Utterance Topic Model for Generating Coherent Summaries

Pradipto Das
University at Buffalo
201 Bell Hall
Buffalo, NY 14228
pdas3@buffalo.edu

Rohini Srihari
University at Buffalo
201 Bell Hall
Buffalo, NY 14228
rohini@cedar.buffalo.edu

ABSTRACT

Generating short multi-document summaries has received a lot of focus recently and is useful in many respects including summarizing answers to a question in an online scenario like Yahoo! Answers. The focus of this paper is to attempt to define a new probabilistic topic model that includes the semantic roles of the words in the document generation process. Words always carry syntactic and semantic information and often such information, for e.g., the grammatical and semantic role (henceforth GSR) of a word like Subject, Verb, Object, Adjective qualifiers, WordNet and VerbNet role assignments etc. is carried across adjacent sentences to enhance local coherence in different parts of a document. A statistical topic model like LDA[5] usually models topics as distributions over the word count vocabulary only. We posit that a document could first be topic modeled over a vocabulary of GSR transitions and then corresponding to each transition, words and and hence sentences can be sampled to best describe the transition. Thus the topics in the proposed model also lend themselves to be distributions over the GSR transitions implicitly. We also later show how this basic model can be extended to a model for query focused summarization where for a particular query, sentences can be ranked by a product of thematical salience and coherence through GSR transitions. We empirically show that the new topic model had lower test set perplexity than LDA and we also analyze the performance of our summarization model using the ROUGE[13] on DUC2005 dataset[1] and PYRAMID[17] on the TAC2008[2] and TAC2009[3] datasets.

Categories and Subject Descriptors

I.5.1 [**Pattern Recognition**]: Models—*statistical*

General Terms

Topic Models, Centering theory, Coherence, Multi-document summarization

[1]http://www-nlpir.nist.gov/projects/duc/duc2005/
[2]http://www.nist.gov/tac/tracks/2008/index.html
[3]http://www.nist.gov/tac/2009/index.html

1. INTRODUCTION

Topic models like LDA[5] have become the cornerstone for understanding the thematic organization of large text corpora in an unsupervised fashion. These models define probabilistic generative process for document generation in a robust manner. The input to such a model are primarily word counts or more recently part-of-speech (POS) tags of words in sentences as popularized by Syntactic Topic Models (STM) [8] in documents. However, in our proposed "utterance topic model" (henceforth UTM), we incorporate the grammatical and semantic information of the words *across a context of utterances* in addition to the word counts.

There are two reasons why we had to come up with a new topic model based on contextual information around the terms - firstly, given a corpus without any thematic structure, the topic modeling framework is a necessary first choice to understanding such organization. Secondly, we wanted to understand the human process of compacting documents into summaries and encode such a process in a statistical framework. Since local coherence is a major factor in understanding various parts of a discourse, it was natural to encode coherence statistically in such a framework.

In the realm of computational linguistics, there has been quite some work in Centering Theory including Grosz et. al. [9]. Their work mainly specifies how discourse interpretation depends on interactions among speaker intentions, attentional state and linguistic form. As such, in our context, we could assume a subset of documents discussing a particular "theme" to be a discourse involving one or more participants. The discourse participants' focus of attention at any given point in time is modeled by their "attentional state". This "state of attention" comprises of a *focus* in the current utterance being understood. This focus within the attentional state helps identify "centers" of utterances that relate different parts of local discourse segments meaningfully and according to [9], the "centers" are semantic objects, not just words, phrases, or syntactic forms. Centering theory helps formalize the constraints on the centers to maximize coherence. In our context, the GSRs along with the explicit realization of the roles through the sentential words approximate the centers. As a simple example on attentional state and centering, consider the following:

- Discourse 1

1. Martha shot her husband Tom.
2. She was abused by Tom for many years.
3. Martha couldn't take it anymore.
4. She shot Tom as he was preparing to have supper.

- Discourse 2

1. Martha shot her husband Tom.
2. Tom had been abusing her for many years
3. Martha couldn't take it anymore.
4. Tom was shot as he was preparing to have supper.

Discourse 1 is an example where the focus of attention is clearly on Martha. If we observe discourse 2, there is a shift of attention from Martha to Tom and vice versa. For e.g. in the first utterance if a reader perceives the focus of attention to be Tom, there is a retention of the focus in the second utterance. If, however, in the first utterance the focus of attention be Martha, then there is a focus shift in the next utterance. In any case, the focus is Martha in the third utterance. Discourse 2 is thus less coherent than discourse 1 in terms of the effort to understand the discourse i.e. discourse 1 has less *inference load*.

In classicial Centering theory as in [9], the term "centers of an utterance" is used to refer to those entities serving to link that utterance to other utterances in the discourse segment that contains it. In the example above, in discourse 1, the pair (Subject, "Martha") approximates a center that is retained through the focus of attention in the utterances. Thus the propagation of these centers of utterances within discourse segments helps maintain the local coherence.

Each word in a sentence of a document has an associated role (syntactic or semantic) with it for e.g., a noun helps identify a concept (abstract or concrete) and thus serves as a part or whole of a center of utterance. If two consecutive sentences contain the same word, then there is a GSR transition (henceforth GSRt) within the context of sentences. We approximate the change in attentional state in local discourse segments through these transitions. If the word is not present in the preceding (or succeeding) sentence then there is still a transition from (to) a *null*, identified by "−−" GSR to (from) the current GSR. A GSRt is thus looked upon as a multinomial distribution over sentences in a document. Although entities (nominal subjects) are only advocated by centering theory, we also used verbs as GSRs of words to understanding the intents in attention. In our context, the GSRs and the explicit realization of these roles through the sentential words approximate the centers.

This paper is organized as follows. The next two section reviews some related work, followed by the mention of how Centering theory was adapted to develop our new topic model. In the next section we show how this model can be extended as a full summarization model. Then we describe the techniques of our proposed method and follow up with results and analysis of the output of our model. The paper is concluded in the last section with some ideas for future work.

2. RELATED WORK

Topic models have been widely applied to text despite a willful ignorance of the underlying linguistic structures that exist in natural language. There have been a lot of work on either applying topic models directly to a certain problem as in [1, 6] or adapting basic LDA style topic modeling as in [16, 15]. In a topic model, the words of each document are assumed to be exchangeable; their probability is invariant to permutation of the positions of the words in a document. A workaround to this inadequacy was posed and addressed in [8]. It is also important to note that although a topic model can suggest documents relevant to a query, finding particularly relevant phrases for question answering is still a challenging task. Our main focus had been to build a new topic model based on the LDA framework that could use linguistic features and semantic roles of words in a discourse segment for representing local coherence.

With regard to extractive multi document summarization, most earlier methods had focussed on clustering of sentence vectors or building graphs of sentences from the relevant documents and then using some graph mining algorithms like Pagerank to select out the most authoritative sentences (as in [19]). Other approaches include algorithmic formulation of summary extraction using greedy, dynamic and integer linear programming methodologies. The work in [14] compares these approaches and also proves that in general the inferring an extractive summary is NP-hard.

