Coherence Based Topic Model for Summarization

Pradipto Das and Rohini Srihari

University at Buffalo, State University of New York

November 16, 2009
TAC 2009, Update Summarization Task

Part I

Motivation

Looking at TAC in a different way

Lets say we have:

- Bunch of docs (let's assume no folder structure)
- Bunch of queries and pointers to a few relevant documents
- So, we kind of "opened up" the TAC dataset a little bit

Having done that, lets do this:

- Try to rank sentences under some model w.r.t queries
- Rank probabilistically so that both words and sentences have some generative aspect
- Once a model is built, quickly return a *multi-document* summary given *any query* - just like search
- **TAC gives us an excellent opportunity to evaluate and get feedback for new models on summarization**

Looking at TAC in a different way

Lets say we have:

- Bunch of docs (let's assume no folder structure)
- Bunch of queries and pointers to a few relevant documents
- So, we kind of "opened up" the TAC dataset a little bit

Having done that, lets do this:

- Try to rank sentences under some model w.r.t queries
- Rank probabilistically so that both words and sentences have some generative aspect
- Once a model is built, quickly return a *multi-document* summary given *any query* - just like search
- TAC gives us an excellent opportunity to evaluate and get feedback for new models on summarization

Looking at TAC in a different way

Lets say we have:

- Bunch of docs (let's assume no folder structure)
- Bunch of queries and pointers to a few relevant documents
- So, we kind of "opened up" the TAC dataset a little bit

Having done that, lets do this:

- Try to rank sentences under some model w.r.t queries
- Rank probabilistically so that both words and sentences have some generative aspect
- Once a model is built, quickly return a *multi-document* summary given *any query* - just like search
- **TAC gives us an excellent opportunity to evaluate and get feedback for new models on summarization**

Looking at TAC in a different way

Lets say we have:

- Bunch of docs (let's assume no folder structure)
- Bunch of queries and pointers to a few relevant documents
- So, we kind of "opened up" the TAC dataset a little bit

Having done that, lets do this:

- Try to rank sentences under some model w.r.t queries
- Rank probabilistically so that both words and sentences have some generative aspect
- Once a model is built, quickly return a *multi-document* summary given *any query* - just like search
- **TAC gives us an excellent opportunity to evaluate and get feedback for new models on summarization**

Looking at TAC in a different way

Lets say we have:

- Bunch of docs (let's assume no folder structure)
- Bunch of queries and pointers to a few relevant documents
- So, we kind of "opened up" the TAC dataset a little bit

Having done that, lets do this:

- Try to rank sentences under some model w.r.t queries
- Rank probabilistically so that both words and sentences have some generative aspect
- Once a model is built, quickly return a *multi-document* summary given *any query* - just like search
- **TAC gives us an excellent opportunity to evaluate and get feedback for new models on summarization**

Looking at TAC in a different way

Lets say we have:

- Bunch of docs (let's assume no folder structure)
- Bunch of queries and pointers to a few relevant documents
- So, we kind of "opened up" the TAC dataset a little bit

Having done that, lets do this:

- Try to rank sentences under some model w.r.t queries
- Rank probabilistically so that both words and sentences have some generative aspect
- Once a model is built, quickly return a *multi-document* summary given *any query* - just like search
- **TAC gives us an excellent opportunity to evaluate and get feedback for new models on summarization**

Looking at TAC in a different way

LDA - A Basic Topic Model

Topic models do a great job in "thematically" structuring unstructred "opened-up" datasets!

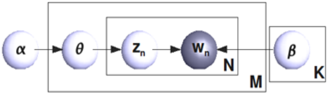

Figure 1: graphical model for Latent Dirichlet Allocation (LDA)

LDA assumes that an N-word document *d* arises from the following generative process:

Draw $\boldsymbol{\theta} \mid \boldsymbol{\alpha} \sim \mathit{Dirichlet}(\boldsymbol{\alpha})$

For each position $n \in \{1, ..., N\}$:

Draw topic assignment $z_n \mid \boldsymbol{\theta} \sim \mathit{Mult}(\boldsymbol{\theta})$

Draw word $w_n \mid \{z_n, \boldsymbol{\beta}_{1:K}\} \sim \mathit{Mult}(\boldsymbol{\beta}_{z_n})$

