Automatic Summary Assessment for Intelligent Tutoring Systems

Abstract

Summary writing is an important part of many English Language Examinations. As grading students' summary writings is a very time-consuming task, computer-assisted assessment will help teachers carry out the grading more effectively. Several techniques such as Latent Semantic Analysis (LSA), n-gram co-occurrence and BLEU have been proposed to support automatic evaluation of summaries. However, their performance is not satisfactory for assessing summary writings. To improve the performance, this paper proposes an ensemble approach that integrates LSA and n-gram co-occurrence. As a result, the proposed ensemble approach is able to achieve high accuracy and improve the performance quite substantially compared with current techniques. A summary assessment system based on the proposed approach has also been developed.

Key words: Summary writing assessment, Intelligent tutoring systems, Latent semantic analysis, N-gram co-occurrence, Adaptive feedback

Preprint submitted to Elsevier *February 26, 2009*

1. Introduction

In today's computerized world, teachers are still required to grade students' written summaries manually. This is a very time-consuming task that reduces the amount of time teachers can devote to other duties. In order to reduce the amount of time they have to spend on grading these summaries, many teachers have chosen to reduce the number of summaries given to their students. However, in doing so, students will have insufficient practices, thereby affecting their summary writing skills. To tackle this problem, one approach is to provide computer-assisted assessment of summary writings.

Computer-assisted assessment is a long-standing problem that has attracted interest from the research community since the sixties and has not been fully resolved yet (Perez et al., 2005). With the recent success of e-learning and the advances in other areas such as Information Extraction (IE) and Natural Language Processing (NLP), automatic assessment of summary writings has become possible. Some of the techniques such as Latent Semantic Analysis (LSA) (Landauer et al., 1997, 1998; Zipitria et al., 2004; Franzke and Streeter, 2006), BLEU (Pérez et al., 2004), n-gram co-occurrence (Lin, 2004) have been proposed. However, most of these techniques are unable to achieve satisfactory performance for assessing summary writings. In this paper, we propose an ensemble approach, that integrates two of the most effective summary evaluation techniques, LSA and n-gram co-occurrence, for improving the accuracy of automatic summary assessment.

Summary writings are usually assessed based on two criteria, content and style. In this paper, the proposed ensemble technique focuses mainly on content assessment. The rest of the paper is organized as follows. Section 2 reviews some of the techniques currently employed in summary evaluation. The proposed approach is presented in Section 3. Performance analysis is discussed in Section 4. Section 5 presents the intelligent summary assessment system developed. Section 6 describes the pedagogical model of our proposed automatic summary assessment system. Finally, Section 7 concludes the paper.

2. Summary Assessment Techniques

This section reviews some of the most popular summary evaluation techniques including those based on Latent Semantic Analysis (LSA) (Landauer et al., 1997, 1998; Zipitria et al., 2004; Franzke and Streeter, 2006) and those based on machine translation evaluation methods (Pérez et al., 2004; Lin and Hovy, 2003; Lin, 2004).

2.1. LSA Based Techniques

Landauer et al. (1998) first developed Latent Semantic Analysis (LSA) in the late '80s with the purpose of indexing documents and information retrieval. Automated assessment of natural text was an interesting problem since that time. Landauer modified LSA to assess natural text. LSA functions by using a matrix to capture words and frequency of the words appearing in a context. The matrix is then transformed using Singular Value Decomposition (SVD). Cosine correlation is used to determine the similarity. Based on the result of Landauer's experiment, LSA is capable of producing results that are approximately as well as experts' assigned scores as the scores correlate with each other. However, LSA does not make use of word order as Landauer claims that word order is not the most important factor in collecting the sense of a passage (Landauer et al., 1997).

A commercial summary evaluation system, Laburpen Ebaluaketa Automatikoa (LEA) (Zipitria et al., 2004), also makes use of LSA to derive summarization scores. LEA is designed to address two types of users, teachers and students. LEA allows teachers to manage summarization exercises and inspect students' answers, and allow students to create their own summaries. There is a support tool that is available to help students write their summaries. LEA evaluates summaries based on the combination of partial scores in cohesion, coherence, adequacy, use of language and comprehension.

Franzke and Streeter (2006) at the University of Colorado at Boulder developed Summary Street$^{\mathrm{TM}}$, an automated tool to evaluate the content of students' summaries. Summary Street grades students writing by comparing it with the actual text, evaluating it based on content knowledge, writing mechanics, redundancy and relevancy. Based on the grading given by Summary Street, feedback is given to help the student know where his/her mistake is. The core of Summary Street is the Knowledge Analysis Technologies$^{\mathrm{TM}}$ (KAT) engine. The KAT engine uses a modified version of Latent Semantic Analysis (LSA).

2.2. Machine Translation Based Techniques

Pérez et al. (2004) modified the BLEU algorithm, which was originally developed for ranking machine translation systems, into one that is capable of marking students' essay. The modified BLEU algorithm is capable of assessing a student's essay for relevant information by matching it with the model essay stored in the system. BLEU's Brevity Penalty factor was modified to increase the performance of the system, the results of the modification showed that it was able to outperform the original BLEU algorithm in terms of correlation. Based on their evaluation, they had concluded that the modified BLEU algorithm is capable of achieving reasonable correlation with the human markers and it is more than sufficient to replace keyword matching techniques in the assessment of students' essays.

