An Ensemble Approach for Semantic Assessment of Summary Writings

Yulan He*, Siu Cheung Hui† and Tho Thanh Quan‡
*School of Engineering, Computing and Mathematics
University of Exeter
North Park Road, Exeter EX4 4QF, UK
y.l.he.01@cantab.net
†School of Computer Engineering
Nanyang Technological University
Nanyang Avenue, Singapore 639798
asschui@ntu.edu.sg
‡Faculty of Computer Science and Engineering
Hochiminh City University of Technology
Hochiminh City, Vietnam
qttho@cse.hcmut.edu.vn

***Abstract*—Computer-assisted assessment of summary writings is a challenging area which has recently attracted much interest from the research community. This is mainly due to the advances in other areas such as information extraction and natural language processing which have made automatic summary assessment possible. Different techniques such as Latent Semantic Analysis, n-gram co-occurrence and BLEU have been proposed for automatic evaluation of summaries. However, these techniques are unable to achieve good performance. 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. The performance of the proposed ensemble approach has shown that it is able to achieve high accuracy and improve the performance quite substantially compared with other existing techniques.**

I. Introduction

Summary writing is an important part of many English examinations. The assessment of summary writings is normally conducted based on content overlapping and linguistic qualities by comparing a student's candidate summary with a reference summary. However, manual assessment is always a tedious and labor-intensive process. Furthermore, there is a problem of inconsistency in human assessment that different markers might give different grades for the same candidate summary. To help alleviate such problems, there has been a growing interest recently in automatic summary assessment or computer-assisted assessment of summaries.

Computer-assisted assessment is a long-standing problem that has attracted much interest from the research community since the sixties and has not been fully resolved yet [1]. 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) [2], [3], [4], [5], BLEU [6], n-gram co-occurrence [7] 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 II reviews some of the techniques currently employed in summary evaluation. The proposed approach is presented in Section III. Performance analysis is discussed in Section IV. Finally, Section V concludes the paper.

II. Summary Assessment Techniques

This section reviews some of the most popular summary evaluation techniques including those based on Latent Semantic Analysis (LSA) [2], [3], [4], [5] and those based on machine translation evaluation methods [6], [8], [7].

A. LSA Based Techniques

Landauer *et al.* [3] 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 [2].

A commercial summary evaluation system, Laburpen Ebaluaketa Automatikoa (LEA) [4], 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 *et al.* [5] at the University of Colorado at Boulder developed Summary Street^{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^{TM} (KAT) engine. The KAT engine uses a modified version of Latent Semantic Analysis (LSA).

B. Machine Translation Based Techniques

Perez *et al.* [6] 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 *et al.* [8] 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 [7] 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.

III. 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.

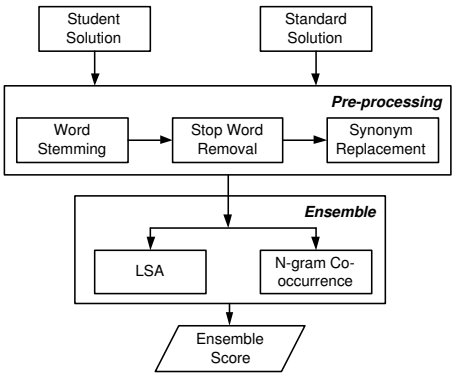

Fig. 1. The proposed ensemble approach.

A. 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 [9] was used. As for the removal of stop words, a list obtained from the University of Glasgow[1] was used.

B. Ensembling

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

1) Modified Latent Semantic Analysis (LSA): When applying 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.

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".

[1]http://ir.dcs.gla.ac.uk/resources.html

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

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

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 Count(n-gram) is the number of n-grams in the reference summary.

The n-gram co-occurrence statistics is defined as $n\text{-gram}(i,j) = \exp\left(\sum_{n=i}^{j} w_n \log C_n\right)$ where $j \geq i$, i and j range from 1 to 4, and $w_n = 1/(j-i+1)$. $n\text{-gram}(1,4)$ is a weighted variable length n-gram match score similar to the IBM BLEU score [10]; when $i = j$, $n\text{-gram}(i,i)$ is simply the average i-gram coverage score C_i.

