Exploring Differential Topic Models for Comparative Summarization of Scientific Papers

Lei He[*#], Wei Li[*], Hai Zhuge[*#]
[#] System Analytics Research Institute, Aston University, Birmingham, UK
[*] Key Lab of Intelligent Information Processing, ICT, CAS, University of Chinese Academy of Sciences, Beijing, China
hel2@aston.ac.uk,weili.ict.kg@gmail.com,zhuge@ict.ac.cn

Abstract

This paper investigates differential topic models (*dTM*) for summarizing the differences among document groups. Starting from a simple probabilistic generative model, we propose *dTM-SAGE* that explicitly models the deviations on group-specific word distributions to indicate how words are used differentially across different document groups from a background word distribution. It is more effective to capture unique characteristics for comparing document groups. To generate *dTM*-based comparative summaries, we propose two sentence scoring methods for measuring the sentence discriminative capacity. Experimental results on scientific papers dataset show that our *dTM*-based comparative summarization methods significantly outperform the generic baselines and the state-of-the-art comparative summarization methods under ROUGE metrics.

1 Introduction

Today, the interconnected nature of real-world applications brings more cross-field research problems leading to a much closer relationship between research areas. Real-world challenges require researchers to quickly get acquainted with knowledge in other areas. For example, imagine a researcher who is familiar with topic models wants to extend her research to opinion summarization. She would be more interested in finding out the current development of sentiment analysis and how topic models can be used in sentiment analysis, rather than the common background knowledge such as topic models and basic NLP technologies. Such a real-world demand encourages the study of multi-document comparative summarization for scientific papers in multiple subject areas. This paper presents the initial study on this problem.

Comparative summarization aims at summarizing the differences among document groups (Wang et al., 2012). The core is to compare different topics and find unique characteristics for each document group. The main motivation of this paper is to apply *dTM* to comparative summarization and to model the group-specific topics to capture the unique word usage for characterising documents in the same group. To our best knowledge, there is no previous study providing in-depth model analysis and detailed experimental results on *dTM* applied for comparative summarization.

We first propose a probabilistic generative model *dTM-Dirichlet* to model the group-specific word distributions to capture the unique word usage for each document group. However, *dTM-Dirichlet* is not a truly differential topic model and it suffers from the problems of high inference cost, over-parameterization and lack of sparsity. Evolving from the idea of *SAGE* (Eisenstein et al., 2011), we develop *dTM-SAGE* to make the word probability distributions for each document group to share a common background word distribution and explicitly models how words are used differently in each group from the background word distribution.

Our main contributions include the following two points: (1) we propose *dTM* to capture unique characteristics of each document group in the application background of comparative summarization for cross-area scientific papers; and (2) we propose two sentence scoring methods to measure the sen-

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 1028–1038, Osaka, Japan, December 11-17 2016.

tence discriminative capacity and a greedy sentence selection method to automatically generate summary for *dTM*-based comparative summarization.

2 Related Work

Multi-document Summarization. Existing multi-document summarization can be either extractive or abstractive (Sekine and Nobata, 2003). Our work focuses on the extractive techniques which involve in assigning saliency scores to sentences and extracting high-scored sentences in a greedy manner to construct a summary (Wan et al., 2007; Cai et al., 2010; Gupta and Lehal, 2010; Celikyilmaz and Hakkani-Tur, 2011). Graph-based ranking techniques such as TextRank (Mihalcea and Tarau, 2004) and LexPageRank (ErKan and Radev, 2004) have been widely used in extractive summarization. A bigram based supervised method was proposed for extractive summarization in ILP framework (Li et al., 2013; Li, 2015). Jha et al. (2015) proposed an extractive algorithm that combines a content model with a discourse model to generate coherent summaries for scientific articles. A multi-dimensional summarization methodology was proposed to transform the paradigm of traditional summarization research through multi-disciplinary fundamental exploration on semantics, dimension, knowledge, computing and cyber-physical society (Zhuge, 2016).

Comparative Summarization. Unlike the generic summarization that summarizes the common information in document collection, the comparative summarization aims to summarize the differences among document groups. Wang et al. (2012) proposed a discriminative sentence selection method to generate summary by selecting sentences in a greedy manner to minimize the generalized variance of a covariance matrix using a multivariate normal model. Shen and Li (2010) proposed a method by building the sentence graph for each document group and extracting a complementary minimum dominating set on each graph to form a discriminative summary.

Update Summarization. The most similar task to comparative summarization is update summarization, which aims to detect and summarize novel information in a document set B under the assumption that users have already learnt the documents in set A, where documents in A chronologically precede the documents in B. The update summarization has been well studied. Most existing methods solve it as a redundancy removal problem by adding functionality to remove redundant sentences using filtering rules (Fisher and Roark, 2008), Maximal Marginal Relevance (Boudin et al., 2008), or graph-based algorithms (Shen and Li, 2010; Li et al., 2008).

More related to this paper is the work of a topic-model based update summarization approach *DualSum* (Delort and Alfonseca, 2012), which learns a general background distribution across the corpus and a document-specific distribution for each document, but also learns two collection-specific distributions for each pair of update collection and base collection: the joint topic distribution and the update topic distribution. This paper revises *DualSum* as a baseline for evaluation in Section 5.2.

