Learning to Summarize using Coherence

Pradipto Das
Department of Computer Science
University at Buffalo
Buffalo, NY 14260
pdas3@buffalo.edu

Rohini Srihari
Department of Computer Science
University at Buffalo
Buffalo, NY 14260
rohini@cedar.buffalo.edu

Abstract

The focus of our paper is to attempt to define a generative probabilistic topic model for text summarization that aims at extracting a small subset of sentences from the corpus with respect to some given query. We theorize that in addition to a bag of words, a document can also be viewed in a different manner. Words in a sentence always carry syntactic and semantic information and often such information (for e.g., the grammatical and semantic role (GSR) of a word like subject, object, noun and verb concepts etc.) is carried across adjacent sentences to enhance coherence in different parts of a document. We define a topic model that models documents by factoring in the GSR transitions for coherence and for a particular query, we rank sentences by a product of thematical salience and coherence through GSR transitions.

1 Introduction

Automatic summarization is one of the oldest studied problems in IR and NLP and is still receiving prominent research focus. In this paper, we propose a new joint model of words and sentences for multi-document summarization that attempts to integrate the coherence as well as the latent themes of the documents.

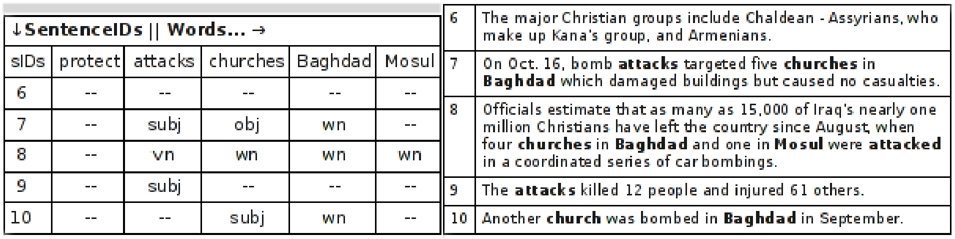
In the realm of computational linguistics, there has been a lot of work in Centering Theory including Grosz et. al. [3]. Their work specifies how discourse interpretation depends on interactions among speaker intentions, attentional state, and linguistic form. In our context, we could assume a subset of documents discussing a particular theme to be a discourse involving one or more participants. Attentional state models the discourse participants' focus of attention at any given point. This focus of attention helps identify "centers" of utterances that relate different parts of local discourse segments meaningfully and according to [3], the "centers" are semantic objects, not just words, phrases, or syntactic forms and centering theory helps formalize the constraints on the centers to maximize coherence. In our context, the GSRs and the explicit realization of these roles through the sentential words approximate the centers.

Essentially, then the propagation of these centers of utterances across utterances helps maintain the local coherence. This local coherence is responsible for the choice of words appearing *across* utterances in a particular discourse segment and helps reduce the *inference load* placed upon the hearer (or reader) to understand the foci of attention.

2 Adapting Centering Theory for Summarization

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 (http://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 Semantex 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.

For further insight, we can construct 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 (in case of multiple occurences of the same word in the same sentence with different GSRs). Figure 1 shows a slice of such a matrix taken from the TAC2008 dataset (http://www.nist.gov/tac/tracks/2008/index.html) which contains documents related to events concerning Christian minorities in Iraq and their current status. Figure 1 suggests, as in [1], that dense columns of the GSRs indicate potentially salient and coherent sentences (7 and 8 here) that present less inference load with respect to a query like "Baghdad attacks".



↓**SentenceIDs || Words... →**					
sIDs	protect	attacks	churches	Baghdad	Mosul
6	--	--	--	--	--
7	--	subj	obj	wn	--
8	--	vn	wn	wn	wn
9	--	subj	--	--	--
10	--	--	subj	wn	--

6	The major Christian groups include Chaldean - Assyrians, who make up Kana's group, and Armenians.
7	On Oct. 16, bomb **attacks** targeted five **churches** in **Baghdad** which damaged buildings but caused no casualties.
8	Officials estimate that as many as 15,000 of Iraq's nearly one million Christians have left the country since August, when four **churches** in **Baghdad** and one in **Mosul** were **attacked** in a coordinated series of car bombings.
9	The **attacks** killed 12 people and injured 61 others.
10	Another **church** was bombed in **Baghdad** in September.

Figure 1: (a) Left: Sentence IDs and the GSRs of the words in them (b) Right: The corresponding sentences

Note that the count for the GSRt "wn→ $--$" for sentenceID 8 is 3 from this snapshot. Inputs to our model are document specific word ID counts and document-sentence specific GSRt ID counts.

