EUSUM: Extracting Easy-to-Understand English Summaries for Non-Native Readers

Xiaojun Wan, Huiying Li and Jianguo Xiao
Institute of Computer Science and Technology, Peking University, Beijing 100871, China
Key Laboratory of Computational Linguistics (Peking University), MOE, China
{wanxiaojun, lihuiying, xiaojianguo}@icst.pku.edu.cn

ABSTRACT

In this paper we investigate a novel and important problem in multi-document summarization, i.e., how to extract an easy-to-understand English summary for non-native readers. Existing summarization systems extract the same kind of English summaries from English news documents for both native and non-native readers. However, the non-native readers have different English reading skills because they have different English education and learning backgrounds. An English summary which can be easily understood by native readers may be hardly understood by non-native readers. We propose to add the dimension of reading easiness or difficulty to multi-document summarization, and the proposed EUSUM system can produce easy-to-understand summaries according to the English reading skills of the readers. The sentence-level reading easiness (or difficulty) is predicted by using the SVM regression method. And the reading easiness score of each sentence is then incorporated into the summarization process. Empirical evaluation and user study have been performed and the results demonstrate that the EUSUM system can produce more easy-to-understand summaries for non-native readers than existing summarization systems, with very little sacrifice of the summary's informativeness.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *abstracting methods*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *text analysis*

General Terms

Algorithms, Experimentation, Design, Human Factors.

Keywords

EUSUM, multi-document summarization, reading easiness

1. INTRODUCTION

Document summarization is a task of producing a condensed version of a document or document set. A summary is usually required to be informative and fluent. Users can easily understand the main content of the document or document set by reading the summary.

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.

To date, various summarization methods and a number of summarization systems have been developed, such as MEAD, NewsInEssence and NewsBlaster. These methods and systems focus on how to improve the informativeness, diversity or fluency of the English summary, and they usually produce the same English summaries for all users, including native readers and non-native readers. However, different users usually have different English reading levels because they have different English education backgrounds and learning environments. And native readers usually have higher English reading levels than non-native readers. In particular, Chinese readers usually have less ability to read English summaries than native English readers. For example, Chinese college students usually have passed the National English Test Band 4 (CET-4), and they have learned English for several years, but they still have more or less difficulty to read original English news and summaries. The difficulty lies in unknown or difficult English words (e.g. "seismographs", "woodbine"), or the complex sentence structure (e.g. "The chairman of the House Agriculture Committee says hearings are planned next year into how the U.S. Forest Service handled last summer's stubborn wildfires that scorched the West, including one-third of Yellowstone National Park."). Therefore, they have to slow down the reading speed in order to understand the news text or summary, or give up the reading process.

In this study, we argue that the English summaries produced by existing methods and systems are not fit for non-native readers (i.e. Chinese readers). We examine a new factor - reading easiness (or difficulty) [1] for document summarization, and the factor can indicate whether the summary is easy to understand by non-native readers or not. The reading easiness of a summary is dependent on the reading easiness of each sentence in the summary. And we propose a novel summarization system – EUSUM (Easy-to-Understand Summarization) for incorporating the reading easiness factor into the final summary. The proposed system first predicts the reading easiness score for each sentence, and then incorporates the reading easiness score into the final sentence ranking process. Both informative and easy-to-understand sentences are selected into the summary. Both automatic evaluation and user study have been performed and the evaluation results verify the effectiveness of the proposed EUSUM system.

The contribution of this paper is summarized as follows: 1) We examine a new factor of reading easiness for document summarization. 2) We propose a novel summarization system – EUSUM for incorporating the new factor and producing easy-to-

[1] In this paper, "reading easiness" and "reading difficulty" refer to the same factor, and we use them interchangeably.

understand summaries for non-native readers. 3) We conduct both automatic evaluation and user study to verify the effectiveness of the proposed system.

The paper is organized as follows: Section 2 introduces related work. Section 3 describes the details of the EUSUM system. Sections 4 and 5 present experimental results and discussions. Lastly we conclude our paper in Section 6.

2. RELATED WORK

2.1 Document Summarization

Document summarization methods can be generally categorized into extraction-based methods and abstraction-based methods. In this paper, we focus on extraction-based methods. Extraction-based summarization methods usually assign each sentence a saliency score and then rank the sentences in a document or document set.

For single document summarization, the sentence score is usually computed by empirical combination of a number of statistical and linguistic feature values, such as term frequency, sentence position, topic signature [19, 22]. The summary sentences can also be selected by using machine learning methods [1, 17] or graph-based methods [8, 23]. Other methods include mutual reinforcement principle [33].

For multi-document summarization, the centroid-based method [27] is a typical method, and it scores sentences based on cluster centroids, position and TFIDF features. NeATS [20] makes use of new features such as topic signature to select important sentences. Machine Learning-based approaches have also been proposed for combining various sentence features [34]. Themes (or topics, clusters) discovery in documents has been used for sentence selection [10]. The influences of input difficulty on summarization performance have been investigated in [25]. Graph-based methods have also been used to rank sentences in a document set. For example, Mihalcea and Tarau [24] extend the TextRank algorithm to compute sentence importance in a document set. Cluster-level information has been incorporated in the graph model to better evaluate sentences [32]. Sentence ordering in summaries has been investigated in [2, 3] to improve the summary fluency.

