Discovery of Topically Coherent Sentences for Extractive Summarization

Asli Celikyilmaz
Microsoft Speech Labs
Mountain View, CA, 94041
asli@ieee.org

Dilek Hakkani-Tür
Microsoft Speech Labs | Microsoft Research
Mountain View, CA, 94041
dilek@ieee.org

Abstract

Extractive methods for multi-document summarization are mainly governed by information overlap, coherence, and content constraints. We present an unsupervised probabilistic approach to model the hidden abstract concepts across documents as well as the correlation between these concepts, to generate topically coherent and non-redundant summaries. Based on human evaluations our models generate summaries with higher linguistic quality in terms of coherence, readability, and redundancy compared to benchmark systems. Although our system is unsupervised and optimized for topical coherence, we achieve a 44.1 ROUGE on the DUC-07 test set, roughly in the range of state-of-the-art supervised models.

1 Introduction

A query-focused multi-document summarization model produces a short-summary text of a set of documents, which are retrieved based on a user's query. An ideal generated summary text should contain the shared relevant content among set of documents *only once*, plus other unique information from individual documents that are directly related to the user's query addressing different levels of *detail*. Recent approaches to the summarization task has somewhat focused on the *redundancy* and *coherence* issues. In this paper, we introduce a series of new generative models for multiple-documents, based on a discovery of hierarchical topics and their correlations to extract topically coherent sentences.

Prior research has demonstrated the usefulness of sentence extraction for generating summary text taking advantage of surface level features such as word repetition, position in text, cue phrases, etc, (Radev, 2004; Nenkova and Vanderwende, 2005a; Wan and Yang, 2006; Nenkova et al., 2006). Because documents have pre-defined structures (e.g., sections, paragraphs, sentences) for different levels of concepts in a hierarchy, most recent summarization work has focused on structured probabilistic models to represent the corpus concepts (Barzilay et al., 1999; Daumé-III and Marcu, 2006; Eisenstein and Barzilay, 2008; Tang et al., 2009; Chen et al., 2000; Wang et al., 2009). In particular (Haghighi and Vanderwende, 2009; Celikyilmaz and Hakkani-Tur, 2010) build hierarchical topic models to identify salient sentences that contain abstract concepts rather than specific concepts. Nonetheless, all these systems crucially rely on extracting various levels of generality from documents, focusing little on redundancy and coherence issues in model building. A model than can focus on both issues is deemed to be more beneficial for a summarization task.

Topical coherence in text involves identifying key concepts, the relationships between these concepts, and linking these relationships into a hierarchy. In this paper, we present a novel, fully generative Bayesian model of document corpus, which can discover topically coherent sentences that contain key shared information with as little detail and redundancy as possible. Our model can discover hierarchical latent structure of multi-documents, in which some words are governed by low-level topics (T) and others by high-level topics (H). The main contributions of this work are:

− construction of a novel bayesian framework to

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 491–499,
Portland, Oregon, June 19-24, 2011. ©2011 Association for Computational Linguistics

capture higher level topics (concepts) related to summary text discussed in §3,
– representation of a linguistic system as a sequence of increasingly enriched models, which use posterior topic correlation probabilities in sentences to design a novel sentence ranking method in §4 and 5,
– application of the new hierarchical learning method for generation of less redundant summaries discussed in §6. Our models achieve comparable qualitative results on summarization of multiple newswire documents. Human evaluations of generated summaries confirm that our model can generate non-redundant and topically coherent summaries.

2 Multi-Document Summarization Models

Prior research has demonstrated the usefulness of sentence extraction for summarization based on lexical, semantic, and discourse constraints. Such models often rely on different approaches including: identifying important keywords (Nenkova et al., 2006); topic signatures based on user queries (Lin and Hovy, 2002; Conroy et al., 2006; Harabagiu et al., 2007); high frequency content word feature based learning (Nenkova and Vanderwende, 2005a; Nenkova and Vanderwende, 2005b), to name a few.

Recent research focusing on the extraction of latent concepts from document clusters are close in spirit to our work (Barzilay and Lee, 2004; Daumé-III and Marcu, 2006; Eisenstein and Barzilay, 2008; Tang et al., 2009; Wang et al., 2009). Some of these work (Haghighi and Vanderwende, 2009; Celikyilmaz and Hakkani-Tur, 2010) focus on the discovery of hierarchical concepts from documents (from abstract to specific) using extensions of hierarchal topic models (Blei et al., 2004) and reflect this hierarchy on the sentences. Hierarchical concept learning models help to discover, for instance, that "baseball" and "football" are both contained in a general class "sports", so that the summaries reference terms related to more abstract concepts like "sports".

Although successful, the issue with concept learning methods for summarization is that the extracted sentences usually contain correlated concepts. We need a model that can identify salient sentences referring to general concepts of documents and there should be minimum correlation between them.