The work by Ye et. al.[20] calculates the semantic similarity among the sentences in the cluster, and between a sentence and the given cluster of documents. The semantic similarity between sentences is determined by the number of sentential concept overlaps calculated from the WordNet synset hierarchy including glosses, hypernyms and meronyms. Another intersting approach taken by [12] where the sentences where scored by a weighted combination of several features including pattern based features which provide clue as to how to interpret an information need. It was shown in [11], that using contextual language models and latent semantic indexing, the resulting summaries were indeed promising based on the results of the ROUGE evaluation tool. Their contextual language model essentially computed a language model within a window of words instead of an explicit n-gram. In yet another unique approach to summarization[10], the syntactic structure of the parse trees was utilized to generate valid triples of basic elements (BEs) or (head|modifier|relation) triples and then summary sentences were extracted using a score directly related to computing important BEs in them. The focus in [18] was more about finding hidden connections among query concepts using textual evidences through sematic cues rather than summarization. However, a final summarization was performed on the evidence trails and was therefore chosen as a system for comparison.

Some of the recent and notable Bayesian topic model approaches to summarization have been presented in [6] and [7]. In both the models the focus had been to model the both the query and the relevant documents together. It is to be noted here that in all these approaches, there has been hardly any focus on the actual process of coherence in a passage. Our extended summarization model is novel in this aspect and also dissociates itself from the queries explicitly.

It only assumes that we have a fixed index and attempts to build a multi document summary given *any query* and its associated relevant documents.

3. BACKGROUND FOR OUR METHOD

In this section we give a brief introduction to centering theory based local coherence and how centering could be statistically represented.

3.1 Centering Theory

As mentioned previously in section 1, the term *centers* of an utterance to refer to those entities serving to link that utterance to other utterances in the discourse segment that contains it. Each utterance, which we approximate by a sentence, S in a discourse segment (DS) is assigned a set of forward-looking centers, $Cf(S, DS)$ and each utterance other than the segment initial utterance is assigned *a single* backward-looking center, $Cb(S, DS)$. The backward-looking center of utterance S_{n+1} connects with one of the forward-looking centers of utterance S_n The Cfs consists of all the referents in the utterance S_n and are ordered according to salience: the subjects are preferred over objects and those over other GSRs. An illustration from [9] below elucidates coherence through such center linkages.

(a) John has been having a lot of trouble arranging his vacation

(b) He cannot find anyone to take over his responsibilities. (he = John) Cb = John; Cf = $\{John\}$

(c) He called up Mike yesterday to work out a plan. (he = John) Cb = John; Cf = $\{John, Mike\}$

For building a statistical topic model that incorportes GSR transitions (henceforth GSRts) across utterances, we attributed words in a sentence with GSRs like subjects, objects, concepts from WordNet synset role assignments(wn), adjectives, VerbNet thematic role assignment(vn), adverbs and "other" (if the feature of the word doesn't fall into the previous GSR categories). Further if a word in a sentence is identified with 2 or more GSRs, only one GSR is chosen based on the left to right descending priority of the categories mentioned. These roles (GSRs) were extracted separately using the text analytics engine Semantex^{TM} (`www.janyainc.com`) . Thus in a window of sentences, there are potentially $(G+1)^2$ GSRts for a total of G GSRs with the additional one representing a null role (denoted by "$--$") as in the word is not found in the contextual sentence. We used anaphora resolution as offered by the product to substitute pronouns with the referent nouns as a preprocessing step. If there are T_G valid GSRts in the corpus, then a sentence is represented as a vector over the GSRt counts only along with a binary vector over the word vocabulary. In the extended model for summarization, we also added one more role called NE (Named Entity), with the highest priority, that encompasses all possible named entity categories.

For further insight on how GSRts were used, we constructed a matrix consisting of sentences as rows and words as columns; the entries in the matrix are filled up with a specific GSR for the word in the corresponding sentence following GSR priorities. Table 1 shows a slice of such a matrix taken from the DUC2005 dataset which contains documents related to events concerning rules imposed on food labeling. Table 1 suggests, as in [2], that dense columns of the GSRs indicate potentially salient and coherent sentences (1 and 2 here) that present less inference load with respect to a query like "Food Labeling".

Table 1: Snapshot of a sentence-word GSR grid view of document

↓SentenceIDs || words... →					
sID	food	consumers	health	confusion	label(ing)
1	nn	$--$	nn	nn	nn
2	nn	$--$	nn	$--$	$--$
3	$--$	subj	$--$	$--$	$--$
4	subj	nn	subj	$--$	$--$

where "nn" is a noun and "ne" is a Named Entity category. The sentences 1 through 4 in the document read as:

1. The_Food_and_Drug_Administration has proposed a stringent set of rules governing the use of health claims on food labels and advertising, ending nearly six years of confusion over how companies may promote the health value of their products.

2. By narrowing standards for what is permissible and strengthening the FDA's legal authority to act against misleading claims , the rules could curtail a trend in food marketing that has resulted in almost 40% of new products and a third of the $ 3.6 billion in food advertising over the last year featuring health-related messages.

3. Most such messages are intended to make the consumer think that eating the product will reduce the risk of heart disease or cancer.

4. The regulations, which will be published next week in the Federal_Register , were criticized by food industry o¡cials , who said they would require health claims to meet an unrealistic standard of scienti↓c proof and would hinder the ability of manufacturers to give consumers new information about nutrition.

Note that the counts for the GSRts "nn→ $--$" and "nn→nn" for sentenceID 1 are 2 from this snapshot. Thus this discourse is dominant in GSRts involving a noun GSR w.r.t. the query words.

4. THE PROPOSED METHOD

In this section we describe our method to model topics not only using word counts but also using contextual GSRts. We also show how an extension transforms the utterance topic model to a "Learning To Summarize" (henceforth LeToS) model.

4.1 Description of the Datasets

The datasets we used for finding topics as well as subsequent summarization are the DUC2005, TAC2008, TAC2009 as well as a more practical real-life data from Yahoo! Answers[4]. The DUC2005 dataset had 50 folders with at least 25 documents in each folder. Each such folder corresponded to a particular "topic focus" or "cluster" representing varying human information needs. The TAC2008 dataset was organized in 48 folders as in DUC2005, however, it also had documents in each folder grouped into two timelines that we merged for the sake of theme detection. The organization for the TAC2009 dataset is also similar with 44 folders. The manually collected Yahoo! Answers dataset consists of 10 such topic focuses with each topic focus pertaining to a particular real-life question. For each such topic focus, we collected 10 relevant answers and each stored each answer in a separate document.

4.2 Utterance Topic Model

We now describe in detail the proposed probabilistic graphical Utterance Topic Model (henceforth UTM). To describe the document generation process, we assume that there are K latent topics, T_G total number of possible GSRts and T GSRts associated with each document. Also denote $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$

[4]http://www.answers.yahoo.com

to be the topic and topic-coupled GSRt proportions in each document. We say topic-coupled GSRt proportions since the expected number of terms per GSRt also depends on their latent topic assignment. Let r_t is the observed GSRt for a particular $GSRt = t$ across a window of 3 sentences; w_n is the observed word in the n^{th} position. Further denote, z_t to be an indicator variable for topic proportions, y_n is the indicator variable for topic-coupled GSRt proportions. At the parameter level, each topic is a multinomial over the vocabulary V of words in the corpus and each topic is also a multinomial over the GSRts following the implicit relation of GSRts to words within sentence windows. Also these GSRts are the output of a separate natural language parsing system.