Topic Models for Documents Comparison. The other type of related work is the comparison of documents. Most existing studies for this goal focus on topic models to discover common and specific themes among document collections, referred to as cross-collection topic models (Paul, 2009). This idea was first explored with an initial topic model *PLSI* (Zhai et al., 2004), and later improved with *LDA* topic model (Blei, 2012; Pual, 2009) which inspires our *dTM-Dirichlet*. There are a number of real-world applications extending cross-collection topic models in different scenarios (Ahmed and Xing, 2010; Li et al., 2011). For example, Paul and Girju (2009) employed cross-collection *LDA* (*cc-LDA*) for cross-cultural analysis of blogs and forums and later they proposed a two-dimensional topic-aspect model (*TAM*) to jointly discover topics and aspects in scientific literature (Paul and Girju, 2010). The common idea behind these cross-collection topic models is that using latent topics capture the common and unique word usage among document collections. Cross-collection topic models neglect the correlations between each collection-specific topic and the common background topic, thus make it insufficient to capture differential word usage. More importantly, the correlations are the essence of the differential topic models.

3 Differential Topic Models

In this section, the differential topic models are explored for comparative summarization. We first develop a simple probabilistic generative model, *dTM-Dirichlet*. Evolved from *dTM-Dirichlet*, *dTM-SAGE* is developed by modelling the correlations as additive relation between the group-specific devi-

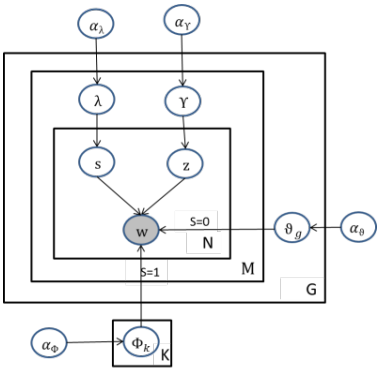

Fig.1 dTM-Dirichlet Model Graph Representation.

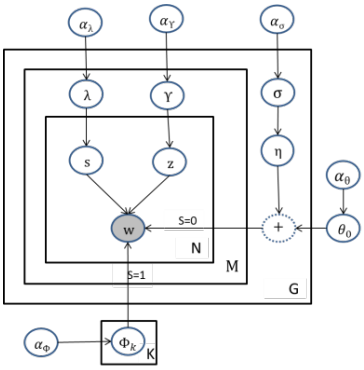

Fig.2 dTM-SAGE Model Graph Representation.

ations and a background word distribution, which enables to capture more salient group-specific words and bypass the problems of high inference cost, over-parameterization and lack of sparsity.

To illustrate *dTM*, we first define some notations to express a document corpus *C*. Let *G* be the number of groups in the corpus, M_g be the number of papers in group g and $N_{g,m}$ be the number of words in paper *m*. A word $w_{g,m,n}$ representing the n^{th} word in paper *m* of group g is a discrete observed variable, defined to be an item in the vocabulary list of the whole corpus.

3.1 *dTM-Dirichlet* Model

dTM-Dirichlet model is a simplified version of cross-collection LDA (*ccLDA*) (Paul and Girju, 2009) for comparing multiple text collections. *dTM-Dirichlet* builds two types of word model. One is for each document group g, in which there is a group-specific content word model ϑ_g that emits discriminative words for the group. The other type is a superset of group-independent word models φ_k that generates either background words shared by all document groups or salient words occurring in several documents of different groups. Reconsidering the example in section 1, the group-independent word model represents two classes of words, i.e. the background words like topic model that are shared by almost all papers; and the salient words like NP chunk and dependency parsing that only occur in several papers of different groups.

We focus on the group-specific word model for comparative summarization. Since background words and salient words provide no group-specific knowledge, they are not distinguished in *dTM-Dirichlet*. Following probabilistic topic models, we assume that word models φ_k and ϑ_g are multinomial distributions over words, drawn from uniform *Dirichlet* distribution (*Dir*) with priors α_φ and α_ϑ.

As shown in Fig. 1, dTM-Dirichlet associates each document a topic distribution $\gamma \sim Dir\ (\alpha_\gamma)$, and the topic assignment variable *z* for each word in the document thus can be multinomially sampled from γ. Besides a topic variable *z*, each word is also assigned with a binary variable *s* that indicates whether the word is a group-independent topic word (*s*=1) or a group-specific content word (*s*=0). Each document has a group-specific word controller $\lambda \sim Beta(\alpha_\lambda)$, which reflects the proportion of group-specific content in a document. s is sampled from a Bernoulli test with the probability of λ.

Formally, the generative process of dTM-Dirichlet model for a corpus *C* divided into *G* document groups is shown in Table 1. When $s_{g,m,n} = 1$, the sample of the group-independent topic word is identical to *LDA*. When $s_{g,m,n} = 0$, the sample of the word $w_{g,m,n}$ is independent from the document's topic distribution $\gamma_{g,m}$ and directly drawn from the group-specific content word distribution θ_g.

dTM-Dirichlet models group-specific word distributions to capture the differential lexicon usage of document groups. However, *dTM-Dirichlet* is not a truly differential topic model, which requires the development of *dTM-SAGE* for comparative summarization.

3.2 *dTM-SAGE* Model

When generating topics for multiple document collections, *LDA*-style generative models associate a multinomial distribution with each document group, which is the same as how we model the group-specific content words in *dTM-Dirichlet* model.

