International Journal of Advances in Computer Science & Its Applications – IJCSIA
Volume 4: Issue 4 [ISSN 2250-3765]
Publication Date : 27 December,2014

Automated summarization assessment system: quality assessment without a reference summary

Asad Abdi [1], Norisma Idris [2]

***Abstract*—this paper presents an algorithm that can be used to assess the quality of the summaries without a gold standard. This algorithm is based on linguistic knowledge. An innovative aspect of our algorithm lies in its ability to improve the performance of existing techniques for evaluation summaries. The evaluation results on the students' summaries demonstrate that the proposed algorithm is able to obtain high accuracy and improve performance compared with the current techniques. The algorithm has also been developed into a learning environment for helping both teachers and students.**

***Keywords*—Summarization Evolution algorithm, linguistic measurement, intelligent tutoring systems, Similarity measure.**

I. Introduction

Summarization is a process of generating a short version of a text by retaining the meaning of the whole text. The main idea of summarizing process is to reduce the size and content of the source text into important information. The process contains the combination of information and the designation of the grade of importance of the information included in a text. In addition, it is a process that merges several activities such as comprehension, selection, interpretation, transformation, and generation. The main goal of summary writing operation is to take an information source, extract content from it, and present the most important content to the user in a condensed form[1]. The output of a summary system may be an extractive or abstractive summarization. An extractive summarization method comprises of selecting important sentences from the original text. The importance of sentences is determined by statistical and linguistic features of sentences. An Abstractive summarization tries to develop a comprehending of the main concepts in a text and then expose those concepts. It uses linguistic methods to analyze and interpret the text and then to find the new concepts and expressions to best describe it by generating a new concise text that takes the most important information from the original text.

Summarization systems can also be categorized as generic and query-based summarization systems. In generic text summarization, the summary is made about whole document. But in query-based text summarization, the provided summary is about the query asked[1, 2]. Text summarization evaluation has been a complex and controversial issue in computational linguistics. Methods for evaluating text summarization can be classified into two categories [3]. The first, an intrinsic evaluation, tests the summarization system in of itself. The intrinsic evaluations can then be divided into content evaluation and text quality evaluation. Whereas content evaluations measure the ability to identify the key topics, text quality evaluations judge the readability, grammar and coherence of automatic summaries.

The second, an extrinsic evaluation, tests the summarization based on how it affects the completion of some other task. Because manual comparison of summaries with model summaries is a costly process, various evaluation methods and measures in the last decade developed.

Early studies used text similarity measures such as cosine similarity to compare summary text and reference summary[4], various vocabulary overlap measures such as set of n-grams overlap or longest common subsequence (LCS)between summary text and reference summary have also been proposed[5-7].

The Bleu machine translation evaluation measure[8] has also been tested in summarization[9]. Latent Semantic Analysis [10-13], is a technique for extracting the hidden dimensions of the semantic representation of terms. ROUGE package for content-based evaluation [14]. It implements a series of recall measures based on N-ram co-occurrence statistics between a summary and a set of reference summaries.

Donaway et al.[4] suggested the idea of using directly the full document for comparison purposes, and proved that content-based measures which compare the document to the summary may be acceptable substitutes for those using reference summaries. Louis and Nenkova [15] proposed a method for evaluation of summarization systems without references. It is based on the direct content-based comparison between summary and its corresponding source document. Louis and Nenkova [15] used the Jensen-Shannon [16] theoretic measure for assessing two summarization tasks query-focused and update summarization. In this paper, we propose an algorithm based on the linguistic knowledge that does not make use of human model summaries at all.

II. Related work

In this section most of the summary assessment systems based on LSA (Latent Semantic Analysis) and machine translation evaluation, such as Summary Street [13], LEA[10], Summary Assessment System[17],Automatic Assessment of Students' free-text Answers[18], ROUGE (Recall-Oriented

Asad Abdi [1] *, Norisma Idris* [2]
[1, 2] Faculty of computer science and information technology,
University of Malaya (UM), Kuala Lumpur, Malaysia

Understudy for Gisting Evaluation)[14] and Automatic Evaluation of Summaries[19] , are introduced.