Our approach differs from the early work, in that, we utilize the advantages of previous topic models and build an unsupervised generative model that can associate each word in each document with three random variables: a sentence $\boldsymbol{S}$, a higher-level topic $\boldsymbol{H}$, and a lower-level topic $\boldsymbol{T}$, in an analogical way to PAM models (Li and McCallum, 2006), i.e., a directed acyclic graph (DAG) representing mixtures of hierarchical structure, where super-topics are multinomials over sub-topics at lower levels in the DAG. We define a tiered-topic clustering in which the upper nodes in the DAG are *higher-level topics* $\boldsymbol{H}$, representing common co-occurence patterns (correlations) between lower-level topics $\boldsymbol{T}$ in documents. This has not been the focus in prior work on generative approaches for summarization task. Mainly, our model can discover correlated topics to eliminate redundant sentences in summary text.

Rather than representing sentences as a layer in hierarchical models, e.g., (Haghighi and Vanderwende, 2009; Celikyilmaz and Hakkani-Tur, 2010), we model sentences as *meta-variables*. This is similar to author-topic models (Rosen-Zvi et al., 2004), in which words are generated by first selecting an author uniformly from an observed author list and then selecting a topic from a distribution over topics that is specific to that author. In our model, words are generated from different topics of documents by first selecting a sentence containing the word and then topics that are specific to that sentence. This way we can directly extract from documents the summary related sentences that contain high-level topics. In addition in (Celikyilmaz and Hakkani-Tur, 2010), the sentences can only share topics if the sentences are represented on the same path of captured topic hierarchy, restricting topic sharing across sentences on different paths. Our DAG identifies tiered topics distributed over document clusters that can be shared by each sentence.

3 Topic Coherence for Summarization

In this section we discuss the main contribution, our two hierarchical mixture models, which improve summary generation performance through the use of tiered topic models. Our models can identify lower-level topics T (concepts) defined as distributions over words or higher-level topics H, which represent correlations between these lower level topics given

sentences. We present our synthetic experiment for model development to evaluate extracted summaries on redundancy measure. In §6, we demonstrate the performance of our models on coherence and informativeness of generated summaries by qualitative and intrinsic evaluations.

For model development we use the DUC 2005 dataset[1], which consists of 45 document clusters, each of which include 1-4 set of human generated summaries (10-15 sentences each). Each document cluster consists $\sim$ 25 documents (25-30 sentences/document) retrieved based on a user query. We consider each document cluster as a corpus and build 45 separate models.

For the synthetic experiments, we include the provided human generated summaries of each corpus as additional documents. The sentences in human summaries include general concepts mentioned in the corpus, the salient sentences of documents. Contrary to usual qualitative evaluations of summarization tasks, our aim during development is to measure the percentage of sentences in a human summary that our model can identify as salient among all other document cluster sentences. Because human produced summaries generally contain non-redundant sentences, we use total number of top-ranked human summary sentences as a qualitative redundancy measure in our synthetic experiments.

In each model, a document d is a vector of N_d words $\mathbf{w}_d$, where each w_{id} is chosen from a vocabulary of size V, and a vector of sentences $\mathbf{S}$, representing all sentences in a corpus of size S_D. We identify sentences as meta-variables of document clusters, which the generative process models both sentences and documents using *tiered* topics. A sentence's relatedness to summary text is tied to the document cluster's user query. The idea is that a lexical word present or related to a query should increase its sentence's probability of relatedness.

4 Two-Tiered Topic Model - TTM

Our base model, the two-tiered topic model (TTM), is inspired by the hierarchical topic model, PAM, proposed by Li and McCallum (2006). PAM structures documents to represent and learn arbitrary, nested, and possibly sparse topic correlations using a directed acyclic graph. Our goals are not so different: we aim to discover concepts from documents that would attribute for the general topics related to a user query, however, we want to relate this information to sentences. We represent sentences $\boldsymbol{S}$ by discovery of general (more general) to specific topics (Fig.1). Similarly, we represent summary unrelated (document specific) sentences as corpus specific distributions θ over background words $\mathbf{w_B}$, (functional words like prepositions, etc.).

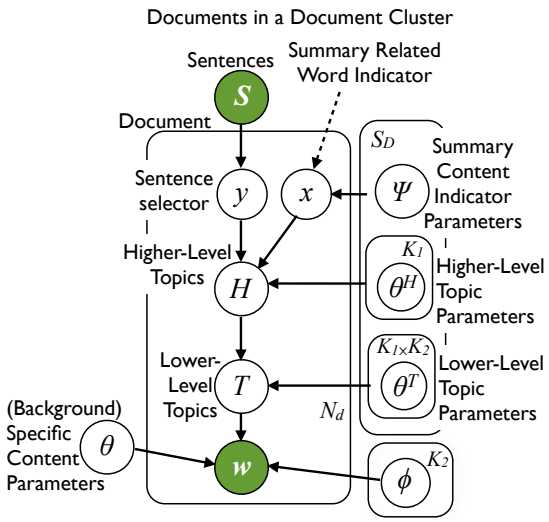

Figure 1: Graphical model depiction of two-tiered topic model (TTM) described in section §4. $\mathbf{S}$ are sentences $s_{i=1..S_D}$ in document clusters. The high-level topics ($H_{k_1=1...K_1}$), representing topic correlations, are modeled as distributions over low-level-topics ($T_{k_2=1...K_2}$). Shaded nodes indicate observed variables. Hyper-parameters for ϕ, θ^H, θ^T, θ are omitted.