Lin and Hovy (2003) conducted a study on using the two machine translation evaluation techniques, BLEU and NIST's n-gram co-occurrence scoring procedures, on the evaluation of summaries. The main idea of the comparison is to measure the closeness of the candidate to the reference summary by using the weighted average of variable length n-gram matches from that of BLEU. Based on the result of their experiments, they had found out that unigram co-occurrence statistics is a good automatic scoring metric as it is capable of constantly achieving high correlation with human assessments.

Lin (2004) also developed an automatic summary evaluation program called Recall-Oriented Understudy for Gisting Evaluation (ROUGE). The current version of ROUGE consists of five different automatic evaluation methods, namely ROUGE-N, ROUGE-L, ROUGE-W, ROUGE-S and ROUGE-SU. ROUGE-N uses n-gram co-occurrences between the candidate and reference summaries, which is similar to the BLEU algorithm in machine translation. N-gram with length greater than one can be used to estimate the fluency of summaries. ROUGE-L consists of matching two sequences by matching their subsequence. The longer the matching subsequence, the more similar the two sequences are. ROUGE-W is similar to ROUGE-L in which they both deal with matching subsequences but in ROUGE-W weights are used. ROUGE-S uses skip-bigram to estimate the similarity between two summaries. Since ROUGE-L and ROUGE-W can only match subsequence, ROUGE-S compensates this by being able to match pairs of word in their sentence order with arbitrary gaps in between them. ROUGE-SU is similar to ROUGE-S with the addition of unigram based co-occurrence statistics. The evaluation of ROUGE had shown that it correlates surprising well with human evaluations.

2.3. Ensemble Techniques

Ensemble techniques (Domingos, 2000; Wheway, 2001) are broadly classified into two categories: unweighted and weighted voting, both with the purpose of combining the strength of weak techniques to produce an accurate final result. Unweighted voting includes method such as Bagging, Error-correcting Output Codes, etc., whereas weighted voting includes Boosting, Stacking, etc. The main difference between them is how they manage the results obtained from the underlying techniques used to build the ensemble approach.

In unweighted voting, the results of the underlying techniques are treated as equals, not placing more value in any one of the techniques. The weakness of this approach is that no priority is placed on any technique, thereby resulting the final score as being a simple average of the underlying techniques' scores.

For weighted voting, each of the underlying techniques are tested to find out their efficiency and accuracy, then a weight based on those criteria is calculated and applied to the results produced by the corresponding technique and an ensemble score will be calculated based on them. A disadvantage of weighted voting is due to the weights, as weights tend to be higher for the more accurate technique, which in turn leads to that technique influencing the final result much more than the others. Thus, the result in the final result produced is similar to a non-ensemble one.

3. Proposed Approach

In semantic assessment of summary writings, student solutions are graded based on the number of content points answered. Apart from those commercial techniques such as LEA and Summary Street, there are mainly four summary assessment techniques, namely LSA, BLEU, n-gram co-occurrence and ROUGE. After evaluating these techniques, we found that the overall performance of ROUGE is quite poor compared with the other three techniques. We then built the ensemble approach using LSA, n-gram co-occurrence and BLEU. However, we found that BLEU produced low scores in its performance when ensembled with other techniques. This might due to the brevity penalty over penalization. As such, the resultant ensemble approach only comprises LSA and n-gram co-occurrence. Furthermore, as the LSA and n-gram co-occurrence techniques have roughly the same performance, both techniques will have very similar weights if the weighted approach is used. Since the weights are similar, the use of the unweighted approach will simplify the amount of processing required by the ensemble approach.

Figure 1 shows the proposed ensemble approach which consists of two major modules: pre-processing and ensembling.

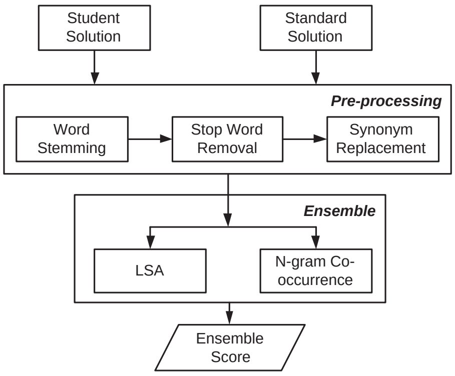


Figure 1: The proposed ensemble approach.

3.1. Pre-processing

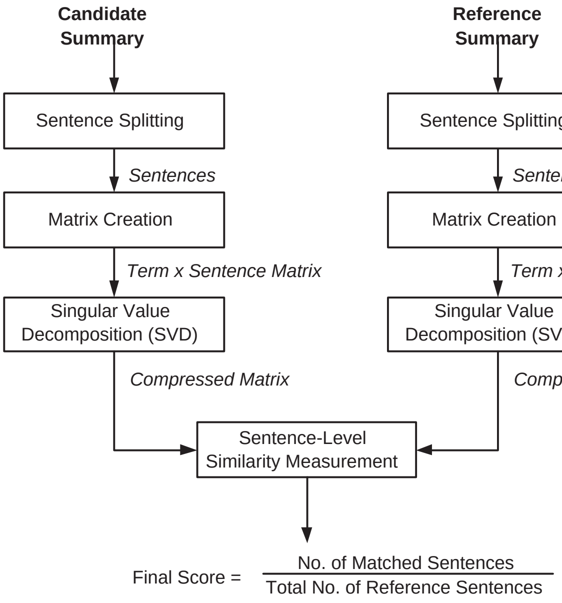
Both the student's candidate solution and the standard (reference or model) solution will first go through the pre-processing module. To avoid the problem that the student's candidate solution uses different words from the reference summary, the pre-processing module aims to create a common basis for comparison by converting all words used by the candidate and reference summaries to a common one. Therefore, the pre-processing module provides text pre-processing functions such as converting synonyms into a common word, eliminating grammatical differences and removing stop words. For the first two functions, WordNet (Miller et al., 2006) was used. As for the removal of stop words, a list obtained from the University of Glasgow[1] was used.