LSA[20] is a technique for extracting the hidden dimensions of the semantic representation of terms, sentences, or documents, on the basis of their contextual use. It has been used in educational applications, such as essay grading[21], as well as in NLP applications containing information retrieval[11] and text segmentation[22].

Laburpen Ebaluaka Automatikoa (LEA)[10], which is based on Latent Semantic Analysis (LSA), has been proposed to evaluate the summary. Summary Street [13], which is based on LSA is a computer-based assessment system that is used to evaluate the content of the summary text.

The Automatic Assessment of Students' free-text answers[18] is based on the BLEU (Bilingual Evaluation Understudy) algorithm and LSA, and was developed for grading students' essays.

Lin [14] proposed an automatic summary assessment system named ROUGE(Recall-Oriented Understudy for Gisting Evaluation), which is used to assess the quality of the summary text. The current system includes a set of statistics (ROUGE-N, ROUGE-L, ROUGE-W, ROUGE-S and ROUGE-SU) that compare summaries with references.

Lin and Hovy [19] proposed a system based on BLEU and N-gram co-occurrence to evaluate summaries with the aim of measuring the closeness of the summary text to the source text. Yuan [17] proposed a summary assessment system based on the modified LSA algorithm and N-gram co-occurrence with the aim of assessing students' written summaries.

Lin et al. [23] suggested a method based on the use of divergences between two probability distributions (the distribution of units in the automatic summary and reference summary). They used two Theoretic measures of divergence,the Kullback-Leibler (KL)[24] and Jensen-Shannon (JS) [16] divergences. Louis and Nenkova [15] proposed a method to compare the distribution of words in full documents with the distribution of words in automatic summaries to derive a content-based evaluation measure.

III. Challenges in the current content evaluation techniques

Since humans may be required to judge the system's output, this may greatly increase the expense of an evaluation. An evaluation which could use a scoring program instead of human judgments is preferable, since it is easily repeatable. One of the important parts in text summarization is to do the assessment of summary, to determine whether an automatic, or even a human-made summary, is appropriate or not. Summary evaluation, either manually or automatically, is a difficult task. The common way to assess the content of the summaries is to compare them with a reference summary. As preparing reference summary is a hard and expensive task. Much effort has to be done in order to have a corpus of texts and their corresponding summaries. On the other hand, different human may chose different sentences, and even, the same human may chose different sentences at different times[25]. The main drawback of the evaluation systems existing so far is that we need at least one reference summary to assess summary, while in our proposed algorithm a reference summary would not be necessary anymore, it takes original text and summary as input to assess summary.

IV. Proposed algorithm

We propose an algorithm which takes syntactic feature and semantic information of words into account to calculate the text similarity. Semantic information is obtained from Word Net, and then syntactic feature are given through analyzing the structure of sentences. Fig. 1 shows the procedure of calculating similarity between source text and summary text. The first stage prepares source text and summary text for further processing. In second stage, for each sentence a word order vector and a semantic vector is created using word group: includes all the distinct words from the pair of sentences, and a lexical database then semantic similarity and word order similarity is computed based on the two semantic vectors and two order vectors respectively. Finally, the sentence similarity is derived by combining semantic similarity and order similarity. In third stage, the result shows how many main ideas of the source text are covered by the summary text. The algorithm consists of three main steps as follows:

