid model.
6. Interpret stability results.

3.1 Topic Modeling

For a *flat* topic structure, we use a Gibbs sampler implementation of Teh et al. (2006) hierarchical Dirichlet processes (HDP). For a *hierarchical* topic structure, we use a Gibbs sampler implementation of a simplified version of Paisley et al. (2015)'s nested hierarchical Dirichlet processes. Our simplified model and Gibbs sampler drops the use of *stay-or-go* stochastic switches at each document Dirichlet process (DP) node. See supplemental notes (Supplemental, 2017b).

3.2 Pairwise Topic Model Alignment

From a set of M topic models, all $M(M-1)/2$ model pairs are aligned based on topic pair assignment costs. Assignment cost between topics from distinct model pairs is calculated as

$$cost_{k,l} = -(m_k/N)(n_l/N) * cosSim(\mathbf{m_k}, \mathbf{n_l}),$$

where (k, l) indexes topics from model pairs, m_k and n_l are topic frequencies, N is corpus size, $\mathbf{m_k}$ and $\mathbf{n_l}$ are vectors of word frequencies for topic pair (k, l), and $cosSim$ calculates the cosine similarity.[4] By using topic frequency ratios in the cost, similar frequency topics are preferred. Since weak similarities are not useful, we censor $cosSim \leq .25$ and substitute zero for their cost.

***Flat* Topic Models** Pairwise costs are assembled into a cost matrix indexed by (k, l) and the optimal cost assignment of the model pair is determined by the Hungarian assignment algorithm. For unequal numbers of topics, vectors of zero (maximum) costs are substituted for nonexistent topics.

***Hierarchical* Topic Models** Hierarchical topic structures are single rooted branching trees of depth L where the root is depth 0. Each tree node includes a topic of word compositions, and each non-leaf tree node includes a Dirichlet process (DP) of topic mixtures. We restrict hierarchical topic structure alignment to require: (1) roots must align, and (2) aligned child branches must align in their ancestors. With these restrictions, we developed Minimize Subtree Cost (algorithm 1) applying the Hungarian algorithm to DP (non-leaf) nodes of the hierarchical topic structure.

Method $minimizeSubtreeCost$ is invoked initially for model pair roots, (σ_0, τ_0) and recursively thereafter for subtree pairs, (σ, τ). If either subtree is a leaf the topic alignment cost is returned. For internal nodes, a cost matrix is constructed between the child nodes for the subtrees, the Hungarian assignment algorithm is invoked to get the optimum cost alignment for the subtrees, the topic cost is added to the subtree costs, and this result is returned. Filling the subtree cost matrix calculates the cost of aligning properties between model pairs of subtree children by minimizing subtree costs for each child pair. Thus calculating subtree costs and filling subtree costs together ***recursively*** span the entire solution space for hierarchical topic alignment. See supplemental java snippets (Supplemental, 2017a).

Time Complexity For *flat* topic structures, topic alignment time complexity is $O(K^2(V + K))$, where K is the number of topics and V is the vocabulary size. Preparation of the cost matrix takes K^2 topic vector cosine similarity calculations over

[4] Alternatively, straight cosine similarity or a divergence measure such as Hellinger distance could be used.

Algorithm 1 Minimize Subtree Cost

Require: Trees σ, τ
  Method: minimizeSubtreeCost(σ, τ)
  if isLeaf(σ) **or** isLeaf(τ) **then**
    return topicCost(σ, τ)
  else
    $\mathbf{costs} \leftarrow$ fillSubtreeCosts(σ, τ)
    return topicCost(σ, τ)
        +HungarianAssignment($\mathbf{costs}$)
  end if

  Method: fillSubtreeCosts(σ, τ)
  for $k = 0$ **to** σ.children.size **do**
    for $l = 0$ **to** τ.children.size **do**
      $\mathbf{costs}[k, l] \leftarrow$ minimizeSubtreeCost
          (σ.children[k], τ.children[l])
    end for
  end for
  return costs

V words giving $O(K^2V)$, and the Hungarian assignment algorithm which minimizes cost has time complexity $O(K^3)$ (Kuhn, 1955).

