A novel contextual topic model for multi-document summarization

Guangbing Yang a, Dunwei Wen b, Kinshuk b, Nian-Shing Chen c, Erkki Sutinen a

a School of Computing, University of Eastern Finland, P.O. Box 111, 80101 Joensuu, Finland
b School of Computing and Information Systems, Athabasca University, 1 University Drive, Athabasca, Alberta T9S 3A3, Canada
c Department of Information Management, National Sun Yat-sen University, Kaohsiung, Taiwan

A R T I C L E   I N F O
Article history:
Available online 28 September 2014

Keywords:
Multi-document summarization
Hierarchical topic model
Contextual topic

A B S T R A C T
Information overload becomes a serious problem in the digital age. It negatively impacts understanding of useful information. How to alleviate this problem is the main concern of research on natural language processing, especially multi-document summarization. With the aim of seeking a new method to help justify the importance of similar sentences in multi-document summarizations, this study proposes a novel approach based on recent hierarchical Bayesian topic models. The proposed model incorporates the concepts of n-grams into hierarchically latent topics to capture the word dependencies that appear in the local context of a word. The quantitative and qualitative evaluation results show that this model has outperformed both hLDA and LDA in document modeling. In addition, the experimental results in practice demonstrate that our summarization system implementing this model can significantly improve the performance and make it comparable to the state-of-the-art summarization systems.

1. Introduction

While the rapid growth of the World Wide Web has resulted in bringing people from different parts of the world much closer and able to access to a vast amount of information on their fingertips, it has also created a serious problem of information overload, which impacts people negatively in comprehending useful information. How to alleviate this problem is of concern to the research on automatic text summarization, especially the extraction based multi-document summarization. The main task of the extraction based multi-document summarization is to extract the most important sentences from multiple documents and format them into a summary. Therefore, finding an appropriate method to justify the importance (or relevance) of a string of text (e.g., a sentence) dominates this research area. Many proposed approaches use statistical methods, lexical chains, graph-based algorithms, or Bayesian language models to produce summaries. For example, a well-known summarizer, SumBasic in statistical methods, specifies the importance of a sentence in a document by counting term frequency or inverse document frequency (TF-IDF) exclusive of stop words (Nenkova & Vanderwende, 2005; Vanderwende, Suzuki, Brockett, & Nenkova, 2007). Others identify the relevance of a sentence by using bigram pseudo sentences for using hybrid statistical sentence-extraction (Ko & Seo, 2008), rhetoric-based multi-document summarization (Atkinson & Munoz, 2013) and semantic document concept technique (Ye, Chua, Kan, & Qiu, 2007) based on the rhetorical structure theory (RST) (Mann & Thompson, 1988) for analyzing grammatical structures in discourses. However, heavy reliance on human expert’s rhetorical roles and linguistic knowledge bases is definitely a bottleneck for RST based approaches.

In recent years, multi-document summarization research has shown increased interest in graph-based approaches (Canhasi & Kononenko, 2014; Ferreira, de Souza Cabral, Freitas, & de Franca Silva, 2013; Ferreira, de Souza Cabral, Freitas, & de Franca Silva, 2014; Glavaš & Snojder, 2014; Mendoza, Bonilla, Noguer, Cobos, & León, 2014; Zhao, Wu, & Huang, 2009) and Bayesian topic model based approaches (Celikyilmaz & Hakkani-Tur, 2011; Daumé III & Marcu, 2006; Eisenstein & Barzilay, 2008; Haghhighi & Vanderwende, 2009). The graph-based approach is a type of bottom-up method. It discusses the similarity problem from the perspective of content structure and becomes popular in web search and text classification. In contrast, a topical model is a type of top-down approach. It considers the same problem based on semantic associations behind the content. Although the graph model can provide a sophisticated structure to represent sentence relations in documents, it does not come up with an algorithm for similarity analysis directly. In the most recent work, the similarity process is done by using either cosine similarity algorithm or square distance measures between vectors or matrices. This leaves
some spaces for improvement in research of graph model based automatic text summarization. For example, Canhasi and Kononenko (2014) use weighted archetypal analysis to compute the similarity. In addition, graph-based natural language processing (NLP) draws highly attention and related methods have been applied in semantic analysis, parsing, and summarization recently. However, a lack of efficient algorithms for better similarity measures is still a big challenge in research of NLP.

In the Bayesian topic model based approaches, similarity is analyzed using advanced methods with respect to probabilistic distributions of topics. In most of the cases, the process of similarity is straightforward as the comparison between probabilistic densities that are usually inference results of the topic models. However, many of those topic model based approaches do not employ hierarchies to represent sentence relations as graph models did. In contrast, they render sentence relations in a flat structure and ignore the hierarchical correlations between sentences or words in many text applications. Only few examples, such as the two-tiered topic model (TTM) in Celikyilmaz and Hakkani-Tur (2011), allocate topics in double layers for representing sentence relations and assign topics in a higher layer to describe the correlations of specific topics indicated in a lower layer. Although this approach demonstrates good performance in the multi-document summarization, the way of arbitrarily choosing topic levels limits its practical utilizations in multi-document summarization. Moreover, either graph-based or Bayesian topic model based approaches fundamentally ignore the lexical co-occurrence that appears in the local context of a string of words. However, in many realistic applications, such as text summarization, speech recognition, syntactic parsing, text mining, and so on, the lexical co-occurrence plays an important role for capturing word associations and analyzing topic correlations (Blei, Griffiths, & Jordan, 2010; Wallach, 2006).

To deal with this challenge and limitations, this research is motivated by the desire for overcoming those limitations in existing approaches and seeking a plausible solution for the problem of multi-document summarization. Therefore, our hypothesis of this research is that a model is deemed to improve the performance of summarization significantly if it is established upon a Bayesian topic model and in the meantime considers some properties of graph models with the lexical co-occurrences and topic hierarchies as well.

In this paper, a novel approach is proposed for the problem of multi-document summarization. Our idea came from the previous research (Celikyilmaz & Hakkani-Tur, 2011) that demonstrated that the correlations between general and specific topics were important features to evaluate the relevance of sentences. With the benefit of the significant achievement of the Bayesian based topic modeling in the research of NLP and the success of graph models in multi-document summarization, our model is built upon the hierarchical Latent Dirichlet Allocation model (hLDA) (Blei et al., 2010) with some concepts from graph models. The metaphor of the nested Chinese Restaurant Process (nCRP) represents a hierarchical prior distribution that provides tree topologies for sampling topics within a hierarchical structure. In addition to the hierarchical structures, our model integrates the concepts of latent topics into n-gram language models (Griffiths, Steyvers, Blei, & Tenenbaum, 2005; Wallach, 2006; Wang, McCallum, & Wei, 2007) to indicate the lexical co-occurrences within the domain of topic modeling. Since both the topic hierarchies and the lexical co-occurrences are related to the context of documents, the topics captured with this contextual information are described as contextual topics in this study, and the corresponding model is called contextual topic model.