Our two-tiered topic model for salient sentence discovery can be generated for each word in the document (Algorithm 1) as follows: For a word w_{id} in document d, a random variable x_{id} is drawn, which determines if w_{id} is query related, i.e., w_{id} either exists in the query or is related to the query[2]. Otherwise, w_{id} is unrelated to the user query. Then sentence s_i is chosen uniformly at random ($y_{s_i} \sim Uniform(s_i)$) from sentences in the document containing w_{id} (deterministic if there is only one sentence containing w_{id}). We assume that if a word is related to a query, it is likely to be summary-related

[1] www-nlpir.nist.gov/projects/duc/data.html

[2] We measure relatedness to a query if a word exists in the query or it is synonymous based on information extracted from WordNet (Miller, 1995).

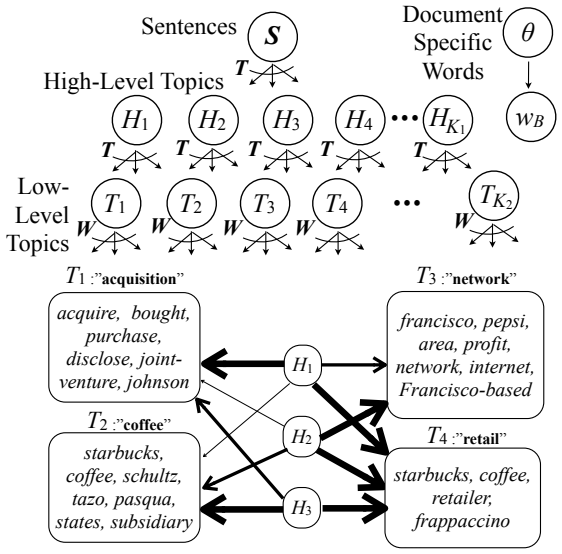

Figure 2: Depiction of TTM given the query *"D0718D: Starbucks Coffee :* ***How has Starbucks Coffee attempted to expand and diversify through joint ventures, acquisitions, or subsidiaries?"***. If a word is query/summary related sentence S, first a sentence then a high-level (H) and a low-level (T) topic is sampled. ($\overset{C}{\curlywedge}$ represents that a random variable is a parent of all C random variables.) The bolded links from $H-T$ represent correlated low-level topics.

Algorithm 1 Two-Tiered Topic Model Generation

Sample: $s_i = 1..S_D$: $\Psi \sim Beta(\eta)$,
$k_1 = 1...K_1$: $\theta^H \sim Dirichlet(\alpha^H)$,
$k_2 = 1...K_1 \times K_2$: $\theta^T \sim Dirichlet(\alpha^T)$,
and $k = 1..K_2$: $\phi \sim Dirichlet(\beta)$.
for documents $d \leftarrow 1, ..., D$ **do**
  for words w_{id}, $i \leftarrow 1, ..., N_d$ **do**
    - Draw a discrete $x \sim Binomial(\Psi_{w_{id}})^{\star}$
    - If $x = 1$, w_{id} is summary related;
      · conditioned on S draw a sentence
        $y_{s_i} \sim Uniform(s_i)$ containing w_i,
      · sample a high-level topic $H_{k_1} \sim \theta^H_{k_1}(\alpha^H)$,
        and a low-level topic $T_{k_2} \sim \theta^T_{k_2}(\alpha^T)$,
      · sample a word $w_{ik_1k_2} \sim \phi_{H_{k_1}T_{k_2}}(\alpha)$,
    - If $x = 0$, the word is unrelated **
      sample a word $w_B \sim \theta(\alpha)$,
      corpus specific distribution.
  end for
end for

$\star$ if w_{id} exists or related to the the query then $x = 1$ deterministic, otherwise it is stochastically assigned $x \sim Bin(\Psi)$.
$\star\star$ w_{id} is a background word.

(so as the sampled sentence s_i). We keep track of the frequency of s_i's in a vector, $DS \in Z^{S_D}$. Every time an s_i is sampled for a query related w_{id}, we increment its count, a degree of sentence saliency.

Given that w_{id} is related to a query, it is associated with two-tiered multinomial distributions: high-level *H* topics and low-level *T* topics. A high-level topic H_{k_i} is chosen first from a distribution over low-level topics *T* specific to that s_i and one low-level topic T_{k_j} is chosen from a distribution over words, and w_{id} is generated from the sampled low-level topic. If w_{id} is *not* query-related, it is generated as a background word w_B.

