2012 12th IEEE International Conference on Advanced Learning Technologies

Chunking and Extracting Text Content for Mobile Learning: A Query-focused Summarizer Based on Relevance Language Model

Guangbing Yang[1], Kinshuk[2], Erkki Sutinen[1] , and Dunwei Wen[2],

[1]School of Computing, University of Eastern Finland, P.O. Box 111 80101 Joensuu, Finland, yguang@cs.joensuu.fi or guangbing@athabascau.ca, erkki.sutinen@cs.joensuu.fi

[2]School of Computing and Information Systems, Athabasca University, 1 University Drive, Athabasca, Alberta T9S 3A3, Canada, kinshuk@athabascau.ca, dunweiw@athabascau.ca

***Abstract*—Millions of text contents and multimedia published on the Web have potential to be shared as the learning contents. However, mobile learners often feel it difficult to extract useful contents for learning. Manually creating content not only requires a huge effort on the part of the teachers but also creates barriers towards reuse of the content that has already been created for e-Learning. In this paper, a text-based content summarizer is introduced to address an approach to help mobile learners to retrieve and process information more quickly by aligning text-based content size to various mobile characteristics. In this work, probabilistic language modeling techniques are integrated into an extractive text summarization system to fulfill the automatic summary generation for mobile learning. Experimental results have shown that our solution is a proper and efficient approach to help mobile learners to summarize important content quickly.**

Keywords-component; content processing; text summarization; mobile learning; relevance modelling.

I. INTRODUCTION

Today's fast evolution of mobile technology has provided great potential to improve the performance of mobile learning [1]. This improvement enables mobile learners to more easily access online information using their mobile devices. However, millions of text contents and multimedia published on the Web everyday make it extremely difficult for mobile learners to extract useful contents for learning. This difficulty not only comes from the oft-decried information overloading problem, but is also caused by generic disadvantages of mobile devices, such as small display screen, limited network bandwidth and storage capability. Moreover, mobile learners usually expect to assimilate the information in a very limited time, such as during their commutes. In this case, it would be helpful if they could obtain the condensed content and important points rather than the entire learning content. Since both situations are caused by the big chunk content, if the content size could be reduced somehow, both problems might be alleviated. However, condensing content may have negative impact on the understanding of the meaning conveyed. Thus, research on how to shorten the text-based contents properly and effectively so as to not lose the meaning would have great potential for effective application of education technology for mobile learning.

Many approaches in mobile learning research have been proposed for revising and reinforcing content to provide appropriate delivery to solve the small screen issue. However, few of these solutions consider learner's characteristics, especially the characteristics of "next generation" or "Net generation" learners. One of the most significant characteristics for next generation learners is that they like multi-tasking and have short attention span [6]. They can perform more tasks simultaneously and shift their attentions quickly from one task to another, but would probably become frustrated if they are asked to read a long report for hours. Some automatic summarization approaches are therefore needed to assist learners in getting the important learning points quickly and easily from larger contents.

Recent research in probabilistic language modeling techniques presents some potential that language modeling techniques are prospected to provide a reliable approach to impose a summarization strategy. Experimental results have presented many advanced techniques in language models, like passage retrieval model [9], query-likelihood language model [10], and so on, which have been applied to perform summarization task successfully.

In this paper, a text-based content summarizer is introduced to help mobile learners to retrieve and process information more quickly by aligning text-based content size to various mobile characteristics. Similar to other language model based summarizers [7, 8], our approach adopts relevance model [11] to perform the retrieval task. First, relevant documents are retrieved by this model based on a given query. Then, top-ranked relevant documents are clustered as a group to perform the sentence similarity evaluation. Finally a maximum similar score is used to reform the sentences into the final summary.

The rest of the paper is organized as follows: Section 2 discusses the system architecture and components based on essential processing. This system is validated through an experiment and Section 3 discusses the experiment with its results, followed by conclusions in Section 4.

II. SYSTEM ARCHITECTURE

Text summarization can be simply defined as a process in which a computer creates a condensed version of the text but still preserves most of the information presented in the original text. Normally, a text summarization system consists of four main components: document pre-processing,

978-0-7695-4702-2/12 $26.00 © 2012 IEEE
DOI 10.1109/ICALT.2012.29

IEEE computer society

relevance model, sentence extraction, and summary generation. A high level view of the system architecture is shown in Figure 1.

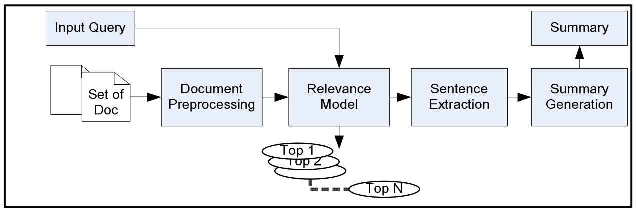

Figure 1. System Architecture

In document pre-processing, all sentences in document(s) need to be pre-processed in order to take out punctuations and stop words, stem words, and index sentences, and mark the sequence number of the sentence that appears in the document.