Furthermore, this contextual topic model (CTM) can also be viewed as a Bayesian network, which is a directed acyclic graph (DAG) whose nodes represent the random variables and edges represent the dependency between nodes. The weight assigned to each node is the conditional probability distribution (CPD) of the topics given both data as input sentences and probability functions as n-gram models for lexical co-occurrence. The weight for each edge is the CPD of the path between nodes. This model has advanced properties of both graph models and Bayesian topic models. Concretely, the entire research reported in this paper is based on the following theoretical foundations: Bayesian probability theory, probabilistic graphical models (Koller & Friedman, 2009), and language modeling framework (e.g., the relevance-based language model from Lavrenko and Croft (2001)) with some concepts from information theory, like entropy and model perplexity. By investigating hierarchical topics and their correlations with respect to the lexical co-occurrences of words, this model is able to determine the similarity of sentences more effectively.

The rest of the paper is organized as follows: Section 2 describes the related Bayesian topic models and graph model based applications in multi-document summarization. Section 3 discusses our proposed approach in details. Section 4 describes the experiment conducted to validate the approach and discusses the findings. Section 5 presents concluding remarks.

2. Related work

The research of automatic text summarization has continued more than 50 years since the publication of Luhn’s pioneer of information science paper (Luhn, 1958). Many approaches have been addressed and many solutions have been evaluated since then. Based on a survey of text summarization (Nenkova & McKeown, 2012), those approaches can be categorized into two classes: topic representation approaches and indicator representation approaches.

The category of topic representation includes many well known approaches, such as word frequency, TF-IDF weighting (like SumBasic (Nenkova & Vanderwende, 2005)), sentence position, title relation, and cue-phrases, all these approaches are built based on the shallow features of documents without considering the semantic associations behind sense. Other approaches take account of semantic associations between words and combine them with those shallow features in the process of sentence similarity. Examples of such approaches are: latent semantic analysis (LSA, Gong & Liu, 2001), topic signatures (Lin & Hovy, 2000), sentence clustering (He, Qin, & Liu, 2012), and Bayesian topic model based approaches, such as BayeSum (Daumé III & Marcu, 2006), topic segmentation (Eisenstein, 2008), and TopicSum from Haghighi and Vanderwende (2009), and so on. Experimental results from both Text Retrieval Conference (TREC) and Document Understanding Conference (DUC) have demonstrated that Bayesian topic model based approaches outperform summarizers implemented with shallow features and can enhance performance of retrieval and document summarization significantly. Although these achievements, these approaches systemically ignore contextual information of words, which can significantly influence overall performance of sentence similarity (Blei et al., 2010; Celikyilmaz & Hakkani-Tur, 2011; Ko & Seo, 2008; Wallach, 2006).

In order to overcome this shortcoming, lexical chains (Barzilay, Elhadad, & McKeown, 1999) and RST based summarization models (Atkinson & Munoz, 2013) have been proposed to interpret word co-occurrences in linguistic context. These approaches address this shortcoming, but bring another disadvantage, in which these methods rely heavily on a particular linguistic knowledge base and constructing these knowledge bases is expensive and time-consuming. Thus, the desire for seeking plausible solutions for this problem motivates investigations of the summarization in a different direction.

As a typical method of indicator representation approaches, the graph model is widely used in automatic text summarization
recently (Canhasi & Kononenko, 2014; Ferreira et al., 2013; Ferreira et al., 2014; Glavas & Šnajder, 2014; Mendoza et al., 2014). In those summarization systems, documents are constructed as a graph, in which its nodes represent sentences and it associated edges are assigned the similarity scores between sentences assigned to those nodes. The most popular method used for measuring sentence similarity is cosine similarity with TF-IDF weighted term frequency. Ferreira et al. (2014) proposed a new sentence-clustering algorithm using such graph model that adapts principles of statistical similarities and linguistic treatment. The statistic similarity from word frequency, semantic similarity from synonyms and hyponym of words in WordNet, and co-reference and discourse relations between sentences are features used in their algorithm. The experiment results reported in this paper over performed the state-of-the-art summarizers by using DUC 2002 dataset. However, due to the coverage of WordNet, heavy reliance on WordNet for semantic similarity limits this approach in practical applications (Nenkova & McKeown, 2012). Another approach presented by Canhasi and Kononenko (2014) uses weighted archetypal analysis of the multi-element graph to construct a query-focused summarizer for multiple documents. In their approach, the input text and query are converted as a multi-element graph (Canhasi & Kononenko, 2014), and the similarity between sentences are measured by using weighted archetypal analysis, in which optimal square distance between two matrixes represents the similarity. The experimental results reported in this paper demonstrate the enhancement of summarization performance and make their summarization system comparable to the state-of-the-art summarization systems. Other recent approaches based on graph models include a graph-based single-document summarizer (Mendoza et al., 2014), event graphs for information retrieval and summarization (Glavas & Šnajder, 2014), and a context-based word indexing approach with respect to graph models (Goyal, Behera, & McGinnity, 2013). Except the context-based word indexing approach, all graph model based approaches discussed above evaluate sentence similarity without investigating contextual information.

Recent research on Bayesian topic model based summarization and hierarchical topic models present some potential to address this problem. The proposed two-tiered topic model (TTM) allocates topics into a hierarchical structure that only takes two levels: the lower level represents the latent topics sampled from words, and the higher level describes the correlations between these topics (Celikyilmaz & Hakkani-Tur, 2011). To distinguish topic levels, this model classifies the topics into two groups using auxiliary criterion from human expert summaries, which is a nuisance widely required by most practical applications of language models. However, restricting topic layers to only two levels and relying on expert summaries for indicating the topic levels limit the utility of the model in practical applications because latent topics captured from multiple documents may cover many ideas so as to not be easily allocated exactly into two layers. On the other hand, human summaries are not always available, especially for online applications, or good enough as golden standards to express important information discussed in documents. Moreover, as mentioned in previous section of this paper, the TTM did not consider the problem of word dependency that can play an important role in practical applications of multi-document summarization. Thus, to deal with these limitations, an alternative approach is essential in multi-document summarization.