The Summarization Problem

Imagine the following scenario:

- You are preparing for a particular exam and you don't know what questions are going to come
- **So firstly, you do a "mental theme structuring" of the materials**
- **Secondly, you want to memorize some key "utterances" corresponding to the themes e.g. sentences that are "most easy to understand"**
- Finally, when you get a question, you want to do the following:
  - **Figure out the specific theme** that the question demands
  - **Choose some relevant utterances** based on the text corresponding to the theme
  - Lastly, **expand** on these utterances **by choosing words that are discussed often in the theme**

The Summarization Problem

Imagine the following scenario:

- You are preparing for a particular exam and you don't know what questions are going to come
- **So firstly, you do a "mental theme structuring" of the materials**
- **Secondly, you want to memorize some key "utterances" corresponding to the themes e.g. sentences that are "most easy to understand"**
- Finally, when you get a question, you want to do the following:
  - **Figure out the specific theme** that the question demands
  - **Choose some relevant utterances** based on the text corresponding to the theme
  - Lastly, **expand** on these utterances **by choosing words that are discussed often in the theme**

The Summarization Problem

Imagine the following scenario:

- You are preparing for a particular exam and you don't know what questions are going to come
- **So firstly, you do a "mental theme structuring" of the materials**
- **Secondly, you want to memorize some key "utterances" corresponding to the themes e.g. sentences that are "most easy to understand"**
- Finally, when you get a question, you want to do the following:
  - **Figure out the specific theme** that the question demands
  - **Choose some relevant utterances** based on the text corresponding to the theme
  - Lastly, **expand** on these utterances **by choosing words that are discussed often in the theme**

The Summarization Problem

Imagine the following scenario:

- You are preparing for a particular exam and you don't know what questions are going to come
- **So firstly, you do a "mental theme structuring" of the materials**
- **Secondly, you want to memorize some key "utterances" corresponding to the themes e.g. sentences that are "most easy to understand"**
- Finally, when you get a question, you want to do the following:
  - **Figure out the specific theme** that the question demands
  - **Choose some relevant utterances** based on the text corresponding to the theme
  - Lastly, **expand** on these utterances **by choosing words that are discussed often in the theme**

The Summarization Problem

Imagine the following scenario:

- You are preparing for a particular exam and you don't know what questions are going to come
- **So firstly, you do a "mental theme structuring" of the materials**
- **Secondly, you want to memorize some key "utterances" corresponding to the themes e.g. sentences that are "most easy to understand"**
- Finally, when you get a question, you want to do the following:
  - **Figure out the specific theme** that the question demands
  - **Choose some relevant utterances** based on the text corresponding to the theme
  - Lastly, **expand** on these utterances **by choosing words that are discussed often in the theme**

The Summarization Problem

Imagine the following scenario:

- You are preparing for a particular exam and you don't know what questions are going to come
- **So firstly, you do a "mental theme structuring" of the materials**
- **Secondly, you want to memorize some key "utterances" corresponding to the themes e.g. sentences that are "most easy to understand"**
- Finally, when you get a question, you want to do the following:
  - **Figure out the specific theme** that the question demands
  - **Choose some relevant utterances** based on the text corresponding to the theme
  - Lastly, **expand** on these utterances **by choosing words that are discussed often in the theme**

The Summarization Problem

Imagine the following scenario:

- You are preparing for a particular exam and you don't know what questions are going to come
- **So firstly, you do a "mental theme structuring" of the materials**
- **Secondly, you want to memorize some key "utterances" corresponding to the themes e.g. sentences that are "most easy to understand"**
- Finally, when you get a question, you want to do the following:
  - **Figure out the specific theme** that the question demands
  - **Choose some relevant utterances** based on the text corresponding to the theme
  - Lastly, **expand** on these utterances **by choosing words that are discussed often in the theme**

A Key Challenge

How do we formalize "sentences" that "are most easy to understand?"

- **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. Her mother taught her to cook chicken potpies for supper.
  2. Tom had been abusing her for many years.
  3. The refrigerator motor started making noise as it's door was left open.
  4. Tom was shot as he was preparing to have supper.

What could be a sentence that is very easy to remember?
"Martha couldn't take it anymore." or "She was abused by Tom for many years."