At a higher level, each document in the corpus has mixing proportions over both the number of latent topics and also over the number of topic-coupled GSRts. In our proposed model, a GSRt along with the topic is also responsible for selecting a word from the vocabulary. The document generation process is shown in Fig. 1 and is explained as a model below:

For each document $d \in 1, ..., M$
  Choose a topic proportion $\boldsymbol{\theta}|\boldsymbol{\alpha} \sim Dir(\boldsymbol{\alpha})$
    Choose topic indicator $z_t|\theta \sim Mult(\boldsymbol{\theta})$
    Choose a GSRt $r_t|z_t = k, \boldsymbol{\rho} \sim Mult(\boldsymbol{\rho}_{z_t})$
  Choose a GSRt proportion $\boldsymbol{\pi}|\boldsymbol{\eta} \sim Dir(\boldsymbol{\eta})$
  For each position n in document d
    Choose $y_n|\pi \sim Mult(\boldsymbol{\pi})$
    Choose a word $w_n|y_n = t, \mathbf{z}, \boldsymbol{\beta} \sim Mult(\boldsymbol{\beta}_{z_{y_n}})$

where $n \in \{1, ..., N_d\}$ is the number of words in document $d \in \{1, ..., M\}$, t is and index into one of the T GSRts and k is an index into one of the K topics; $\boldsymbol{\beta}$ is a KxV matrix and $\boldsymbol{\rho}$ is a KxT_G matrix. The model can be viewed as a generative process that first generates the GSRts and subsequently generates the words that describes the GSRts. For each document, we first generate T_G GSRts using a simple LDA model and then for each of the N_d words, a GSRt is chosen and a word w_n is drawn conditioned on the same factor that generated the chosen the GSRt. Instead of influencing the choice of the GSRt to be selected from an assumed distribution (e.g. uniform or poisson) of the number of GSRts, the document specific proportions are used i.e. $\boldsymbol{\pi} - \boldsymbol{\eta}$ is the expected number of terms assigned to a GSRt influenced by the generating topics.

Direct posterior inference over the latent variables is intractable because of coupling of the parameters and the latent factors given the observed variables. We thus resort to approximate inference through Variational Bayes [3]. Variational Bayes breaks the edges between coupled random variables and parameters, removes the observed variables that lead to coupling and introduces free variational parameters that act as surrogate to the causal distribution of the original latent variables. The resulting simpler tractable distribution is shown in Fig. 2. In the variational setting, for each document we have $\sum_{k=1}^{K} \phi_{tk} = 1$ and $\sum_{t=1}^{T} \lambda_{nt} = 1$. Note that $\boldsymbol{\theta}$ is K-dimensional and $\boldsymbol{\pi}$ is T_G-dimensional.

4.3 Parameter Estimation and Inference

This section outlines the various updates of the latent variables and the parameters. In this paper we have resorted to mean field variational inference [5, 3] to find as tight as possible an approximation to the log likelihood of the data (the joint distribution of the observed variables given the parameters) by minimizing the KL divergence of the posterior distribution of the latent variables over the variational parameters to likelihood of the data. The details can be found in [5, 3]. For tractability, we assume a fully factorized variational distribution

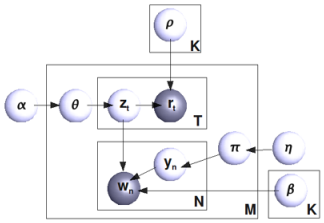

Figure 1: Graphical model representation of UTM

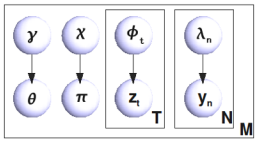

Figure 2: The variational dual of UTM

$$q(\boldsymbol{\theta}, \boldsymbol{\pi}, \mathbf{z}, \mathbf{y}|\boldsymbol{\gamma}, \boldsymbol{\chi}, \boldsymbol{\phi}, \boldsymbol{\lambda}) = q(\boldsymbol{\theta}|\boldsymbol{\gamma}) q(\boldsymbol{\pi}|\boldsymbol{\chi}) \prod_{t=1}^{T} q(z_t|\boldsymbol{\phi_t}) \prod_{n=1}^{N} q(y_n|\boldsymbol{\lambda_n}) \quad (1)$$

and then maximize the lowerbound on $p(\mathbf{r}, \mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\eta}, \boldsymbol{\rho}, \boldsymbol{\beta})$

The variational functional to optimize can be shown to be [3]

$$\mathcal{F} = E_q[\log p(\mathbf{r}, \mathbf{w}, \boldsymbol{\theta}, \boldsymbol{\pi}, \mathbf{z}, \mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\eta}, \boldsymbol{\rho}, \boldsymbol{\beta})] - E_q[\log q(\boldsymbol{\theta}, \boldsymbol{\pi}, \mathbf{z}, \mathbf{y}|\boldsymbol{\gamma}, \boldsymbol{\chi}, \boldsymbol{\phi}, \boldsymbol{\lambda})] \quad (2)$$

where $E_q[f(.)]$ is the expectation of $f(.)$ under the q distribution.

4.4 Latent variable inference

The key inferential problem that we are trying to solve here is to infer the posterior distribution of the latent variables given the observations and parameter values. We essentially convert the intractable integration problem to a tractable lower bound optimization problem. From Fig. 2, we obtain the variational parameters to be $\boldsymbol{\gamma}, \boldsymbol{\chi}, \boldsymbol{\phi}, \boldsymbol{\lambda}$. The maximum likelihood estimations of these indicator variables are as follows:

$$\gamma_i = \alpha_i + \sum_{t=1}^{T_d} \phi_{ti}$$
$$\chi_t = \eta_t + \sum_{n=1}^{N_d} \lambda_{nt}$$
$$\lambda_{nt} \propto \exp\{(\Psi(\chi_t) - \Psi(\sum_{f=1}^{T} \chi_f)) + (\sum_{i=1}^{K} \phi_{ti} \log \beta_{z_{(y_n=t)}=i,n})\}$$
$$\phi_{ti} \propto \exp\{\log \rho_{it} + (\Psi(\gamma_i) - \Psi(\sum_{k=1}^{K} \gamma_k)) + (\sum_{n=1}^{N_d} \lambda_{nt} \log \beta_{z_{(y_n=t)}=i,n})\}$$

4.5 Maximum Likelihood Parameter estimation

We now write down the expressions for the maximum likelihood of the parameters of the original graphical model using derivatives w.r.t the parameters of the functional $\mathcal{F}$ in Equ.

(2). We have the following results:

$$\rho_{ig} \propto \sum_{d=1}^{M} \sum_{t=1}^{T_d} \phi_{dti} r_{dt}^{g}$$
$$\beta_{ij} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} (\sum_{t=1}^{T_d} \lambda_{nt} \phi_{ti}) w_{dn}^{j}$$

where g and t are index variables for all possible GSRts and document specific GSRts respectively. and r_{dt}^{g} is 1 iff $t = g$ and 0 otherwise. The updates of $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$ are exactly the same as mentioned in [5].