[1] `http://ir.dcs.gla.ac.uk/resources/linguistic_utils/stop_words`

3.2. Ensembling

The ensemble approach comprises the modified LSA algorithm and n-gram co-occurrence which are discussed in this subsection.

3.2.1. Modified Latent Semantic Analysis (LSA)



Candidate Summary
Reference Summary
Sentence Splitting
Sentences
Matrix Creation
Term x Sentence Matrix
Singular Value Decomposition (SVD)
Compressed Matrix
Sentence-Level Similarity Measurement
Final Score = No. of Matched Sentences / Total No. of Reference Sentences

Figure 2: Modified Latent Semantic Analysis (LSA).

Figure 2 shows the overall process of applying the modified Latent Semantic Analysis (LSA) to summary assessment. First, a reference summary and a student's candidate summary is split into a set of sentences. For any summary, suppose there are m distinct terms in n sentences. The summary can be represented as a term-sentence $(m \times n)$ matrix X, whose component x_{ij} is the weighted frequencies for how often a term t_i occurs in a sentence d_j. The original matrix X is then broken into the product of three new matrices $X = U\Sigma V^T$ where U and V are the matrices of the left and right singular vectors for terms and sentences respectively. Σ comprises a diagonal of scaling factors. Some number k of the scaling factors is retained and the matrices are recombined using only the retained factors. Thus, the original matrix X is approximated with a rank-k matrix $X_k = U_k\Sigma V_k^T$ by setting the smallest $r - k$ singular values to zero (r is the rank of X).

The result is a compressed form of the original matrix in which frequency values are approximated (raised or lowered) depending on the number of factors used. After generating the compressed matrix for a reference summary, a vector for each sentence can be constructed by taking values in the matrix for each term found in that sentence. A vector for each sentence in the candidate summary can also be computed in a similar way. The cosine distance between the reference vector and the candidate vector can then be calculated as an indication of their semantic similarity. A candidate sentence can be considered as matched with a reference sentence if their cosine distance is within an empirically determined threshold. The final score is computed as the total number of matched sentences out of the total number of sentences in the reference summary.

3.2.2. n-gram Co-occurrence

An n-gram is a subsequence of n items from a given sequence. In our application here, n-gram refers to a subsequence of n words in a sentence. An n-gram of size 1 is a “unigram”; size 2 is a “bigram”; size 3 is a “trigram”; and size 4 or more is simply called an “n-gram”.

N-gram co-occurrence measures how well a candidate summary overlaps with a reference summary using a weighted average of variable length n-gram matches. The calculation of n-gram co-occurrence statistics for summary assessment is shown in Figure 3.

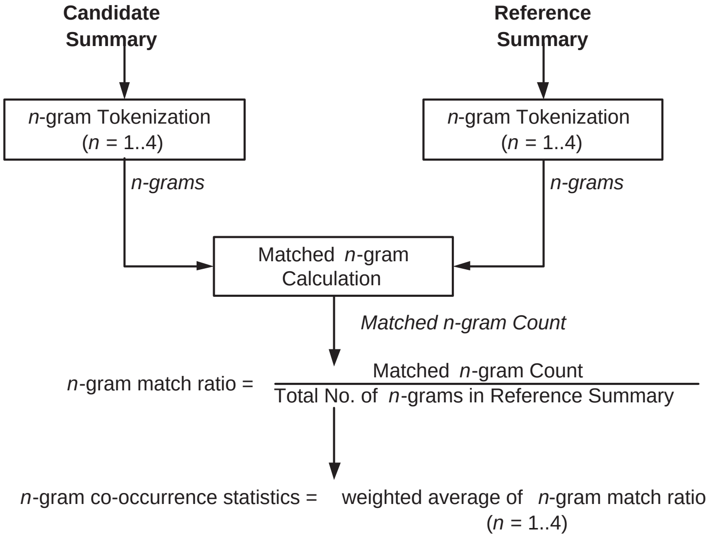

Candidate Summary
Reference Summary
n-gram Tokenization (n = 1..4)
n-gram Tokenization (n = 1..4)
n-grams
n-grams
Matched n-gram Calculation
Matched n-gram Count
n-gram match ratio = Matched n-gram Count / Total No. of n-grams in Reference Summary
n-gram co-occurrence statistics = weighted average of n-gram match ratio (n = 1..4)

Figure 3: n-gram Co-occurrence.

First, the n-gram match ratio is calculated as follows:

$$C_n = \frac{\sum_{S_r \in \mathcal{S}} \sum_{n\text{-gram} \in S_r} \text{Count}_{\text{match}}(n\text{-gram})}{\sum_{S_r \in \mathcal{S}} \sum_{n\text{-gram} \in S_r} \text{Count}(n\text{-gram})} \tag{1}$$

where $\mathcal{S} = \{S_1, S_2, ..., S_R\}$ comprises all the sentences in a reference summary. $\text{Count}_{\text{match}}(n\text{-gram})$ is the maximum number of n-grams co-occurring in a candidate summary and a reference summary and $\text{Count}(n\text{-gram})$ is the number of n-grams in the reference summary.

