A Contextual Query Expansion Based Multi-document Summarizer for Smart Learning

Guangbing Yang∗, Kinshuk†, Dunwei Wen†, and Erkki Sutinen∗

∗School of Computing, University of Eastern Finland
P.O. Box 111 80101 Joensuu, Finland
Email: yguang@cs.joensuu.fi, erkki.sutinen@cs.joensuu.fi
†School of Computing and Information Systems, Athabasca University
1 University Drive, Athabasca, Alberta T9S 3A3, Canada
Email: dunweiv@athabascau.ca, kinshuk@athabascau.ca

Abstract—Smart learning environments have recently emerged as education solutions that integrate digital devices, digital learning content, and software for more effective and interactive learning settings. However, problems of information overload and duplication restrict students’ interactivity with the smart learning environment and limit its development. An approach in contextual query expansion based multi-document summarization is proposed in this paper to provide a solution to alleviate these problems. In our approach, a multi-document summarization system is built upon a topical n-grams model with a query expansion algorithm to capture the contextual information conveyed by word order and abstract topics in documents for enhancing sentence ranking in text summarization. The experimental results show that our proposed approach can significantly improve the summarization performance and eventually benefit students in the smart learning environment.

Keywords-smart learning environment, contextual topics, multi-document summarization.

I. INTRODUCTION

Smart schools and smart learning environments have recently emerged as education solutions that combine digital devices, digital learning content (especially Web based learning content), and software for a more effective and interactive learning setting. The successful establishment of this interactive learning environment depends on how information can be shared and correlated smoothly among students, teachers, and smart devices. However, due to the rapid increase of enormous amount of online information, the problem of the oft-decried information overload and duplication negatively impacts students’ interactivity with the smart learning environments and correlation between each other. The concerns of how to alleviate these problems have given rise to interest in research on multi-document summarization. A main task of multi-document summarization is to determine the importance and similarity between sentences in a set of documents. Many solutions have been addressed to justify the importance of sentences, normally by using either statistical methods or Bayesian language models to produce summaries. Most of them have been established based on the “bag-of-words” simplification and the term-frequency measure to generate summaries without considering word order in a string of text. Word order, however, not only conveys important grammatical information and lexical meaning, but also specifies the context in which words appear. This contextual information is helpful to improve the evaluation of the importance of sentences in multi-document summarization [3].

Recent research has benefited from Bayesian topic models to represent the corpus concepts as abstract topics [1], [2], [13], [16]. These topic models identify important sentences that contain such abstract topics rather than specified topics. Experiment results have shown that these models have improved summarization performance significantly. Nonetheless, all these models were built upon the “bag-of-words” simplification without considering the contextual information conveyed by word order. Since both the Bayesian topic models and the contextual information are useful for multi-document summarization, an approach that can deal with both is deemed to improve the evaluation of sentence importance and eventually benefit summarization task. Therefore, questions arise here: what performance will be when the contextual information is used in multi-document summarization? Furthermore, how to use topic models to represent the contextual information, and how to informatively use “contextual topics” in multi-document summarization to identify salient sentences more effective. Eventually, from the perspective of education, what will be an implication to the smart learning environment?

This paper introduces a novel approach that is built upon the topical n-grams model [22] for contextual topics within the query-likelihood (QL) language modeling framework [8] for query expansion and sentence ranking. The contextual topics can be described as topics investigated within the context that mainly refers to the proceedings of words in a sentence of a document.

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

II. RELATED WORK

Recent research in multi-document summarization has focused on using Bayesian statistic approaches in machine learning and language modeling to find important sentences and words in multiple documents. BayeSum model [13] applied Bayesian principled topic model to summarize sentences within the query expansion technique in information retrieval. TopicSum [2] imposed Latent Dirichlet Allocation (LDA), a topic model [9], to estimate and allocate word distributions during the sentence selection. Two-Tiered Topic Model (TTM) [1] used hierarchical topic model to deal with both issues of sentence ranking and coherence. Experimental results in ROUGE [7] measurement and manual evaluation in DUC have shown these models can improve summarization performance significantly. However, none of these approaches have considered the order of words that conveys important grammar structure and lexical meaning, which can significantly affect accurate meaning of a sentence [3].