Recently developed Bayesian collection models incorporated the concept of latent topics into n-gram language models, such as the LDA-HMM model integrating topics and syntax (Griffiths et al., 2005), structured topic model (Wallach, 2006), and topical n-grams (Wang et al., 2007). These models have demonstrated excellent results on tasks related to parts-of-speech, document classifications, clustering, information retrieval, and so on. Experimental results have shown that the application of these models resulted in significant improvement in information retrieval on TREC data collections. As a common and effective approach for better results in practical applications, such as the information retrieval and extractive based multi-document summarization, most of these models chose a language model, such as the query likelihood model (Ponte & Croft, 1998), as the framework to integrate the features of topic modeling into the practical applications. In our case, a relevance-based language model (Lavrenko & Croft, 2001) has been selected as the language-modeling framework to utilize features provided by our contextual topic model in the application of multi-document summarization.

Relevance-based language model (Lavrenko & Croft, 2001) is a Bayesian language model mainly focusing on applications of information retrieval and document classification. It addresses the problem that the automatic query expansion is conceptually difficult to be integrated into the language-modeling framework (Ponte & Croft, 1998) and provides a framework to support query expansion techniques in language modeling. Experiment results have demonstrated that relevance-based language model outperforms query likelihood model (Lavrenko & Croft, 2001) and have justified that relevance-based language modeling is a suitable solution for integrating different models into the language-modeling framework. In addition, research in literature has also reported that relevance-based models are better suited to summarization task (Berger & Mittal, 2000) and text segmentation (Beeferman, Berger, & Lafferty, 1999).

The success of these models and applications suggest that the mechanism of incorporating the concept of latent topics into n-grams is a plausible approach that is helpful for the problems discussed in this paper. In addition, due to the generative properties of Bayesian topic models, contextual information can be simply modeled as a probabilistic distribution to be inferred from input data.

3. Methods

3.1. Core concept

An article typically focuses on some important content for a reader by providing connections from the topic of one paragraph into the topic of next paragraph or from a main topic to subtopics in the same section. The content of the articles may be helpful in exposing levels of these topics. Barzilay, Elhadad, and McKeown (2002) reported that sentences sharing relevant topics tend to appear together (normally, an important idea is described in a topic sentence and supported by following details in a paragraph). In addition, a hierarchical structure of latent topics captured from the content of articles may reveal certain correlations between words or sentences (Celikyilmaz & Hakkani-Tur, 2011). The latent topics allocated at a higher level of this hierarchy may represent general concepts in a document, and topics assigned with a lower level may describe more specific ideas of the document. However, it is difficult to automatically construct the hierarchical structures of latent topics and precisely indicate topic levels via the content analysis. Although previous research in multi-document summarization showed that both general and specific topics and their correlations were helpful in reducing the redundancy and increasing the coherence (Celikyilmaz & Hakkani-Tur, 2011), few plausible approaches have been addressed to solve this problem effectively. Thus, the hypothesis for this study is that a Bayesian hierarchical topic model, which can distinguish general and specific topics and indicate their correlations via analyzing the topic hierarchies, is a plausible approach to improve the performance of summarization. Based on this hypothesis, the problem of how to determine
the general and specific topics and their correlations from content can be transferred to the problem of how to learn the topic hierarchies from documents. In order to clearly explain our idea described above, a topic path with a sample summary that was generated by a summarization system implementing our model is illustrated in Fig. 1. The particular topic assignments and level arrangements illustrated in the figure are taken at the approximate mode of the contextual topic model posterior conditioned on articles from DUC 2006 corpus (Dang, 2006).

A sample summary was generated from 25 news articles in article collection D0614D of DUC 2006 corpus. The generated summary is associated with a path that is sampled from an nCRP prior distribution. Topics in the highest level (level 0) are more functional terms and common words that often appear in news articles, such as year(s), percent, etc., and topics in lower levels have more specific meanings. In addition, the levels assigned to topics reflect certain correlations between topics. For example, names of several Canadian politicians are assigned to level 3 and are descendants of topics Quebec and Canada (as shown in Fig. 1).

A well-known Bayesian hierarchical topic model, namely hierarchical Latent Dirichlet Allocation (hLDA) model (Blei et al., 2010) uses the nested Chinese Restaurant Process (nCRP) as the prior for sampling topic path distributions. In this hierarchical structure, each node is associated with a topic sampled over words. Along the path from the root node to a leaf node, the model assigns general topics closer to the root and allocates the specific ones down to the tree leaves. Following this tree topology, a general concept normally has a shorter path and a higher level for the corresponding topic. Although hLDA can model topics as paths down to a tree and can allocate topics at multiple levels due to the nCRP prior, it is still established based on the strategy of “bag-of-words” and ignores the lexical co-occurrence, an important feature for text applications. In order to take into account the lexical co-occurrence or word co-occurrence, our model incorporates the concepts of n-grams into hierarchically latent topics for capturing the word dependency that appears in the local context of a word. Before introducing details of the algorithm for allocating general topics and specific topics, a brief discussion for theoretical comparisons with the related work is given as follows.

As discussed previously, the idea of this proposed model comes from hLDA model (Blei et al., 2010), topical N-grams (Wang et al., 2007), and TTM (Celikyilmaz & Hakkani-Tur, 2011) with some concepts of graph models. Fig. 2 illustrates graphical representations of hLDA, topical N-grams, TTM (Celikyilmaz & Hakkani-Tur, 2011).
and our proposed model, the CTM. Based on these diagrams, one can see the CTM is developed from hLDA model and topical n-grams model, but the n-gram indicator (denote as $x_i$ in diagram b and d) in the CTM is a multivariate Bernoulli distribution rather than a latent Dirichlet distribution, which is used for sampling n-gram indicator in the topical N-grams model. The theoretical support for choosing the Bernoulli distribution comes from the definition of the independencies of Bayesian Network in graphic model. More details will be discussed in next section. This simplification is established upon an assumption of independency between topic distributions and n-gram indicator distributions. It makes the joint distribution of this model (see Eq. (1)) can be factorized into two parts. This step simplifies mathematical analysis and model inference, but lack for topic correlations during topic sampling. The significant difference between the CTM and hLDA (see diagram a in Fig. 2) model is the n-gram mechanism included in the CTM. The n-gram property in the CTM model increases the inference precision. This conclusion is proven by the experiment results both in model perplexity evaluation and ROUGE-2 evaluation discussed in Section 4. Furthermore, the hierarchy of topics in the CTM is automatically learned from the data. The length of a path, which is defined as the number of topic nodes traversed for a sentence, is determined by topic distributions rather than arbitrarily assigned, like the higher-level and lower-level topics in TTM (see diagram c in Fig. 2). Therefore, our CTM model presents more sophisticated approach to determine topic structures.