The n-gram co-occurrence statistics used are based on the following equation:

$$n\text{-gram}(i, j) = \exp\,(\sum_{n=i}^{j} w_n \log C_n) \tag{2}$$

where $j \geq i$, i and j range from 1 to 4, and $w_n = 1/(j - i + 1)$. n-gram$(1, 4)$ is a weighted variable length n-gram match score similar to the IBM BLEU score (Papineni et al., 2002); when $i = j$, n-gram(i, i) is simply the average i-gram coverage score C_i.

3.2.3. Ensemble Approach

In the ensemble approach, the scores of the individual techniques of LSA and n-gram co-occurrence are used for the unweighted voting by taking an unweighted average, i.e.,

$$\text{Ensemble Score} = \frac{\text{LSA Score} + n\text{-gram co-occurrence Score}}{2} \tag{3}$$

A threshold value will be assigned to determine if the averaged score is considered as a positive or negative solution.

4. Performance Analysis

In this section, we present the performance of the proposed ensemble approach in comparison with other assessment techniques. As LEA and Summary Street$^{\text{TM}}$ are commercial and patented techniques, we were unable to obtain their programs for testing. However, the other techniques such as LSA, BLEU, ROUGE and n-gram co-occurrence are compared. The following six different types of tests are used to compare the performance. The objectives of these tests are given below:

- *Exact test* - It is used to judge if the technique is capable of providing a high score for totally related candidate summary and reference solution.

- *Opposite test* - It is used to judge if the technique is capable of providing extremely low score when the candidate summary and reference solution are totally unrelated.

- *Content test* - It is used to determine whether the technique is capable of producing a score that is proportional to the number of content points present in the candidate summary.

- *Synonym test* - It is used to determine if the technique is able to evaluate the candidate summary based on their content and not be influenced by the different synonyms used in the summaries.

- *Grammar test* - It is used to determine if the technique is able to evaluate the candidate summary based on their content and not be affected by the different grammar used in the summaries.

- *Student test* - It aims to determine if the technique is capable of producing score that is closely related to the one that is given by a human expert. The candidate summaries used in this test are written by current students, as opposed to those that are generated artificially used for the above tests. Therefore, this allows us to test if the technique is capable of accurately assessing real-life summaries.

The six tests are used to evaluate the performance of the ensemble approach in comparison with other base techniques. All reference summary solutions used in the tests are obtained from Cambridge O-Level English Language Examination (Rajamanikum, 2000; Lee, 2005). The performance evaluation was conducted on 50 test samples (or student summaries) with 1 being the most accurate and 0 being the most inaccurate for all the tests. All the test samples were collected from a class of students taking the Mid-Year Examination 2007 of Hillgrove Secondary School in Singapore. These candidate summaries had been graded by their O-Level English teacher.

Figure 4 shows the accuracy of the ensemble approach for each of the six tests versus the different settings of the threshold value that defines the matching criteria. It can be observed that the optimal threshold value is 0.7 as the accuracy starts to decline beyond this value.

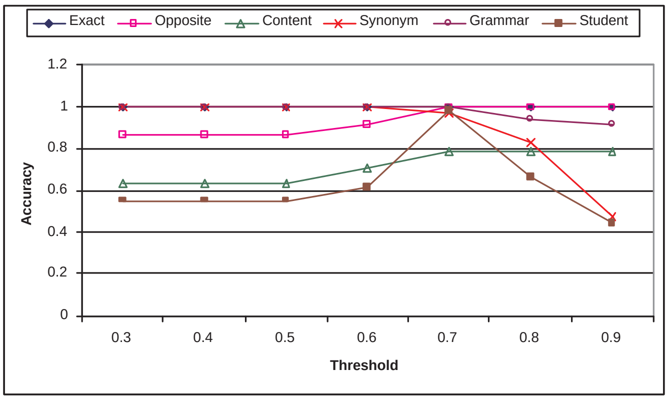


Figure 4: Performance of the proposed ensemble approach vs threshold values.

Table 1 and Figure 5 show a comparison of the ensemble system and the base techniques on LSA, *n*-gram co-occurrence, BLEU and ROUGE using their best performance parameters and thresholds. It can be observed that the ensemble system is able to outperform all the base techniques in all the tests except for the content test. For the other tests, the ensemble approach is capable of outperforming the other techniques by at least 0.003 and at most 0.774 in terms of accuracy. Based on the results of the tests, the proposed approach is capable of producing equal or higher accuracy compared to the existing techniques in all tests except for one. Even though the proposed approach did not perform as well in the content tests, its overall accuracy of 96% is still much higher than that of the existing techniques.

Table 1: Cross comparison between the ensemble approach and other techniques on the six tests.

Test	*LSA*	*N-gram*	*BLEU*	*ROUGE*	*Ensemble*
Exact	0.800	0.997	0.942	0.528	1
Opposite	0.787	0.966	1	1	1
Content	0.981	0.932	0.948	0.706	0.793
Synonym	0.716	0.917	0.649	0.194	0.968
Grammar	0.739	0.946	0.789	0.346	1
Student	0.849	0.822	0.871	0.476	0.978

When comparing the chances of producing false positives with the existing base techniques as shown in Table 2, the ensemble approach is slightly worse than the other techniques as it had the highest chances of producing them while having the lowest probability for producing false negatives. On the whole, the ensemble approach proves to be superior to the base techniques.

5. System Implementation

A summary assessment system has been developed based on the proposed approach. It comprises eight main components as depicted in Figure 6.