5. EXTENDING UTTERANCE TOPIC MODEL FOR SUMMARIZATION

In this section, we investigate how a topic model like UTM that incorporates word level contextual features could be extended to a model for summarization. The motivation to model the summarization process as generative model arises from the following example: Suppose in an exam, a student is asked to write an essay type answer based out of a large amount of preparatory reading materials. Now, under usual circumstances, he would not memorize the entire set of materials. Instead, for possible question scenarios, the student remembers only selected sentences (be directly extracted from text or manufactured through natural language generation techniques) which are much like those found in the summary slide(section) of a lecture(chapter) about a particular topic. Then a coherent answer is constructed out of those by expanding on the summary sentences.

From table 1, we have observed that dense columns (non "$--$" entries) of the document level sentence-term GSR grid identify potential coherent informative sentences w.r.t particular query words. Thus to extend UTM into a summarization model, we treat each GSRt as distributions over sentences. We thus have a topic-word multinomial simplex as well as a GSRt-sentence multinomial simplex and the sentences and the words are related through underlying topics and contextual GSRts. This line of thought again is influenced by the work carried out in [2]

To define a simplistic summarization model, we describe the document generation process as follows:

For each document $d \in 1, ..., M$
  Choose a topic proportion $\boldsymbol{\theta}|\boldsymbol{\alpha} \sim Dir(\boldsymbol{\alpha})$
    Choose topic indicator $z_t|\theta \sim Mult(\boldsymbol{\theta})$
    Choose a GSRt $r_t|z_t = k, \boldsymbol{\rho} \sim Mult(\boldsymbol{\rho}_{z_t})$
  Choose a GSRt proportion $\boldsymbol{\pi}|\boldsymbol{\eta} \sim Dir(\boldsymbol{\eta})$
  For each position n in document d:
    For each instance of utterance s_p for which w_n occurs in s_p in document d:
      Choose $v_p|\pi \sim Mult(\boldsymbol{\pi})$
      Choose $y_n \sim v_p \delta(w_n \in s_p)$
      Choose a sentence $s_p \sim Mult(\boldsymbol{\Omega}_{v_p})$
      Choose a word $w_n|y_n = t, \mathbf{z}, \boldsymbol{\beta} \sim Mult(\boldsymbol{\beta}_{z_{y_n}})$

where N is the number of words in document $d \in 1, ..., M$, P is the number of sentences in the same document and t is and index into one of the T GSRts. $\delta(w_n \in s_p)$ is the delta function which is 1 iff the n^{th} word belong to the p^{th} sentence and 0 otherwise. Under this extension, $\boldsymbol{\pi}_t - \boldsymbol{\eta}_t$ to be the expected number of words and sentences per topic-coupled GSRt in each document. Each topic-coupled GSRt is also treated as a multinomial $\boldsymbol{\Omega_t}$ over the total number U of sentences in the corpus. Thus when we select a GSRt using $\boldsymbol{\pi}$ and choose a word w_n to describe it, we also sample a sentence s_p containing w_n. In disjunction, $\boldsymbol{\pi}$ along with v_p, s_p and $\boldsymbol{\Omega}$ focus mainly on coherence among the coarser units - the sentences. However, the influence of a particular GSRt like "subj→subj" on coherence may be discounted if that is not the dominant trend in the transition topic. This fact is enforced through the coupling of empirical GSRt proportions to topics of the sentential words. Figure 3 give the depiction of the above process as a graphical model. The variational Bayesian counterpart of the model is exactly the same as in figure 2 but with an additional independent P plate inside of the M plate for sentence-GSRt multinomials i.e a plate with a directed arc from variational ζ_p to indicator v_p.

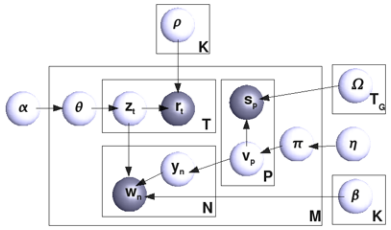

Figure 3: Graphical model representation of LeToS by extending UTM

For obtaining summaries, we order sentences w.r.t query words by accumulating the sentence-query word pair probability scores by computing:

$$p(s_u|\mathbf{q}) = \sum_{l=1}^{Q} (\sum_{t=1}^{T} \sum_{i=1}^{K} \zeta_{ut} \phi_{ti} (\lambda_{lt} \phi_{ti}) \gamma_{di} \chi_{dt}) \delta(w_l \in s_u) \quad (3)$$

where Q is the number of the query words in query vector $\mathbf{q}$ and s_u is the u^{th} sentence in the corpus that belongs to all such document d's which are relevant to the query, w_l is the l^{th} query word, and i and t and topic and GSRt indices respectively. Normally, under this model we can enforce that each sentence in the summary be actually extracted form a unique document only, however, if we need larger more coherent summaries, we can include the sentences in the window of each most probable sentence. Further, whenever possible, the sentences are scored over only "rich" GSRts which lack any "$--$" GSRs.

5.1 Parameter Estimation and Inference in the Extended Model

The set of equations in section 4.4 is augmented by the updates of the variational sentence multinomial and the posterior Dirichlet update for the proportions of GSRts as:

$$\chi_t = \eta_t + \sum_{n=1}^{N_d} \lambda_{nt} + \sum_{p=1}^{P_d} \zeta_{tp}$$
$$\zeta_{pt} \propto \Omega_{pt} \exp\{\Psi(\chi_t) - \Psi(\sum_{j=1}^{T} \chi_j)\}$$

Note that these are again per-document updates. The only addition to the set of equations given in section 4.5 is:

$$\Omega_{tu} \propto \sum_{d=1}^{M} \sum_{p=1}^{P_d} \zeta_{dpt} s_{dp}^{u}$$

where u is an index into one of the S sentences in the corpus and $s^u_{dp} = 1$ if the p^{th} sentence in document d is one among S.

6. RESULTS AND DISCUSSIONS

In this section, we first study some topics generated by both LDA, Correlated topic model - CTM[4] and UTM for the DUC2005 dataset and then we try to measure the test set perplexity as in [5].

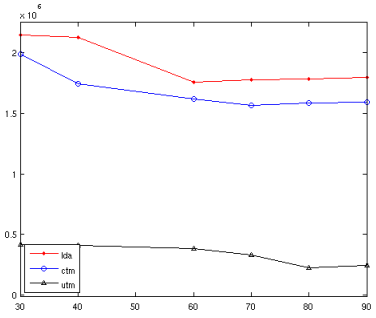

Figure 4: Perplexities of LDA, CTM and UTM

Table 2: Some topics from LDA for TAC2008

topic19	**topic0**	**topic2**	**topic7**
mines	company	Armstrong	ice
coal	Fannie_Mae	samples	glaciers
safety	executive	tested	years
accidents	financial	Tour	sea
China	officer	L'Equipe	warming
coal_mine	chief	EPO	scientists
years	account	doping	Antarctica
produced	Raines	times	Alaska
officials	billion	years	global
killed	top	French	levels

Table 3: Some topics from UTM for TAC2008

topic35	**topic5**	**topic58**	**topic47**
mine	Fannie_Mae	Armstrong	planet
coal	company	steroids	Pluto
China	account	tested	ice
safety	Raines	samples	scientists
year	Fannie	years	glaciers
accident	financial	drugs	objects
officials	executive	Tour	Earth
panda	Howard	doping	warming
state	chief	L'Equipe	sea
coal_mine	years	EPO	melting

We note that perplexity of a held-out test set is calculated as $exp\{-\frac{\sum_{d=1}^{M} log\, p(\mathbf{w}_d)}{\sum_{n=1}^{N_d} N_d}\}$, where probability of document d with N_d words is given by $p(\mathbf{w}_d)$ with $\mathbf{w}_d$ being the only observed variable. In the UTM, however, we have two observed variables - $\mathbf{w}_d$ and $\mathbf{r}_d$ where the latter is a vector over GSRts. Thus for UTM, we only calculate the conditional word perplexity given the GSRts. From the figure 4 we note that the held-out perplexity is least for 80 topics for UTM and 60 and 70 for LDA (red graph) and CTM (blue graph) respectively. Also, UTM had the least test-set perplexity w.r.t. to the other two models. Similar trends have been observed for other datasets too and are not reproduced here due to space constraints.