Fig. 3 shows a sample graph using the CTM. Referring to Fig. 1, a snapshot of hierarchical topic path with level assignments, it shows sentence sample path $P(c | w, z)$, which is defined as the path-allocation part in Eq. (1) without showing the hyperparameters here. This graph presents how sentences will be constructed with the topic distributions over words. This figure shows two paths for sentence S1 and S2. Each node (denote as $c$) is associated with a topic distribution over terms. Each edge (denote as $c$) is weighted by the path distribution. Discrete distributions between these two sentence graphs are topic mixtures for each sentence $P(z, x | w, c)$, which is defined as the level-allocation part in Eq. (1). Sentence similarity can be indicated by the Kullback–Leibler (KL) divergence between distributions of two sentences. Comparing the sentence similarity graph discussed in Fig. 2 of Canhasi and Kononenko (2014, p. 538), the method for computing the similarity is totally different. Canhasi and Kononenko (2014) uses archetypal analysis by computing the Euclidean distance between matrixes. Other graph based approaches, like Ferreira et al. (2014), normally use shallow features, such as word frequency, TF-IDF, etc. or latent semantic analysis (LSA), to compute the similarity directly from

Fig. 2. Graphical representations for the hLDA model, the topic n-grams model, the two-tiered topic model (TTM), and the contextual topic model (CTM) are illustrated in diagram a, b, c, and d respectively.
input data. In contrast, Bayesian topic model based approaches (e.g., the CTM reported in this paper) usually evaluate the similarity by either computing the likelihood or approximating posterior distributions, or by performing model comparison with trained data.

3.2. Contextual topic model

As we discussed previously, the CTM proposed in this research is a Bayesian network and its nodes are random variables representing conditional probability distributions of topics with respect to input text and its contextual information. The purpose of developing this model is to establish a mechanism to incorporate the hierarchical concepts of latent topics into n-grams to indicate the topic hierarchies with respect to the word dependencies. One of the main applications for this model is the multi-document summarization. As we discussed previously, the latent topics with n-gram model (Griffiths et al., 2005; Wallach, 2006; Wang et al., 2007) evaluated word dependencies within latent topics, but ignored their hierarchies; the hLDA model justified topic hierarchies, but was short of the consideration for word dependencies. Therefore, a model that can capture both the hierarchies and word dependencies over latent topics is deemed to be more useful for improving performance of summarization. In our model, the idea of the original topical n-gram model is extended to include the hierarchical topics for capturing topic hierarchies and word dependencies. In addition, our model simplifies the sampling process of the probability distributions for the lexical co-occurrence of words by integrating the level distributions with a bigram distribution indicator from a Bernoulli distribution, rather than bring another latent topic distribution for the bigram modeling, which is the approach adapted in the topical n-grams (Wang et al., 2007).

The collapsed Gibbs sampling algorithm is employed in this model to sample the path $c$ and the level allocations to topic $z_{c,n}$ in those paths where $z_{c,n}$ is allocated (Liu, 1994). The process of a sentence generation looks like this: first, the model randomly chooses an L-level path through the topic hierarchies, and then samples the words from the L-dimensional topics that are associated with nCRP prior along this path. One can refer to the original work of Blei et al. (2010) for more details about this generative process. The graphical representation for our contextual topic model is presented in Fig. 2. In this graphic model, $c = \{c_1, c_2, \ldots, c_L\}$ are latent variables that represent the nodes in L-level trees and associate with the nCRP prior. Each level is sampled from the latent topic variables denoted as $z$, which are distributions of parameters $\theta$ as the Dirichlet distribution. In addition, the word dependency is sampled from another hidden variable denoted as $x$ in a Bernoulli distribution, indicating whether the distribution of the observed word is modeled based on the n-gram or a unigram. The theoretical support for choosing the Bernoulli distribution comes from the definition of the independencies of Bayesian Network in the graphic model, variables $w_{1,1}$, $x_i$, and $w_i$ forms a $\nu$-structure where $w_{i-1} \rightarrow x_i \leftarrow w_i$ determines that $w_{i-1}$ can influence $w_i$ via $x_i$ if and only if $x_i$ is observed (Koller & Friedman, 2009). The probabilistic distribution of this structure is $p(x_i|w_{i-1}, \cdot)$. Here variables $x$ and $c$ are supposed as independent since the word variables $w$ are observed. Thus, the posterior conditional distribution for $c$, which is associated with the L topics sampled from word variables $w$ in the sentences, allocates the topics along the path of the sentence generation. The graphical model of the contextual topic model is shown in Fig. 2. The joint distribution $P(w, c, x, z; \alpha, \beta, \delta)$ can be factorized as a product of two conditional probability distributions listed as follows:

$$P(w, c, x, z; \alpha, \beta, \delta) = P(z, x|w, c; \alpha, \beta, \delta) \times P(c|w, z, \gamma, \beta)$$

where variables $w$, $c$, $z$, and $x$ are vectors of distributions of words, paths, topics (or levels of paths) and word dependencies.

In the right side of Eq. (1), two distributions represent the level allocation and path allocation respectively as the results of the collapsed Gibbs sampling. For details about the process of sampling for the path allocation, one can refer the original work of Blei et al. (2010). Here the level allocation is discussed in detail since it includes the word dependency distributions, and is different with the original work of Blei et al. (2010). Furthermore, the probabilistic inference, hyper-parameter estimation, and the method to compute the sentence ranking are discussed in this section as well.

For the level allocation, the joint distribution is listed as follows:

$$P(z_{c,n}|x_{c,n}, X_{-c,n}, W, C, x, \alpha, \beta, \delta) = P(z_{c,n}|x_{c,n}, X_{-c,n}, W, C, x, \alpha, \beta, \delta) \times P(x_{c,n}|Z_{c,n}, X_{-c,n}, \beta, \delta) \times P(z_{c,n}|X_{-c,n}, \beta) \times P(w_{c,n}|z_{c,n}, W, C, \alpha, \beta, \delta)$$

where $z_{c,n}$ denotes the topic associated with the $n$th word in the document $d$, $x_{c,n}$ is the variable of dependency status for the $n$th word in the document $d$. $X_{-c,n}$ denotes the vectors of level allocations or the topic assignments except topic $x_{c,n}$. $X_{-c,n}$ is the vectors of status allocations leaving out $x_{c,n}$, $X_{-c,n} = 1$ indicates the $n$th word sampled from a n-gram model (here we use bigram for a simple description), and $X_{-c,n} = 0$ indicates $n$th word sampled from a unigram. $W$ is the vector of all unique words in the corpus, also called vocabulary in some literature, $C$ is the vector of paths, and $\alpha$, $\beta$, $\gamma$ are hyper-parameters, which are normally initialized as constant numbers.