- *Main User Interface* is a container GUI that allows user to choose

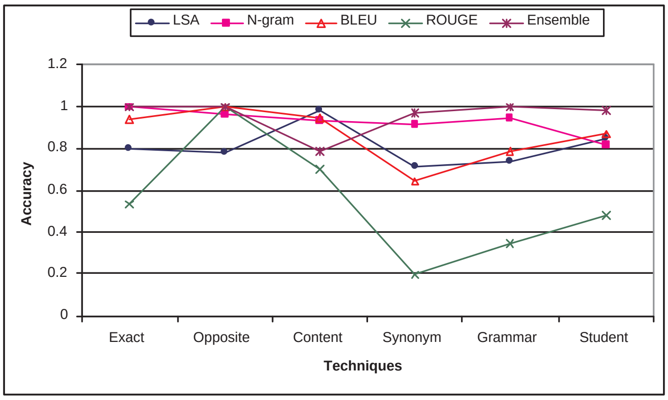


Figure 5: Performance comparison of the existing techniques and the proposed ensemble approach.

Table 2: Cross comparison between the ensemble approach and other techniques in FP and FN.

Method	*False Positive*	*False Negative*
LSA	0.094	0.228
N-gram	0.033	0.093
BLEU	0.026	0.194
ROUGE	0.003	0.476
Ensemble	0.124	0.046

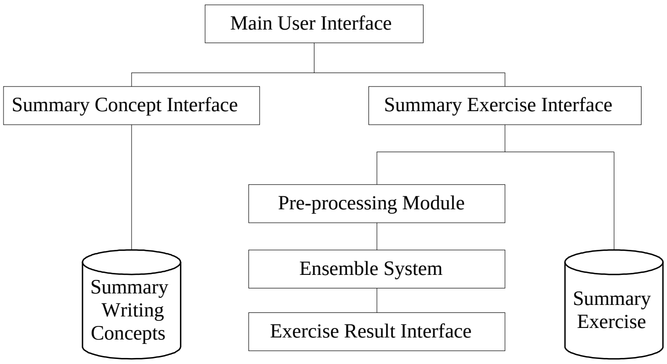


Figure 6: The summary assessment system.

between viewing the Summary Concept Interface or Summary Exercise Interface.

- *Summary Concept Interface* allows the user to view summary writing concepts in the form of web pages.

- *Summary Writing Concepts* contains a set of web pages that contains summary writing concepts.

- *Summary Exercise Interface* allows the user to practise his/her summary writing skills by practising on one of the exercises present in the Summary Exercise.

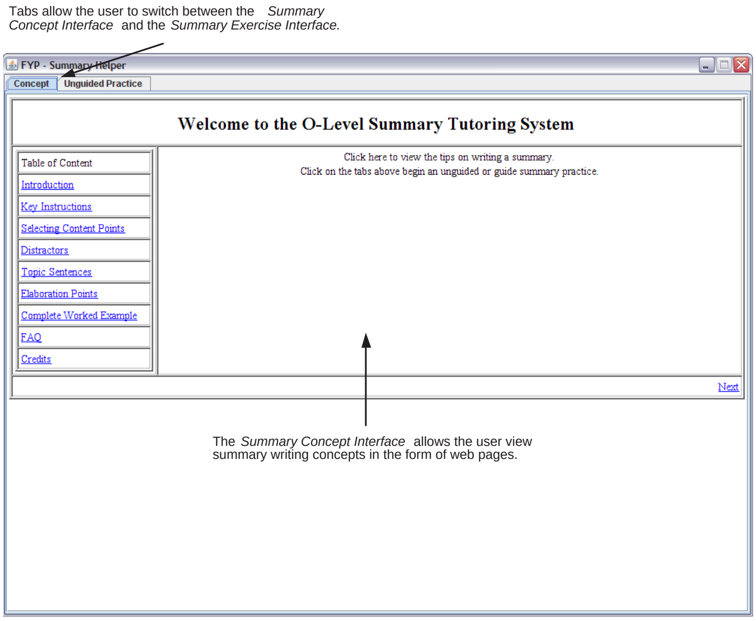
- *Summary Exercise* contains a set of practise summaries for the user to practise on.

- *Pre-processing Module* pre-processes the student's candidate summary and the reference solution before passing them onto the assessment technique for scoring.

- *Ensemble System* computes the co-relationship between the candidate summary and reference solution.

- *Exercise Result Interface* is used to display the results from the ensemble system to the user. It also informs the user which content points had been left out in the user's solution.

Figure 7 illustrates the *Main User Interface* and *Summary Concept Interface* components. It allows users to select between viewing of summary writing concepts and practising summary writing using the *Summary Exercise Interface* component by changing the tab.


Tabs allow the user to switch between the *Summary Concept Interface* and the *Summary Exercise Interface.*

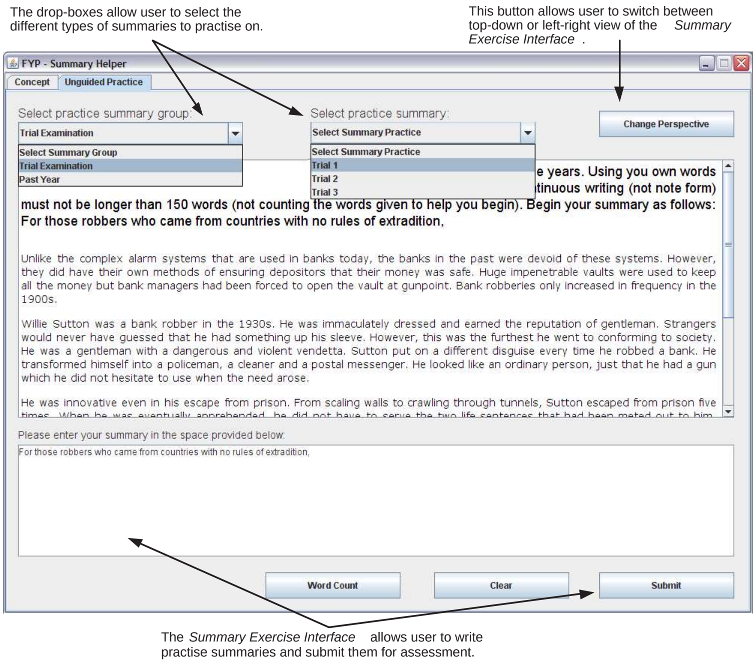
Welcome to the O-Level Summary Tutoring System

Click here to view the tips on writing a summary.
Click on the tabs above begin an unguided or guide summary practice.