The resulting tiered model is shown as a graph and plate diagrams in Fig.1 & 2. A sentence sampled from a query related word is associated with a distribution over K_1 number of high-level topics H_{k_i}, each of which are also associated with K_2 number of low-level topics T_{k_j}, a multinomial over lexical words of a corpus. In Fig.2 the most confident words of four low-level topics is shown. The bolded links between H_{k_i} and T_{k_j} represent the strength of correlation between T_{k_j}'s, e.g., the topic *"acquisition"* is found to be more correlated with *"retail"* than the "network" topic given H_1. This information is used to rank sentences based on the correlated topics.

4.1 Learning and Inference for TTM

Our learning procedure involves finding parameters, which likely integrates out model's posterior distribution $P(\mathbf{H}, \mathbf{T}|\mathbf{W_d}, \mathbf{S})$, $d \in D$. EM algorithms might face problems with local maxima in topic models (Blei et al., 2003) suggesting implementation of approximate methods in which some of the parameters, e.g., θ^H, θ^T, ψ, and θ, can be integrated out, resulting in standard Dirichlet-multinomial as well as binomial distributions. We use Gibbs sampling which allows a combination of estimates from several local maxima of the posterior distribution.

For each word, x_{id} is sampled from a sentence specific binomial ψ which in turn has a smoothing prior η to determine if the sampled word w_{id} is (query) summary-related or document-specific. Depending on x_{id}, we either sample a sentence along with a high/low-level topic pair or just sample background words w_B. The probability distribution over sentence assignments, $P(y_{s_i} = s|\mathbf{S})$ $s_i \in \mathbf{S}$, is assumed to be uniform over the elements of **S**, and deterministic if there is only one sentence in the docu-

ment containing the corresponding word. The optimum hyper-parameters are set based on the training dataset model performance via cross-validation [3].

For each word we sample a high-level H_{k_i} and a low-level T_{k_j} topic if the word is query related ($x_{id} = 1$). The sampling distribution for TTM for a word given the remaining topics and hyper-parameters $\alpha^H, \alpha^T, \alpha, \beta, \eta$ is:

$$p_{\texttt{TTM}}(H_{k_1}, T_{k_2}, x = 1 | \mathbf{w}, \mathbf{H}_{-k_1}, \mathbf{T}_{-k_2}) \propto \frac{\alpha^H + n_d^{k_1}}{\sum_{H\prime} \alpha^{H\prime} + n_d} * \frac{\alpha^T + n_d^{k_1 k_2}}{\sum_{T\prime} \alpha^{T\prime} + n_d{}^H} * \frac{\eta + n_x^{k_1 k_2}}{2\eta + n_{k_1 k_2}} * \frac{\beta_w + n^w_{k_1 k_2 x}}{\sum_{w\prime} \beta_{w\prime} + n_{k_1 k_2 x}}$$

and when $x = 0$ (a corpus specific word),

$$p_{\texttt{TTM}}(x = 0 | \mathbf{w}, \mathbf{z}_{H-k}, \mathbf{z}_{t-k}) \propto \frac{\eta + n^x_{k_1 k_2}}{2\eta + n_{k_1 k_2}} * \frac{\alpha_w + n^w}{\sum_{w\prime} \alpha_{w\prime} + n}$$

The $n_d^{k_1}$ is the number of occurrences of high-level topic k_1 in document d, and $n_d^{k_1 k_2}$ is the number of times the low-level topic k_2 is sampled together with high-level topic k_1 in d, $n^w_{k_1 k_2 x}$ is the number of occurrences of word w sampled from path H-T given that the word is query related. Note that the number of tiered topics in the model is fixed to K_1 and K_2, which is optimized with validation experiments. It is also possible to construct extended models of TTM using non-parametric priors, e.g., hierarchal Dirichlet processes (Li et al., 2007) (left for future work).

4.2 Summary Generation with TTM

We can observe the frequency of draws of every sentence in a document cluster **S**, given it's words are related, through $DS \in \mathbb{Z}^{S_D}$. We obtain DS during Gibbs sampling (in §4.1), which indicates a saliency score of each sentence $s_j \in \mathbf{S}$, $j = 1..S_D$:

$$score^{\texttt{TTM}}(s_j) \propto \#\left[w_{id} \in s_j, x_{id} = 1\right] / nw_j \quad (1)$$

where w_{id} indicates a word in a document d that exists in s_j and is sampled as summary related based on random indicator variable x_{id}. nw_j is the number of words in s_j and normalizes the score favoring sentences with many related words. We rank sentences based on (1). We compare TTM results on synthetic experiments against PAM (Li and McCallum, 2006) a similar topic model that clusters topics in a hierarchical structure, where super-topics are distributions over sub-topics. We obtain sentence scores for PAM models by calculating the sub-topic significance (TS) based on super-topic correlations, and discover topic correlations over the entire document space (corpus wide). Hence; we calculate the TS of a given sub-topic, $k = 1, .., K_2$ by:

$$TS(z_k) = \frac{1}{D} \sum_{d \in D} \frac{1}{K_1} \sum_{k_1}^{K_1} p(z^k_{sub} | z^{k_1}_{sup}) \quad (2)$$

where z^k_{sub} is a sub-topic $k = 1..K_2$ and $z^{k_1}_{sup}$ is a super-topic k_1. The conditional probability of a sub-topic k given a super-topic k_1, $p(z^k_{sub}|z^{k_1}_{sup})$, explains the variation of that sub-topic in relation to other sub-topics. The higher the variation over the entire corpus, the better it represents the general theme of the documents. So, sentences including such topics will have higher saliency scores, which we quantify by imposing topic's significance on vocabulary:

$$score^{\texttt{PAM}}(s_i) = \frac{1}{K_2} \sum_{k}^{K_2} \prod_{w \in s_i} p(w | z^k_{sub}) * TS(z_k) \quad (3)$$