- **Discourse 2 - Attention is digressing!**

Borrowing Centering from Linguistics

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 2a**
  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.

Contextual Utterances and Attentional State - an example

• Discourse 1 is an example where the **focus of attention is clearly on Martha**
• If we observe discourse 2a, there could be a **retention of attention from Tom to Tom**
• **If**, *however,* **in the first utterance** *the* **focus of attention be Martha**, *then there is a* **focus shift in the next utterance**
• **Discourse 2a is thus less coherent** than discourse 1 in terms of the effort to understand the discourse i.e. <u>discourse 1 has less **inference load**</u>

GSRts (Grammatical & Semantic Role transitions)

Contextual Utterances and GSRts

• we attributed words in a sentence with G GSRs like subjects, objects, concepts from WordNet synset role assignments(wn) [e.g. **WNdevice, WNdisease, WNdrug**], adjectives, VerbNet thematic role assignment(vn) [e.g., **VNcleaning, VNcovering, VNexamininig**], adverbs and "other" (if the feature of the word doesn't fall into the previous GSR categories)

• Further if a word in a sentences is identified with 2 or more features, only one feature is chosen based on the left to right descending priority of the G tags mentioned

• Thus in a window of sentences, there are potentially $(G+1)^2$ GSRts for a total of G GSRs with the additional GSR representing a null role (denoted by "−−") as in the word is not found in the contextual sentence

we can construct a **sentence-term matrix for a document** with these GSRs as shown in Table 1

GSRts (Grammatical & Semantic Role transitions)

Contextual Utterances and GSRts

• we attributed words in a sentence with G GSRs like subjects, objects, concepts from WordNet synset role assignments(wn) [e.g. **WNdevice, WNdisease, WNdrug**], adjectives, VerbNet thematic role assignment(vn) [e.g., **VNcleaning, VNcovering, VNexamininig**], adverbs and "other" (if the feature of the word doesn't fall into the previous GSR categories)

• Further if a word in a sentences is identified with 2 or more features, only one feature is chosen based on the left to right descending priority of the G tags mentioned

• Thus in a window of sentences, there are potentially $(G+1)^2$ GSRts for a total of G GSRs with the additional GSR representing a null role (denoted by "−−") as in the word is not found in the contextual sentence

we can construct a **sentence-term matrix for a document** with these GSRs as shown in Table 1

GSRts (Grammatical & Semantic Role transitions)

Contextual Utterances and GSRts

• we attributed words in a sentence with G GSRs like subjects, objects, concepts from WordNet synset role assignments(wn) [e.g. **WNdevice, WNdisease, WNdrug**], adjectives, VerbNet thematic role assignment(vn) [e.g., **VNcleaning, VNcovering, VNexamininig**], adverbs and "other" (if the feature of the word doesn't fall into the previous GSR categories)

• Further if a word in a sentences is identified with 2 or more features, only one feature is chosen based on the left to right descending priority of the G tags mentioned

• Thus in a window of sentences, there are potentially $(G+1)^2$ GSRts for a total of G GSRs with the additional GSR representing a null role (denoted by "$--$") as in the word is not found in the contextual sentence

we can construct a **sentence-term matrix for a document** with these GSRs as shown in Table 1

GSRts (Grammatical & Semantic Role transitions)

Contextual Utterances and GSRts

• we attributed words in a sentence with G GSRs like subjects, objects, concepts from WordNet synset role assignments(wn) [e.g. **WNdevice, WNdisease, WNdrug**], adjectives, VerbNet thematic role assignment(vn) [e.g., **VNcleaning, VNcovering, VNexamininig**], adverbs and "other" (if the feature of the word doesn't fall into the previous GSR categories)

• Further if a word in a sentences is identified with 2 or more features, only one feature is chosen based on the left to right descending priority of the G tags mentioned

• Thus in a window of sentences, there are potentially $(G+1)^2$ GSRts for a total of G GSRs with the additional GSR representing a null role (denoted by "$--$") as in the word is not found in the contextual sentence

we can construct a **sentence-term matrix for a document** with these GSRs as shown in Table 1

Understanding Coherence

A sample set of sentences

- (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.