Other summarization tasks include topic-focused (query-biased) document summarization [31], update summarization [18]. All these summarization tasks do not consider the reading easiness factor of the summary for non-native readers.

2.2 Reading Difficulty Prediction

A reading difficulty measure can be originally described as a function or model that maps a text to a numerical value corresponding to a difficulty or grade level [12]. And reading difficulty prediction can be viewed as a regression of difficulty grade level based on a set of features derived from the text. Earlier work on reading difficulty prediction is conducted for the purpose of education or language learning. For example, one purpose is to find appropriate reading materials of the appropriate difficulty level, in terms of both vocabulary and grammar, for English as a First or Second Language students. And almost all earlier work focuses on document-level reading difficulty prediction.

A variety of features have been investigated in reading difficulty measures. Average sentence length and word length are simple proxies for grammatical and lexical complexity of a text, as in the Dale-Chall model [5]. The Flesch-Kincaid meaure [15] is probably the most common reading difficulty in use in earlier days. The Lexile Framework [29] uses individual word frequency estimates as a measure of lexical difficulty, and it uses a Rasch model based on the features of word frequency and sentence length. In recent years, more sophisticated features and models are used. Smoothed unigram language modeling has been developed to capture the predictive ability of individual words based on their frequency at each reading difficulty level [7]. A statistical approach is proposed to infer the distribution of a word's likely acquisition age automatically from authentic texts collected from the Web, and then an effective semantic component for predicting reading difficulty of news texts is provided by combining the acquisition age distributions for all words in a document [14]. Schwarm and Ostendorf [28] incorporate syntactic features derived from syntactic parses of text, and their system performs better than the Flesch-Kincaid and Lexile measures. The frequency of grammatical constructions has been used as a measure of grammatical difficulty, and the final prediction function is a linear function of the lexical and grammatical components [11, 12]. Pitler and Nenkova [26] combine lexical, syntactic and discourse features to produce a highly predictive model of text readability. In addition to English language, François [9] presents an approach to assessing the readability of French texts. More recently, a machine learning approach is used for predicting the readability of web search summaries or snippets [13].

In this study, we investigate the reading difficulty (or easiness) prediction of English sentences for Chinese readers, i.e. whether an English sentence is easy to understand by Chinese readers or not.

Note that sentence ordering in a long summary also has influences on the reading difficulty or readability of the summary, and proper order of extracted sentences can improve their readability [2]. However, sentence ordering is another research problem and we do not take into account this factor in this study.

3. THE EUSUM SYSTEM

3.1 System Overview

The main idea of the proposed EUSUM system is to incorporate the sentence-level reading easiness factor into the summary extraction process. Each sentence is associated with two factors: informativeness and reading easiness. The informativeness of a sentence is computed by using previous summarization methods. The reading easiness of a sentence is measured by an EU (easy-to-understand) score, which is predicted by using statistical regression methods. The two scores are then combined and both informative and easy-to-understand sentences are chosen into the summary. The three steps of the EUSUM system will be described in details in next two sections.

As mentioned in Section 2.2, we do not consider the fluency factor of the whole summary in this study, which has been investigated in related research areas (e.g. sentence ordering [2, 3]).

3.2 Sentence-Level Reading Easiness Prediction

In this study, reading easiness refers to how easily a text can be understood by non-native readers. Reading easiness prediction is a task of mapping a text to a numerical value corresponding to a reading easiness. The larger the value is, the more easily the text can be understood. We focus on predicting the reading easiness score of an English sentence for Chinese college students.

As mentioned earlier, Chinese college students usually have studied English for several years and they usually have passed the CET-4 test[2] or above, which means that they have some ability to read ordinary English articles. However, because of different English learning environments and different learning abilities, these students may have different English reading levels. Many students have some difficulty to read original English news or summaries. The two factors most influencing the reading process are as follows:

1) Unknown or difficult English words: for example, most Chinese college students do not know the words such as "seismographs", "woodbine".

2) Complex sentence structure: for example, a sentence with two or more clauses introduced by a subordinating conjunction is usually difficult to read.

As introduced in Section 2.2, various regression methods have been used for reading difficulty prediction. In this study, we adopt the ε-support vector regression (ε-SVR) method [30] for the reading easiness prediction task. The SVR algorithm is firmly grounded in the framework of statistical learning theory (VC theory). The goal of a regression algorithm is to fit a flat function to the given training data points.

In the experiments, we use the LIBSVM tool [6] with the RBF kernel for the regression task, and we use the parameter selection tool of 10-fold cross validation via grid search to find the best parameters with respect to mean square error (MSE), and then use the best parameters to train the whole training set.

We use the following two groups of features for each sentence: the first group includes surface features, and the second group includes parse based features.

The four surface features are as follows:

1) **Sentence length**: It refers to the number of words in the sentence. A long sentence may be more difficult to understand than a short sentence.
2) **Average word length**: It refers to the average length of words in the sentence. Usually, an English word with few characters is more easily recognized and remembered than that with many characters.
3) **CET-4 word percentage**: It refers to the percentage of how many words in the sentence appear in the CET-4 word list (690 words). As mentioned earlier, most Chinese college students have passed CET-4, and the words appearing in the CET-4 word list are likely to be recognized by the students.
4) **Number of peculiar words**: It refers to the number of infrequently occurring words in the sentence. We collect all words in the experimental corpus, and choose the top 2000 words with low frequency as the peculiar words. The frequency of each word is extracted from the Google Web 1T 1-gram database [4].