Several Bayesian topic models have been developed recently to incorporate the concept of latent topics into n-gram language models, such as structured topic model [14], HMM-LDA model [20], and topical n-grams [22]. Structured topic model [14] is a bigram topic model that combines LDA [9] with the hierarchical Dirichlet language model [12] by integrating latent topic variables into a hierarchical Bayesian language model. HMM-LDA model [20] combined LDA [9] model with a hidden Markov model (HMM) to capture both semantic dependencies indicated by word co-occurrences and syntactic dependencies described by the order of words. Two components are built in this model, in which the syntactic component is an HMM and the semantic component is a topic model. HMM-LDA model has demonstrated excellent results on tasks of part-of-speech and document classification although its bigram component does not carry any topic assignment. Topical n-grams model [22] has addressed an approach to automatically determine whether a topic assignment is essential for a bigram. Experiment results have shown this model can improve performance significantly in ad-hoc information retrieval on TREC collections. Although these models have demonstrated their success in various tasks in information retrieval and document classification, finding a suitable method to adapt these techniques to the multi-document summarization is indeed a meaningful work due to the difference between the task of information retrieval and multi-document summarization.

III. CONTEXTUAL QUERY EXPANSION BASED SUMMARIZATION

A. Core concept

Generally, the contextual information helps understand the content. It can be the identity of things named in the content, or topics, keywords, etc., as part of interpretive information. In this study, the contextual information is restricted to those immediately preceding terms of a word and their associations with topics. This contextual information not only represents word order, but also indicates semantic correlation between words in a document. Topics that are modeled across this contextual information are called contextual topics.

Topical n-grams model [22] is a mixture model of the bigram model and latent Dirichlet allocation (LDA) [9] topic model. It can generate the contextual topics by automatically determining whether or not two successive terms are indeed a bigram if they are meaningful phrases or collocations within the context of the document. For instance, a phrase “apple computer”, if assigned a topic to each individual word, the term “apple” will be mostly assigned to “fruit” topic, but two terms together as a collocation will be assigned to topic “personal computer” or “Mac” if the context of the document is relevant to computer or technology. The idea of employing the topical n-grams model in multi-document summarization is for generating a contextual topic from a bigram distribution if a meaningful phrase or a collocation exists within the context of the content.

The key process in our approach relies on the topical n-grams model and the query likelihood (QL) language model. More specifically, the topical n-grams model generates contextual topics, which perceive the contextual information of the words and present it as key phrases. The QL model provides a language-modeling framework to integrate contextual topics with query likelihood for recognizing semantic correlations between words, and to collect candidate sentences for representing the most relevance to the query and contextual topics.

B. Contextual topics as query expansion

The topical n-grams model uses Gibbs sampling to approximate the joint posterior distribution of parameters. One can refer the original work [22] for details of how to approximate the conditional probability of topics using Gibbs sampling. Here we focus on the integration of the topical n-grams model [22] into the query likelihood model [8] at the level of query term. The mixture distribution of the topical n-grams and QL model is given as follows:

\[
p(q_i|q_{i-1}, d) = \lambda P_{QL}(q_i|d) + (1 - \lambda)P_{topical-n}(q_i|q_{i-1}, d) \quad (1)
\]

where \( \lambda \) is an empirical value for weighting purpose, its optimal value is set to 0.4 for the best performance in our experiment. \( P_{QL} \) represents the query likelihood model and its probability value is computed by using the Dirichlet smoothing method [8]. \( P_{topical-n} \) can be represented as
follows [22]:

\[ P_{\text{topical} – n}(q_i \mid q_{i-1}, d) = \sum_{z_i=1}^{T} \left( (P(x_i = 0 \mid \hat{\psi}_{q_i-1}) P(q_i \mid \hat{\phi}_{z_i}) + (P(x_i = 1 \mid \hat{\psi}_{q_i-1}) P(q_i \mid \hat{\psi}_{z_i}) \right) P(z_i \mid \hat{\theta}^{(d)}) \]  