From Eq. (2), the posterior of the conditional joint distribution for the level consists of four terms. The first term $P(z_{c,n}|x_{c,n}, X_{-c,n}, \beta)$ is a distribution over levels. The second term is the distribution of a given word (i.e., the $n$th word) based on a possible level (topic) assignment. For the details about these distributions, one can refer the original work of Blei et al. (2010). Here we focus on the last two terms in Equation (2).

The third term $P(w_{c,n}|x_{c,n}, \alpha, \beta, \delta)$ is a distribution over a Bernoulli process $\psi_{c,n}$. This distribution has in theory infinite number of components, but the number of words in a corpus is finite in practice and the Bernoulli process is memoryless. Thus, the probabilistic distribution of this term is as follows:

$$p(x_{c,n} = 1|X_{c,n-1}, \delta) = \psi_{c,n}^{x_{c,n}}(1 - \psi_{c,n})^{1-x_{c,n}}$$

where the parameter $\psi_{c,n}$ is the posterior of a Beta prior $\delta|a_0, a_1$, and is estimated as:

$$\hat{\psi}_{c,n} = \frac{a_0 + q_{c,n}}{\sum_{k=1}^{L}(a_k + q_{c,k})}$$

Fig. 3. A sample graph for a 3-level tree using the CTM. S1 and S2 represent sentences and each of them is associated with a path $c = \{c_1, c_2, c_3 \ldots \}$ through the hierarchy, where each node $z$ is associated with topic distribution over terms ($w_1$–$w_{10}$ in this sample). Discrete distributions on the middle of two graphs are topic mixtures for sentence S1 and S2 correspondingly.
where $\alpha_k$ (or $\theta_k$ when $k$ is either 0 or 1) is the hyper-parameter of the Beta prior, and $q_0$ is the number of the status variable $x_{d,i} = k$ ($0$ or $1$) given the $i$th word.

The distribution expressed in the fourth term shows in two different ways: one is for unigram modeling when $x_{d,i} = 0$, another is for $n$-gram modeling when $x_{d,i} = 1$. The distribution for bigram model $p(w_{d,i}|w_{d,i-1}) p(w_{d,i-1}|Z_C, X_{(d-1),i}) W_{(d-1),i}$ is the probability of given word based on a possible topic assignment. The parameters $\phi$ are sampled from a symmetric Dirichlet distribution with hyper-parameter $\beta$. Thus, the probability can be given as follows:

$$p(w_{d,i}|w_{d,i-1}) p(w_{d,i-1}|Z_C, X_{(d-1),i}) W_{(d-1),i} \beta$$

where $n_{d,i}$ is the number of times of the $(i-1)$ th word assigned topic $z$ in document $d$ and $n_{i} = n_{d,i}$ is the count of words assigned topic $z$ excluding the $i$th word. The bigram $p(w_{d,i}|w_{d,i-1})$ is given as a Dirichlet multinomial model, and its value is as follows:

$$p(w_{d,i}|w_{d,i-1}) = \frac{n_i}{N} + (1 - \lambda_{i-1}) \left( \frac{n_{i-1} + \beta n_{d,i-1}}{n_{i-1} + \beta} \right)$$

where $n_{d,i}$ is the number of times the $i$th word immediately follows the $(i-1)$th word, $n_i$ is the number of the $i$th word in document $d$, $n_{i-1}$ is the number of the $(i-1)$th word, and $N$ is the total number of words in the corpus; while $\lambda_{i-1}$ is given as $\beta/(\beta + n_{d,i-1})$. For details about the Gibbs sampler of this part, one can refer to the original works of Wallach (2008) and MacKay and Peto (1995). The distribution for unigram model $p(w_{d,i}|Z_C, X_{(d-1),i}) W_{(d-1),i}$ is the probability of a given word based on the corresponding topic assignment, and its estimated value by Gibbs sampling is as follows:

$$p(w_{d,i}|Z_C, X_{(d-1),i}) W_{(d-1),i} = \frac{n_i}{N} + (1 - \lambda_{i-1}) \left( \frac{n_{i-1} + \beta n_{d,i-1}}{n_{i-1} + \beta} \right)$$

where the notations are same as those described in Eq. (6). Based on Eqs. (6) and (7), the posterior parameter $\phi_{d,i}$ is estimated as:

$$\phi_{d,i} = \begin{cases} \frac{n_{d,i+1}}{n_{d,i} + \beta} \times \frac{n_{d,i-1} + \beta n_{d,i-1}}{n_{i-1} + \beta n_{d,i-1}} & \text{if } x_{d,i} = 1 \\ \frac{n_{i-1} + \beta n_{d,i-1}}{n_{i-1} + \beta n_{d,i-1}} & \text{if } x_{d,i} = 0 \end{cases}$$

Fig. 2 diagram d shows the graphical representation of our contextual topic model. It illustrates the hierarchical relationships between parameters.

The way to estimate the posterior parameter $\phi_{d,i}$ can be found in Blei et al. (2010). In addition, for the estimation of hyper-parameters from data, one can refer to the method discussed in Wallach (2008), which uses the Gibbs EM algorithm (Andrieu, Doucet, & Jordan, 2003).

3.3. Summary generation with contextual topic model

To generate a query-focused summary in a particular summarization application, a relevance-based language model (Lavrenko & Croft, 2001) is used to involve user’s query in the settings of the multi-document summarization. A relevance-based sentence model is built for retrieving candidate sentences based on their relevance to the given query. In addition, this relevance-based sentence model acts as a language-modeling framework to integrate our contextual topic model with a query-focused mechanism. The details of our contextual topic model and the integration with the relevance-based sentence model are discussed in the following sections.