We use the Stanford Lexicalized Parser [16] with the provided English PCFG model to parse a sentence into a parse tree. The output tree is a context-free phrase structure grammar representation of the sentence. The four parse features are as follows:

1) **Depth of the parse tree**: It refers to the depth of the generated parse tree. Usually the higher the parse tree is, the more complex the sentence is.
2) **Number of SBARs in the parse tree**: SBAR is defined as a clause introduced by a (possibly empty) subordinating conjunction. It is an indictor of sentence complexity, especially for Chinese readers.
3) **Number of NPs in the parse tree**: It refers to the number of noun phrases in the parse tree.
4) **Number of VPs in the parse tree**: It refers to the number of verb phrases in the parse tree.

All the above feature values are scaled by using the provided svm-scale program.

At this step, each sentence s_i can be associated with a reading easiness score $EaseScore(s_i)$ predicted by the ε-SVR method. The larger the score is, the more easily the sentence is understood. The score is finally normalized by dividing by the maximum score.

3.3 Sentence-Level Informativeness Evaluation

In this study, we adopt two typical methods for evaluating the informativeness of each sentence in a document set. The two methods are described briefly in the following sections.

3.3.1 Centroid-Based Method

The centroid-based method is the algorithm used in the MEAD system. The method uses a heuristic and simple way to sum the sentence scores computed based on different features. In our implementation, the score for each sentence is a linear combination of the weights computed based on the following three features: 1) **Centroid-based Weight.** The weight $C(s_i)$ of sentence s_i is calculated as the cosine similarity between the sentence text and the concatenated text for the whole document set D. The weight is then normalized by dividing by the maximal weight. 2) **Sentence Position.** The weight $P(s_i)$ is calculated for sentence s_i to reflect its position priority as $P(s_i)=1-(pos_i-1)/n_i$, where pos_i is the position number of sentence s_i in a particular document and n_i is the total number of sentences in the document. Obviously, pos_i ranges from 1 to n_i. 3) **First Sentence Similarity.** The weight $F(s_i)$ is computed as the cosine similarity value between sentence s_i and the corresponding first sentence in the same document.

After all the above weights are calculated for each sentence, we sum the three weights and get the overall score $InfoScore(s_i)$ for sentence s_i. After the scores for all sentences are computed, the

[2] CET-4 is College English Test Band 4, which is a national English level test in China, and all college students are required to pass this test before graduation.

score of each sentence is normalized by dividing by the maximum score.

3.3.2 Graph-Based Method

The basic idea of the graph-based method is that of "voting" or "recommendation" between sentences. Formally, given a document set D, let $G=(V, E)$ be an undirected graph to reflect the relationships between sentences in the document set. V is the set of vertices and each vertex s_i in V is a sentence in the document set. E is the set of edges. Each edge e_{ij} in E is associated with an affinity weight $f(s_i, s_j)$ between sentences s_i and s_j ($i \neq j$). The weight is computed using the standard cosine measure between the two sentences. Here, we have $f(s_i, s_j)=f(s_j, s_i)$ and let $f(s_i, s_i)=0$ to avoid self transition.

We use an affinity matrix M to describe G with each entry corresponding to the weight of an edge in the graph. $M = (M_{i,j})_{|V|\times|V|}$ is defined as $M_{i,j}=f(s_i,s_j)$. Then M is normalized to $\tilde{M}$ to make the sum of each row equal to 1.

Based on matrix $\tilde{M}$, the saliency score $InfoScore(s_i)$ for sentence s_i can be deduced from those of all other sentences linked with it and it can be formulated in a recursive form as in the PageRank algorithm:

$$InfoScore(s_i) = \mu \cdot \sum_{all\ j \neq i} InfoScore(s_j) \cdot \tilde{M}_{j,i} + \frac{(1-\mu)}{|V|}$$

where μ is the damping factor usually set to 0.85, as in the PageRank algorithm.

After the scores for all sentences are computed, the score of each sentence is normalized by dividing by the maximum score.

3.4 Summary Extraction

After we obtain the reading easiness score and the informativeness score of each sentence in the document set, we linearly combine the two scores to get the combined score of each sentence.

Formally, let $EaseScore(s_i) \in [0,1]$ and $InfoScore(s_i) \in [0,1]$ denote the reading easiness score and the informativeness score of sentence s_i, the combined score of the sentence is:

$$CombinedScore(s_i) = InfoScore(s_i) + \lambda \times EaseScore(s_i)$$

where $\lambda \geq 0$ is a parameter controlling the influences of the reading easiness factor. If λ is set to 0, the summary is extracted without considering the reading easiness factor. Usually, λ is not set to a large value because we must maintain the content informativeness in the extracted summary. Therefore, we choose the parameter value empirically in order to balance the two factors of content informativeness and reading easiness.

For multi-document summarization, some sentences are highly overlapping with each other, and thus we apply the same greedy algorithm in [31] to penalize the sentences highly overlapping with other highly scored sentences, and finally the informative, novel, and easy-to-understand sentences are chosen into the summary.