Fig. 4 illustrates the average salience sentence selection performance of TTM and PAM models (for 45 models). The x-axis represents the percentage of sentences selected by the model among all sentences in the DUC2005 corpus. 100% means all sentences in the corpus included in the summary text. The y-axis is the % of selected human sentences over all sentences. The higher the human summary sentences are ranked, the better the model is in selecting the salient sentences. Hence, the system which peaks sooner indicates a better model.

In Fig.4 TTM is significantly better in identifying human sentences as salient in comparison to PAM. The statistical significance is measured based on the area under the curve averaged over 45 models.

5 Enriched Two-Tiered Topic Model

Our model can discover words that are related to summary text using posteriors $\hat{P}(\theta^H)$ and $\hat{P}(\theta^T)$,

[3] An alternative way would be to use Dirichlet priors (Blei et al., 2003) which we opted for due to computational reasons but will be investigated as future research.

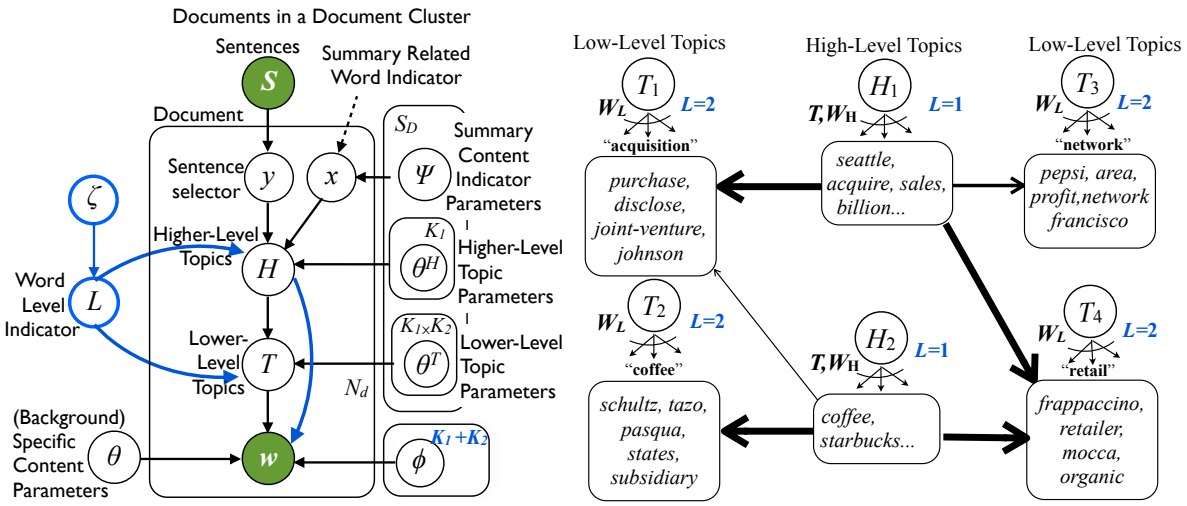

Figure 3: Graphical model depiction of sentence level enriched two-tiered model (ETTM) described in section §5. Each path defined by **H/T** pair $k_1 k_2$, has a multinomial ζ over which level of the path outputs a given word. **L** indicates which level, i.e, high or low, the word is sampled from. On the right is the high-level topic-word and low-level topic-word distributions characterized by ETTM. Each $\mathbf{H_{k_1}}$ also represented as distributions over general words $\mathbf{W_H}$ as well as indicates the degree of correlation between low-level topics denoted by boldness of the arrows.

as well as words w_B specific to documents (via $\hat{P}(\theta)$) (Fig.1). TTM can discover topic correlations, but cannot differentiate if a word in a sentence is more general or specific given a query. Sentences with general words would be more suitable to include in summary text compared to sentences containing specific words. For instance for a given sentence: *"Starbucks Coffee has attempted to expand and diversify through joint ventures, and acquisitions."*, *"starbucks"* and *"coffee"* are more general words given the document clusters compared to *"joint"* and *"ventures"* (see Fig.2), because they appear more frequently in document clusters. However, TTM has no way of knowing that *"starbucks"* and *"coffee"* are common terms given the context. We would like to associate general words with high-level topics, and context specific words with low-level topics. Sentence containing words that are sampled from high-level topics would be a better candidate for summary text. Thus; we present enriched TTM (ETTM) generative process (Fig.3), which samples words not only from low-level topics but also from high-level topics as well.