3.3.1. Relevance-based sentence model

Relevance-based language model is a Bayesian generative language model that presents a theoretical justification for estimating relevance based only on user’s query and document corpus (Lavrenko & Croft, 2001). The model is built based on the assumption that both the query and documents are samples from the same unknown relevance model, denoted as $R$. Based on this theorem, given the query $Q$, the probabilistic distribution of sampling a word $w$ in our relevance-based sentence model is as follows:

$$p(w|Q, R) = p(Q|w, R)p(w|R)/p(Q|R)$$

(9)

For more details about the relevance-based language model, one can refer to the original work (Lavrenko & Croft, 2001). Here we focus on how to integrate it with our contextual topic model to rank the sentences over their relevance to the query.

3.3.2. Model Integration and Sentence Ranking

Suppose given the contextual topic model and the relevance sentence model, represented as $C_i$ and $R$, respectively, the process of measuring the relevance of sentences (or ranking the sentences) is the process of predicting likelihood of sentences with the given user query over the models: $C_i$ and $R$. This can be modeled as a joint distribution of the sentence and the query given the model $C_i$ and $R$ in Equation (10).

$$p(w, q|C_i, R) = p(w|C_i)p(q|w, R) \propto p(w|C_i)p(w|q)$$

(10)

where variable $w$ represents the sentence to be ranked, and $C_i$ and $R$ represent the trained contextual topic model and relevance models respectively. The above joint probability distribution can be factorized as a product of two distributions. The probabilistic values of this distribution can be approximated as the measure of the sentence relevance weighted by its predictive likelihood from the contextual topic model.

To facilitate the sentence ranking process, each sentence is assigned a ranking score, which is defined as a log summation of the predictive likelihood of a sentence represented by Eq. (10).

$$\text{score}(w, q) = \sum_{j=1}^{N_q} \log p(w_j|C_i)p(w_j|q, R)$$

(11)

where $N_q$ is the number of words in the sentence $s$, and $N_q$ is the number of terms in the query $q$. For calculating the probabilistic values of $p(w_j|q, R)$, one can refer the work of Croft, Metzler, and Strohman (2010). The probabilistic values of $p(w_j|C_i)$ can in principle be calculated by integrating all parameters from the joint distribution expressed in Eq. (1).

4. Experiments and discussions

4.1. Perplexity of contextual topic model

In order to evaluate our contextual topic model quantitatively, a widely accepted measure, namely perplexity of a held-out test set, is used to measure the quality of the model proposed in this paper. The perplexity is a monotonically decreasing of its value in the quality of the model proposed in this paper. The perplexity is given as:

$$\text{perplexity}(P_{\text{emp}}, q) = \exp \left( -\frac{\sum_{i=1}^{M} \log P_{\text{emp}}(w_i)}{N_d} \right)$$

(12)

Here $P_{\text{emp}}(w_d)$ is the test data likelihood that is estimated by the contextual topic model that was generated using training data set.
Table 1
Perplexity values for various levels against different number of topics on Seafood data set.

<table>
<thead>
<tr>
<th># of topics</th>
<th>Number of levels</th>
<th>Perplexity values</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>50</td>
<td>46.86</td>
<td>84.21</td>
</tr>
<tr>
<td>100</td>
<td>39.28</td>
<td>71.3</td>
</tr>
<tr>
<td>150</td>
<td>34.33</td>
<td>61.84</td>
</tr>
<tr>
<td>200</td>
<td>29.9</td>
<td>48.33</td>
</tr>
<tr>
<td>250</td>
<td>26.62</td>
<td>47.27</td>
</tr>
</tbody>
</table>

Based on this definition, a lower perplexity value indicates better predictive performance of the model.

In our experiment, the NIPS corpus is used to perform the model training and held-out test. The NIPS data set contains 1732 conference papers with 46,874 unique terms from NIPS conferences between the years 1987 and 1999. 10% of the data is held out for test purpose, and remaining 90% of the data is used for training the contextual topic model. AQUINT dataset in DUC 2006 corpus is also used for evaluating the model perplexity. The DUC 2006 corpus contains 750 news articles with 23,663 unique terms from the Associated Press and New York Times (1998–2000) and Xinhua News Agency (1996–2000). The Seafood dataset is a smaller data set, and includes 156 text documents with 13,031 unique terms. The contents of that data are about seafood industries.

We compared our contextual topic model to hLDA model and LDA model described in the literature (Blei, Ng, & Jordan, 2003; Blei et al., 2010). All models ran for 1000 iterations of the Gibbs sampler with \(\alpha, \beta, \gamma, \delta\) hyper-parameters set as 10.0, 1.0, 1.0, and 0.95 respectively. The number of hierarchy levels was set to 3 for both contextual topic model and hLDA model. In addition, all models were trained based on the same data and using the Gibbs EM with exactly the same stopping criteria as well. Lower perplexity values indicate better generalization performance and higher possibility to avoid overfitting.

4.2. Topic hierarchy estimation

For both the CTM and hLDA, the number of hierarchical levels assigned to the model influences its hierarchical structure. Although the CTM does not need to know the topic numbers a priori, the number of topics allocated for each level is different when various levels are assigned to a model. In order to estimate hierarchical structures of topics and find a suitable hierarchical level for our summarization application, an experiment was conducted using Seafood data set with the same hyper-parameters \((\alpha, \beta, \gamma, \delta)\) but various level settings from the extreme case as the number of levels is set to 1, to 5, which is the level number when perplexity of the model showed no significant sensitive to the level numbers. Table 1 reports the results of the estimation for hierarchical structure with different topic number and level settings.

Fig. 5 illustrates the comparison results listed in Table 1. One can see the perplexity values decrease almost linearly against the number of the topics when the hierarchical level is set to 3. This experiment for topic hierarchy estimation reported the same findings for the hierarchy level settings given by hLDA model (Blei et al., 2010).

Moreover, for better understanding of the hierarchical structures in an application, Fig. 5 illustrates the results of estimating hierarchies on AQUINT data set, which is the corpus used for evaluating summarization performance in this paper. One can see that the topics were separated properly in a hierarchy when the number of levels was 3.

The integer codes listed in Fig. 6 show the properties of the path and levels in a topic hierarchy. The first number indicates the level of a node in a path, followed by total number of topics for that path inside a bracket. The last number separated by a forward slash shows the number branches (child nodes) in that path. For example, 0(1125)/363 means that this node is the root with 1125 topics and 363 child nodes in the entire tree. Since this example came from a 3 level hierarchy model, the level 3 nodes all have no child node (all show/0). In this example, each node contains the top ten words including phrases, which are underlined words, from the corresponding topic distribution with respect to the word dependency modeled by the n-gram.