In contrast, *SAGE* (Sparse Additive Generative model) (Eisenstein et al., 2011) provides an alternative way to *LDA* by endowing each document group with a model of the deviation in log-frequency

The generative process of dTM-Dirichlet	The generative process of dTM-SAGE
1. For each topic k, where $1 \le k \le K$ a. Draw $\Phi_k \sim Dir(\alpha_\Phi)$ 2. For each document group *g*, where $1 \le g \le G$ a. Draw $\theta_g \sim Dir(\alpha_\theta)$ b. For each document *m* in group *g*, where $1 \le m \le M_g$ 1) Draw $\lambda_{g,m} \sim Beta(\alpha_\lambda)$ 2) Draw $\gamma_{g,m} \sim Dir(\alpha_\gamma)$ 3) For each word *n*, where $1 \le n \le N_{g,m}$ a) Draw $s_{g,m,n} \sim Bern(\lambda_{g,m})$ b) If $s_{g,m,n} = 1$ (a group-independent topic word) A. Draw a topic assignment $z_{g,m,n} \sim \gamma_{g,m}$ B. Draw a word $w_{g,m,n} \sim \Phi_{z_{g,m,n}}$ c) If $s_{g,m,n} = 0$ (a group-specific content word) A. Draw $w_{g,m,n} \sim \theta_g$	1. Draw $\theta_0 \sim Dir(\alpha_\theta)$ 2. For each topic *k*, where $1 \le k \le K$ a. Draw $\Phi_k \sim Dir(\alpha_\Phi)$ 3. For each document group *g*, where $1 \le g \le G$ a. For each term *v*, where $1 \le v \le V$ 1) Draw $\sigma_{g,v} \sim Exponential(\alpha_\sigma)$ 2) Draw $\eta_{g,v} \sim N(0, \sigma_{g,v})$ b. Set $\boldsymbol{\vartheta_g} \propto exp(\boldsymbol{\vartheta_0} + \boldsymbol{\eta_g})$ c. For each document *m* in group *g*, where $1 \le m \le M_g$ 1) Draw $\lambda_{g,m} \sim Dir(\alpha_\lambda)$ 2) Draw $\gamma_{g,m} \sim Dir(\alpha_\gamma)$ 3) For each word *n*, where $1 \le n \le N_{g,m}$ a) Draw $s_{g,m,n} \sim Bern(\lambda_{g,m})$ b) If $s_{g,m,n} = 1$, Draw $z_{g,m,n} \sim \gamma_{g,m}$, Draw $w_{g,m,n} \sim \Phi_{z_{g,m,n}}$ c) If $s_{g,m,n} = 0$, Draw $w_{g,m,n} \sim \theta_g$

Table 1: The generation process of *dTM-Dirichlet* and *dTM-SAGE*.

from a constant background distribution, which brings three advantages: First, a sparsity-inducing prior can be applied to limit the number of terms whose probability diverges from the background term frequencies. Second, multi-facets latent variables can be easily combined by adding each facet component together to reduce the inference cost. Third, it is redundant to learn unique probabilities for high-frequency background words of each group. Modelling the deviation of each group-specific word distribution cancels the relearn process for the background words.

We propose *dTM-SAGE* which explicitly models the deviation in log-frequency of each group-specific word distribution from a background lexical distribution. *dTM-SAGE* also builds word models for group-independent topic words and group-specific content words. The group-independent topic words consist of background topic words and salient topic words.

dTM-SAGE models two types of group-independent words separately: as shown in Fig. 2, the salient topic words captured by φ_k and the background topic words captured by ϑ_0. The word models φ_k and ϑ_0 are multinomial distributions drawn from uniform Dirichlet prior with parameter α_φ and α_ϑ. To enable ϑ_0 to capture real background topic words shared by all document groups, we replace the constant background distribution in *SAGE* with a latent distribution learnt by MAP estimation using a Newton optimization.

The major difference between *dTM-SAGE* and *dTM-Dirichlet* is how the group-specific topics are generated. In Fig.2, each document group g has a group component vector η_g representing the deviations in log-frequencies from the background distribution ϑ_0. The group-specific topic is represented by log frequency deviations rather than word probabilities. Given the background distribution ϑ_0 and the group component vector η_g, the group-specific topic distribution ϑ_g for each word in a document in the group g, denoted by $\boldsymbol{\vartheta_g} \propto exp(\boldsymbol{\vartheta_0} + \boldsymbol{\eta_g})$, is computed by Equation (1):

$$p(w \mid \theta_0, \eta_g) \;/\; \sum_v \exp(\theta_{0,v}, \eta_{g,v}) \tag{1}$$

In Equation (1), g indexes the group component vector and v indexes the term in the corpus vocabulary. Following *SAGE*, we ignore covariance between terms. For each term v, $\eta_{g,v}$ is drawn from a zero-mean Gaussian distribution $N(0, \sigma_{g,v})$, where the variance $\sigma_{g,v}$ is drawn from the Exponential distribution parameterized by α_σ. The compound model $\int N(\eta; 0, \sigma) Exponential(\sigma; \alpha_\sigma) d\sigma$ is equivalent to a zero-mean Laplace prior on η inducing sparsity and meanwhile permitting large degrees of deviations. In *dTM-SAGE*, we treat the variance σ as a latent variable and develop variational inference on it, which is the same as *SAGE*. The remaining part of *dTM-SAGE* is the same as *dTM-Dirichlet* model. Formally, the generative process of *dTM-SAGE* is shown in Table 1. See Appendix A for inference details of $\boldsymbol{\vartheta_0}$ and $\boldsymbol{\eta_g}$.

4 Comparative Summary Generation

To summarize differences among document groups, we rely on group-specific topics ϑ_g to select most discriminative sentences for summary generation. This section introduces the sentence scoring and the sentence selection techniques developed for *dTM*-based comparative summarization.