Table 1: Sentence Term GSR grid view of document - APW_ENG_20041204.0231 of D0808B from TAC2008

↓SentenceIDs ‖ words... →					
sID	protect	attacks	churches	Baghdad	Mosul
6	−−	−−	−−	−−	−−
7	−−	subj	obj	wn	−−
8	−−	vn	wn	wn	wn
9	−−	subj	−−	−−	−−
10	−−	−−	subj	wn	−−

Borrowing Centering from Linguistics

An **"utterance"** is *an uttering of a sequence of words (usually meaningful) at a certain point in the discourse.*

Contextual Utterances and Attentional state

- 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
- **Attentional state** contains information about the objects, properties, relations, and discourse intentions that are most salient at any given point
    - Speaking loosely, attentional state models the discourse participants' foci of attention at any given point in the discourse.
- **Essentially, then the centers of utterances help identify the attentional state**

Summarizing

Key facts

- Attentional state models the relationship between foci to centers of utterances
- Centers are semantic objects, not necessarily words, phrases, or syntactic forms
- The most coherent parts of the discourse are the one that comprise of utterances that have a propagation of the foci of attention
- **The main goal of centering theory is to formulate constraints on the centers of utterances so as to maximize coherence**

Summarizing

Key facts

- Attentional state models the relationship between foci to centers of utterances
- Centers are semantic objects, not necessarily words, phrases, or syntactic forms
- The most coherent parts of the discourse are the one that comprise of utterances that have a propagation of the foci of attention
- **The main goal of centering theory is to formulate constraints on the centers of utterances so as to maximize coherence**

Summarizing

Key facts

- Attentional state models the relationship between foci to centers of utterances
- Centers are semantic objects, not necessarily words, phrases, or syntactic forms
- The most coherent parts of the discourse are the one that comprise of utterances that have a propagation of the foci of attention
- **The main goal of centering theory is to formulate constraints on the centers of utterances so as to maximize coherence**

Summarizing

Key facts

- Attentional state models the relationship between foci to centers of utterances
- Centers are semantic objects, not necessarily words, phrases, or syntactic forms
- The most coherent parts of the discourse are the one that comprise of utterances that have a propagation of the foci of attention
- **The main goal of centering theory is to formulate constraints on the centers of utterances so as to maximize coherence**

Summarizing

Key facts

- Attentional state models the relationship between foci to centers of utterances
- Centers are semantic objects, not necessarily words, phrases, or syntactic forms
- The most coherent parts of the discourse are the one that comprise of utterances that have a propagation of the foci of attention
- **The main goal of centering theory is to formulate constraints on the centers of utterances so as to maximize coherence**

Part II

The proposed method

Extending LDA to include GSRts

- For each document, first generate T_G GSRts using a simple LDA model
- 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 proportions are used
- $\boldsymbol{\pi}$ is topic-coupled latent factor for the empirical proportion of the GSRts

The document generation process is shown in Fig. 2:

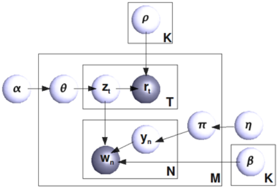

Figure 2: graphical model for the extended LDA

For each document $d \in 1, \ldots, M$
  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 $n \in \{1, \ldots, N_d\}$ is the number of words in a document $d \in \{1, \ldots, M\}$

Adding Sentences as Observed Variables

in our LeToS (Learning To Summarize) model, additionally, we have:

- The sentences are sampled from $\mathbf{\Omega}_t$ by choosing a GSRt proportion that is coupled to the factor that generates r_t through the observed w_n
- Strong coupling of $\boldsymbol{\pi}$ to $\boldsymbol{\theta}$ through the observed w_n enforces the correspondence between the GSRts, words and sentences
- The coupling tries to group together most salient sentences interms of coherence determined from an holistic view of the corpus

The document generation process is shown in Fig. 3:

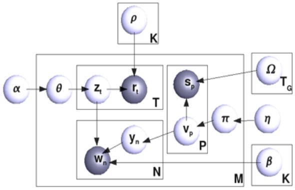

Figure 3: graphical model for LeToS

For each document $d \in 1, \ldots, M$
  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 $s_p \sim Mult(\mathbf{\Omega}_{v_p})$
      Choose a word $w_n|y_n = t, \mathbf{z}, \boldsymbol{\beta} \sim Mult(\boldsymbol{\beta}_{z_{y_n}})$

where $p \in \{1, \ldots, P_d\}$ is the number of sentences in a document $d \in \{1, \ldots, M\}$; T_G is the number of all possible GSRts in the corpus.