Level 1 in the *hierarchical* structure is similar to the *flat* topic structure. Time complexity is $O(B^2(V+B))$, with branching factor, B, in place of number of topics, K. Each increment in level increases by a factor of B^2 the tree node pairs from the parent level. The resulting time complexity for level l beyond the root is then $O(B^{2l}(V+B))$. For $B > 1$ the final level dominates the order calculation, and so the time complexity for a *hierarchical* structure of depth L is $O(B^{2L}(V + B))$.

We compare this with the time complexity for the *flat* structure alignment problem by expressing K as though from a *flattened hierarchical* structure, $K = (1 - B^{L+1})/(1 - B)$.[5] Then, $O(K^2(V+K)) = O([(1-B^{L+1})/(1-B)]^2(V+[(1 - B^{L+1})/(1 - B)]))$. For $B > 1$ the terms with B in the ratio dominate, and so expressing *flat* structure in *hierarchical* terms gives time complexity $O(B^{2L}(V + B^L))$. Cost of assignment for *flat* is greater by a factor of B^{L-1} versus a comparable *hierarchical* structure.

This is a surprising result! We had expected hierarchical structure to add time complexity, but instead it reduces time complexity with increasing level compared to a corresponding *flat* structure. Alignment of topics between *hierarchical* structures is less time complex than for *flat* structures.

3.3 Pairwise Stability

Given the topic model alignment, we calculate alignment, similarity, and divergence measures. Table 1 gives *a priori* and preliminary calibration study interpretations of the stability measures.

Proportion Aligned Alignment is calculated as, $pAlign = K'/[(K_\sigma + K_\tau)/2]$, where K' is the number of aligned topics, and K_σ and K_τ are the number of topics for each model.

Weighted Similarity Similarity is calculated as topic frequency weighted similarity of the topic word compositions of the (σ, τ) model pair, [6]

$$wtSim_{\sigma,\tau} = \sum_{\substack{(k,l)\in \\ aligned}} \frac{m_k + n_l}{2N} cosSim(\mathbf{m_k}, \mathbf{n_l}),$$

where (k, l) indexes topics from the *flat* or *hierarchically* aligned model pair, m_k and n_l are topic frequencies, N is the corpus size, $\mathbf{m_k}$ and $\mathbf{n_l}$ are vectors of word frequencies for topic pair (k, l), and $cosSim$ calculates the cosine similarity. Only aligned topics are added to the $wtSim$, but the corpus size includes all observations, so the fewer aligned topics, the lower the weighted similarity. For the *hierarchical* model we require that ancestors are also aligned.

Divergence Divergence is calculated as the Jensen-Shannon divergence (JSD) between topic frequency distributions for model pairs. Distributions are calculated as follows: (1) model σ topic frequency counts are assembled in array $\mathbf{s}$ by topic index k, (2) frequencies of unaligned topics from σ are set to zero with the sum of frequencies of unaligned topics set in $\mathbf{s}_K$ where K is the maximum number of topics for the (σ, τ) model pair, (3) model τ topic frequency counts are assembled in array $\mathbf{t}$ by topic index l, (4) frequencies of unaligned topics from τ are set to zero with the sum of frequencies of unaligned topics set in $\mathbf{t}_{K+1}$, and (5) topic frequencies in $\mathbf{t}$ are reordered according to the alignment mapping between (σ, τ). Thus, aligned topics coincide with respect to their positions in $\mathbf{s}, \mathbf{t}$ and unaligned frequencies are kept separate between models. Divergence is calculated as

$$JSD(\mathbf{s}||\mathbf{t}) = 1/2(KLD(\mathbf{s}||\mathbf{m}) + KLD(\mathbf{t}||\mathbf{m})),$$

[5] Sum of geometric series, $\sum_{l=0}^{L} B^l$, for a branching tree.

[6] Unweighted or other weighting could be used as well.