(2)

The \( \hat{\psi}_{q_i-1}, \hat{\phi}_{z_i}, \hat{\psi}_{z_i}, \hat{\theta}^{(d)} \) are posterior of parameter \( \psi, \phi, \sigma, \) and \( \theta \). The detailed mathematical approach for estimating these posterior parameters can be found in [22]. Table 1 shows an example of the generated contextual topics represented as key words and phrases from a document cluster D0611B in Document Understanding Conferences (DUC) 2006 corpus. In this example, the overall topic for this document cluster is about “organic methods of pest control,” and the weighting score is the quantity value of the product of \( P_{\text{topical} – n} \) over all query terms using Equation (2). In this example, the topical n-grams model selected

Table 1
EXAMPLE OF THE KEY WORDS AND PHRASES FROM DOCUMENT CLUSTER D0611B IN DUC 2006 CORPUS

<table>
<thead>
<tr>
<th>Topics</th>
<th>Score</th>
<th>Phrases</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>farming</td>
<td>0.0995</td>
<td>flea beetles</td>
<td>0.0216</td>
</tr>
<tr>
<td>beetles</td>
<td>0.0077</td>
<td>organic farming</td>
<td>0.0138</td>
</tr>
<tr>
<td>pest</td>
<td>0.0085</td>
<td>pest control</td>
<td>0.0083</td>
</tr>
<tr>
<td>garden</td>
<td>0.0149</td>
<td>organic gardeners</td>
<td>0.0083</td>
</tr>
<tr>
<td>test</td>
<td>0.0112</td>
<td>Methyl Bromide</td>
<td>0.0066</td>
</tr>
<tr>
<td>tree</td>
<td>0.0112</td>
<td>eucalyptus trees</td>
<td>0.0049</td>
</tr>
<tr>
<td>fertilizers</td>
<td>0.0019</td>
<td>synthetic fertilizers</td>
<td>0.0049</td>
</tr>
<tr>
<td>plant</td>
<td>0.0087</td>
<td>plant pathologist</td>
<td>0.0038</td>
</tr>
</tbody>
</table>

some meaningful phrases and collocations. For instance, the phrase ‘flea beetles’, ‘Methyl Bromide’, and ‘organic farming’, etc. are terms strongly related to the topic of the document cluster. Other phrases, such as ‘pest control’, ‘synthetic fertilizers’, and ‘plant pathologist’, etc., are very detailed information revealed by the topical n-grams model to capture strong cohesion with the overall topics of the document cluster.

C. System implementation

An overview of the implementation of our contextual query expansion based multi-document summarizer is described as follows. In addition, to help better understand the implementation, an overall system architecture is presented in Figure 1.

First, before generating the contextual topics and ranking sentences, the system preprocesses each document in the corpus to label the stop words and segment it to a set of sentences. In addition, all the stop words in the corpus are labeled. If the immediately preceding term of a word is marked as a stop word or this word itself is the first term in the sentence, the word will be sampled as a unigram rather than a bigram. Second, the contextual topics are generated within the context of word sequence in a sentence, and are represented as a set of key words and phrases, which will be used as query expansion for sentence ranking. In this step, a mixture model, which is the combination of topical n-grams model and the query likelihood model, is generated for sentence ranking.