From both tables 2 and 3, we display a few sample topics as distributions over the vocabulary by manually matching the topics from LDA and UTM. Majority of the words in these topics come from documents that spoke about "Describe the coal mine accidents in China and actions taken," "Give an account of the criminal investigation of Franklin Raines," "Describe accusations that seven-time Tour de France winner Lance Armstrong used the performance-enhancing drug EPO" and "Describe the developments and impact of the continuing Arctic and Antarctic ice melts" respectively.

There is one important observation that comes out of observing table 3. We see two *intrusive* words "panda" and "Pluto" in topics 35 and 47 respectively. If we recall our Utterance topic model, we are not directly generating topics from word counts. Instead, the words are chosen to fit a GSRt that a topic generates. Thus, word co-occurence is not the only factor in determining the thematical structure of the documents for UTM. The word Pluto has been observed in topic 47 because "Pluto" was found to describe the same GSRts as the other words in the topic. This phenomenon happened because there were a smaller set of documents that dealt with scientists discovering another planet outside of Pluto's orbit. Similarly there were a few documents reporting shortage of bamboos as food for pandas in China. Thus UTM could discover new related concepts in a topic through the principles of centering whereas the words in LDA topics are derived from cooccurence. In CTM, this relation is captured through correlation *between* topics. Thus, intuitively, for UTM, topic 47 is more about "crisis events in China." From a summarization point of view, there are no downsides to the effects of such inclusion of a few *intrusive* words. In terms of an information need like "effects of global warming," there would likely be no relevant documents containing Pluto and so these words don't affect sentence probability calculation w.r.t. the query words.

Recall that LeToS is a fixed index (i.e no train-test split because of unique sentenceIDs) summarization system where the data is not allowed to change but the queries are i.e., once the model parameters are calculated on the fixed data, we use variational inference to determine the topics of the free form queries, perform query expansion using related topical words and finally select the top sentences which are weighted by products of other variational parameters. To determine the number of topics fitted to the data, one way is to run UTM on the dataset, and decide the best number of topics which from Figure 4 is 80 for the DUC2005 dataset.

Table 4 shows the performance of our summarization system to some top systems in DUC2005. Each summary was input to ROUGE as one single line. The top 4 rows in the table report the ROUGE scores of 4 gold-standard human summaries. The description of the systems named NUS, HKPoly, IIITH, BFQS, BE-ISI and UIR can be found in [20, 12, 11, 7, 10, 18]. Different runs of our system have been named LeToS-[kk]-[NE]-qE-[U/NU] where *kk* denotes

Table 4: Comparison of DUC2005 ROUGE Results

Systems	Rouge-2 Recall	95% Conf. Interval	Rouge-SU4 Recall	95% Conf. Interval
ModelA	0.34367	0.30939 - 0.37816	0.39876	0.36682 - 0.43142
ModelB	**0.36794**	0.33509 - 0.40440	**0.43518**	0.40642 - 0.46766
ModelC	0.30019	0.26992 - 0.33272	0.37335	0.34434 - 0.40416
ModelD	0.31269	0.28657 - 0.34182	0.38028	0.35551 - 0.40693
NUS	0.14632	0.12305 - 0.17200	0.23557	0.21646 - 0.25593
HK Poly	0.13984	0.11951 - 0.16282	0.23066	0.21194 - 0.25070
IIITH	0.14127	0.11740 - 0.16612	0.22849	0.20762 - 0.25163
LeToS-60-qE-U	0.13213	0.11064 - 0.15452	0.21425	0.19610 - 0.23395
LeToS-70-qE-U	0.12799	0.10648 - 0.14990	0.21448	0.19711 - 0.23455
LeToS-80-qE-U	**0.13888**	0.11332 - 0.16617	**0.22302**	0.20023 - 0.24589
LeToS-90-qE-U	0.12318	0.10329 - 0.14607	0.21242	0.19394 - 0.23263
LeToS-60-NE-qE-U	0.12556	0.10551 - 0.14537	0.21409	0.20009 - 0.22944
LeToS-70-NE-qE-U	**0.12904**	0.10692 - 0.15211	**0.21747**	0.20005 - 0.23662
LeToS-80-NE-qE-U	0.12481	0.10604 - 0.14501	0.21166	0.19586 - 0.22867
LeToS-90-NE-qE-U	0.12512	0.10679 - 0.14575	0.21385	0.19699 - 0.23102
LeToS-60-qE-NU	0.11320	0.09531 - 0.13337	0.19659	0.17934 - 0.21604
LeToS-70-qE-NU	0.11198	0.09233 - 0.13352	**0.19710**	0.18001 - 0.21641
LeToS-80-qE-NU	**0.11767**	0.09757 - 0.13863	0.20317	0.18336 - 0.22364
LeToS-90-qE-NU	0.11586	0.09764 - 0.13678	0.20264	0.18524 - 0.22224
LeToS-60-NE-qE-NU	**0.10837**	0.08754 - 0.13308	**0.19365**	0.17555 - 0.21414
LeToS-70-NE-qE-NU	0.08939	0.07229 - 0.10976	0.18461	0.16862 - 0.20149
LeToS-80-NE-qE-NU	0.09289	0.07617 - 0.11173	0.18546	0.17052 - 0.20204
LeToS-90-NE-qE-NU	0.09252	0.07710 - 0.10863	0.18788	0.17356 - 0.20317
BFQS	0.12976	0.10834 - 0.15281	0.21703	0.19938 - 0.23647
BE-ISI	0.11973	0.09801 - 0.14425	0.21084	0.19337 - 0.22957
UIR	0.09622	0.07994 - 0.11504	0.17894	0.16544 - 0.19240

the number of topics; a presence of NE indicates use of a separate Named Entity role as a GSR that is ranked higher than the "Subject" role and groups all Named Entity categories together; U/NU means that each sentence of the summary is from a unique document or sentences could belong to the same document; qE denoted that the query words were expanded using topic inference prior to summarization. From table 4, we observe from the ROUGE2 and ROUGE SU4 scores that for 80 topics, the scores were highest and thus 80 is the best value of the number of topics which also agrees from that obtained from UTM. This is so, since in LeToS, $\boldsymbol{\pi}_t - \boldsymbol{\eta}_t$ is the expected number of words *plus sentences* per topic-coupled GSRt in each document, the relative document specific topic-coupled GSRt proportions remain the same. From the ROUGE scores, it is observed that using a separate NE category as a GSR did not improve the ROUGE scores and hence this category was not included while evaluating summaries using the PYRAMID method. Also, we observe that if we reject a document once a sentence is selected, the ROUGE scores are much higher reflecting the phenomenon that a human would probably pack information from as many documents as possible into a single summary. Imparting "a final coherent flow" into such a summary involves a lot of editing to obey centering constraints whilst not sacrificing information. This post-processing part is important but outside the purview of this model. However, table 4 clearly highlights how far automatic summaries are to the human counterparts.