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.,

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

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

IV. 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™ 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 [11], [12]. 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 2 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.

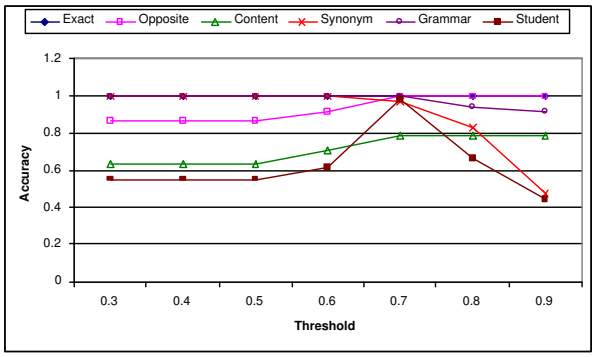

Fig. 2. Performance of the proposed ensemble approach vs threshold values.

Figure 3 shows 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.

When comparing the chances of producing false positives with the existing base techniques as shown in Table I, 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.

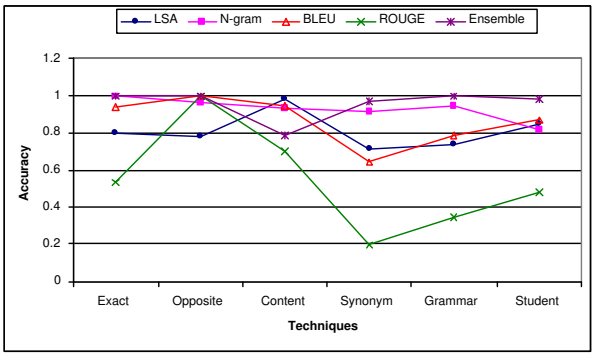

Fig. 3. Performance comparison of the existing techniques and the proposed ensemble approach.

TABLE I
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

V. CONCLUSIONS

In this paper, we propose an ensemble approach which integrates two of the most effective assessment techniques of LSA and n-gram co-occurrence into an efficient technique for automatic summary assessment. Performance comparison between the proposed ensemble approach with other existing techniques has also been conducted. The proposed approach has achieved an overall accuracy of 96% as compared to the best existing technique, BLEU, which has an overall accuracy of 87%. For future work, as the techniques used and proposed 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.

REFERENCES

[1] D. Perez, A. Gliozzo, C. Strapparava, E. Alfonseca, P. Rodriguez, and B. Magnini, "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, May 2005.

[2] T. Landauer, D. Laham, B. Rehder, and M. E. Schreiner, “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*, 1997.

[3] T. Landauer, P. Foltz, and D. Laham, “Introduction to latent semantic analysis,” *Discourse Processes*, vol. 25, pp. 259–284, 1998.

[4] I. Zipitria, J. Elorriaga, A. Arruate, and A. de IIarraza, “From human to automatic summary evaluation,” in *7th International Conference on Intelligent Tutoring System*, 2004.

[5] M. Franzke and L. Streeter, “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, 2006.

[6] D. Pérez, E. Alfonseca, and P. Rodríguez, “Upper bounds of the BLEU algorithm applied to assessing student essays,” in *Proceedings of the 30th International Association for Educational Assessment (IAEA) Conference*, 2004.

[7] C.-Y. Lin, “ROUGE: a package for automatic evaluation of summaries,” in *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*, Barcelona, Spain, July 2004.

[8] C.-Y. Lin and E. Hovy, “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*, 2003, pp. 71–78.

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

[10] K. Papineni, S. Roukos, T. Ward, and W. jing Zhu, “BLEU: A method for automatic evaluation of machine translation,” in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[11] K. Rajamanikum, *English language (Yearly) Worked Solutions*. Redspot Publishing Singapore, 2000.

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