In relevance model [11], the original query is reformulated by the highest-ranked relevant words, which are initially retrieved by Ponte's heuristic query expansion approach [10]. The following formula is used to rank all the words (denoted as *w*):

$$avg(w,R) = \sum_{D \in R} \log(\frac{p(w \mid D)}{p(w)}) \quad (1)$$

Where *R* is the collection of top-ranked documents (usually the top five documents are chosen) from query likelihood (QL) model [10], $p(w \mid D)$ is the probability of word *w* in relevant document *D*, and $p(w)$ is the prior distribution of word *w*. The outcome of this step is a set of ranked words. *K* (normally 5 based on [11]) highest-ranked words are selected and added to the original query.

After this step, all documents are re-ranked by relevance model over the extended query. At this step, N (an experimental number that indicates the maximum number of the most relevant documents will be selected for the final sentence extraction) highest-ranked documents are selected and their probabilistic distribution is used for processing of the sentence extraction.

Sentence extraction is processed based on two hypotheses. One is that the top-ranked documents estimated by relevance model are the most relevant documents over the given query. Another one is that the expended query words represent the most important keywords for this summarization processing. Based on these two hypotheses, a sentence ranking model has been designed.

First, the sentences in these top-ranked relevant documents are segmented and marked with their order numbers according to their original sequence in each document.

Second, the query likelihood language model approach is adopted here to build sentence models for each sentence. The Jelinek-Mercer [5] smoothing approach is used but is modified for sentence rather than for document. The modified Jelinek-Mercer smoothing approach is given as the following formula:

$$p_{\lambda}(t \mid S) = (1-\lambda) p_{ml}(t \mid S) + \lambda P(t \mid C) \quad (2)$$

where S represents the sentence model built, variable *t* represents the expended query terms, $p_{ml}(t \mid S)$ is the maximum likelihood of the sentence given the query term *t*, $P(t \mid C)$ is the general proportion of the term *t* in the entire collection (C) of the top-ranked documents, and parameter λ is an experimental value, which is set as 0.7 for long queries based on [5].

Given $Q' = \{q_1, q_2, ..., q_m, m_1, ...m_k\}$ as the extended query, the final measure of sentence similarity can be computed using following formula, which is adopted from Zhai's risk minimization language model [4]:

$$P(Q' \mid S) = \prod_{t \in Q'} P(t \mid S) \times \prod_{t \notin Q'} (1.0 - P(t \mid S)) \quad (3)$$

Then, sentences are ranked based on the value of the equation (3).

The process of the summary generation is very simple. The candidate sentences had been ranked previously. Therefore, the process only needs to select the top ranked sentences until reaching the allowed size of the summary. To make the summary more readable, the original order of the sentence in the documents is followed in the summary. If two or more sentences have the same order number (since they may come from different documents but with the same position), these sentences are ordered based on the rank of the documents they belong to.

III. Experiments and Result evaluation

A. Performance measurement

The sentence based precision and recall ratios [2] are used to measure the summarization performance. The precision measures the correctness of the sentences in the summary. The recall measures the effectiveness of the system in the summary.

A traditional e-learning course, 'Environmental Studies' is used in this experiment under a mobile learning environment. This course includes 144 external web pages as external reading materials. All reading materials are text based contents with total 2,461 sentences and 36,054 words. The average number of sentences is around 17 per article. These reading articles were summarized at 4 different summarization levels which are represented by the number of sentences retrieved: 3, 5, 10, and 15 sentences in the article. The human generated summaries were obtained from students who previously studied this course. First, 'Environmental protection' is used as the original query to retrieve the most relevant words from these articles. Table I lists 20 of the highest-ranked words. Top 5 words, namely 'pollution', 'environmentalist', 'emission', 'earth', and 'forest' are then selected to be combined with original query terms for second retrieval using relevance model. Finally, 5 highest ranked articles (listed as A1 to A5 in table II) are selected from the new generated rank list. Based on these five articles, summaries are generated. Average precision and recall are then calculated without or with expended query terms, by comparing with human generated summaries. In table II, the column title 'number of sentences in summary' represents the number sentences retrieved and used in the summary.