ETTM discovers three separate distributions over words: (*i*) high-level topics *H* as distributions over corpus general words $\mathbf{W_H}$, (*ii*) low-level topics *T* as distributions over corpus specific words $\mathbf{W_L}$, and

Level Generation for Enriched TTM

Fetch $\zeta_k \sim Beta(\gamma)$; $k = 1...K_1 \times K_2$.
For w_{id}, $i = 1, ..., N_d$, $d = 1, ..D$:
  If $x = 1$, sentence s_i is summary related;
    - sample H_{k_1} and T_{k_2}
    - sample a level L from $Bin(\zeta_{k_1 k_2})$
    - If $L = 1$ (general word); $w_{id} \sim \phi_{H_{k_i}}$
    - else if $L = 2$ (context specific); $w_{id} \sim \phi_{H_{k_1} T_{k_2}}$
  else if $x = 0$, do Step 14-16 in Alg. 1.

(*iii*) background word distributions, i.e,. document specific $\mathbf{W_B}$ (less confidence for summary text). Similar to TTM's generative process, if w_{id} is related to a given query, then $x = 1$ is deterministic, otherwise $x \in \{0, 1\}$ is stochastically determined if w_{id} should be sampled as a background word (w_B) or through hierarchical path, i.e., *H-T* pairs. We first sample a sentence s_i for w_{id} uniformly at random from the sentences containing the word $y_{s_i} \sim Uniform(s_i)$). At this stage we sample a *level* $\mathbf{L}_{w_{id}} \in \{1, 2\}$ for w_{id} to determine if it is a high-level word, e.g., more general to context like *"starbucks"* or *"coffee"* or more specific to related context such as *"subsidiary"*, *"frappucino"*. Each path through the DAG, defined by a H-T pair (total of $K_1 K_2$ pairs), has a binomial $\zeta_{K_1 K_2}$ over which

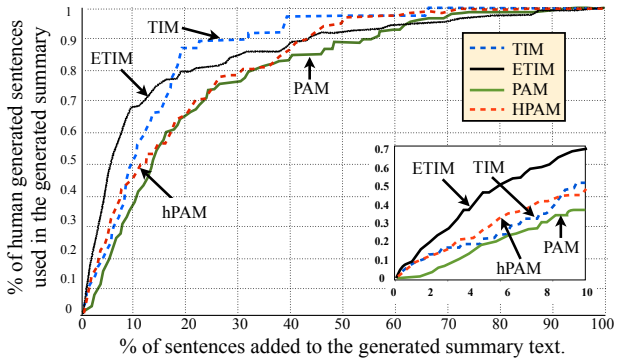

Figure 4: Average saliency performance of four systems over 45 different DUC models. The area under each curve is shown in legend. Inseam is the magnified view of top-ranked 10% of sentences in corpus.

level of the path outputs sampled word. If the word is a specific type, $x = 0$, then it is sampled from the background word distribution θ, a document specific multinomial. Once the level and conditional path is drawn (see level generation for ETTM above) the rest of the generative model is same as TTM.

5.1 Learning and Inference for ETTM

For each word, x is sampled from a sentence specific binomial ψ, just like TTM. If the word is related to the query $x = 1$, we sample a high and low-level topic pair $H - T$ as well as an additional level $\boldsymbol{L}$ is sampled to determine which level of topics the word should be sampled from. L is a corpus specific binomial one for all $H - T$ pairs. If $L = 1$, the word is one of corpus general words and sampled from the high-level topic, otherwise ($L = 2$) the word is corpus specific and sampled from a the low-level topic. The optimum hyper-parameters are set based on training performance via cross validation.

The conditional probabilities are similar to TTM, but with additional random variables, which determine the level of generality of words as follows:

$$p_{\text{ETTM}}(T_{k_1}, T_{k_2}, L|\mathbf{w}, \mathbf{T}_{-k_1}, \mathbf{T}_{-k_2}, L_{)} \propto$$
$$p_{\text{TTM}}(T_{k_1}, T_{k_2}, x = 1|.) * \frac{\gamma + N^L_{k_1 k_2}}{2\gamma + n_{k_1 k_2}}$$

5.2 Summary Generation with ETTM

For ETTM models, we extend the TTM sentence score to be able to include the effect of the general words in sentences (as word sequences in language models) using probabilities of K_1 high-level topic distributions, $\phi^w_{H_{k=1..K_1}}$, as:

$$score^{\text{ETTM}}(s_i) \propto \#\left[w_{id} \in s_j, x_{id} = 1\right] / nw_j * \frac{1}{K_1} \sum_{k=1..K_1} \prod_{w \in s_i} p(w|T_k)$$

where $p(w|T_k)$ is the probability of a word in s_i being generated from high-level topic H^k. Using this score, we re-rank the sentences in documents of the synthetic experiment. We compare the results of ETTM to a structurally similar probabilistic model, entitled hierarchical PAM (Mimno et al., 2007), which is designed to capture topics on a hierarchy of two layers, i.e., super topics and subtopics, where super-topics are distributions over abstract words. In Fig. 4 out of 45 models ETTM has the best performance in ranking the human generated sentences at the top, better than the TTM model. Thus; ETTM is capable of capturing focused sentences with general words related to the main concepts of the documents and much less redundant sentences containing concepts specific to user query.