Third, a very important step for effective expansion is to choose the topic words that are appropriate for the context [21](p. 201). The generated topics are selected as the expansion terms. But there are some redundancies between original query terms and the topic terms. To alleviate this issue, the expected mutual information measure (EMIM) [21] is used in this study. The method is described as follows: (1) Each original query term compares with the generated topic words or phrases, if they are exactly the same word, the topic word or phrase is kept as the expansion query term. Note that the second term in the phrase is used to compare with the query term because normally the second term in a phrase takes the main meaning in the entire phrase. (2) If the terms are different, the EMIM is used to estimate their co-occurrence (how close they are within a context of the document cluster). If the EMIM estimation shows that they are close and the threshold of total number of terms for the query expansion is not reached, both terms in the original query and generated topics are kept in the expansion list. If their co-occurrence is low, the term in topics is kept when the threshold limit is not reached. The threshold is set to the summation of the number of terms in the original query (no stop words count) plus a constant. The algorithm for this expansion term selection is listed in Figure 2. In this algorithm, \( N_q \) is the number of windows (documents) containing word \( Q_{q_i} \); \( N_{q(i)} \) is the number of windows containing both words \( Q_{q_i} \) and \( T_i \).

After the co-occurrence evaluation, the original query is then updated with the contextual topics, and the QL-based sentence-ranking algorithm is performed over the updated query and each candidate sentence is assigned a ranking
Equation (3) shows a way to compute the ranking score of topic n-grams and QL model is used for summarization, the score, which is a log summation of probability values applied in our approach to reduce redundancy during the summary generation [15].

For sentence ranking, each sentence is assigned a ranking score. Thus, the Maximal Marginal Relevance (MMR) is applied in our approach to reduce redundancy during the summary generation [15].

D. Query likelihood based sentence ranking

For sentence ranking, each sentence is assigned a ranking score, which is a log summation of probability values calculated using Equation (1). If the combination of the topic n-grams and QL model is used for summarization, the Equation (3) shows a way to compute the ranking score of a sentence:

\[ rs(s, q) = \sum_{i=1}^{N_s} \log \left( \prod_{j=1}^{N_q} p(q_i|q_{i-1}, w_j) \right) \]  

(3)

Where \( N_s \) is the number of words in the sentence \( s \), and \( N_q \) is the number of words in the query \( q \) (including query expansion terms). The first term of the query or the sentence is always sampled by the unigram. If only the QL model is used for summarization (Note: the QL model is used as the baseline summarization system in our experiment for system comparison), the Equation (4) is used to compute the ranking score of a sentence:

\[ rs(s, q) = \sum_{i=1}^{N_s} \log \left( \prod_{j=1}^{N_q} p(q_i|w_j) \right) \]  

(4)

E. Redundancy detection

After the sentence ranking process, the top-ranked sentences are selected and extracted for a summary. In order to reduce redundancy, the MMR algorithm [15] is applied in our approach to remove the redundant sentences from the summary. The MMR approach has been widely used in query-focused summarization and is easy to fit into our ranking process. In the process of the redundancy detection, the similarity between the sentence and the query and the similarity between the sentence and the summary sentences, which are already selected in the summary, are computed by using the MMR algorithm to determine the most suitable sentence as the one that is most similar to the query and least similar to the text already in the summary. This process continues until the length of the summary reached the predefined limit, which is 250 words in this case. After this process, the sentences with the highest scores and the lowest redundancy are selected to create a final summary.

IV. EXPERIMENT AND DISCUSSION

A. Experimental data and evaluation metrics

The experiment used two data sets from the DUC 2005 and DUC 2006, in which the query-focused multi-document summarization was the main task, to evaluate the performance of our summarization solution. The document clusters in the DUC 2005 was used for training the topical n-grams model and tuning parameters. The data from the DUC 2006 corpus were used for testing. There are 50 document clusters in each corpus, and each cluster includes 25 news articles selected from Text RETrieval Conference (TREC) and AQUAINT.

This experiment used ROUGE [7], 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 recall, precision, and F measures of these metrics were reported in this experiment.