3 The Proposed Method

To describe the document generation process under our proposed "Learning To Summarize" (henceforth LeToS), we assume that there are K latent topics and T topic-coupled GSRts associated with each document; r_t is the observed GSRt, w_n is the observed word and s_p is the observed sentence. Denote $\boldsymbol{\theta}_k - \boldsymbol{\alpha}_k$ to be the expected number of GSRts per topic; $\boldsymbol{\pi}_t - \boldsymbol{\eta}_t$ to be the expected number of words and sentences per topic-coupled GSRt in each document. Further denote, z_t to be a K dimensional indicator for $\boldsymbol{\theta}$, v_p be the T dimensional indicator for $\boldsymbol{\pi}$ and y_n is an indicator for the same topic-coupled GSRt proportions as v_p, each time a word w_n is associated with a particular sentence s_p. At the parameter level, each topic is a multinomial $\boldsymbol{\beta_k}$ over the vocabulary V of words and each topic is also a multinomial $\boldsymbol{\rho_k}$ over the GSRts following the implicit relation of GSRts to words within sentence windows. Each topic-coupled GSRt is also treated as a multinomial $\boldsymbol{\Omega_t}$ over the total number U of sentences in the corpus. $\delta(w_n \in s_p)$ is the delta function which is 1 iff the n^{th} word belong to the p^{th} sentence. The document generation process is shown in Fig. 3 and is explained as a pseudocode in Fig. 2.

The model can be viewed as a generative process that first generates the GSRts and subsequently generates the words that describe the GSRt and hence an utterance unit (a sentence in this model). For each document, we first generate 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 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 topic-coupled proportions are used. Finally the sentences are sampled from $\boldsymbol{\Omega}_t$ by choosing a GSRt proportion that is coupled to the factor that generates r_t through the constituent 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

```
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 $w_n|y_n = t, \mathbf{z}, \boldsymbol{\beta} \sim Mult(\boldsymbol{\beta}_{z_{y_n}})$
```

Figure 2: Document generation process of LeToS

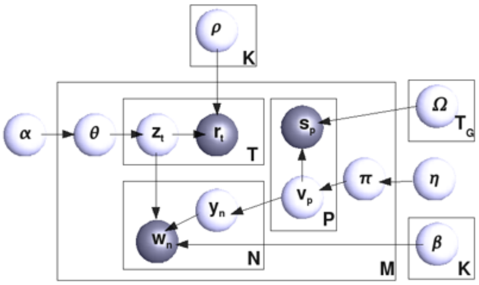


Figure 3: Graphical model representation of LeToS

dominant trend in the transition topic. This fact is enforced through the coupling of empirical GSRt proportions to topics of the sentential words.

3.1 Parameter Estimation and Inference

In this paper we have resorted to mean field variational inference [2] 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 approximate factorized mean field distribution to the posterior distribution of the latent variables given the data. In the variational setting, for each document we have $\sum_{k=1}^{K} \phi_{tk} = 1$, $\sum_{t=1}^{T} \lambda_{nt} = 1$ and $\sum_{t=1}^{T} \zeta_{pt} = 1$ and the approximating distribution is factorized as:

$$q(\boldsymbol{\theta}, \boldsymbol{\pi}, \mathbf{z}, \mathbf{y}, \mathbf{v}|\gamma, \boldsymbol{\chi}, \boldsymbol{\phi}, \boldsymbol{\lambda}, \boldsymbol{\zeta}) = q(\boldsymbol{\theta}|\gamma)q(\boldsymbol{\pi}|\boldsymbol{\chi}) \prod_{t=1}^{T} q(z_t|\boldsymbol{\phi_t}) \prod_{n=1}^{N} q(y_n|\boldsymbol{\lambda_n}) \prod_{p=1}^{P} q(v_p|\boldsymbol{\zeta_p}) \tag{1}$$