We used TAC2008 as a development set to measure performance on the TAC2009 dataset. In the A timeline of the TAC2008 dataset, the average Pyramid scores for very short 100 word summaries over 48 queries were obtained as **0.3089** with a rank of 14 out of 58 submissions. For TAC2009 also, using the manual Pyramid [17] scoring for summaries, the average Pyramid scores for the 100 word summaries over 44 queries were obtained as **0.3024** for the A timeline and **0.2601** for the B timeline for LeToS and ranked 13^{th} and 9^{th} of 52 submissions. Note that the score is lower overall due to the extractive nature of summarization and a short 100 word limit. The scores for the system in [18] that uses coherence to some extent and a baseline returning all the leading sentences (up to 100 words) in the most recent document are (0.1756 and 0.1601) and (0.175 and 0.160) respectively for the A and B timelines. The score for the B timeline is lower due to redundancy which was not addressed in our model. These scores indicate that performance of our model was consistent with the development (TAC2008) dataset and test (TAC2009) datasets.

Here we also present a sample summarized ≈120 words answer for a Yahoo! Answers question using our proposed extension of UTM to LeToS. The questions were fed into the model with standard stopwords removal. For the question "**Are Sugar substitutes bad for you?**," the summary was derived as " *The show stated Aspartame turns into METHANOL in your body and is like drinking FORMALDEHYDE! Splenda is another popular one, but because the body doesn't recognize, the body won't digest it, and can actually make you GAIN weight. The FDA has approved it for 5 mg/Kg body weight, which is the least of all the sweeteners and comes out to 6 cans of diet cola per day. Aspartame is at the root of diseases such as: aspartame fibromaylagia, aspartame restless leg syndrome, aspartame and migraines, aspartame and tumors, aspartame allergy, aspartame multiple sclerosis,*

bladder cancer aspartame, aspartame and central nervous system, aspartame and infertility, aspartame and weight gain,...." We note here that each sentence is forced to be selected from a different document to maximize information content. The word "aspartame" was selected here due to topic based query expansion.

We also generated baseline summaries that take two sentences with at least one query word overlap from the beginning and end of each document till the length constraint was satisfied. The baseline was obtained as " *Fructose is another extract of sugar. Hope that shed a little light on your questions. It's made by substituting three atoms of chlorine for three hydroxyl groups on the sugar molecule. Honey enters the bloodstream slowly, 2 calories per minute, while sugar enters quickly at 10 calories per minute, causing blood sugars to fluctuate rapidly and wildly. This sugar substitute, sold commercially as Equal and NutraSweet, was hailed as the savior for dieters who for decades had put up with saccharine's unpleasant after taste. Too much phenylalanine causes seizures, elevated blood plasma, is dangerous for pregnancy causing retardation, PMS caused by phenylalanine's blockage of serotonin, insomnia, and severe mood swings. Sugar substitues, turn into formaldehyde in the body...*" Clearly for the Yahoo! Answers dataset, even though quantitative evaluation was not available, our summary reads much better than the baseline. Indeed, length constraints do not allow us to include the context of sentences around each sentence in each document, but implicitly the context is clear. Also, under our summarization model, because of the way we define GSRts, we can also relax the constraint that an utterance is a full sentence by defining any other meaningful sequence of words to be an utterance.

7. CONCLUSION

The possibility of building a statistical generative model for documents using lexical and semantic bundles in context has been explored in this paper. A key extension to this model would be to find a representation to understand the **meaning** of the query. Another interesting aspect to explore would be Bayesian non-parameteric methods that eliminate the dependance on the number of topics.

8. REFERENCES

[1] Rachit Arora and Balaraman Ravindran. Latent dirichlet allocation and singular value decomposition based multi-document summarization. In *Proceedings of the Eighth IEEE International Conference on Data Mining (ICDM 2008)*, pages 91–97. IEEE Press, 2008.

[2] Regina Barzilay and Mirella Lapata. Modeling local coherence: an entity-based approach. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 141–148. Association for Computational Linguistics, 2005.

[3] Matthew J. Beal. *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

[4] David Blei and John Lafferty. Correlated topic models. In *Advances in Neural Information Processing Systems*, volume 18, 2005.

[5] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[6] Dewei Chen, Jie Tang, Limin Yao, Juanzi Li, and Lizhu Zhou. Query-focused summarization by combining topic model and affinity propagation. In *APWeb/WAIM*, pages 174–185, 2009.

[7] Hal Daumé III and Daniel Marcu. Bayesian query-focused summarization. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, Sydney, Australia, 2006.

[8] Jordan Boyd Graber and David Blei. Syntactic topic models. In *Advances in Neural Information Processing Systems*, volume 21, 2009.

[9] Barbara J. Grosz, Scott Weinstein, and Arvind K. Joshi. Centering: A framework for modeling the local coherence of discourse. In *Computational Linguistics*, volume 21, pages 203–225, 1995.

[10] Eduard Hovy, Chin-Yew Lin, and Liang Zhou. A be-based multi-document summarizer with query interpretation. In *Proceedings of DUC2005*, 2005.

[11] Jagadeesh J, Prasad Pingali, and Vasudeva Varma. A relevance-based language modeling approach to duc 2005. `http://duc.nist.gov/pubs.html#2005`.

[12] Wenjie Li, Wei Li, Baoli Li, Qing Chen, and Mingli Wu. The hong kong polytechnic university at duc 2005. `http://duc.nist.gov/pubs.html#2005`.

[13] Chin-Yew Lin and Eduard Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 71–78. Association for Computational Linguistics, 2003.

[14] Ryan T. McDonald. A study of global inference algorithms in multi-document summarization. In *ECIR*, volume 4425 of *Lecture Notes in Computer Science*, pages 557–564. Springer, 2007.

[15] Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and Chengxiang Zhai. Topic sentiment mixture: modeling facets and opinions in weblogs. In *In Proc. of the 16th Int. Conference on World Wide Web*, pages 171–180, 2007.

[16] Ramesh Nallapati and William Cohen. Link-plsa-lda: A new unsupervised model for topics and influence of blogs. In *International Conference for Weblogs and Social Media*, 2008.

[17] Ani Nenkova and Rebecca Passonneau. Evaluating content selection in summarization: The pyramid method. In Daniel Marcu Susan Dumais and Salim Roukos, editors, *HLT-NAACL 2004: Main Proceedings*, pages 145–152. Association for Computational Linguistics, 2004.

[18] Rohini Srihari, Li Xu, and Tushar Saxena. Use of ranked cross document evidence trails for hypothesis generation. In *Proceedings of the 13th International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 677–686, San Jose, CA, 2007.

[19] Furu Wei, Wenjie Li, Qin Lu, and Yanxiang He. A document-sensitive graph model for multi-document summarization. *Knowledge and Information Systems*.

[20] Shiren Ye, Tat-Seng Chua, Min-Yen Kan, and Long Qiu. Nus at duc 2005: Understanding documents via concept links. `http://duc.nist.gov/pubs.html#2005`.