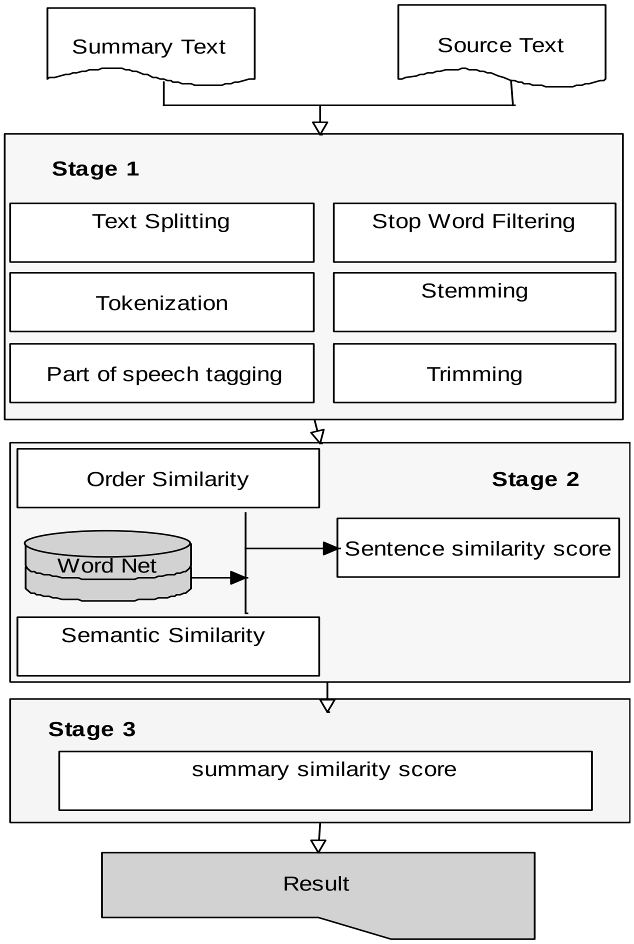

Figure 1. The Architecture of the proposed Algorithm.

A. Stage 1: Linguistics

This stage performs a basic linguistic analysis on both source text and summary text. Thus, it prepares them for further processing. It consists of sentence splitting, trimming, tokenization, stemming, stop word removal and part of speech tagging.

B. Stage 2: the combined semantic and syntactic similarity measures

Given two texts, summary text and source text, this stage aims to identify the similarity score for each sentence of summary text. The similarity measures between each sentence from the summary text and whole sentences from the source text are determined using the composition of order similarity and semantic similarity. The maximum value is assigned as a similarity score for each sentence from the summary. It includes a few components as follows:

1) Semantic similarity

We use the semantic-vector approach [17, 18] to measure the semantic similarity between sentences. The semantic vector is derived from the word group and corresponding sentence. The dimension of the semantic vector equals the number of words in the word group. The value of a cell of the semantic vector is determined using the corresponding sentence and word group and the weight of each cell in semantic vector is determined through these steps:

1. If a word in word group appears in sentence, then the weight of cell in semantic vector is set to 1.
2. If the word does not appear in the sentence, then the weight of cell in the semantic vector is set to 0.

A semantic-vector is created for each of the two sentences. The semantic similarity measure is computed based on the two semantic vectors. We use the cosine-vector based method to calculate the similarity, the formula is as follows:

$$Similarity_{sem} = \frac{d_1 \cdot d_2}{\|d_1\| \cdot \|d_2\|} \qquad (1)$$

Where d_1 and d_2 are the semantic vectors of sentences S_1 and S_2, respectively.

2) Order similarity

We use the syntactic-vector approach [26] to measure the word order similarity between sentences. The syntactic-vector is derived from the word group and corresponding sentence, so the dimension is equal to the number of words in the word set. Unlike the semantic-vector, each cell of the syntactic-vector is weighted using a unique index. The unique index can be the index position of the words that appear in the corresponding sentence. the weight of each cell in order vector is determined through these steps:

1. If a word in word group appears in sentence, then the weight of cell in order vector is set to the corresponding index number from sentence.
2. If a word in word group is not present in sentence, the weight of cell is set to 0.

A syntactic-vector is created for each of the two sentences. The syntactic similarity measure is computed based on the two syntactic-vectors. The following equation is used to calculate word order similarity between sentences:

$$Similarity_{wo} = 1 - \frac{\|d_{o1} - d_{o2}\|}{\|d_{o1} + d_{o2}\|} \qquad (2)$$

Where do_1 and do_2 are the syntactic vectors of sentences S_1 and S_2, respectively.