In the algorithm, the final rank score $RankScore(s_i)$ of each sentence s_i is initialized to its combined score $CombinedScore(s_i)$. And at each iteration, the highly ranked sentence (e.g. s_i) is selected into the summary, and the rank score of each remaining sentence s_j is penalized by using the following formula:

$$RankScore(s_j) = RankScore(s_j) - \omega \cdot \tilde{M}_{j,i} \cdot CombinedScore(s_i)$$

where $\omega > 0$ is the penalty degree factor. The larger ω is, the greater penalty is imposed to the rank score. If $\omega=0$, no diversity penalty is imposed at all. The iteration is stopped after the summary length limit is reached.

4. EXPERIMENTS

4.1 Reading Easiness Prediction

4.1.1 Experimental Setup

In the experiments, we first constructed the gold-standard dataset in the following way.

DUC2001 provided 309 news articles for document summarization tasks, and the articles were grouped into 30 document sets. The news articles were selected from TREC-9. We chose five document sets (d04, d05, d06, d08, d11) with 54 news articles out of the DUC2001 test set. The documents were then split into sentences and there were totally 1736 sentences.

Two college students (one undergraduate student and one graduate student) manually labeled the reading easiness score for each sentence separately. The score ranges between 1 and 5, and 1 means "very hard to understand", and 5 means "very easy to understand", and 3 means "mostly understandable". The final reading easiness score was the average of the scores provided by the two annotators.

After annotation, we randomly separated the labeled sentence set into a training set of 1482 sentences and a test set of 254 sentences. We then used the LIBSVM tool for training and testing.

Two standard metrics were used for evaluating the prediction results. The two metrics are as follows:

Mean Square Error (MSE): This metric is a measure of how correct each of the prediction values is on average, penalizing more severe errors more heavily.

Pearson's Correlation Coefficient (ρ): This metric is a measure of whether the trends of prediction values matched the trends for human-labeled data.

4.1.2 Experimental Results

Table 1 shows the prediction results. For comparison, the result for the Flesch-Kincaid measure[3] is also reported in the table. We can see that the overall results of our method are very promising. And the correlation is high. The results guarantee that the use of reading easiness scores in the summarization process is feasible.

Table 1. Reading easiness prediction results

Method	MSE	ρ
Flesch-Kincaid	0.704	0.377
SVR (Surface features)	0.121	0.929
SVR (Parse features)	0.227	0.853
SVR (Surface features + Parse features)	**0.112**	**0.931**

[3] The reading easiness scores based on the FK measure are directly obtained by accessing the following web service: http://www.standards-schmandards.com/exhibits/rix/index.php

We can also see that either the surface feature set or the parse feature set can achieve good prediction result, and the two feature sets can contribute to the overall prediction results. However, the Flesch-Kincaid measure does not perform well.

4.2 Document Summarization

4.2.1 Experimental Setup

In this study, we used the multi-document summarization task (task 2) in DUC2001 for evaluation. As mentioned in Section 4.1.1, DUC2001 provided 30 document sets. Because we have used five document sets (d04-d11) for training and testing in the task of reading easiness prediction, we used the remaining 25 document sets for summarization evaluation, and the average document number per document set is 10. The sentences in each article have been separated and the sentence information has been stored into files. A summary was required to be created for each document set and the summary length was 100 words. Generic reference summaries were provided by NIST annotators for evaluation.

We used the LIBSVM tool with the learned model to predict the reading easiness score for each sentence in the documents, and then used the scores for summary extraction.

Different from traditional summarization tasks, our task is to incorporate the reading easiness factor into multi-document summary, and the easy-to-understand summarization can be considered as a novel summarization task. Therefore, we evaluate a summary from the following two aspects:

Content Informativeness: This aspect is widely evaluated in almost all traditional summarization tasks. It refers to how much a summary reflects the major content of the document set. Usually, it can be measured by comparing the system summary with the reference summary.

We used the ROUGE-1.5.5 toolkit for automatic evaluation of the content informativeness, and the toolkit was officially adopted by DUC for automatic summarization evaluation. The toolkit measures summary quality by counting overlapping units such as the n-gram, word sequences and word pairs between the candidate summary and the reference summary [21]. The ROUGE-1.5.5 toolkit reports separate F-measure scores for 1, 2, 3 and 4-gram, and also for longest common subsequence co-occurrences. In this study, we show four ROUGE F-measure scores in the experimental results: ROUGE-1 (unigram-based), ROUGE-2 (bigram-based), ROUGE-W (based on weighted longest common subsequence, weight=1.2), and ROUGE-SU* (based on skip bigram with unigram)[4].

Reading Easiness: This aspect is not evaluated by previous summarization tasks. We aim to evaluate the reading easiness level of the whole summary. Because the reading easiness level of a summary is dependent on the reading easiness scores of the sentences in the summary, we use the average reading easiness score of the sentences in a summary as the summary's reading easiness level. The overall reading easiness score is the average across all 25 document sets.

[4] We also used the option "-l 100" for truncating the summary and used the option "-m" for word stemming when using the ROUGE-1.5.5 toolkit.

In addition to the above automatic evaluation procedures, we also performed pilot user studies for evaluation. Four Chinese college students participated in the user studies. We have developed a user study tool for facilitating the subjects to evaluate each summary from the two aspects of content informativeness and reading easiness. Each subject can assign a score from 1 to 5 on each aspect for each summary. For reading easiness, 1 means "very hard to understand", and 5 means "very easy to understand". For content informativeness, 1 means "least informative", and 5 means "very informative". During each user study procedure, we compared two summarization systems' results. And the two summaries produced by the two systems for the same document set were presented in the same interface, and then the four subjects assigned scores to each summary after they read and compared the two summaries. The final score of a summary on one aspect was the average of the scores assigned by the four subjects. And the overall scores were averaged across all subjects and all 25 document sets.