4.1 Sentence Scoring

Both *dTM-Dirichlet* and *dTM-SAGE* model the group-specific word distributions ϑ_g to capture the unique content in each document group. For *dTM-SAGE*, we can also get a corpus background topic distribution ϑ_0 that reflects the common themes shared by all groups. To measure the sentence discriminative capacity, we develop two sentence scoring methods: one is based on the word discriminative scores and the other is measured by the difference of the probabilities that a sentence is generated from a group-specific topic distribution and the background topic distribution.

First, given a set of group-specific word distributions ϑ_g ($1 \le g \le$ G), we define the *w*ord *d*iscriminative score $DSW(v, g)$ of a term v to a group g as $DSW(v,g) = \sum_{g' \neq g}(\vartheta_{g,v} - \vartheta_{g',v}) / \sqrt{\sum_g \vartheta_{g,v}^2 + \epsilon}$, where ϵ is a small number (set to 0.05) to avoid the error of division by zero. Larger value of the word discriminative score indicates more discriminative ability the word has. The intuition is that a word more likely to occur in a particular group and less likely to occur in other groups tends to be more discriminative. The *d*iscriminative *c*apacity of a sentence s to a group g $DCS\_dsw(s, g)$ is the average over the word discriminative scores:

$$DCS\_dsw(s, g) = \sum_{w \in s} DSW(w, g) \,/\, len(s) \tag{2}$$

The other method to measure the discriminative capacity of a sentence relies on the likelihood that the sentence is generated from a group-specific distribution and the background topic distribution. Its design is motivated by the idea that a word is more discriminative if it occurs more often in a group-specific topic and occurs rarely in the shared background topic. Given a topic-word distribution ϑ, the probability of a sentence *s* generated from ϑ:

$$\log P(s \mid \theta) = \sum_{w \in s} \log \theta_w \tag{3}$$

Given a set of group-specific word distributions ϑ_g ($1 \le g \le$ G) and the background topic distribution ϑ_0, the *d*iscriminative *c*apacity of a sentence s to a group g is defined as the *d*ifference of sentence *g*enerative *p*robabilities $DCS\_dgp(s, g)$:

$$DCS\_dgp(s, g) = u \log P(s \mid \theta_g) - (1 - u) \log P(s \mid \theta_0) \tag{4}$$

where u is a balance factor trading off between group-specific words and background words.

4.2 Sentence Selection

To select discriminative sentences to form group summary, we use different sentence selection methods according to sentence scoring techniques.

For the sentence scoring based on the word discriminative scores, we first rank the sentences according to the sentence discriminative capacity score DCS_dsw. Then we select a sentence with the highest score if it satisfies the redundancy constraint that indicated by a cosine similarity threshold (empirically set to 0.8).

For the scoring based on difference sentences generative probabilities, suppose we have a set of candidate sentences S to form a summary for group g and we want to select k sentences denoted as S_k. A greedy sentence selection schema is proposed to build S_k by iteratively choosing a j$^{\text{th}}$ sentence that currently has the maximum sentence discriminative capacity score DCS_dgp:

$$s_j^* = \arg\max_{s_j \in S \setminus S_{j-1}} DCS\_dgp(s, g) \tag{5}$$

In order to discourage redundancy, after select one sentence, we update the group-specific topic distribution ϑ_g by setting $\vartheta_{g,w} \propto \vartheta_{g,w}^2$ for each word w in the selected sentence s_j^*. Sentences are selected in this manner until reaching the summary limit.

5 Experiments and Results

5.1 Data Collection and Annotation

Comparative summarization is not a new task. However, to our best knowledge there is no public benchmark data set available. For collecting experiment data, we choose three tasks in NLP: summarization (*SUMMA*), sentiment analysis (*SA*) and geographical NLP tasks (*GEO*) to form three document groups. To make different groups share more salient themes, we focus on papers using probabilistic

Group	Keywords		|D|	|S|
	Title	Plain Text		
SUMMA	summarization	topic model	35	6636
SA	Sentiment	topic model	45	10239
GEO	N/A	topic model, geographical	49	8249

Table 2: General Information of the Dataset

Measures	LDA	SAGE	Variant of DualSum	dTM-Dirichlet	dTM-SAGE
Perplexity	2218.37	2177.29	*1564.04	2052.78	**1891.10**
C_A (Wiki)	0.098	0.143	0.130	0.138	**0.147**
C_V (Wiki)	0.321	0.334	0.344	**0.360**	0.355
C_UCI (Wiki)	-2.116	-1.917	-1.272	-1.495	**-0.905**
C_UCI (Intra)	-0.895	-0.849	-0.662	-0.661	**-0.608**

Table 3: Comparisons of Perplexity and Topic Coherences for Different Topic Models.

topic models. We collect 129 papers in total for the three groups from ACL Anthology Searchbench providing full-text search for 28,000 papers in the ACL Anthology. For each group, we search with two types of keyword filters: plain text filter and title filter. Table 2 shows the general information of each document group, including the keywords, the number of documents |D| and the number of sentences |S|. To pre-process the dataset, we exclude all tables, figures and formulas, remove stop words, perform stemming with Porter Stemmer, and prune words less than 5 times across the corpus. There are 3720 tokens after pre-processing.

We hire three PhD students in Aston University to annotate the dataset. After reading papers in each group, each annotator is asked to first pick out all discriminative sentences in each paper and then write reference summaries delivering the major differences for each group. Additional instructions are given to annotators: Each reference summary should be no more than 300 words; and the discriminative sentences should enable the judgment of which group the paper belongs to. Equipped with the annotated dataset, two parts of evaluations are performed: evaluation of differential topic models and evaluation of the summarization methods.