9. APPENDIX

This section gives partially complete derivations to find out the optimal settings of the hidden variables and the parameters of the utterance topic model. Note that the inference part i.e. inferring variational distributions for hidden variables (E-step) is document specific, while the model parameter estimation (M-step) is corpus wide. We start out with some initial values of the parameters and following [5, 3], we find the posterior distribution over the latent variables parameterized by the free variational parameters in the VBE step and holding this distribution fixed, optimize the parameters of the model in the VBM step. In each of these steps, we select out only those terms from $\mathcal{F}$ that depend on the variable being optimized.

$$(\boldsymbol{\gamma}^*, \boldsymbol{\chi}^*, \boldsymbol{\phi}^*, \boldsymbol{\lambda}^*) = \underset{(\boldsymbol{\gamma},\boldsymbol{\chi},\boldsymbol{\phi},\boldsymbol{\lambda})}{\arg\min}\, KL(q(\boldsymbol{\theta}, \boldsymbol{\pi}, \mathbf{z}, \mathbf{y}|\boldsymbol{\gamma}, \boldsymbol{\chi}, \boldsymbol{\phi}, \boldsymbol{\lambda}) \\ ||p(\boldsymbol{\theta}, \boldsymbol{\pi}, \mathbf{z}, \mathbf{y}|\mathbf{r}, \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\eta}, \boldsymbol{\rho}, \boldsymbol{\beta})) \tag{4}$$

By Jensen's inequality, we have

$$\log p(\mathbf{r}, \mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\eta}, \boldsymbol{\rho}, \boldsymbol{\beta}) \geq \{E_q[p(\mathbf{r},\mathbf{w}, \boldsymbol{\theta}, \boldsymbol{\pi}, \mathbf{z}, \mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\eta}, \boldsymbol{\rho}, \boldsymbol{\beta})] \\ -E_q[q(\boldsymbol{\theta}, \boldsymbol{\pi}, \mathbf{z}, \mathbf{y}|\boldsymbol{\gamma}, \boldsymbol{\chi}, \boldsymbol{\phi}, \boldsymbol{\lambda})]\} = \mathcal{F} \tag{5}$$

We thus have, $\mathcal{F}(\boldsymbol{\gamma}, \boldsymbol{\chi}, \boldsymbol{\phi}, \boldsymbol{\lambda}; \boldsymbol{\alpha}, \boldsymbol{\eta}, \boldsymbol{\rho}, \boldsymbol{\beta}) =$

$$E_q[log\, p(\boldsymbol{\theta}|\boldsymbol{\alpha})] + E_q[log\, p(\boldsymbol{\pi}|\boldsymbol{\eta})] + E_q[log\, p(\mathbf{z}|\boldsymbol{\theta})] + \\ E_q[log\, p(\mathbf{r}|\mathbf{z}, \boldsymbol{\rho})] + E_q[log\, p(\mathbf{w}|\mathbf{y}, \mathbf{z}, \boldsymbol{\beta})] - E_q[log\, q(\boldsymbol{\theta}|\boldsymbol{\gamma})] \\ -E_q[log\, q(\boldsymbol{\pi}|\boldsymbol{\chi})] - E_q[log\, q(\mathbf{z}|\boldsymbol{\phi})] - E_q[log\, q(\mathbf{y}|\boldsymbol{\lambda})] \tag{6}$$

Each of the terms in the equation (6) expands out to:

$$\log\Gamma(\sum_{j=1}^{K}\alpha_j) - \sum_{i=1}^{K}\log\Gamma(\alpha_i) + \sum_{i=1}^{K}(\alpha_i - 1)(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{K}\gamma_j)) \tag{7}$$

$$+\log\Gamma(\sum_{f=1}^{T}\eta_f) - \sum_{t=1}^{T}\log\Gamma(\eta_t) + \sum_{t=1}^{T}(\eta_t - 1)(\Psi(\chi_t) - \Psi(\sum_{f=1}^{T}\chi_f)) \tag{8}$$

$$+\sum_{t=1}^{T}\sum_{i=1}^{K}(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{K}\gamma_j))\phi_{ti} \tag{9}$$

$$+\sum_{t=1}^{T_d}\sum_{i=1}^{K}\sum_{g=1}^{T_G}\phi_{ti}\log\rho_{it}r_{dt}^{g} \tag{10}$$

$$+\sum_{n=1}^{N}\sum_{t=1}^{T}(\Psi(\chi_t) - \Psi(\sum_{j=1}^{T}\chi_t))\lambda_{tn} \tag{11}$$

$$+\sum_{n=1}^{N}\sum_{i=1}^{K}\sum_{j=1}^{V}(\sum_{t=1}^{T}\lambda_{nt}\phi_{ti})\log\beta_{z_{(y_n=t)=i}j}w_n^j \tag{12}$$

$$-\log\Gamma(\sum_{j=1}^{K}\gamma_j) + \sum_{i=1}^{K}\log\Gamma(\gamma_i) - \sum_{i=1}^{K}(\gamma_i - 1)(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{K}\gamma_j))\ldots \tag{13}$$

$$-\log\Gamma(\sum_{j=1}^{T}\chi_j) + \sum_{t=1}^{T}\log\Gamma(\chi_t) - \sum_{t=1}^{T}(\chi_t - 1)(\Psi(\chi_t) - \Psi(\sum_{j=1}^{T}\chi_j)) \tag{14}$$

$$-\sum_{t=1}^{T}\sum_{i=1}^{K}\phi_{ti}\log\phi_{ti} \tag{15}$$

$$-\sum_{n=1}^{N}\sum_{t=1}^{T}\lambda_{nt}\log\lambda_{nt} \tag{16}$$

Where, each term in a document is represented as a binary vector w_n^j , j $\in \{1,\ldots,V\}$, V being the number of terms in the vocabulary. The total number of GSR transitions is fixed at T_G and Ψ is the digamma function. It is to be understood that the t index for variational parameter updates is specific to the GSRt IDs in a document d and that for the global parameters like $\boldsymbol{\rho}$, g is a global index into one of the possible T_G GSRts. N is the maximum of the document lengths interms of unique terms.

9.1 Inference on Variational Parameters

Here we estimate the free variational parameters for the variational model depicted in Fig. 2 following the constraints on ϕ and λ.