Basis	Value	Interpretation
a priori	$alignment = 1$	full alignment
calibration	$alignment \approx 0.6$	useful alignment
a priori	$similarity = 1$	full similarity
calibration	$similarity \approx 0.6$	useful similarity
calibration	$similarity \approx 0.25$	marginal similarity
a priori	$divergence = 0$	full convergence
calibration	$divergence \approx 0.1$	strong convergence
calibration	$divergence \approx 0.4$	strong divergence

Table 1: Preliminary interpretation of stability

where $\mathbf{m} = (\mathbf{s} + \mathbf{t})/2$ and KLD is the Kullback-Leibler divergence. For the *hierarchical* model we require that ancestors are also aligned.

3.4 Cluster Topic Models

There are multiple ways in which topics can be organized and assigned - whether performed automatically or by human experts. So we test whether model pairs align to a single stable model group, or if multiple stable groups can be identified.

We use group-average agglomerative clustering (Manning et al., 2008) on pairwise weighted similarity, $wtSim$, to form model clusters. This results in compact clusters maximizing separation between clusters while minimizing the distance between the cluster centroid and its members. Clustering begins with each model forming its own cluster and ends when either all models form a single cluster or no more clusters can be formed that meet $\overline{wtSim} > cutPoint$, where $\overline{wtSim}$ is the average weighted similarity. Output is a list of clusters where each cluster includes a list of models ordered by entry into the cluster and $\overline{wtSim}$.

Agglomerative clustering is fast and simple; pairwise similarity scores do not have to be recalculated after each clustering step. However, we don't know what are the similarities or differences between clusters without inspecting them.

3.5 Form Topic Centroid Models

With only one cluster, no unclustered models, and good similarity, the models seem stable. We form topic centroids and report this centroid model as the representative topic model. With multiple clusters, we should consider the appropriateness of multiple solutions – perhaps corresponding to multiple human solutions. We form centroids for each topic and report centroid models as representative of the clusters. The occurrence of many unclustered models would indicate instability.

Controls specify a censor limit for similarity below which topics do not merge into a centroid, and a minimum number of models and minimum topic frequency below which topics drop from the centroid topic model. While a cluster may have several models, not all topics need not be aligned across all models.

Form Topic Centroid Model (algorithm 2) forms cluster centroid models by copying the cluster centroid from the initial model and then aligning and entering individual models into the centroid iteratively based on their order of entry into the cluster. The method `optimizeSubtreeMap`, a variation on the previous `minimizeSubtreeCost` (algorithm 1), returns the topic correspondence mapping. Topics which do not meet the topic similarity censor limit ($wtSim < .25$) are not aligned. Unaligned topics are provisionally added to the centroid model in case subsequent models in the list have similar topics. After the centroid model is formed, topics which to not meet a minimum topic frequency limit or minimum number of topic models limit are dropped.

Algorithm 2 Form Topic Centroid Model

Require: Cluster list of trees λ
Method: formCentroidModel(λ)
$\mu \leftarrow \lambda_0$
for $i = 1$ **to** λ.size **do**
  $mapping \leftarrow$ optimizeSubtreeMap(μ, λ_i)
  for all $topic \in \lambda_i$ **do**
    if $topic \in mapping$ **then**
      $index \leftarrow mapping.indexOf(topic)$
      aggregateTopic($\mu, \lambda_i, index, topic$)
    else
      addTopic($\mu, \lambda_i, topic$)
    end if
  end for
end for
for all $topic \in \mu$ **do**
  if failsDropLimits($topic$) **then**
    drop($\mu, topic$)
  end if
end for

3.6 Centroid Model Stability

For each cluster's centroid model, we align individual models with the centroid model and estimate stability. The method is similar to that for pairwise stability with the exception that the centroid model is always one member of the pair and so only M (centroid, model) pairs are analyzed.

3.7 Use in Hierarchical Summarization

The final product is a single stable centroid model, when one exists. The stable centroid model shows the topic structure, the proportional importance of each topic, and the word composition of each topic as a discrete probability distribution. In our hierarchical summarization process, this centroid model would be further transformed (nested, pruned, aggregated) by taking into account extrinsic requirements of summary size, and paragraph and subparagraph structure. The resulting topic structure model would be used to extract information proportionally for each topic, and organize the section and paragraph structured summary.

If the centroid model is not stable, then hierarchical summarization would not be credible. If there are multiple identifiable stable clusters, then their centroid models become candidates for organizing the hierarchical summary.