The *Summary Concept Interface* allows the user view summary writing concepts in the form of web pages.

Figure 7: The main user interface of the summary assessment system.

By clicking on the practise tab, the user will be brought to the *Summary Exercise Interface* component where he/she can choose the group and type of summary which he/she wants to practise on. After the selection, the instructions and the passage to be summarized will be displayed to the user. The user can input his/her solution before having his/her summary assessed by the system. This is illustrated in Figure 8. After the user's summary is assessed by the system, the results will be displayed to the user using the *Exercise Result Interface* component as shown in Figure 9.



The drop-boxes allow user to select the different types of summaries to practise on.

This button allows user to switch between top-down or left-right view of the *Summary Exercise Interface* .

e years. Using you own words
tinuous writing (not note form)
must not be longer than 150 words (not counting the words given to help you begin). Begin your summary as follows: For those robbers who came from countries with no rules of extradition,

Unlike the complex alarm systems that are used in banks today, the banks in the past were devoid of these systems. However, they did have their own methods of ensuring depositors that their money was safe. Huge impenetrable vaults were used to keep all the money but bank managers had been forced to open the vault at gunpoint. Bank robberies only increased in frequency in the 1900s.

Willie Sutton was a bank robber in the 1930s. He was immaculately dressed and earned the reputation of gentleman. Strangers would never have guessed that he had something up his sleeve. However, this was the furthest he went to conforming to society. He was a gentleman with a dangerous and violent vendetta. Sutton put on a different disguise every time he robbed a bank. He transformed himself into a policeman, a cleaner and a postal messenger. He looked like an ordinary person, just that he had a gun which he did not hesitate to use when the need arose.

He was innovative even in his escape from prison. From scaling walls to crawling through tunnels, Sutton escaped from prison five times. When he was eventually apprehended, he did not have to serve the two life sentences that had been meted out to him.

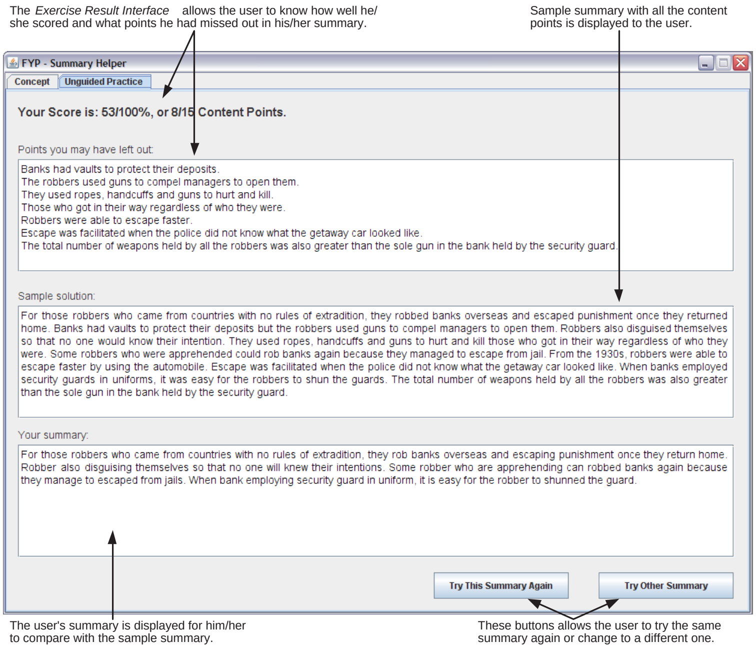
For those robbers who came from countries with no rules of extradition,

The *Summary Exercise Interface* allows user to write practise summaries and submit them for assessment.

Figure 8: The Summary Exercise Interface of the summary assessment system.

6. Pedagogical Model for Automatic Summary Assessment

The pedagogical model of our proposed automatic summary assessment system is a "scaffolding" model which is able to give students adaptive feedbacks based on their previous learning experience. This section describes



The *Exercise Result Interface* allows the user to know how well he/she scored and what points he had missed out in his/her summary.

Sample summary with all the content points is displayed to the user.

Your Score is: 53/100%, or 8/15 Content Points.

Points you may have left out:

Banks had vaults to protect their deposits.
The robbers used guns to compel managers to open them.
They used ropes, handcuffs and guns to hurt and kill.
Those who got in their way regardless of who they were.
Robbers were able to escape faster.
Escape was facilitated when the police did not know what the getaway car looked like.
The total number of weapons held by all the robbers was also greater than the sole gun in the bank held by the security guard.