TABLE I. HIGHEST RANKED WORDS

Rank #	1,2, 3…20
Words	Pollution, environmentalist, emission, earth, forest, pollute, liable, recycle, conservation, carbon, rain, urban, industrialise, tree, toxic, world, Rio, nafta (NAFTA), dioxide, warm

TABLE II. EVALUATION WITHOUT AND WITH QUERY EXPENSION

Selected Articles		***A1***	***A2***	***A3***	***A4***	***A5***	Total # of relevant sentences by human	
Total # sens		22	44	39	28	41		
# of relevant sents vs. human jud.		5	5	7	5	7	29	
		Number of relevant sentences retrieved by sys.					Avg Precision	Avg Recall
number of sentences in summary	**Without expended query terms**							
	3	0	1	0	1	0	2/(3x5)	2/29
	5	1	1	0	1	0	3/(5x5)	3/29
	10	2	2	2	2	1	9/(10x5)	9/29
	15	4	3	3	2	3	15/(5x15)	15/29
	With expended query terms							
	3	**2**	1	**1**	1	0	5/(3x5)	5/29
	5	**2**	1	1	1	**1**	6/(5x5)	6/29
	10	**3**	**4**	3	**3**	**2**	15/(5x10)	15/29
	15	4	**4**	**4**	**3**	4	19/(5x15)	19/29

B. Discussion

Experimental results show that the average precision improved significantly when expended query terms were applied in the retrieval processing. In particular, when the retrieved sentences were limited to 3 for each article, the precision gained 2.5 times improvement compared with the value without expended query in table II. This is because the expended query terms, like 'pollution', 'environmentalist', 'emission', and so on, are highly relevant to articles retrieved by the relevance model. This result has also verified our hypothesis that the expended query words discovered by relevance model are able to represent the most important information for this summarization processing. In addition, the recall value has increased when more sentences are retrieved. That is because the more retrieved sentences are allowed in the summaries; the more relevant sentences are included. However, the evaluation is not accurate enough since the base sentence relevance comes from humans' judgments, which might include humans' preferences and various comprehensions of the content.

A practical scenario has been created to evaluate the performance of the approached summarization solution in mobile learning environment. Based on the experiments, 3 to 5 sentences (normally around 100 words in total) can be displayed properly in the most of mobile device's screens.

IV. CONCLUSIONS AND FUTURE WORK

This paper has presented a statistical language modeling based summarization system for mobile learning. The experimental results have demonstrated that the system is able to extract important information effectively from a practical document collection in education. The good performance is due to the higher relevancy topics and keywords explored by the relevance model. Although many relevant topics and keywords are retrieved, there are still a few irrelevant terms picked by this model. This irrelevance brings 'noise' to the retrieval processing and eventually affects the effectiveness of the summarization. This would be one of main limitations of the system, providing a direction for the future improvements. For future work, the system can be improved by integrating it with statistical topic modeling approaches in machine learning that provide certain learning capability to the system for a better summarization where the summarizing patterns are expected to match the ones in human generated summaries. In addition to imposing more advanced statistic modeling approaches in the system, an automatic summary evaluation, such as ROUGE [3], and a formal methodology in education will be applied to further evaluate the effectiveness of the summarization and learning performance.

ACKNOWLEDGMENT

This study is supported in part by the NSERC, iCORE, Xerox, and the research related funding by Mr. A. Markin.

REFERENCES

[1] A. DeGani, G. Martin, G. Stead, and F. Wade (2010). "Mobile Learning Shareable Content Object Reference Model (m-SCORM) Limitations and Challenges [N09-35]", Retrieved Feb 12, 2012 from http://www.m-learning.org/knowledge-centre/research.

[2] A. Nenkova, (2006). "Summarization evaluation for text and speech: Issues and approaches". In Processdings of Interspeech'06, Pittsburgh, PA, 2006.

[3] C. Y. Lin, (2004). "Rouge: A package for automatic evaluationof summaries". In Processdings of ACL workshop on Text Summarization Branches Out.

[4] C. Zhai, (2002). "Risk Minimization and Language Modeling in Text Retrieval". PhD Thesis, Carnegie Mellon University, 2002.

[5] C. Zhai and J. Lafferty, (2001). "A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval". In Proceedings of the 24th ACM SIGIR, 334-342

[6] D. G. Oblinger, and J. L. Oblinger, (2005). "Educating the net generation". Retrieved November, 24, 2011 from http://www.educause.edu/educatingthenetgen

[7] H. Daumé III and D. Marcu, (2006). "Bayesian query-focused summarization". In Proceedings of the Conference of the Association for Computational Linguistics (ACL).

[8] J.-C. Ying, S.-J. Yen, Y.-S. Lee, Y.-C. Wu, J.-C. Yang, (2007). "Language Model Passage Retrieval for Question-Oriented Multi Document Summarization". In Proceedings of the Document Understanding Conference 2007. http://duc.nist.gov/pubs.html#2007.

[9] J. P. Callen, (1994). "Passage-level evidence in document retrieval". In Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 302-310.

[10] J. Ponte, and W. B. Croft, (1998). "A Language Modeling Approach to Information Retrieval". In Proceedings of the 21st ACM SIGIR Conference on Research & Development on Information Retrieval, 275-281.

[11] V. Lavrenko and W. B. Croft, (2001). "Relevance-Based Language Models," In the 24th ACM SIGIR Conference on Research & Development on Information Retrieval, 120-127.