4 Stability Experiments

The purpose of the stability experiments is to demonstrate the methodology over corpora for *flat* and *hierarchical* structures. When stable centroid models result from replicate topic analyses, they can credibly be transformed to take into account extrinsic summarization requirements, and carried forward to the information extraction phase of our hierarchical summarization process.

4.1 Corpora

Corpora used in this study are Journal of the ACM (JACM) abstracts from years 1987-2000, Global Warming (GW) articles for the year 2015 (Live Science, 2015), Proceedings of the National Academy of Sciences (PNAS) abstracts for years 1991-2001 (Ponweiser et al., 2015), Neural Information Processing Systems (NIPS) proceedings for years 1988-1999 from (Lichman, 2013). PNAS and GW texts were lemmatized. Stop words and words with frequency less than ten were removed. JACM and GW are small corpora; JACM has very small abstracts while GW has short articles; PNAS has numerous abstracts and NIPS has longer articles.

4.2 Experimental Design

An 18 run factorial design (3 corpora x 3 levels x 2 growth rates) crosses JACM, GW, and PNAS corpora, with *flat* (L=0) and *hierarchical* (L=2,3) topic structures, and topic *growth* rates to achieve two different topic count ranges. Four replicate topic analyses were run at each factorial setting. For training, our simplified Gibbs sampler used α=1.0 and η=0.01 with optimization. The growth parameter γ was set to create topic counts at low (L), medium (M), and high (H) ranges.

Corpus	J	V	N	D
JACM	534	1,328	33,517	62.8
GW	116	970	31,894	274.9
PNAS	27,688	9,685	2,713,006	98.0
NIPS	1,491	6,149	1,813,400	1,216.2

Table 2: Corpora Characteristics.

J=document count, V=vocabulary size, N=corpus size, D=average document size.

Separately, an *ad hoc* experiment was performed on a set of 16 trials on the NIPS corpus with hierarchical (L=3) model using similar training control settings. This experiment demonstrates the occurrence of multiple clusters.

4.3 Results - Factorial Design

Stability analysis was performed for each experimental group of replicates. Topics were not aligned when $wtSim < .25$, clustering terminated when when $avgWtSim < cutPoint = .5$,[7] and topics were dropped from the cluster centroid model when $nModel_k < 2$.

Table 3 shows the results for the factorial design with corpus, hierarchical topic structure (L), and growth rate (γ). Results reported are number of topics in training model (K), and stability measures of number (K') and proportion of topics aligned (pAlign) in centroid model, average weighted similarity (wtSim), and hierarchical Jensen-Shannon divergence (hJSD). Ideal results based on *a priori* values (table 1) would be $pAlign \approx 1$, $wtSim \approx 1$, $hJSD \approx 0$.

We expected simpler would be more stable (Ockham's razor), such that more levels and topics give poorer stability. This is largely confirmed by stability measures in that greater hierarchy levels and greater topic count models generally had poorer stability measures. Hierarchical L=3 models and with the JACM corpus especially showed poorer stability.

[7] JACM $L = 3$ model used .4 for cut point.

Model		Train	Stability			
L	γ	K	K'	pAlign	wtSim	hJSD
JACM						
0	M	70.3	70.5	1.00	0.867	0.028
2	M	78.0	66.0	0.85	0.839	0.052
3	M	84.8	48.2	0.57	0.682	0.128
0	H	106.8	106.8	1.00	0.851	0.034
2	H	104.5	87.2	0.83	0.831	0.062
3	H	108.5	46.7	0.43	0.700	0.157
GW						
0	M	65.8	65.8	1.00	0.869	0.030
2	M	73.8	72.0	0.98	0.894	0.028
3	M	82.3	59.8	0.73	0.762	0.100
0	H	99.0	98.2	0.99	0.871	0.023
2	H	108.0	89.8	0.83	0.824	0.081
3	H	105.8	62.8	0.59	0.726	0.133
PNAS						
0	L	86.8	86.5	0.99	0.930	0.013
2	L	76.8	72.3	0.94	0.905	0.052
3	L	76.3	58.8	0.77	0.732	0.137
0	M	135.0	134.0	0.99	0.920	0.017
2	M	140.3	122.5	0.87	0.875	0.071
3	M	134.3	92.2	0.69	0.752	0.143

Table 3: Experimental results - stability.