Sample solution:

For those robbers who came from countries with no rules of extradition, they robbed banks overseas and escaped punishment once they returned home. Banks had vaults to protect their deposits but the robbers used guns to compel managers to open them. Robbers also disguised themselves so that no one would know their intention. They used ropes, handcuffs and guns to hurt and kill those who got in their way regardless of who they were. Some robbers who were apprehended could rob banks again because they managed to escape from jail. From the 1930s, robbers were able to escape faster by using the automobile. Escape was facilitated when the police did not know what the getaway car looked like. When banks employed security guards in uniforms, it was easy for the robbers to shun the guards. The total number of weapons held by all the robbers was also greater than the sole gun in the bank held by the security guard.

Your summary:

For those robbers who came from countries with no rules of extradition, they rob banks overseas and escaping punishment once they return home. Robber also disguising themselves so that no one will knew their intentions. Some robber who are apprehending can robbed banks again because they manage to escaped from jails. When bank employing security guard in uniform, it is easy for the robber to shunned the guard.

The user's summary is displayed for him/her to compare with the sample summary.

These buttons allows the user to try the same summary again or change to a different one.

Figure 9: The Exercise Result Interface of the summary assessment system.

the method used to update the student model in the summary assessment system in order to facilitate adaptive feedback provision.

6.1. Adaptive Feedback

Modern theories of learning put an emphasis on the critical role of practice and they also highlight the importance of feedback because inherent risks exist in unguided environments (Clark, 2004; Kirschner et al., 2006). Examples of the unguided risks include the development of incomplete or fragmented knowledge and the formation of misconceptions, etc. Chi et al. (2001) argue that an intelligent tutoring system should implement ways to elicit constructive responses from students as students' constructive activities from interaction is important for learning. Lane (2006) observed that a great deal of attention has been given to the modeling of the affective state of learners recently as evidenced from the learning sciences that expert human tutors do manage the motivational and emotional states of learners, and also instruction can be adjusted according to motivation in ways that improve learning. In computer environment, feedback frequency could be adjusted based on the motivational state estimate of the students which could be gathered by some highly detailed measurements such as time between keystrokes by the students.

Thus, in view of the above, one essential feature of an intelligent tutoring system is its ability of adapting to an individual student's knowledge skill levels during his/her interactions with the system. That is, depending on an individual student's knowledge skill and perhaps some of the other relevant features, the system should adapt dynamically to learning interactions and be able to give personalized advices to the student. A student model stores the knowledge skill levels and other relevant features about a student. It is important that such a student model could be updated dynamically to reflect the current student learning capabilities. The system could then decide on what should be the next action to be taken based on the current student model.

Recently, a number of e-learning systems that use adaptive feedback have been developed. In (Arroyo et al., 2000; Shute et al., 2007), Piagetian tests and Bayesian networks were applied to determine students' cognitive abilities and background knowledge which are subsequently used to choose the next best problem for students to work on. This is so-called adaptive sequencing. Timms (2007) proposed Item Response Theory models to adapt help to students' ability such that more explicit hints with step-by-step instructions are provided for slow learners whilst more conceptual hints are provided for fast learners. In the automatic summary assessment system proposed here, students' knowledge skill levels are determined by their performance on practicing exercises through the system. The system could then adapt its interactivity to individual students.

Adaptive feedback could also be best described under the scaffolding learning theory. Scaffolding learning originates from Lev Vygotsky's sociocultural theory (Vygotsky, 1962) and centers around his concept of the Zone of Proximal Development (ZPD). The scaffolding teaching strategy provides individualized support based on the learner's ZPD (Chang et al., 2002). Scaffolding must begin from what is near to a learner's experience and build to what is just beyond the level of what the learner can do alone (Olson and Platt, 2000). Scaffolds may include models, cues, prompts, hints, partial solutions, think-aloud modeling and direct instruction (Hartman, 2002). The scaffolds prompt learners to complete the next step of the task, thus helping them through the ZPD (Bransford et al., 2000). Questions may also be used as scaffolds to help students solve a problem or complete a task. The level of questioning or specificity could be increased until the student is able to provide a correct answer.

6.2. Student Model Update

In our proposed summary assessment system, a student model has been developed to facilitate adaptive feedback provision. Information stored in the student model include the personal information about each student such as name, sex, age, etc. Students' learning records are also stored in the student model. Each student has his/her own learning record which tracks individual performance.

When using the summary assessment system, teachers can group the essays into several categories such as the following:

- Narrative essays (stories and events)
- Descriptive essays
- Analytical essays (argumentative and reflective essays)
- Situational essays (reports and articles, letters and speech)
- General essays

First of all, the student model should be able to reflect each student's summary writing skills, with respect to the summary assessment results and the number of exercises taken thus far. In addition, the model is expected to show the relative performance of a student in various essay categories. The system will identify the potential weakness in the student's summary writing ability based on the relative performance. It will then suggest a checklist of points to look out and more targeted exercises that the student could carry out in order to improve his/her writing skills.

At this point, an assumption is made that a student's summary writing performance can be measured by his/her assessment results. A student's performance for each essay category and the overall performance (for all the essay categories) will be recorded. If a student's corresponding performance measure in a particular essay category is below a pre-defined threshold, the system will then suggest more essays from that category to the student in upcoming exercises.

On the other hand, teachers could also group the summary exercises according to the difficulty levels. Students could start with the easy exercises to build up their confidence and gradually progress to intermediate and more difficult levels.

We believe that an automatic summary assessment system that combines adaptive feedback and the scaffolding teaching strategy would be able to provide much better learning experience for students.

7. Conclusion

In this paper, various techniques on summary evaluation are first reviewed. We then propose an ensemble approach which integrates the best techniques into a single efficient assessment technique that is capable of producing high accuracy. The proposed approach is capable of achieving an overall accuracy of 96% as compared to the best existing technique, BLEU, which has an overall accuracy of 87%. In addition, we have implemented the proposed ensemble approach into a summary assessment system for automatic grading of English summary writings. The pedagogical model of the proposed automatic summary assessment system has also been presented.