5.2 Evaluations on *dTM*

In this section, we compare *dTM-Dirichlet* and *dTM-SAGE* with other three topic models in terms of model perplexity and topic coherences: (1) the standard *LDA* topic model, which we run across the corpus and perform Newton optimization to update hyper-parameters; (2) *SAGE*, which a sparse additive generative model proposed in (Eisenstein et al., 2011), and the non-parametric Jeffrey's prior make it parameter-free; (3) the variant of *DualSum*, which is proposed for update summarization (Delort and Alfonseca, 2012) and revised to perform comparative summarization by replacing pairs of collection-specific distributions with group-specific distributions. We implement the variant of *DualSum*, *dTM-Dirichlet* and *dTM-SAGE* models. Experimental settings are detailed below.

Settings for the variant of *DualSum*. The dirichlet priors for word distributions are empirically set to 0.1 and $\alpha_\lambda = (2.0,2.0,1.0)$ to encourage more words generated from the group-specific distributions and document-specific distributions.

Settings for *dTM-Dirichlet*. The dirichlet priors for word distributions α_Φ and α_θ are set to 0.1. For other papramenters, we set the number of group-independent topics $K = 20$, the prior for the topic distribution $\alpha_\gamma = 50/K$, and the prior for the group-specific word controller $\alpha_\lambda = 2.0$. *Beta*(2.0, 2.0) yields equal probabilities that words sampling from the group-specific distribution and the group-independent distributions.

Settings for *dTM-SAGE*. Parameters are set the same as those in dTM-Dirichlet: $\alpha_\Phi = \alpha_\theta = 0.1$, $K = 20$, $\alpha_\gamma = 50/K$ and $\alpha_\lambda = 2.0$. The variational distribution of the variance $\boldsymbol{\sigma}$ is $Gamma(\tilde{\alpha}, \tilde{b})$ which is initialized as $\tilde{\alpha} = 10.0$ and $\tilde{b} = 5.0$. The initialization for $\boldsymbol{\vartheta_0}$ and $\boldsymbol{\eta}$ are from the Uniform distribution $U(0,1)$ and the Normal distribution $N(0,0.5)$ respectively.

First, we investigate the model perplexity. Perplexity is a general measure for evaluating the generative ability of a probabilistic topic model. We compute the perplexity on a held-out test set, 20% of the original dataset. Note that we calculate the perplexity for all models except the variant of *DualSum*, since it models the document-specific distribution for each document and thus there is no natural way to assign probability to new document. For the variant of *DualSum*, we train the model on the whole dataset and report the results on the test set, though it by no means can reflect the generalization capacity of the model.

Perplexity results are shown in the first row in Table 3, from which we can see that the perplexity scores decrease by 7% and 13% respectively between *dTM-Dirichlet* and standard *LDA* and between *dTM-SAGE* and standard *SAGE*. The better results of differential topic models over the standard ones are due to the discrimination between group-specific topics and group-independent topics. Both *SAGE* methods outperform their counterparts of the Dirichlet-multinomial, because the sparsity-inducing prior enables *SAGE* to control sparsity adaptively without over-fitting (Eisenstein et al., 2011).

SAGE	dTM-Dirichlet	dTM-SAGE
sentence, topic, query document, summary, word, generative, model, vertice, distribution,	sentence, summary, document, topic, rouge, extract, score, select, multi, system	sentence, rouge, ilp, duc, tac, summary, timeline, lexrank, redundant, mead

Table 4: Top 10 Words for the Group *SUMMA*.

	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-SU4	Precision
Baselines					
Centroid	0.23084	0.01867	0.21739	0.05672	0.383
LexPageRank	0.25334	0.02092	0.23822	0.06767	0.417
MMR	0.28272	0.02817	0.26333	0.08094	0.433
State-of-the-arts					
DSS	0.30898	0.03766	0.29346	0.09239	0.600
CDS	0.31749	0.03717	0.29047	0.09340	0.549
TM-based Methods					
Basic LDA (dsw)	0.29812	0.03625	0.27940	0.08865	0.517
Variant of DualSum (dsw)	0.37445	0.04584	0.34542	0.11245	0.650
dTM-Dirichlet (dsw)	0.33024	0.06047	0.31388	0.12363	0.700
dTM-SAGE (dsw)	0.39173	0.06800	0.35764	0.12716	0.717
dTM-SAGE (dgp)	**0.42266**	**0.08801**	**0.38519**	**0.16205**	**0.750**

Table 5: Comparison of Rouge Scores (F-score) and Precision.

Summary by dTM-Dirichlet.

①Most of the existing multi-document summarization methods decompose the documents into sentences and work directly in the sentence space using a term-sentence matrix.

②Bayesian sentences-based topic model, every sentences in a document is assumed to be associated to a unique latent topic.

③While previous MDS systems have focused primarily on salience and coverage but not coherence, G-Flow generates an ordered summary by jointly optimizing coherence and salience.

④Markov Random Walk Model (MRW) Graphs methods have been successfully applied to weighting sentences for generic and query-focused summarization.

⑤The topic distributions are used to get the sentence scores and rank sentences.

Summary by dTM-SAGE.

①In recent years, three major techniques have emerged to perform multi-document summarization: graph-based methds such as LexRank, Biased-LexRank for query-focused summarization, language models such as KLSum and variants based on topic models, such as BayeSum and TopicSum.