4.4 Results - *Ad hoc* Design - NIPS

We analyzed a set of 16 trials on the NIPS corpus run under somewhat similar conditions with topic counts in the 90 to 200 range with hierarchical L=3. Given the corpus size, non-equality of conditions, and diversity of topic counts, we weren't surprised to find multiple distinct clusters.

Stability analysis was performed with control settings: topics not aligned for $wtSim < .25$, clustering terminated for $\overline{wtSim} < cutPoint = .5$ or .6, and topics dropped from the cluster centroid model for $nModel_k < 2$. Results are reported in table 4. At $cutPoint = 0.5$, all models formed one cluster; at $cutPoint = 0.6$, three separate clusters were identified and six models were not joined to any cluster. Proportion of aligned topics declined ($nModel_k < 2$ is a more stringent test when there are only 2 or 3 models in the cluster), but similarity and divergence measures were substantially improved for each of the three separate clusters.

4.5 Impact on Hierarchical Summarization

For corpora in the factorial design, both *flat* and *hierarchal* L=2 topic structures resulted in good stability (high alignment and similarity with little divergence), so the centroid topic model can credibly be carried forward for use in our hierarchical summarization process. The hierarchical L=3 models are generally less stable.

Cluster	nModels	pAlign	wtSim	hJSD
cut point=0.5				
0	16	0.81	0.592	0.246
cut point=0.6				
0	5	0.66	0.783	0.073
1	2	0.31	0.829	0.140
2	3	0.50	0.821	0.086
*	6 models were not clustered			

Table 4: *Ad hoc* stability experiment on NIPS.

The NIPS stability analysis for a single cluster shows moderate similarity of models and moderate divergence of topic distributions, while more restrictive clustering reveals three separate clusters and six unassigned models. This bears further investigation.

5 Discussion

We have:

- placed modeling hierarchical topic structure in the analysis phase of our hierarchical text summarization process;
- established the importance of a stable topic model for use in the analysis phase;
- developed a methodology for aligning and measuring stability of topic models;
- defined innovative and simple *hierarchical* topic structure model alignment via a recursive algorithm applying the Hungarian algorithm to individual Dirichlet processes;
- quantified time complexity of our hierarchical alignment algorithm and showed reduced time complexity at increasing *hierarchical* level versus *flat* topic structures;
- developed alignment, similarity, and divergence stability measures for *hierarchical* topic structures;
- applied agglomerative clustering to form coherent groups of topic models:
  - constructed representative cluster centroid models, and

  - calculated centroid model stability;

- demonstrated the methodology, finding credible models for *flat* and *hierarchical* L=2 structures;

- demonstrated the methodology on a large set of *hierarchical* L=3 topic models run on the NIPS corpus, finding multiple coherent clusters plus unclustered models;

- mentioned parenthetically work on a pilot calibration study for stability measures;

Future Work There is work to be done on topic model stability, model alignment, and stability measurement:

- apply our methodology to larger, more varied models and different inference methods;

- improve, expand, and publish calibration studies beyond our pilot;

- explore other topic model alignment cost measures;

- further improve topic alignment including options other than *up-to-one* matching;

- improve hierarchical structure topic model stability.

Summarization - Next Step We further transform the hierarchical topic structure taking into account extrinsic summarization requirements. The product from the analysis phase is a hierarchical structure topic model where each topic includes its proportional representation of the corpus and a composition of words given as a discrete probability distribution. This structure is used in information extraction, where topic compositions match information from the corpus, e.g., sentences, and proportional representation budgets the quantity of information to be extracted for each topic. The transformed topic structure organizes summary topic and paragraph structure.

Conclusion Our topic model stability methodology lets us diagnose and compute "usable" hierarchical topic models for collections of long documents. This is an essential and "attractive starting point towards hierarchical text summarization." [8]

[8] Thanks to reviewer for this concise statement of benefit.

References

Nikolaos Aletras and Mark Stevenson. 2013. Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 13–22. Association for Computational Linguistics.