For future work, as the techniques used and reviewed in this paper are mainly based on latent semantic analysis or machine translation based evaluation techniques, we will investigate the effectiveness of using machine learning or statistical approaches for the assessment of summary writings. In addition, as our current approach only focuses on semantic assessment of contents, we also intend to develop a complete summary assessment system by incorporating an English language assessor and style checker.

The current assessment system is mainly targeted for secondary education and for the assessment of summary writings in English only. It is possible to extend the system to cover other languages by investigating the language specific adaptations to our proposed ensemble approach.

References

Arroyo, I., Beck, J., Woolf, B., Beal, C., Schultz, K., 2000. Macroadapting animalwatch to gender and cognitive differences with respect to hint interactivity and symbolism. In: Gauthier, G., Frasson, C., VanLehn, K. (Eds.), Intelligent Tutoring Systems. LNCS 1839. pp. 574–583.

Bransford, J., Brown, A., Cocking, R., 2000. How People Learn: Brain, Mind, and Experience & School. National Academy Press, Washington, DC.

Chang, K., Chen, I., Sung, Y., 2002. The effect of concept mapping to enhance text comprehension and summarization. The Journal of Experimental Education 71 (1), 5–23.

Chi, M., Siler, S., Jeong, H., Yamaguchi, T., Hausmann, R., 2001. Learning from human tutoring. Cognitive Science 25, 471–533.

Clark, R., 2004. Design document for a guided experiential learning course. Final report on contract DAAD 19-99-D-0046-0004 from TRADOC to the Institute for Creative Technologies and the Rossier School of Education.

Domingos, P., 2000. Bayesian averaging of classifiers and the overfitting problem. In: Proceedings of the 17th International Conference on Machine Learning. Morgan Kaufmann, San Francisco, CA, pp. 223–230.

Franzke, M., Streeter, L., 2006. Building student summarization, writing and reading comprehension skills with guided practice and automated feedback. Highlights From Research at the University of Colorado, A white paper from Pearson Knowledge Technologies.

Hartman, H., 2002. Human Learning and Instruction. City College of City University of New York, New York, Ch. Scaffolding & Cooperative Learning, pp. 23–69.

Kirschner, P., Sweller, J., Clark, R., 2006. Why minimally guided learning does not work: An analysis of the failure of discovery learning, problem-based learning, experiential learning and inquiry-based learning. Educational Psychologist 41 (2), 75–86.

Landauer, T., Foltz, P., Laham, D., 1998. Introduction to latent semantic analysis. Discourse Processes 25, 259–284.

Landauer, T., Laham, D., Rehder, B., Schreiner, M. E., 1997. How well can passage meaning be derived without using word order? a comparison of latent semantic analysis and humans. In: Proceedings of the 19th Annual Conference of the Cognitive Science Society.

Lane, H., August 2006. Intelligent tutoring systems: Prospects for guided practice and efficient learning. In: Army's Science of Learning Workshop. Hampton, VA.

Lee, J., 2005. O-Level English. Singapore Asian Publications (S) Pte Ltd.

Lin, C.-Y., July 2004. ROUGE: a package for automatic evaluation of summaries. In: Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004). Barcelona, Spain.

Lin, C.-Y., Hovy, E., 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. pp. 71–78.

Miller, G. A., Fellbaum, C., Tengi, R., Wakefield, P., Poddar, R., Langone, H., Haskell, B., 2006. WordNet: a lexical database for the English language. Princeton University Cognitive Science Laboratory.

Olson, J., Platt, J., 2000. Teaching Children and Adolescents with Special Needs. Prentice-Hall, Inc., Upper Saddle River, NJ, Ch. The Instructional Cycle, pp. 170–197.

Papineni, K., Roukos, S., Ward, T., jing Zhu, W., 2002. BLEU: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318.

Pérez, D., Alfonseca, E., Rodríguez, P., 2004. Upper bounds of the BLEU algorithm applied to assessing student essays. In: Proceedings of the 30th International Association for Educational Assessment (IAEA) Conference.

Perez, D., Gliozzo, A., Strapparava, C., Alfonseca, E., Rodriguez, P., Magnini, B., May 2005. Automatic assessment of students' free-text answers underpinned by the combination of a bleu-inspired algorithm and latent semantic analysis. In: Proceedings of the 18th International FLAIRS Conference. Clearwater Beach, Florida.

Rajamanikum, K., 2000. English language (Yearly) Worked Solutions. Redspot Publishing Singapore.

Shute, V., Hansen, E., Almond, R., 2007. Evaluating ACED: The impact of feedback and adaptivity on learning. In: Luckin, Koedinger (Eds.), Artificial Intelligence in Education. IOS Press, pp. 230–237.

Timms, M., 2007. Using item response theory (IRT) to select hints in an ITS. In: Luckin, Koedinger (Eds.), Artificial Intelligence in Education. IOS Press, pp. 213–221.

Vygotsky, L., 1962. Thought and Language. MIT Press, Cambridge, MA.

Wheway, V., 2001. Using boosting to simplify classification models. In: Proceedings of the 2001 IEEE International Conference on Data Mining. Washington, DC, USA, pp. 558–565.

Zipitria, I., Elorriaga, J., Arruate, A., de IIarraza, A., 2004. From human to automatic summary evaluation. In: 7th International Conference on Intelligent Tutoring System.