② Bayesian Query-Focused Summarization, we present BayeSum (Bayesian summarization), a model for sentence extraction in query-focused summarization.

③Sentence Selection Strategy, The task of timeline summarization aims to produce a summary for each time and the generated summary should meet criteria such as relevance, coverage and coherence.

④Models that use more structure in the representation of documents have also been proposed for generating more coherent and less redundant summaries, such as HierSUM and TTM.

⑤In generating a summary, we use MMR (Maximal Marginal Relevance for multi-document) to avoid redundancy in a summary.

Table 6: Comparison of 5-Sentence Summary Generated by *dTM-Dirichlet* and *dTM-SAGE*.

To check the quality of the generated group-specific topics, we investigate various topic coherence measures. The intuition behind the topic coherence measures is that words clustering into a single topic tend to co-occur in the same document. It has been previously verified that topic coherence score is highly correlated with human-judged topic coherence in many works. We rely on Palmetto library (Röder et al., 2015), an online open source implementation, which offers a framework to calculate many coherence measures within a reference corpus of the English Wikipedia.

In our experiment, we compare three widely-used coherence scores over the five topic models: (1) *C_A*, which is the pairwise comparison of the top words based on a context window of size 5, and proposed in (Aletras and Stevenson, 2013); (2) *C_V*, which is a one-set segmentation of the top words based on a sliding window of size 110, and proposed in (Röder et al., 2015); (3) *C_UCI*, which is the pointwise mutual information (PMI) of all word pairs of the top words based on a sliding window of size 10, and proposed in (Newman et al., 2010).

We focus on the group-specific topics. For each group-specific topic-word distribution we get a word list containing the top-20 words and calculate the coherence scores for each word list. The topic coherence results shown in Table 3 are the average coherence scores of the three group word lists. The coherence scores are calculated within two reference corpus: the English Wikipedia (*Wiki*) and the original dataset (*Intra*). Table 4 shows the top 10 words selected by *SAGE*, *dTM-SAGE* and *dTM-Dirichlet* for the group *SUMMA*. Main observations found in Table 3 include:

(1) The three differential topic models generally perform much better than the standard *LDA* and *SAGE* models on all coherence measures, which shows the superiority of our *dTM* models by distinguishing group-specific words and group-independent words;

(2) *dTM-SAGE* consistently performs the best among all the five models in terms of *C_A* and *C_UCI* with the significant increase at least by 6.5% over *dTM-Dirichlet* and 8.2% over the variant of *DualSum*, which shows the advantage of *dTM-SAGE* in accurately ranking the group-specific words due to the essence of the differential word model;

(3) *dTM-Dirichlet* outperforms the variant of *DualSum* with *C_A* and *C_V*, however, it performs nearly the same or even worse when measured with *C_UCI*.

As shown in Table 4, words selected by *dTM-SAGE* (like rouge, lexrank, redundant) are more informative and discriminative than words selected by *SAGE* and *dTM-Dirichlet*.

5.3 Evaluations on Summarization

To evaluate the quality of the generated summaries, we compare our *dTM*-based comparative summarization methods with five other typical methods under ROUGE metrics (Lin and Hovy, 2003). Further, to check the discriminative ability of the comparative summaries, following the evaluation method of (Wang et al., 2012), we investigate the precision of the discriminative sentence selection.

In our experiment, we implement three types of summarization methods: (1) Generic baseline methods, including the centroid-based method (Radev et al. 2004), the graph-based method LexPag-

eRank (Radev et al., 2004) and the MMR-based method (Carbonell and Goldstein, 1998); (2) State-of-the-art comparative summarization methods, including the discriminative sentence selection (*DSS*) method (Wang et al., 2012) and the complementary dominating set (*CDS*) method (Shen and Li, 2010); (3) TM-based comparative summarization methods, which combine four different TMs with two sentence scoring strategies DCS_dsw and DCS_dgp defined in section 4.1, including the basic $LDA\ (dsw)$, the variant of $DualSum\ (dsw)$, *dTM-Dirichlet* (dsw), *dTM-SAGE* (dsw) and *dTM-SAGE* (dgp). For each group, we select 20 sentences to form the final summary.

First, we examine the precision of the discriminative sentence selection. For each group we have 20 sentences in a summary and count how many sentences belong to the annotated discriminative sentence set. Comparisons of the precision results of discriminative sentence selection by different methods are shown at the last column in Table 5. From the precision results, we find that (1) our *dTM*-based comparative summarization methods can select over 70% discriminative sentences, which significantly outperform the state-of-the-art methods with a nearly 20% increase on the precision score; (2) All generic summarization methods perform rather worse due to different concerns on summarization resulting in the lack of discriminative ability of summaries.

We use ROUGE-1.5.5 toolkit to evaluate the quality of generated summaries by comparing them with human-written reference summaries. In our experiment, we limit the length of all summaries to 250 words and report the average ROUGE scores (F-Scores) on various summarization methods in Table 5. According to the results, we observe that: (1) our *dTM*-based comparative summarization methods perform much better (paired t-test with $p<0.05$) than all the baselines, which demonstrates that targeting at a different goal for summarizing the general information among document groups, generic summarization methods are less applicable for comparative summarization, though by removing redundancy, MMR performs better than the other two baselines but still lags behind other summarization methods specifically proposed for comparative summarization; (2) our *dTM*-based comparative summarization methods significantly outperform (paired t-test with $p<0.05$) the other two state-of-the-art comparative summarization methods, which shows that summarizing differences by extracting group-specific topics is more effective than directly summarizing at the sentence level; (3) Both *dTM-SAGE* methods achieve better ROUGE scores than *dTM-Dirichlet*, which is ascribe to the advantage of a differential word model contributing to more informative and discriminative group-specific topics (discussed in section 5.2); and, (4) For *dTM-SAGE*, the greedy sentence selection schema based on DCS_dgp is more effective than simply ranking sentence with DCS_dsw.