David M. Blei, Thomas L. Griffiths, and Michael I. Jordan. 2010. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *J. ACM*, 57(2):7:1–7:30.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Asli Celikyilmaz and Dilek Hakkani-Tur. 2010. A hybrid hierarchical model for multi-document summarization. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 815–824, Stroudsburg, PA, USA. Association for Computational Linguistics.

Asli Celikyilmaz and Dilek Hakkani-Tur. 2011. Discovery of topically coherent sentences for extractive summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 491–499, Portland, Oregon, USA. Association for Computational Linguistics.

Janara Christensen, Stephen Soderland, Gagan Bansal, and Mausam. 2014. Hierarchical summarization: Scaling up multi-document summarization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Jason Chuang, Margaret E Roberts, Brandon M Stewart, Rebecca Weiss, Dustin Tingley, Justin Grimmer, and Jeffrey Heer. 2015. Topiccheck: Interactive alignment for assessing topic model stability. In *Proceedings of NAACL-HLT*, pages 175–184.

Jean-Yves Delort and Enrique Alfonseca. 2011. Description of the google update summarizer at TAC-2011. In *Proceedings of the Fourth Text Analysis Conference, TAC 2011, Gaithersburg, Maryland, USA, November 14-15, 2011*. NIST.

Derek Greene, Derek O'Callaghan, and Padraig Cunningham. 2014. How many topics? stability analysis for topic models. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part I*, volume 8724 of *Lecture Notes in Computer Science*, pages 498–513. Springer Berlin Heidelberg.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235.

Aria Haghighi and Lucy Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 362–370, Stroudsburg, PA, USA. Association for Computational Linguistics.

Harold W. Kuhn. 1955. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden*, pages 530–539.

M. Lichman. 2013. UCI machine learning repository.

Live Science. 2015. Live Science. Online at livescience.com.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK.

Rebecca Mason and Eugene Charniak. 2011. Extractive multi-document summaries should explicitly not contain document-specific content. In *Proceedings of the Workshop on Automatic Summarization for Different Genres, Media, and Languages*, WASDGML '11, pages 49–54, Stroudsburg, PA, USA. Association for Computational Linguistics.

John William Paisley, Chong Wang, David M. Blei, and Michael I. Jordan. 2015. Nested hierarchical dirichlet processes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(2):256–270.

Martin Ponweiser, Bettina Grün, and Kurt Hornik. 2015. Finding scientific topics revisited. In Maurizio Carpita, Eugenio Brentari, and El Mostafa Qannari, editors, *Advances in Latent Variables*, Studies in Theoretical and Applied Statistics, pages 93–100. Springer International Publishing.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, pages 399–408, New York, NY, USA. ACM.

U.S. Department of Commerce: National Institute of Standards and Technology. 2010. Text analysis conference 2010 – summarization track.

U.S. Department of Commerce: National Institute of Standards and Technology. 2011. Text analysis conference 2011 – summarization track.

Supplemental. 2017a. Hierarchicaltopicagreementxtra.java, hierarchicalmodelstorextra.java. Supplemental material for EMNLP Summarization workshop 2017 - java snippets on topic model alignment. Request from author by email.

Supplemental. 2017b. Topicmodeltheoryxtra.pdf. Supplemental material for EMNLP Summarization workshop 2017 - Topic model theory. Request from author by email.

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. 2005. Sharing clusters among related groups: Hierarchical Dirichlet processes. In *Advances in Neural Information Processing Systems*, volume 17.

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2006. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Yee Whye Teh, Kenichi Kurihara, and Max Welling. 2007. Collapsed variational inference for hdp. In *NIPS*, pages 1481–1488. Curran Associates, Inc.

U.S. Department of Commerce: National Institute of Standards and Technology. 2015. Document understanding conferences.

Alta de Wall and Etienne Barnard. 2008. Evaluating topic models with stability. In *19th Annual Symposium of the Pattern Recognition Association of South Africa*. Pattern Recognition Association of South Africa.

Yi Yang, Shimei Pan, Jie Lu, Mercan Topkara, and Yangqiu Song. 2016. The stability and usability of statistical topic models. *ACM Trans. Interact. Intell. Syst.*, 6(2):14:1–14:23.