As shown in Fig. 6, our model discovered interesting topics and their structures. In the root of the hierarchy, the most of discovered topics are function words. In the cases of 3 levels hierarchical structure, the model has not only captured the function words but also delineated subtrees within various domains (such as politics and climate in the example shown in Fig. 6). Furthermore, our model captured interesting phrases including the name entities. These results suggest that contextual topic model can be an effective tool in text applications and benefit tasks of multi-document summarization.

4.3. ROUGE evaluation

4.3.1. Experimental data

Two data sets from the DUC 2005 and DUC 2006 tasks, in which the query-focused multi-document summarization was the main task, were used to perform the evaluation task. There were 50 document clusters in each corpus. Each document cluster contained 25 news articles selected from TREC and AQUINT, and each cluster associated with a topic description. The topic description in each cluster was used as the original query for training the relevance model and ranking the sentences. The corpus from DUC 2005 was used for training and parameter tuning and the corpus from DUC 2006 was used for testing.

4.3.2. Evaluation metrics

This experiment used ROUGE (Lin, 2004; Lin & Hovy, 2003), the official evaluation toolkit for text summarization in DUC, to evaluate the performance of our summarization system. The evaluation metrics: ROUGE-1, ROUGE-2, and ROUGE-SU4 were investigated during the experiment, and the average recall values of these metrics were reported in this experiment. The ROUGE-1 measurement evaluates how well the testing summary is consistent with human judgments (Lin & Hovy, 2003), the ROUGE-2 metrics examines the performance of the bigram overlap, and the ROUGE-SU4 metrics evaluates the performance of the skip-bigram.

4.3.3. Experimental results

In this experiment, the summarization system that implemented the CTM and a relevance-based sentence model was compared with other three summarizers to demonstrate the effectiveness of our proposed method for multi-document summarization. Those three systems were: a baseline system, which was a relevance model based summarizer without query expansion (namely RMSUM and Run ID as RM shown in Table 2). The RMSUM was built upon the relevance based language model (Lavrenko & Croft, 2001) with Dirichlet smoothing approach (Zhai & Lafferty, 2001). The second system expanded the baseline system with query expansion that used topic words and phrases generated from the CTM. This summarizer was called ERMSUM and Run ID was ER. The third system employed the topical n-gram model (Wang et al., 2001).
2007) with the relevance based sentence model (namely TopicalN
and Run ID as TN shown in Table 2). The system implemented the
CTM was named as CTMSUM and Run ID was CTM, as shown in
Table 2. In order to compare the performance of the multi-
document summarization solution addressed in this study, the
results of three summarizers with the highest ROUGE scores in
DUC 2006 are also listed in Table 2.

In Table 2, Run-ID 24, 12, and 13 are for systems IIITH-Sum
(Jagarlamudi, Pingali, & Varma, 2006), OnModer (Ye & Chua,
2006), and SFU_v36 (Melli et al., 2006), respectively. Comparing
the results of CTMSUM with baseline system RMSUM, ERESUM,
which is the summarizer based on relevance model with query
expansion, and TopicalN, which is the topical n-grams model based
summarizer with query expansion, one can see the significant per-
formance gains achieved in CTMSUM system. It has achieved better
results in recall measures for all major ROUGE evaluation metrics
recommended by DUC 2006 (Dang, 2006). Fig. 7 illustrates the
performance comparisons of those seven systems. This result
indicates that the contextual topic model proposed in this research
is a plausible approach that can improve the performance of
summarization.

Comparing the ROUGE evaluation results of CTMSUM with the
baseline RESUM, one can see the significant performance gains
achieved in CTMSUM system. The reported average recall values
in CTMSUM are increased by 8.9%, 33.8%, and 20.3% in ROUGE-1,
ROUGE-2, and ROUGE-SU4 respectively over the baseline system
RMSUM. Fig. 8 presents this performance gain. This result suggests
that the system CTMSUM has achieved significant improvements
in ROUGE evaluation, especially in ROUGE-2.

5. Discussion

As shown in Fig. 4, the CTM is somewhat sensitive to the topic
numbers due to the symmetric Dirichlet prior for the topic sam-
pling. The comparison between the CTM and the LDA have shown
that for all data sets in three corpora, the CTM obtains much lower
perplexity values when the number of topics is more than 40.
hierarchical structure for topics, and the probabilistic values of topics are weighted with their level distributions. Therefore, when the number of the hierarchical levels of a sampling path (each path represents a document) increases, the number of topics allocated to each level decreases. Although the number of topics can be theoretically infinite in a Bayesian hierarchical topic model, in realistic case, it is finite because the observed document always contains certain number of words (Blei et al., 2010). This is also the reason that the perplexity values have less variance when the number of sampling topics is greater than 150.

In the NIPS data set (upper-left panel of Fig. 4), contextual topic model demonstrates lower perplexity values in a wide range of topics than both the hLDA and LDA models. It shows similar perplexities as the hLDA model in the range of 10–30 topics, but much better than the LDA model in the entire range of topics. Similar results were reported in the experiments for other two data sets. All three models reported larger perplexity values in the ACQUIT data set, but the CTM outperformed the hLDA and LDA when the number of topics was above 50. In the Seafood data set, the CTM outperformed other two models in the entire range of topics. In addition, all perplexity values were smaller than values evaluated in other three corpora since the Seafood corpus is the smallest data set in size and has only 156 documents with 13,031 terms in total.

As shown in Fig. 6, the CTM shows an interesting phenomenon of “rich-get-richer” process. Some tree branches have more and more children and the descants branches and numbers of leaves can become quite large due to the stochastic process called nCRP prior employed in the CTM. This phenomenon suggests that the CTM can produce word distributions following the power-law distribution. Another interesting phenomenon is that the number of levels can significantly influence the hierarchical structure of latent topics. As shown in Fig. 5, when the value of \( L \) (the variable of level) increases, for example \( L = 5 \), the amount of topic numbers should also increase in order to obtain the similar performance. Although the CTM is a non-parameter Bayesian topic model, it still needs a finite number for sampling latent topics in each level. When the value of \( L \) is set to 3 or 4, which depends on the size of the corpus (for example, for NIPS dataset and ACQUIT, the CTM obtained

**Table 2**

Experimental results of ROUGE evaluation on DUC 2006 data.