5.4 Summary Example

We show an example of the summary generated for the group *SUMMA* by our *dTM-SAGE* and *dTM-Dirichlet* in Table 6. Looking into the summaries, we find that all sentences in both summaries are related to summarization but different in the degree of their discriminative ability. Apparently, the summary generated by *dTM-SAGE* is more specific and unique to summarization, while the summary generated by *dTM-Dirichlet* still contains some general information about topic models in sentence ② and ⑤. Another observation is that the summary of *dTM-SAGE* tends to contain more salient group-specific terms that may not occur in most of group documents but still possess high discrimination, like '*query-focused*', '*MMR*' and '*HierSUM*'. In contrast, the summary by *dTM-Dirichlet* covers more background group-specific words, like '*summarization*' and '*MDS*'. Although these background group-specific terms are discriminative for the group, they are relatively less informative than the salient terms for the purpose of summarization.

6 Conclusions

This paper studies the differential topic models for comparative summarization on cross-area scientific papers. A differential topic model *dTM-SAGE* is proposed to capture the unique characteristics of each document group and generate coherent group-specific topics. A greedy sentence selection method with two sentence discriminative capacity scoring schemas is designed to automatically generate summary for *dTM*-based comparative summarization methods, which achieve significant improvements with various ROUGE metrics. The analysis on experiment results shows that the summaries generated by our *dTM-SAGE* method can cover major differences for each group.

Acknowledgements

Research was partially supported by National Science Foundation of China and National Science Foundation of Jiangsu (No. BK20140895).

Reference

Ahmed, A., and Xing, E. P. 2010. Staying informed: supervised and semi-supervised multi-view topical analysis of ideological perspective. In *Proceedings of EMNLP*, 1140-1150.

Aletras, N., and Stevenson, M. 2013. Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics*, 13-22.

Boudin, F., and El-Bèze, M. 2008. A scalable MMR approach to sentence scoring for multi-document update summarization. In *Proceedings of COLING*, 23-26.

Blei, D. M. 2012. Probabilistic topic models. *Communications of the ACM*, 55(4), 77-84.

Cai, X., Li, W., Ouyang, Y., and Yan, H. 2010. Simultaneous ranking and clustering of sentences: a reinforcement approach to multi-document summarization. In *Proceedings of COLING*, 134-142.

Celikyilmaz, A., and Hakkani-Tur, D. 2010. A hybrid hierarchical model for multi-document summarization. In *Proceedings of ACL*, 815-824.

Celikyilmaz, A., and Hakkani-Tür, D. 2011. Discovery of topically coherent sentences for extractive summarization. In *Proceedings of ACL*, 491-499.

Delort, J. Y., and Alfonseca, E. 2012. DualSum: a topic-model based approach for update summarization. In *Proceedings of the European Chapter of ACL*, 214-223.

Erkan, G., and Radev, D. R. 2004. LexPageRank: Prestige in Multi-Document Text Summarization. In *Proceedings of EMNLP*, Vol. 4, 365-371.

Eisenstein, J., Ahmed, A., and Xing, E. P. 2011. Sparse additive generative models of text. In *Proceedings of ICML*, 1041-1048.

Fisher, S., and Roark, B. 2008. Query-focused supervised sentence ranking for update summaries. In *Proceedings of the TAC*.

Gupta, V., and Lehal, G. S. 2010. A survey of text summarization extractive techniques. *Journal of Emerging Technologies in Web Intelligence*, 2(3), 258-268.

Jha, R., Coke, R. and Radev, D. 2015. Surveyor: A system for generating coherent survey articles for scientific topics. In *Proceedings of AAAI*, 2167-2173.

Lin, C. Y., and Hovy, E. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *HLT-NAACL*, 71-78.

Li, W., Wei, F., Lu, Q., and PNR, Y. H. 2008. Ranking sentences with positive and negative reinforcement for query-oriented update summarization. In *Proceedings of COLING*, 489-496.

Li, P., Wang, Y., Gao, W., and Jiang, J. 2011. Generating aspect-oriented multi-document summarization with event-aspect model. In *Proceedings of EMNLP*, 1137-1146.

Li, C., Qian, X., and Liu, Y. 2013. Using Supervised Bigram-based ILP for Extractive Summarization. In *Proceedings of ACL*, 1004-1013.

Li, W. 2015. Abstractive Multi-document Summarization with Semantic Information Extraction. In *Proceedings of EMNLP*,1908-1913.

Mihalcea, R., and Tarau, P. 2004. TextRank: Bringing order into texts. In *Proceedings of EMNLP, 404-411*.

Newman, D., Lau, J. H., Grieser, K., and Baldwin, T. 2010. Automatic evaluation of topic coherence. In *Proceedings of NAACL*, 100-108.

Paul, M. 2009. Cross-collection topic models: Automatically comparing and contrasting text. *Master's thesis, University of Illinois Urbana Champaign, IL, USA.*

Paul, M., and Girju, R. 2009. Cross-cultural analysis of blogs and forums with mixed-collection topic models. In *Proceedings of EMNLP*, 1408-1417.

Paul, M., and Girju, R. 2010. A two-dimensional topic-aspect model for discovering multi-faceted topics. In *Proceedings of AAAI*, 545-550.