6 Final Experiments

In this section, we qualitatively compare our models against state-of-the art models and later apply an intrinsic evaluation of generated summaries on topical coherence and informativeness.

For a qualitative comparison with the previous state-of-the models, we use the standard summarization datasets on this task. We train our models on the datasets provided by DUC2005 task and validate the results on DUC 2006 task, which consist of a total of 100 document clusters. We evaluate the performance of our models on DUC2007 datasets, which comprise of 45 document clusters, each containing 25 news articles. The task is to create max. 250 word long summary for each document cluster.

6.1. ROUGE Evaluations: We train each document cluster as a separate corpus to find the optimum parameters of each model and evaluate on test document clusters. ROUGE is a commonly used measure, a standard DUC evaluation metric, which computes recall over various n-grams statistics from a model generated summary against a set of human generated summaries. We report results in R-1 (recall against unigrams), R-2 (recall against bigrams), and R-SU4

ROUGE	w/o stop words			w/ stop words		
	R-1	R-2	R-4	R-1	R-2	R-4
PYTHY	**35.7**	8.9	**12.1**	42.6	**11.9**	**16.8**
HIERSUM	33.8	**9.3**	11.6	42.4	11.8	16.7
HybHSum	35.1	8.3	11.8	**45.6**	11.4	**17.2**
PAM	32.1	7.1	11.0	41.7	9.1	15.3
hPAM	31.9	7.0	11.1	41.2	8.9	15.2
TTM*	**34.0**	8.7	11.5	**44.7**	10.7	16.5
ETTM*	32.4	8.3	11.2	**44.1**	10.4	16.4

Table 1: ROUGE results of the best systems on DUC2007 dataset (best results are **bolded**.) * indicate our models.

(recall against skip-4 bigrams) ROUGE scores w/ and w/o stop words included.

For our models, we ran Gibbs samplers for 2000 iterations for each configuration throwing out first 500 samples as burn-in. We iterated different values for hyperparameters and measured the performance on validation dataset to capture the optimum values.

The following models are used as benchmark: (***i***) PYTHY (Toutanova et al., 2007): Utilizes human generated summaries to train a sentence ranking system using a classifier model; (***ii***) HIERSUM (Haghighi and Vanderwende, 2009): Based on hierarchical topic models. Using an approximation for inference, sentences are greedily added to a summary so long as they decrease KL-divergence of the generated summary concept distributions from document word-frequency distributions. (***iii***) HybHSum (Celikyilmaz and Hakkani-Tur, 2010): A semi-supervised model, which builds a hierarchial LDA to probabilistically score sentences in training dataset as summary or non-summary sentences. Using these probabilities as output variables, it learns a discriminative classifier model to infer the scores of new sentences in testing dataset. (***iv***) PAM (Li and McCallum, 2006) and hPAM (Mimno et al., 2007): Two hierarchical topic models to discover high and low-level concepts from documents, baselines for synthetic experiments in §4 & §5.

Results of our experiments are illustrated in Table 6. Our unsupervised TTM and ETTM systems yield a 44.1 R-1 (w/ stop-words) outperforming the rest of the models, except HybHSum. Because HybHSum uses the human generated summaries as supervision during model development and our systems do not, our performance is quite promising considering the generation is completely unsupervised without seeing any human generated summaries during training. However, the R-2 evaluation (as well as R-4) w/ stop-words does not outperform other models. This is because R-2 is a measure of bi-gram recall and neither of our models represent bi-grams whereas, for instance, PHTHY includes several bi-gram and higher order n-gram statistics. For topic models bi-grams tend to degenerate due to generating inconsistent bag of bi-grams (Wallach, 2006).

6.2. Manual Evaluations: A common DUC task is to manually evaluate models on the quality of generated summaries. We compare our best model ETTM to the results of PAM, our benchmark model in synthetic experiments, as well as hybrid hierarchical summarization model, hLDA (Celikyilmaz and Hakkani-Tur, 2010). Human annotators are given two sets of summary text for each document set, generated from either one of the two approaches: best ETTM and PAM or best ETTM and HybHSum models. The annotators are asked to mark the better summary according to five criteria: *non-redundancy* (which summary is less redundant), *coherence* (which summary is more coherent), *focus and readability* (content and no unnecessary details), *responsiveness* and *overall* performance.

We asked 3 annotators to rate DUC2007 predicted summaries (45 summary pairs per annotator). A total of 42 pairs are judged for ETTM vs. PAM models and 49 pairs for ETTM vs. HybHSum models. The evaluation results in frequencies are shown in Table 6. The participants rated ETTM generated summaries more coherent and focused compared to PAM, where the results are statistically significant (based on t-test on 95% confidence level) indicating that ETTM summaries are rated significantly better. The results of ETTM are slightly better than HybHSum. We consider our results promising because, being unsupervised, ETTM does not utilize human summaries for model development.