4.2.2 Experimental Results

4.2.2.1 Automatic Evaluation Results

In this section, we report the automatic evaluation results of EUSUM from both two aspects. Though the combination weight λ in EUSUM can be set to any non-negative value, it ranges from 0 to 1 in our experiments, because a much larger λ will lead to a big sacrifice of the content informativeness for the summary. The penalty degree factor ω for EUSUM is set to 10, as in [31]. Table 2 shows the ROUGE scores and the reading easiness score of the EUSUM system with the centroid-based method, which is denoted as EUSUM(Centroid). Table 3 shows the ROUGE scores and the reading easiness score of the EUSUM system with the graph-based method, which is denoted as EUSUM(Graph).

Seen from the tables, the ROUGE scores of EUSUM(Centroid) and EUSUM(Graph) are decreased with the increase of the combination weight λ, and the reading easiness scores of them are increased with the increase of λ. And we can see that with the increase of λ, the summary's reading easiness can be more quickly becoming significantly different from that of the summary with λ=0, while the summary's content informativeness is not significantly affected when λ is set to a small value. Moreover, For EUSUM(Graph), even the ROUGE scores with λ=0.1 are better than that with λ=0. The results demonstrate that when λ is set to a small value, the content informativeness aspect of the extracted summary are almost not affected, but the reading easiness aspect of the extracted summary can be significantly improved.

By comparing the performance values in the two tables, we can see that when λ is fixed, the ROUGE-1, ROUGE-W and ROUGE-SU* scores of EUSUM(Graph) are higher than the corresponding scores of EUSUM(Centroid), which verifies the effectiveness of the graph-based summarization method. We can also see that when λ is fixed, the reading easiness scores of EUSUM(Graph) are always higher than the corresponding scores of EUSUM(Centroid), which demonstrates that EUSUM(Graph) can extract more easy-to-understand summaries than EUSUM(Centroid). We explain the results by that the graph-based sentence extraction method tends to extract sentences with good feature values for indicating reading easiness. For example, sentence length is one of the important features for reading

easiness prediction, and a shorter sentence is more likely to be easy to understand. We compare the average sentence length (average word number per sentence) in the summaries extracted by EUSUM(Graph) and EUSUM(Centroid) in Figure 1. We can see that EUSUM(Graph) usually extracts shorter sentences than EUSUM(Centroid), which verifies the results from one perspective. Overall, the results show that EUSUM(Graph) is more suitable than EUSUM(Centroid) for extracting easy-to-understand summaries[5].

Table 2. EUSUM (Centroid) results vs. λ

λ	ROUGE-1 Average_F	ROUGE-2 Average_F	ROUGE-W Average_F	ROUGE-SU* Average_F	Reading Easiness score
0	0.31682	0.05896	0.13532	0.09405	3.45681
0.1	0.31468	0.05529	0.13346	0.09335	3.53835
0.2	0.31419	0.05488	0.13367	0.09254	**3.58633**
0.3	0.31153	0.05379	0.13311	0.09085	**3.65514**
0.4	0.30938	0.05255	0.13203	0.08878	**3.74286**
0.5	**0.30754**	0.05029	0.13113	0.08802	**3.82912**
0.6	**0.30402**	**0.04920**	**0.13011**	**0.08648**	**3.85643**
0.7	**0.30228**	**0.04792**	**0.12955**	**0.08506**	**3.96930**
0.8	**0.30227**	**0.04700**	**0.12959**	**0.08353**	**4.09463**
0.9	**0.29344**	**0.04233**	**0.12579**	**0.07808**	**4.22541**
1	**0.29242**	**0.04033**	**0.12514**	**0.07762**	**4.33986**

Table 3. EUSUM (Graph) results vs. λ

λ	ROUGE-1 Average_F	ROUGE-2 Average_F	ROUGE-W Average_F	ROUGE-SU* Average_F	Reading Easiness score
0	0.32010	0.05205	0.13701	0.09488	3.91532
0.1	0.32286	0.05364	0.13928	0.09612	3.98131
0.2	0.31928	0.05155	0.13693	0.09414	**4.07323**
0.3	0.31751	0.04920	0.13587	0.09235	**4.16455**
0.4	0.31598	0.04677	0.13518	0.09100	**4.31125**
0.5	**0.30828**	**0.04544**	**0.13231**	**0.08773**	**4.40275**
0.6	**0.30723**	**0.04516**	**0.13308**	**0.08753**	**4.47232**
0.7	**0.30608**	**0.04525**	**0.13306**	**0.08656**	**4.56735**
0.8	**0.30197**	**0.04396**	**0.13165**	**0.08430**	**4.63991**
0.9	**0.29936**	**0.04343**	**0.13131**	**0.08324**	**4.67847**
1	**0.29498**	**0.04166**	**0.12993**	**0.07964**	**4.72482**

(The bolded scores indicate that the difference between the scores and the corresponding scores when λ=0 is statistically significant by using t-test.)