Radev, D. R., Jing, H., and Budzikowska, M. 2000. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *Proceedings of the NAACL-ANLP*, 21-30.

Radev, D. R., Jing, H., Styś, M., and Tam, D. 2004. Centroid-based summarization of multiple documents. *Information Processing & Management*, *40*(6), 919-938.

Röder, M., Both, A., and Hinneburg, A. 2015. Exploring the space of topic coherence measures. In *Proceedings of the 8th ACM international conference on Web search and data mining*, 399-408. ACM.

Sekine, S., and Nobata, C. 2003. A survey for multi-document summarization. In *Proceedings of the HLT-NAACL*, Vol.5, 65-72.

Shen, C., and Li, T. 2010. Multi-document summarization via the minimum dominating set. In *Proceedings of COLING*, 984-992.

Wan, X., Yang, J., and Xiao, J. 2007. Towards an iterative reinforcement approach for simultaneous document summarization and keyword extraction. In *Proceedings of ACL*, Vol.7, 552-559.

Wang, D., Zhu, S., Li, T., and Gong, Y. 2012. Comparative document summarization via discriminative sentence selection. *ACM Transactions on Knowledge Discovery from Data* (*TKDD*), 6(3), 12.

Zhai, C., Velivelli, A., and Yu, B. 2004. A cross-collection mixture model for comparative text mining. In *Proceedings of ACM SIGKDD*, 743-748. ACM.

Zhuge, H. 2016. Multi-Dimensional Summarization in Cyber-Physical Society, Morgan Kaufmann.

Appendix A. Variational Inference of ϑ_0 and η_g

Generally, we take MAP (maximum a posterior) estimation for the background word distribution ϑ_0 and the group component vectors η and develop variational inference techniques for all other variables.

In dTM-SAGE, the lower bound L with regarding to ϑ_0, η and σ is:

$$L = \log P(\theta_0 \mid \alpha_\theta) + \sum_g \sum_m \sum_n E_Q[\log P(w_{g,m,n} \mid s_{g,m,n} = 0, \theta_0, \boldsymbol{\eta_g})]$$

$$+\sum_g E_Q[\log P(\boldsymbol{\eta_g} \mid 0, \boldsymbol{\sigma_g})] + \sum_g E_Q[\log P(\boldsymbol{\sigma_g} \mid \alpha_\sigma)] - \sum_g E_Q[\log Q(\boldsymbol{\sigma_g})]$$

Maximize L with respect to $\boldsymbol{\vartheta_0}$:

$$L(\theta_0) = \sum_v (\alpha_g^v - 1) * \log \vartheta_0^v + \sum_g \sum_m \sum_n \tilde{\lambda}_{gmn}^0 * \{\vartheta_0^{w_{g,m,n}} - \log(\sum_{v'} \exp(\eta_g^{v'} + \theta_0^{v'}))\}$$

By Assuming $T(v) = \frac{\exp(\eta_g^v + \theta_0^v)}{\sum_j^V \exp(\eta_g^j + \theta_0^j)}$, taking derivatives with respect to ϑ_0^v:

$$\frac{\partial L}{\partial \vartheta_0^v} = \frac{\alpha_g^v - 1}{\vartheta_0^v} + \sum_g \sum_m \sum_n \tilde{\lambda}_{gmn}^0 * \{I(w_{g,m,n} = v) * (1 - T(v)) + I(w_{g,m,n} \neq v) * (-T(v))\}$$

We use Newton-Raphson method to optimize $\boldsymbol{\vartheta_0}$. First, we derive the Hessian matrix by setting:

$$H_{vv}(\theta_0) = \frac{\partial^2 L}{\partial \vartheta_0^{v2}} = -\frac{\alpha_g^v - 1}{(\vartheta_0^v)^2} + \sum_g \sum_m \sum_n \tilde{\lambda}_{gmn}^0 * (T(v)^2 - T(v))$$

$$H_{vv'}(\theta_0) = \frac{\partial^2 L}{\partial \vartheta_0^v \partial \vartheta_0^{v'}} = \sum_g \sum_m \sum_n \tilde{\lambda}_{gmn}^0 * T(v)T(v')$$

After getting Hessian matrix, we invert it with Sherman-Morrison formula and compute the Newton step: $\Delta\boldsymbol{\vartheta_0} = H^{-1}(\boldsymbol{\vartheta_0})\nabla_{\boldsymbol{\vartheta_0}} L(\boldsymbol{\vartheta_0})$. Same procedure on η:

$$\frac{\partial L}{\partial \eta_g^v} = -\eta_g^v * [(\tilde{a}_g^v - 1)\tilde{b}_g^v]^{-1} + \sum_m \sum_n \tilde{\lambda}_{gmn}^0 * \{I(w_{g,m,n} = v) * (1 - T(v)) + I(w_{g,m,n} \neq v) * (-T(v))\}$$

$$H_{vv}(\boldsymbol{\eta}_g) = \frac{\partial^2 L}{\partial \eta_g^{v2}} = -[(\tilde{a}_g^v - 1)\tilde{b}_g^v]^{-1} + \sum_g \sum_m \sum_n \tilde{\lambda}_{gmn}^0 * (T(v)^2 - T(v))$$

$$H_{vv'}(\boldsymbol{\eta}_g) = \frac{\partial^2 L}{\partial \eta_g^v \partial \eta_g^{v'}} = \sum_g \sum_m \sum_n \tilde{\lambda}_{gmn}^0 * T(v)T(v')$$