Summaries from LeToS

Scoring sentences

- Iterate over the data to fit the parameters
- For each query and a set of relevant docs, infer the hidden variables very quickly using Variational Bayes
- Score sentences as, $p(s_{dp}|\mathbf{w_q}) \propto \sum_{l=1}^{Q}(\sum_{t=1}^{T}\sum_{i=1}^{K}\zeta_{dpt}\phi_{dti}(\lambda_{dlt}\phi_{dti})\gamma_{di}\chi_{dt})\delta(w_l \in s_{dp})$
- Further, the sentences are scored over only "rich" GSRts which lack the "$-- \rightarrow --$" transition

Key observation

- Sentences are selected in the vicinity of highly factual sentences - topic based query expansion proved helpful
- An important future direction will be to incorporate entity topic contexts that affects selection of sentences

Part III

Results and Conclusions

Topics under LDA and LeToS

- "Follow the **legal proceedings in the criminal trials of Tyco Chief Dennis Kozlowski**"
- "Trace developments in the **trial of Ripudaman Singh Malik and Ajaib Singh Bagri**, two suspects in the Air India bombings"
- "Describe **preparations for Hurricane Rita and estimates of its strength**"

Table 2: Some most probable words under LDA topics for TAC2009

topic16	topic36	topic22
Kozlowski	bombings	Hurricane
million	Malik	Rita
Tyco	Sikhs	evacuations
company	Bagri	Texas
trial	India	Louisiana
Swartz	case	area
loans	killed	state
charges	year	category
Chief	flight	Katrina
prosecutor	witnesses	Wednesday

Table 3: Some most probable words under LeToS topics for TAC2009

topic58	topic1	topic28
Kozlowski	Malik	Hurricane
Tyco	bombs	Rita
million	India	evacuated
company	Sikh	storms
loan	killing	Texas
trial	Flight	Louisiana
Swartz	Bagri	area
employees	Air_India	Katrina
Prosecutors	murder	Gulf
bonuses	trial	Wednesday

A Summary Sentence Validating Coherence

- An example summary sentence from folder D0906B-A of TAC2009 "A" timeline is: "A fourth day of thrashing thunderstorms began to take a heavier toll on **southern California** on Sunday with at least three **deaths** blamed on the **rain**, as flooding and **mudslides** forced road closures and emergency crews carried out harrowing rescue operations."
- The **next two contextual sentences** in the document of the previous sentence are: *"In Elysian_Park, just north of downtown, a 42-year-old homeless man was* ***killed*** *and another injured when a* ***mudslide*** *swept away their makeshift encampment."* **AND** *"Another man was* ***killed*** *on Pacific_Coast_Highway in Malibu when his sport utility* ***vehicle*** *skidded into a* ***mud*** *patch and plunged into the Pacific Ocean."*.

▷ If the query is *"Describe the effects and responses to the heavy* **rainfall** *and* **mudslides** *in* **Southern California**," **observe the focus of attention on mudslides as subject in the first two sentences in Table. 4**

southern	California	mudslides	mud	rain	man	vehicle	deaths	killed
nn	*wn*	*subj*	−−	*wn*	−−	−−	*wn*	−−
−−	−−	*subj*	−−	−−	*subj*	−−	−−	*vb*
−−	−−	−−	*wn*	−−	*subj*	*subj*	−−	*vb*

Table 4: sentence-GSR grid for a sample summary document slice

Overall Pyramid Performance (Corrected)

Overall Pyramid scores for LeToS Summaries for TAC2009
A timeline Pyramid scores Average: 0.3024 (Modified Rank = 13^{th} of 52)
B timeline Pyramid scores Average: 0.2601 (Modified Rank = 9^{th} of 52)

Overall Pyramid scores for Baseline Summaries[1] for TAC2009
A timeline Pyramid scores Average: 0.175
B timeline Pyramid scores Average: 0.160

[1] Baseline returns the first 100 words from the most recent document

Part IV

The End

That's all folks - Thanks!

Questions?