In the above experiments, the penalty weight ω is fixed to 10. We now take EUSUM(Graph) as an example to show how the penalty weight ω influences the two aspects of the proposed summarization system. Figures 2 and 3 show the reading easiness score curves and the ROUGE-SU* F-score curves of EUSUM(Graph) with different settings, respectively. We can see that the reading easiness scores of EUSUM(Graph) with different settings have a tendency to increase with the increase of ω. And after ω is larger than 10, the reading easiness scores for most settings do not change any more, which shows that the penalty weight has no significant influences on the reading easiness of the summaries when the weight is set to a moderately large value. The ROUGE-SU* scores are firstly increasing with the increase of ω and then decreased with the increase of ω, which demonstrates that less or much penalty will lower the performance instead of content informativeness.

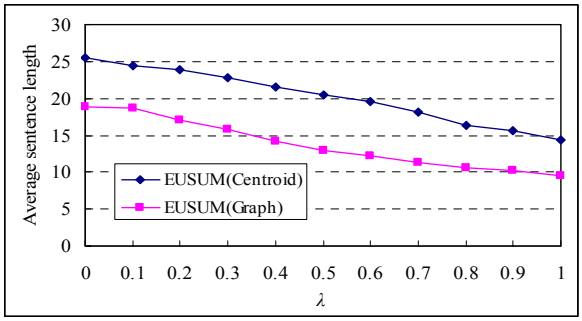

Figure 1. EUSUM average sentence length (sentence word number) comparison.

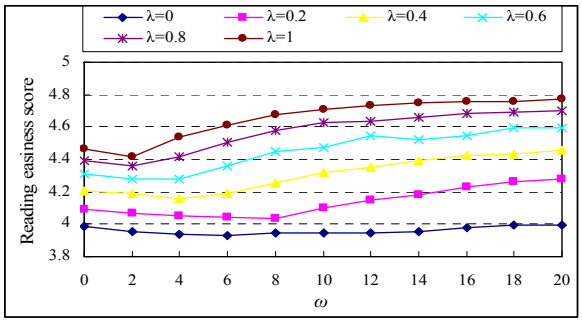

Figure 2. EUSUM(Graph) reading easiness vs. penalty weight.

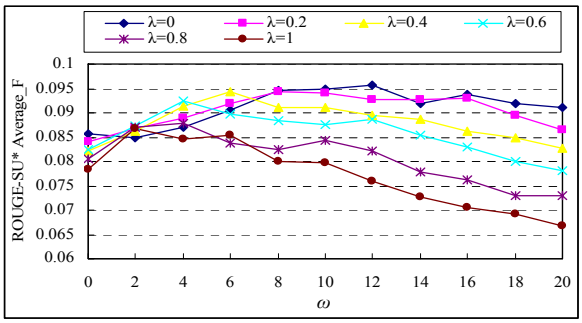

Figure 3. EUSUM(Graph) content informativeness vs. penalty weight.

4.2.2.2 User Study Results

In order to validate the effectiveness of the system by real non-native readers, two user study procedures were performed:

User study 1: The summaries extracted by EUSUM(Centroid) (λ=0) and EUSUM(Graph) (λ=0.2) are compared and scored by subjects. Seen from Tables 2 and 3, most ROUGE scores of the two systems are very similar, and the reading easiness score of EUSUM(Graph) (λ=0.2) is higher than that of EUSUM(Centroid) (λ=0). Table 4 gives the averaged subjective scores of the two systems. The user study results verify that the summaries by EUSUM(Graph) (λ=0.2) are indeed significantly easy to understand by non-native readers, while the content informativeness of the two systems are not significantly different.

User study 2: The summaries extracted by EUSUM(Graph) (λ=0) and EUSUM(Graph) (λ=0.3) are compared and scored by subjects.

[5] Actually, we can improve the centroid-based method by incorporating some useful features such as sentence length, but it is not the focus of this paper.

Seen from Tables 2 and 3, most ROUGE scores of the two systems are not significantly different, and the reading easiness score of EUSUM(Graph) (λ=0.3) is higher than that of EUSUM(Graph) (λ=0). Table 5 gives the averaged subjective scores of the two systems. The user study results verify that the summaries by EUSUM(Graph) (λ=0.3) is indeed significantly easy to understand by non-native readers, while the content informativeness of the two systems are not significantly different.

Table 4. Results for user study 1

	Content Informativeness	Reading Easiness
EUSUM(Centroid) (λ=0)	3.47	3.33
EUSUM(Graph) (λ=0.2)	3.63	4.01*

Table 5. Results for user study 2

	Content Informativeness	Reading Easiness
EUSUM(Graph) (λ=0)	3.71	3.73
EUSUM(Graph) (λ=0.3)	3.52	4.02*

(* indicates that the performance difference is statistically significant by using t-test.)

4.2.2.3 Running Examples

In order to better compare the results, we give several typically extracted summaries for two document sets D14 and D59. The predicted reading easiness score of each sentence is also given in brackets.