For γ:

$$\mathcal{F}_{[\gamma]} = -\log\Gamma(\sum_{j=1}^{K}\gamma_j) + \sum_{i=1}^{K}\log\Gamma(\gamma_i) \\ +\sum_{i=1}^{K}(\alpha_i + \sum_{t=1}^{T}\phi_{ti} - \gamma_i)(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{K}\gamma_j)) \tag{17}$$

$$\frac{\partial\mathcal{F}_{[\gamma]}}{\partial\gamma_i} = (\alpha_i + \sum_{t=1}^{T}\phi_{ti} - \gamma_i)(\Psi'(\gamma_i) - \Psi'(\sum_{j=1}^{K}\gamma_j)) \\ -(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{K}\gamma_j)) + (\Psi(\gamma_i) - \Psi(\sum_{j=1}^{K}\gamma_j)) \tag{18}$$

Setting the above derivative to 0, we get,

$$\gamma_i = \alpha_i + \sum_{t=1}^{T}\phi_{ti} \tag{19}$$

For χ

Taking derivative of $F[\chi]$ w.r.t. χ_t, we have

$$\chi_t = \eta_t + \sum_{n=1}^{N}\lambda_{nt} \tag{20}$$

For λ:

$$\mathcal{F}_{[\lambda]} = \sum_{n=1}^{N}\sum_{t=1}^{T}(\Psi(\chi_t) - \Psi(\sum_{j=1}^{T}\chi_j))\lambda_{tn} + \\ \sum_{n=1}^{N}\sum_{i=1}^{K}\sum_{j=1}^{V}(\sum_{t=1}^{T}\lambda_{nt}\phi_{ti})\log\beta_{z_{y_n}}w_n^j \\ -\sum_{n=1}^{N}\sum_{t=1}^{T}\lambda_{nt}\log\lambda_{nt} + \mu(\sum_{t=1}^{T}\lambda_{nt} - 1) \tag{21}$$

where μ is the Lagrange multiplier in (21)

$$\frac{\partial F}{\partial \lambda_{nt}} = 0 \Rightarrow (\Psi(\chi_t) - \Psi(\sum_{j=1}^{T} \chi_j) + (\sum_{t=1}^{T} \phi_{ti} \log \beta_{z_{y_n} n}) - 1 - \log \lambda_{nt} + \mu = 0$$

$$\Rightarrow \lambda_{nt} = \exp\{(\Psi(\chi_t) - \Psi(\sum_{f=1}^{T} \chi_f) + (\sum_{i=1}^{K} \phi_{ti} \log \beta_{z_{y_n} n}) - 1 + \mu\}$$

$$\Rightarrow exp\{\mu - 1\} = \frac{1}{\sum_{t=1}^{T} exp\{(\Psi(\chi_t) - \Psi(\sum_{f=1}^{T} \chi_f) + (\sum_{i=1}^{K} \phi_{ti} \log \beta_{z_{y_n} n})\}}$$

Setting the derivative $\frac{\partial F}{\partial \lambda_{nt}}$ to 0 gives us,

$$\lambda_{nt} \propto \exp\{(\Psi(\chi_t) - \Psi(\sum_{f=1}^{T} \chi_f)) + (\sum_{i=1}^{K} \phi_{ti} \log \beta_{z_{y_n} n})\} \quad (22)$$

For $\boldsymbol{\phi}$:

$$\begin{aligned} \mathcal{F}_{[\phi]} = & \sum_{t=1}^{T} \sum_{i=1}^{K} (\Psi(\gamma_i) - \Psi(\sum_{j=1}^{K} \gamma_j)) \phi_{ti} + \sum_{t=1}^{T} \sum_{i=1}^{K} \phi_{ti} \log \rho_{ti} \\ & + \sum_{n=1}^{N} \sum_{i=1}^{K} \sum_{j=1}^{V} (\sum_{t=1}^{T} \lambda_{nt} \phi_{ti}) \log \beta_{z_{(y_n)} n} w_n^j \\ & - \sum_{t=1}^{T} \sum_{i=1}^{K} \phi_{ti} \log \phi_{ti} + \mu (\sum_{i=1}^{K} \phi_{ti} - 1) \end{aligned} \quad (23)$$

where μ is the Lagrange multiplier in $\mathcal{F}_{[\phi]}$. As before,

$$\frac{\partial F}{\partial \phi_{ti}} = 0 \Rightarrow \phi_{ti} \propto \exp\{\log \rho_{ti} + (\Psi(\gamma_i) - \Psi(\sum_{k=1}^{K} \gamma_k)) + (\sum_{n=1}^{N} \lambda_{nt} \log \beta_{z_{(y_n)} n})\} \quad (24)$$

9.2 Model Parameter Estimation

Here we calculate the maximum likelihood settings of the parameters that do not grow with the data. So we need to need to take into account the contribution of these for all the documents and not just a single document.

For $\boldsymbol{\rho}$:

$$\mathcal{F}_{[\rho]} = \sum_{d=1}^{M} \sum_{t=1}^{T_d} \sum_{i=1}^{K} \sum_{g=1}^{T_G} \phi_{dti} \log \rho_{it} r_{dt}^g + \sum_{i=1}^{K} \mu_i (\sum_{g=1}^{T_G} \rho_{ig} - 1) \quad (25)$$

wherethe μ_t's are the K Lagrange multipliers in (25)

$$\begin{aligned} \frac{\partial F}{\partial \rho_{ig}} &= \sum_{d=1}^{M} \sum_{t=1}^{T_d} \phi_{dti} r_{dt}^g \frac{1}{\rho_{ig}} + \mu_i \\ \frac{\partial F}{\partial \rho_{ig}} = 0 \Rightarrow \rho_{ig} &= -\frac{\sum_{d=1}^{M} \sum_{t=1}^{T_d} \phi_{dti} r_{dt}^g}{\mu_i} \\ \Rightarrow \mu_i &= -\sum_{g=1}^{T_G} \sum_{d=1}^{M} \sum_{t=1}^{T_d} \phi_{dti} r_{dt}^g \\ \therefore \frac{\partial F}{\partial \rho_{ig}} = 0 \Rightarrow \rho_{ig} &\propto \sum_{d=1}^{M} \sum_{t=1}^{T_d} \phi_{dti} r_{dt}^g \end{aligned} \quad (26)$$

For $\boldsymbol{\beta}$:

$$\mathcal{F}_{[\beta]} = \sum_{n=1}^{N} \sum_{i=1}^{K} \sum_{j=1}^{V} (\sum_{t=1}^{T} \lambda_{nt} \phi_{ti}) \log \beta_{z_{y_n} n} w_n^j + \sum_{i=1}^{K} \mu_i (\sum_{j=1}^{V} \beta_{ij} - 1) \quad (27)$$

where μ_is are the K Lagrange Multipliers in (27)

$$\frac{\partial F}{\partial \beta_{ij}} = 0 \Rightarrow \beta_{ij} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} (\sum_{t=1}^{T_d} \lambda_{nt} \phi_{ti}) w_{dn}^j \quad (28)$$

For $\boldsymbol{\alpha}$:

$$\begin{aligned} \mathcal{F}_{[\alpha]} = & \sum_{d=1}^{M} (\log \Gamma(\sum_{i=1}^{K} \alpha_i) - \sum_{i=1}^{K} \log \Gamma(\alpha_i) \\ & + \sum_{i=1}^{K} (\alpha_i - 1)(\Psi(\gamma_{di}) - \Psi(\sum_{k=1}^{K} \gamma_{dk}))) \end{aligned} \quad (29)$$

$$\frac{\partial F}{\partial \alpha_i} = M(-\Psi(\alpha_i) + \Psi(\sum_{j=1}^{K} \alpha_j)) + \sum_{d=1}^{M} (\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{K} \gamma_{dj})))$$

$$\frac{\partial F}{\partial \alpha_i \alpha_j} = \partial(i,j) M \Psi'(\alpha_i) - \Psi'(\sum_{j=1}^{K} \alpha_j)$$

The derivative w.r.t. α_i depends on α_j and thus we can resort to Newton's iterative method to find out the maximal α using the gradient and Hessian vector and matrix respectively as in [5]. Ψ' is the trigamma function.

For $\boldsymbol{\eta}$:

The update is similar to $\boldsymbol{\alpha}$ update