7 Conclusion

We introduce two new models for extracting topically coherent sentences from documents, an important property in extractive multi-document summarization systems. Our models combine approaches from the hierarchical topic models. We empha-

	PAM	ETTM	Tie	HybHSum	ETTM	Tie
Non-Redundancy	13	**26**	3	12	**18**	19
Coherence	13	**26**	3	15	**18**	16
Focus	14	**24**	4	14	**17**	18
Responsiveness	15	**24**	3	**19**	12	18
Overall	15	**25**	2	17	**22**	10

Table 2: Frequency results of manual evaluations. Tie indicates evaluations where two summaries are rated equal.

size capturing correlated semantic concepts in documents as well as characterizing general and specific words, in order to identify topically coherent sentences in documents. We showed empirically that a fully unsupervised model for extracting general sentences performs well at summarization task using datasets that were originally used in building automatic summarization system challenges. The success of our model can be traced to its capability of directly capturing coherent topics in documents, which makes it able to identify salient sentences.

Acknowledgments

The authors would like to thank Dr. Zhaleh Feizollahi for her useful comments and suggestions.

References

R. Barzilay and L. Lee. 2004. Catching the drift: Probabilistic content models with applications to generation and summarization. *In Proc. HLT-NAACL'04*.

R. Barzilay, K.R. McKeown, and M. Elhadad. 1999. Information fusion in the context of multi-document summarization. *Proc. 37th ACL*, pages 550–557.

D. Blei, A. Ng, and M. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*.

D. Blei, T. Griffiths, M. Jordan, and J. Tenenbaum. 2004. Hierarchical topic models and the nested chinese restaurant process. *In Neural Information Processing Systems [NIPS]*.

A. Celikyilmaz and D. Hakkani-Tur. 2010. A hybrid hierarchical model for multi-document summarization. *Proc. 48th ACL 2010*.

D. Chen, J. Tang, L. Yao, J. Li, and L. Zhou. 2000. Query-focused summarization by combining topic model and affinity propagation. *LNCS– Advances in Data and Web Development*.

J. Conroy, H. Schlesinger, and D. OLeary. 2006. Topic-focused multi-document summarization using an approximate oracle score. *Proc. ACL*.

H. Daumé-III and D. Marcu. 2006. Bayesian query focused summarization. *Proc. ACL-06*.

J. Eisenstein and R. Barzilay. 2008. Bayesian unsupervised topic segmentation. *Proc. EMNLP-SIGDAT*.

A. Haghighi and L. Vanderwende. 2009. Exploring content models for multi-document summarization. *NAACL HLT-09*.

S. Harabagiu, A. Hickl, and F. Lacatusu. 2007. Satisfying information needs with multi-document summaries. *Information Processing and Management*.

W. Li and A. McCallum. 2006. Pachinko allocation: Dag-structure mixture models of topic correlations. *Proc. ICML*.

W. Li, D. Blei, and A. McCallum. 2007. Nonparametric bayes pachinko allocation. *The 23rd Conference on Uncertainty in Artificial Intelligence*.

C.Y. Lin and E. Hovy. 2002. The automated acquisition of topic signatures fro text summarization. *Proc. CoLing*.

G. A. Miller. 1995. Wordnet: A lexical database for english. *ACM, Vol. 38, No. 11: 39-41*.

D. Mimno, W. Li, and A. McCallum. 2007. Mixtures of hierarchical topics with pachinko allocation. *Proc. ICML*.

A. Nenkova and L. Vanderwende. 2005a. Document summarization using conditional random fields. *Technical report, Microsoft Research*.

A. Nenkova and L. Vanderwende. 2005b. The impact of frequency on summarization. *Technical report, Microsoft Research*.

A. Nenkova, L. Vanderwende, and K. McKowen. 2006. A composition context sensitive multi-document summarizer. *Prof. SIGIR*.

D. R. Radev. 2004. Lexrank: graph-based centrality as salience in text summarization. *Jrnl. Artificial Intelligence Research*.

M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. 2004. The author-topic model for authors and documents. *UAI*.

J. Tang, L. Yao, and D. Chens. 2009. Multi-topic based query-oriented summarization. *SIAM International Conference Data Mining*.

K. Toutanova, C. Brockett, M. Gamon, J. Jagarlamudi, H. Suzuki, and L. Vanderwende. 2007. The phthy summarization system: Microsoft research at duc 2007. In *Proc. DUC*.

H. Wallach. 2006. Topic modeling: Beyond bag-of-words. *Proc. ICML 2006*.

X. Wan and J. Yang. 2006. Improved affinity graph based multi-document summarization. *HLT-NAACL*.

D. Wang, S. Zhu, T. Li, and Y. Gong. 2009. Multi-document summarization using sentence-based topic models. *Proc. ACL 2009*.