EUSUM(Centroid)(λ=0) for D14:

A U.S. Air Force F-111 fighter-bomber crashed today in Saudi Arabia, killing both crew members, U.S. military officials reported. (3.97397)
A jet trainer crashed Sunday on the flight deck of the aircraft carrier Lexington in the Gulf of Mexico, killing five people, injuring at least two and damaging several aircraft (3.182)
U.S. Air Force war planes participating in Operation Desert Shield are flying again after they were ordered grounded for 24 hours following a rash of crashes. (3.41654)
A U.S. military jet crashed today in a remote, forested area in northern Japan, but the pilot bailed out safely and was taken by helicopter to an American military base, officials said. (3.42433)

EUSUM(Graph)(λ=0) for D14:

The U.S. military aircraft crashed about 800 meters northeast of a Kadena Air Base runway and the crash site is within the air base's facilities. (3.84771)
Two U.S. Air Force F-16 fighter jets crashed in the air today and exploded, an air force spokeswoman said. (4.35604)
West German police spokesman Hugo Lenxweiler told the AP in a telephone interview that one of the pilots was killed in the accident. (3.79754)
Even before Thursday's fatal crash, 12 major accidents of military aircraft had killed 95 people this year alone. (3.92878)
Air Force Spokesman 1st Lt. Al Sattler said the pilot in the Black Forest crash ejected safely before the crash and was taken to Ramstein Air Base to be examined. (3.70656)

EUSUM(Graph)(λ=0.3) for D14:

Two U.S. Air Force F-16 fighter jets crashed in the air today and exploded, an air force spokeswoman said. (4.35604)
The U.S. military aircraft crashed about 800 meters northeast of a Kadena Air Base runway and the crash site is within the air base's facilities. (3.84771)
West German police spokesman Hugo Lenxweiler told the AP in a telephone interview that one of the pilots was killed in the accident. (3.79754)
Even before Thursday's fatal crash, 12 major accidents of military aircraft had killed 95 people this year alone. (3.92878)
However, suspension of training flights indicated otherwise. (4.97415)
Listed as dead from the 433rd were Maj. (4.99479)

EUSUM(Centroid)(λ=0) for D59:

The Northwest Airlines jet that crashed Sunday in Detroit, killing at least 154 people, was involved in two incidents of engine failure in the past two years. (3.54385)
A French DC-10 jetliner with 171 people aboard experienced a powerful high-altitude explosion, possibly from a terrorist bomb, before crashing in a remote desert region of Niger in northern Africa, officials in France said Wednesday. (2.60945)
Freshman congressman Larkin Smith (R-Miss). died in a light plane crash in Mississippi, authorities said Monday, making him the second member of the House killed in an aviation accident in a week. (3.18245)
Local news reporters quoted witnesses as saying that the plane appeared to nosedive into the earth. (3.9432)

EUSUM(Graph)(λ=0) for D59:

Sunday's crash was the first time in 24 years that passengers were killed in an accident involving a Northwest plane. (3.98315)
FAA officials at the crash scene wouldn't speculate on the reasons for the crash or comment on the plane's engines. (4.16897)
There were reports from passengers and observers that the plane's right-wing engine also failed before the crash. (3.94093)
In July 1988, a United DC-10 crashed in Sioux City, Iowa, after an engine broke apart in flight, killing 112 people. (4.2899)
The worst airline accident ever in the U.S. was the 1979 crash of an American Airlines jet in Chicago. (4.38547)
FAA records show that besides those incidents that involved the plane that crashed, problems with the turbine sections of JT8D-200 series engines occurred on three Republic flights in the past four years. (2.97711)

EUSUM(Graph)(λ=0.3) for D59:

Sunday's crash was the first time in 24 years that passengers were killed in an accident involving a Northwest plane. (3.98315)
FAA officials at the crash scene wouldn't speculate on the reasons for the crash or comment on the plane's engines. (4.16897)
There were reports from passengers and observers that the plane's right-wing engine also failed before the crash. (3.94093)
A team of National Transportation Safety Board investigators left Washington Wednesday night for Sioux City. (4.06095)
The DC-10 operated by the French airline UTA crashed Tuesday after taking off from N'Djamena, Chad, on a flight that originated in Brazzaville, Congo. (3.63393)
It can smash an airplane into the ground. (4.86026)

5. DISCUSSION

In this study, the experiments were performed by Chinese college students. Because college students in different countries may have different English reading levels, the experimental results may be slightly changed if we use non-native students in other countries for evaluation. Even for Chinese readers, college students and high school students may have different English reading levels, and thus the experimental results may be slightly changed if we use high school students for evaluation. That's to say, the reading easiness level of a summary should be adjusted with the particular reader. In practice, we can let readers to tune the combination weight λ in the proposed EUSUM system, and they can select the best weight for extracting summaries best suitable for reading.

Similarly, for native English readers, different persons may have different English reading levels, and thus the framework proposed in this paper is also applicable. However, the reading easiness score of each sentence may be different because of the differences between the English reading abilities and behaviors of Chinese readers and native English readers.

6. CONCLUSION AND FUTURE WORK

In this study, we investigate the new factor of reading easiness for document summarization, and we propose a novel summarization system - EUSUM for producing easy-to-understand summaries for non-native readers. We performed automatic evaluation and user study to verify the effectiveness of the proposed system.

In future work, we will further improve the summary's reading easiness in the following two ways: 1) The summary fluency (e.g. sentence ordering in a summary) has influences on the reading easiness of a summary, and we will consider the summary fluency factor in the summarization system. 2) More sophisticated sentence reduction and sentence simplification techniques will be investigated for improving the summary's readability.