B. Experimental results

In this experiment, our contextual query expansion based multi-document summarizer was compared with the baseline system, which was the query likelihood language model based summarizer without query expansion (namely QL-Sum), to demonstrate the effectiveness of our proposed method for multi-document summarization. In order to compare the performance of the summarization solution addressed in this study, the results of three benchmark summarizers with the highest ROUGE scores in DUC 2006 were also listed in Table 2.

Table 2 shows the system comparison results of ROUGE-1 on DUC 2006 data. Run-ID 24, 12, and 13 are system IIITH-Sum [17], OnModer [19], and SFU-v36 [11], respectively, and they were the benchmark models in DUC 2006. Run-ID T1 and Q1 are our systems: Topical-N and QL-Sum, respectively. QL-Sum is the baseline system in which only the query likelihood model is used for choosing and ranking

\[
Q \leftarrow \text{the original query} \\
T \leftarrow \text{the list of generated topics} \\
E \leftarrow \text{the expansion list} \\
\text{while } d \leq |D| \text{ do} \\
\quad \text{while } q \leq |Q| \text{ do} \\
\quad \quad \text{while } t \leq |T| \text{ do} \\
\quad \quad \quad \text{if } Q_t = T_t \text{ then} \\
\quad \quad \quad \quad E \leftarrow E + T_t \\
\quad \quad \quad \text{else} \\
\quad \quad \quad \quad p = p(Q_q, T_t) \times \log \frac{p(Q_q, T_t)}{p(Q_q)p(T_t)} \\
\quad \quad \quad \text{end if} \\
\quad \quad \text{end while} \\
\quad \text{end while} \\
\text{end while} \\
\text{end while} \\
\text{end if} \\
\text{end while} \\
\text{end while} \\
\text{Where } p(Q_q) = N_q/N, p(Q_q, T_t) = N_{q|t}/N, \text{ and } p(T_t) \text{ is the weight score calculated using [20]. Compare } p \text{ for each term, for maximum } p, \text{ add the corresponding term to } E \text{ if length of } E \text{ is less than the threshold. } |D| \text{ is the number of documents, } |Q| \text{ is the number of terms in the query, and } |T| \text{ is the number of the topic generated.}
\]
Figure 3. Performance comparison between different systems

Table II
EXPERIMENTAL RESULTS OF ROUGE-1 ON THE DATA OF DUC 2006

<table>
<thead>
<tr>
<th>Run-ID</th>
<th>System</th>
<th>ROUGE-1</th>
<th>ROUGE-2</th>
<th>ROUGE-SU4</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>Topical-N</td>
<td>0.4010</td>
<td>0.0893</td>
<td>0.1459</td>
</tr>
<tr>
<td>24</td>
<td>IIITH-Sum</td>
<td>0.4098</td>
<td>0.0951</td>
<td>0.1546</td>
</tr>
<tr>
<td>12</td>
<td>OnModer</td>
<td>0.4049</td>
<td>0.0899</td>
<td>0.1476</td>
</tr>
<tr>
<td>13</td>
<td>SFU-v36</td>
<td>0.3846</td>
<td>0.0799</td>
<td>0.1353</td>
</tr>
<tr>
<td>Q1</td>
<td>QL-Sum</td>
<td>0.3541</td>
<td>0.0659</td>
<td>0.1186</td>
</tr>
</tbody>
</table>

Comparing the results of Topical-N with the baseline QL-Sum, one can see the significant performance gains achieved in Topical-N system. The reported recall in Topical-N system increase 13.2%, 35.5%, and 22% on ROUGE-1, ROUGE-2, and ROUGE-SU4 evaluations over the baseline QL-Sum, respectively. This result indicates that the contextual information conveyed by word order can positively impact the performance of summarization. Figure 4 presents this performance gain. In this diagram, the upper panel illustrates the recall gains in ROUGE-1, ROUGE-2, and ROUGE-SU4 evaluations, and the lower panel shows the precision gains. From this diagram, one can see the system Topical-N achieve significant improvements in ROUGE evaluation, especially in ROUGE-2.
C. Discussion