<table>
<thead>
<tr>
<th>Run-ID</th>
<th>System</th>
<th>Recall</th>
<th>ROUGE-1</th>
<th>ROUGE-2</th>
<th>ROUGE-SU4</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTM</td>
<td>CTMSUM</td>
<td>0.4175</td>
<td>0.0986</td>
<td>0.1548</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>IIITH-Sum</td>
<td>0.4098</td>
<td>0.0951</td>
<td>0.1546</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>OnModer</td>
<td>0.4049</td>
<td>0.0899</td>
<td>0.1476</td>
<td></td>
</tr>
<tr>
<td>TN</td>
<td>TopicalN</td>
<td>0.4032</td>
<td>0.0908</td>
<td>0.1502</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>SFU_v36</td>
<td>0.3846</td>
<td>0.0799</td>
<td>0.1353</td>
<td></td>
</tr>
<tr>
<td>ER</td>
<td>ERMSUM</td>
<td>0.3914</td>
<td>0.0796</td>
<td>0.1365</td>
<td></td>
</tr>
<tr>
<td>RM</td>
<td>RMSUM</td>
<td>0.3836</td>
<td>0.0742</td>
<td>0.1287</td>
<td></td>
</tr>
</tbody>
</table>
better results when \( L \) was set to 3), the CTM can get a better hierarchical structure as shown in Fig. 6. Another interesting phenomenon is observed that the variance of perplexity values is very small when the number of topics is greater than 160 in NIPS data set and 200 in Seafood data set. This result shows that the corpus’s size can influence the model performance as well. Since the number of hierarchical levels and the amount of topics sampled for each level can significantly affect the overall performance of the model, the number of levels for the hierarchical structure and the amount of topics sampled in the model are set to 3 and 150 respectively in our experiment for the particular application in multi-document summarization.

As shown in Figs. 7 and 8, these experimental results validate that our proposed approach can improve the summarization performance significantly. Our system CTMSUM outperforms the top performing system (Run-ID as shown in the figure, which is the one that performed the best ROUGE evaluation in DUC 2006) on all major recall metrics of ROUGE-1, ROUGE-2, and ROUGE-SU4 (increased by 1.4%, 4.5%, 0.07%, respectively). From Fig. 8, one can see that our system CTMSUM has achieved significant improvements in ROUGE evaluation, especially in ROUGE-2 evaluation. In addition, the ROUGE-2 evaluation recorded in (Canhasi & Kononenko, 2014) is 0.0917 for DUC2006 corpus. The same evaluation obtained in our experiment is 0.0986 and the gained 7.5% increase. With respect to both topic hierarchies and lexical co-occurrence of words, this result demonstrates that contextual topics produced by the CTM indeed impact the performance of summarization significantly. Another interesting phenomenon revealed in this experiment is that the summarizer ERMSUM, which is a relevance model based summarizer with query expansion, has not achieved too much improvement over the baseline system. A possible reason is that the CTM generates the expanded query over a bigram language model; however, the unigram-based model might not be in favor of a bigram-based query expansion. This limitation has resulted in the identification of a future study for more sophisticated method to integrate the expended queries with the bigram and the unigram models effectively.

6. Conclusions and future work

This paper has presented a contextual topic model for multi-document summarization in which each sentence is viewed as hierarchical topics with respect to contextual information, and has shown how this model can be used to gain insight into multiple documents and use it to determine the sentence similarity. The extensive experimental results for the perplexity of our model have demonstrated that the number of hierarchical levels can significantly affect topic hierarchy and the number of topics allocated to each level of this hierarchy. The experimental results on several data sets have demonstrated our model has gained better performance in model prediction. The results of experiment on the DUC2006 corpus for a summarization system implementing our model have demonstrated the effectiveness of the proposed approach and justified it as a comparable summarizer to the state-of-the-art summarization systems.

One of the main contributions of this work is that a new model based on a Bayesian based hierarchical topic model has been proposed to solve the problem of multi-document summarization. The addressed approach has verified that the contextual information conveyed by the lexical co-occurrence is helpful to justify the relevance and similarity of sentences in multi-document summarization. In practice, the contextual information is represented as model parameters that are trained from input data. This makes summarization systems that employ this model have context feature embedded systemically. Another main contribution is the creation of a new method for model integration that allows summarization systems to easily take account of user query when ranking sentences. In addition, a fully mathematical basis for this integration was established by using Bayes’ rule and chain rules in probability theory. Another contribution of this work is that it performs a complete system comparison to evaluate the performance. The analysis results indicate our model is comparable to the state-of-the-art summarization systems, and outperforms the benchmark systems on all major recall metrics of ROUGE-1, ROUGE-2, and ROUGE-SU4.

Practical implications of this work can be categorized within the perspectives of text summarization, word sense disambiguation and hierarchical clustering in natural language processing, and intelligent tutoring system. Our proposed approach may provide an appropriate mechanism to apply contextual information into particular applications, like events for event summarization,
personal interests or preferences for personalized automatic summarization, and opinions for opinion summarization. Those events, interests, or opinions are all kinds of contextual information. From the perspective of word sense disambiguation, indicating a word in context is an essential step to determine word senses. Some particular applications include automatic translations, thesauri, etc. From the perspective of a clustering problem, identifying the context of a word can reduce the uncertainty about the next word. A particular application is the language model for information retrieval. From the perspective of intelligent tutoring system, the results of this work can be applied to improve students’ learning performance in online learning environments, especially in mobile and ubiquitous learning settings.

Although our proposed model has shown potentials to be used in a variety of tasks, several limitations have to be pointed out for further improvement. First limitation is the performance issue. The model has to take longer time to be trained under a larger data set in order to keep certain level of accuracy in prediction. Second, a Bernoulli distribution is used for indicating word co-occurrence in the model. It is a simple solution and is poor to cover co-occurrence among multiple words. Other limitations include the problem of sentence coherence and a lack of online settings for stream text because the model has been trained using a large data corpus before summarizing documents.

In future study, hierarchical Bayesian nonparametric topic models can be considered to reduce the parameter space and eventually to improve the performance. In addition, to catch co-occurrence among multiple words, a multivariate Gaussian distribution can replace the Bernoulli distribution for covariance. For the coherence problem, other types of contextual information, like the time for an event occurred, similarity between sentences to the main topics, etc., can be used as extra features for sentence coherence in the model. Nonetheless, the findings of this work also encourage us to extend this model to different tasks, such as the emotion recognition, image processing and event detection.

Acknowledgements

The authors acknowledge the support of NSERC, ICORE, Xerox, and the research related gift funding by Mr. A. Markin. This study is also partially supported by the College of Management, National Sun Yat-sen University, TAIWAN under grant CMNSYSU-SRS-2013-02 and the National Science Council, Taiwan under Grants NSC 102-2911-I-110-501 and NSC 101-2511-S-110-003-MY3.

References