7. ACKNOWLEDGMENTS

This work was fully supported by NSFC (60873155), and partially supported by RFDP (20070001059), Beijing Nova Program (2008B03), NCET (NCET-08-0006) and National High-tech R&D Program (2008AA01Z421).

8. REFERENCES

[1] M. R. Amini, P. Gallinari. The Use of Unlabeled Data to Improve Supervised Learning for Text Summarization. In Proceedings of SIGIR2002, 105-112.

[2] R. Barzilay, N. Elhadad and K. McKeown, Inferring strategies for sentence ordering in multidocument news summarization, Journal of Artificial Intelligence Research 17, 2002.

[3] D. Bollegala, N. Okazaki and M. Ishizuka. A bottom-up approach to sentence ordering for multi-document summarization. In Proceedings of ACL2006.

[4] T. Brants, A. Franz. *Web 1T 5-gram Version 1*. Linguistic Data Consortium, Philadelphia, 2006.

[5] J. S. Chall and E. Dale. Readability revisited: the new Dale-Chall readability formula. Brookline Books. Cambridge, MA, 1995.

[6] C.-C. Chang and C.-J. Lin. LIBSVM : a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[7] K. Collins-Thompson and J. Callan. Predicting reading difficulty with statistical language models. Journal of the American Society for Information Science and Technology, 56(13), 2005.

[8] G. ErKan, D. R. Radev. LexPageRank: Prestige in Multi-Document Text Summarization. In Proceedings of EMNLP2004.

[9] T. L. François. Combining a statistical language model with logistic regression to predict the lexical and syntactic difficulty of texts for FFL. In Proceedings of the EACL2009 Student Research Workshop, 2009.

[10] S. Harabagiu and F. Lacatusu. Topic themes for multi-document summarization. In Proceedings of SIGIR-05.

[11] M. Heilman, K. Collins-Thompson, J. Callan and M. Eskenazi. Combining lexical and grammatical features to improve readability measures for first and second language texts. In Proceedings of HLT-2007.

[12] M. Heilman, K. Collins-Thompson and M. Eskenazi. An analysis of statistical models and features for reading difficulty prediction. In Proceedings of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications, 2008.

[13] T. Kanungo and D. Orr. Predicting the readability of short web summaries. In Proceedings of WSDM2009.

[14] P. Kidwell, G. Lebanon and K. Collins-Thompson. Statistical estimation of word acquisition with application to readability prediction. In Proceedings of EMNLP2009.

[15] J. Kincaid, R. Fishburne, R. Rodgers and B. Chissom. Derivation of new readability formulas for navy enlisted personnel. Branch Report 8-75. Chief of Naval Training, Millington, TN, 1975.

[16] D. Klein and C. D. Manning. Fast Exact Inference with a Factored Model for Natural Language Parsing. In Proceedings of NIPS-2002.

[17] J. Kupiec, J. Pedersen, F. Chen. A.Trainable Document Summarizer. In Proceedings of SIGIR1995, 68-73.

[18] W. Li, F. Wei, Q. Lu and Y. He. PNR2: ranking sentences with positive and negative reinforcement for query-oriented update summarization. In Proceedings of COLING-08.

[19] C. Y. Lin, E. Hovy. The Automated Acquisition of Topic Signatures for Text Summarization. In Proceedings of the 17th Conference on Computational Linguistics, 495-501, 2000.

[20] C..-Y. Lin and E.. H. Hovy. From Single to Multi-document Summarization: A Prototype System and its Evaluation. In Proceedings of ACL-02.

[21] C.-Y. Lin and E.H. Hovy. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In Proceedings of HLT-NAACL -03.

[22] H. P. Luhn. The Automatic Creation of literature Abstracts. IBM Journal of Research and Development, 2(2), 1969.

[23] R. Mihalcea, P. Tarau. TextRank: Bringing Order into Texts. In Proceedings of EMNLP2004.

[24] R. Mihalcea and P. Tarau. A language independent algorithm for single and multiple document summarization. In Proceedings of IJCNLP-05.

[25] A. Nenkova and A. Louis. Can you summarize this? Identifying correlates of input difficulty for generic multi-document summarization. In Proceedings of ACL-08:HLT.

[26] E. Pitler and A. Nenkova. Revisiting readability: a unified framework for predicting text quality. In Proceedings of EMNLP2008.

[27] D. R. Radev, H. Y. Jing, M. Stys and D. Tam. Centroid-based summarization of multiple documents. Information Processing and Management, 40: 919-938, 2004.

[28] S. Schwarm and M. Ostendorf. Reading level assessment using support vector machines and statistical language models. In Proceedings of ACL2005.

[29] A. J. Stenner. Measuring reading comprehension with the Lexile framework. Fourth North American Conference on Adolescent/Adult Literacy, 1996.

[30] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[31] X. Wan, J. Yang and J. Xiao. Using cross-document random walks for topic-focused multi-documetn summarization. In Proceedings of WI2006.

[32] X. Wan and J. Yang. Multi-document summarization using cluster-based link analysis. In Proceedings of SIGIR-08.

[33] X. Wan, J. Yang and J. Xiao. Towards an Iterative Reinforcement Approach for Simultaneous Document Summarization and Keyword Extraction. In Proceedings of ACL2007.

[34] K.-F. Wong, M. Wu and W. Li. Extractive summarization using supervised and semi-supervised learning. In Proceedings of COLING-08.