These experimental results evaluate that our approach proposed in this study can improve the performance of summarization although our system, Topical-N, only outperforms the third benchmark system on recall metrics of ROUGE-1 and 2, and presents a little poor over IIITH-Sum (Run-ID 24) and OnModer (Run-ID 12) on ROUGE-1, ROUGE-2, and ROUGE-SU4, respectively. Our system presents a mediocre performance in average recall evaluation. This result implies the ranking algorithm (see Equation (3)) applied in our solution might be biased in favor of bigram rather than the unigram because the difference between the average recall of ROUGE-2 evaluation in Topical-N and IIITH-Sum (Run-ID 24) is 0.6% rather than 0.9% in ROUGE-1 and 1.0% in ROUGE-SU4 evaluation. Furthermore, the increase in average precision of ROUGE-2 evaluation in Topical-N is much bigger than the increase in average precision of ROUGE-1 and ROUGE-SU4 evaluations if compared with the baseline system. Nonetheless, our summarizer can produce more coherent summaries because the meaningful phrases and collocations are specified explicitly as contextual topics during the summarizing process.

D. Implications for the smart learning environment

Recent research shows that the Net Generation students learn most through interactivity [4]. The growing requests for the immediately useful learning content from students increasingly demand the interactivity of the teaching and classroom environment. Thus, a learning environment shall have capability to provide adaptive learning content to students for their immediate use and right away interactivity. This immediateness determines that the size of learning content has to be small enough but the information covered by this condensed content need to satisfy students’ learning request. Research in summarization for mobile learning [12] shows that a summary with around 30 percent compression rate can reach 90 percent learning achievement but only needs 44 percent learning time. Based on their conclusion, a properly summarized learning content is not only able to satisfy learning achievements, but also able to align content size with the unique characteristics and affordances of mobile devices [12]. Furthermore, from the perspective of the concept of “learning spaces”, employing technology of summarization in smart learning environment promotes connection of social spaces because the condensed learning contents not only facilitate learners by reducing reading time but also make ubiquitous mobile devices even more attractive to both educators and learners by aligning unique characteristics of mobile devices [18], [23]. Study of the effectiveness of the summarization in mobile learning settings has investigated that the properly summarized learning content can improve learning performance significantly [12]. In addition, from the perspective of collaborative learning, delivering properly summarized learning contents in smart learning environment helps learners to share information and enhances the strengths of peer review and critical thinking among learners [6].

Therefore, if the properly summarized learning content combined with a smart device, which is one of the key components in the smart learning environment, is able to support learning significantly, we have enough reason to believe the multi-document summarization is deemed to improve the adaptivity of learning content and benefit the interactivity of teaching and learning in smart learning environment.

V. Conclusion and Future Work

In this paper, our proposed compound model of the top-ical n-grams model and query-likelihood model has shown significant performance gains over the baseline system. Our experimental results have not only demonstrated that the proposed approach presents our summarizer is comparable to the benchmark systems in DUC2006, but also justified that the method for integrating two models is suitable for addressing the problem of multi-document summarization in this study. The evaluation of ROUGE of our approach has shown improvements when the summarization problem is tackled with the contextual information conveyed by word order. Nevertheless, the proposed approach in multi-document summarization is plausible to enhance the adaptivity of learning content and increase the learning experience of students in smart learning environment. In the future, more sophisticated ranking algorithms, such as a hierarchical topic model, could be considered so as to catch the associations between topics. In addition, a statistical analysis of the effectiveness of the contextual query expansion based multi-document summarization in smart learning environment will be evaluated to indicate a proper method to improve the precision of the sentence selection.

ACKNOWLEDGMENT

This research was supported by the NSERC, iCORE, Xerox, and the research related funding by Mr. A. Markin.

REFERENCES