The variational functional to optimize can be shown to be

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

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

The maximum likelihood estimations of these indicator variables for the topics and the topic-coupled GSRts are as follows:

$$\begin{aligned}
&\gamma_i = \alpha_i + \textstyle\sum_{t=1}^{T_d} \phi_{ti}\,; \qquad \chi_t = \eta_t + \sum_{n=1}^{N_d} \lambda_{nt} + \sum_{p=1}^{P_d} \zeta_{pt} \\
&\lambda_{nt} \propto \exp\{(\Psi(\chi_t) - \Psi(\textstyle\sum_{f=1}^{T} \chi_f)) + (\sum_{i=1}^{K} \phi_{ti} \log \beta_{z_{(y_n=t)}=i,n})\} \\
&\phi_{ti} \propto \exp\{\log \rho_{it} + (\Psi(\gamma_i) - \Psi(\textstyle\sum_{k=1}^{K} \gamma_k)) + (\sum_{n=1}^{N_d} \lambda_{nt} \log \beta_{z_{(y_n=t)}=i,n})\} \\
&\zeta_{pt} \propto \Omega_{pt} \exp\{\Psi(\chi_t) - \Psi(\textstyle\sum_{j=1}^{T} \chi_j)\}
\end{aligned}$$

We now write 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:

$$\begin{aligned}
&\rho_{ig} \propto \textstyle\sum_{d=1}^{M} \sum_{t=1}^{T_d} \phi_{dti} r_{dt}^{g}\,; \qquad \beta_{ij} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} (\sum_{t=1}^{T_d} \lambda_{nt}\phi_{ti}) w_{dn}^{j}\,; \\
&\Omega_{tu} \propto \textstyle\sum_{d=1}^{M} \sum_{p=1}^{P_d} \zeta_{dpt} s_{dp}^{u}
\end{aligned}$$

where r_{dt}^{g} is 1 iff $t = g$ and 0 otherwise with g as an index variable for all possible GSRts; u is an index into one of the U sentences in the corpus and $s_{dp}^{u} = 1$ if the p^{th} sentence in document d is one among U. $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$ can be optimized as mentioned in [2].

For obtaining summaries, we order sentences w.r.t query words by computing the following:

$$p(s_{dp}|\mathbf{w_q}) \propto \sum_{l=1}^{Q} (\sum_{t=1}^{T} \sum_{i=1}^{K} \zeta_{dpt}\phi_{dti}(\lambda_{dlt}\phi_{dti})\gamma_{di}\chi_{dt})\delta(w_l \in s_{dp}) \tag{3}$$

where Q is the number of the query words 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; and w_l is the l^{th} query word. Further, the sentences are scored over only “rich” GSRts which lack any “$-- \rightarrow --$” transitions whenever possible. We also expand the query by a few words while summarizing in real time using topic inference on the relevant set of documents. This is done to better “understand” the query.

4 Results and Discussions

Tables 1 and 2 show some topics learnt from the TAC2009 dataset (http://www.nist.gov/tac/2009/Summarization/index.html) . From table 2, we observe that the topics under both models are the same qualitatively. Moreover, it has been observed that constraining LeToS to words and GSRts as the only observed variables shows *lower* word perplexity than LDA on heldout test data. Empirically, it has been seen though that the time complexity for LeToS is sligthly higher than LDA due to the extra iterations over the GSRts and sentences.

topic16	**topic36**	**topic38**	**topic22**
Kozlowski	bombings	solar	Hurricane
million	Malik	energy	Rita
Tyco	Sikhs	power	evacuations
company	Bagri	BP	Texas
trial	India	company	Louisiana
Swartz	case	year	area
loans	killed	panel	state

Table 1: Some topics under LDA for TAC2009

topic58	**topic1**	**topic42**	**topic28**
Kozlowski	Malik	solar	Hurricane
Tyco	bombs	energy	Rita
million	India	power	evacuated
company	Sikh	electricity	storms
loan	killing	systems	Texas
trial	Flight	government	Louisiana
Swartz	Bagri	production	area

Table 2: Some topics under LeToS for TAC2009

For TAC2009, using the more meaningful Pyramid [4] scoring for summaries, the average Pyramid scores for very short 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. The scores for a state-of-the-art summarization system [5] 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.

5 Conclusion

Overall, we have integrated centering theory based coherence into topic model. Models like LeToS tend to capture “what is being discussed” by selecting sentences that have low reader “inference load”. On the other hand, the model gets penalized if the summaries need to be very factual. This could probably be avoided by defining finer GSR categories such as named entities. Another drawback of the model is its lack of understanding the **meaning** of the query. However, generating specific summaries w.r.t. an information need using topic modeling is akin to answering natural language questions. That problem is hard, albeit an open one under the topic modeling umbrella.

References

[1] 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.

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

[3] 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.

[4] Aaron Harnly, Ani Nenkova, Rebecca Passonneau, and Owen Rambow. Automation of summary evaluation by the pyramid method. In *Recent Advances in Natural Language Processing (RANLP)*, 2005.

[5] 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.