3) Overall sentence similarity measurement

Semantic similarity represents the lexical similarity. On the other hand, word order similarity provides information about the relationship between words. Based on the notion that both the semantic and syntactic information have an important role in understanding of a sentence, we calculate the sentence similarity using the composition of semantic similarity and syntactic similarity:

$$Sim_{(S_1, S_2)} = \lambda Sim_{sem} + (\lambda - 1) Sim_{wo} \qquad (3)$$

Where $0.5<\lambda<1$ is the weighting parameter, specifying the relative contributions to the overall similarity measure from the semantic and syntactic similarity measures.

C. Stage 3: summary similarity score

This stage displays the results from the system to the user. It shows the similarity measure as a score to the user. This result indicates that how many main ideas of the source text are covered by the summary text. We use the following equations to calculate the Final Score (FS) for any student's written summary:

$$FinalScore = \left(\frac{\sum_{S \in S_{summary}} MSS(S)}{N}\right) \times 100 \qquad (4)$$

Where $S_{summary}= (S_1,S_2,\ldots,S_N)$ includes all the sentences in the summary text, where N is the total number of sentences in the summary text. MSS is the Maximum Similarity Score between a sentence from the summary text and all the sentences from the source text.

V. Experiments

We conduct our analysis and assess the algorithm based on the student's written summary datasets. The performance of the algorithm is compared with other evaluation techniques, such as LSA, N-gram, BLEU and ROUGE. We use two various types of tests to compare the performance. The objectives of these tests are as follows:

- Similar test –to determine the ability of the proposed algorithm to provide a high similarity score for related summary text and source text.
- Dissimilar test – to determine the ability of the proposed algorithm to provide a low similarity score for unrelated summary text and source text.

Table 1 and Fig. 2 display a comparison of the algorithm and existing assessment techniques, such as LSA, N-gram, BLEU and ROUGE. The practical tests prove that the algorithm outperforms the other examined methods and that it is also more accurate than the other methods. The algorithm is able to obtain an accuracy of (93%) in comparison with the best existing method, (ROUGE), which has an accuracy of (89%).

Table 1. Performance comparison between Algorithm and other techniques.

	Various Tests			
	Similar		**Dissimilar**	
Method	AR_{sim}	Stdev	AR_{diss}	Stdev
Algorithm	0.9152	0.0748	0.9355	0.0340
LSA	0.9032	0.0876	0.3871	0.0798
BLEU N (1.. 4)	0.8065	0.1893	0.6129	0.0694
N-gram N (1.. 4)	0.7097	0.2170	0.7742	0.0830
ROUGE	0.871	0.1555	0.9251	0.0361

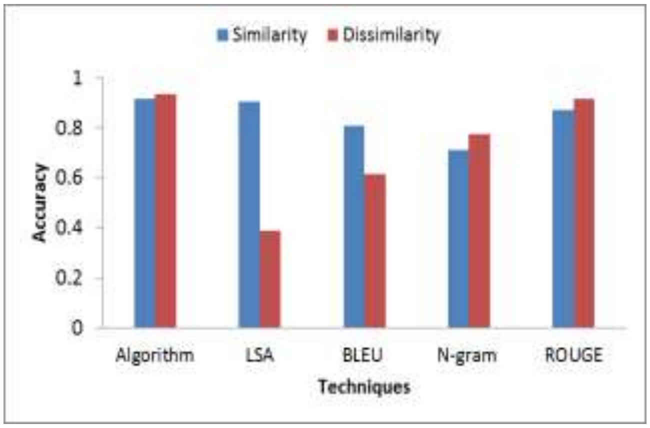

Figure 2. Comparison accuracy rate between the *Algorithm* and other techniques.

VI. Conclusion

This paper presented an algorithm for measuring the similarity between source text and summary text, based on syntactic feature and semantic information of words. The algorithm is able to obtain an accuracy of (93%) in comparison with the best existing technique, (ROUGE), which has an accuracy of (89%). Moreover, we implemented the algorithm into an automatic summarization assessment system to grade student written summaries in the English language

References

[1] Mani I, Maybury MT. Advances in automatic text summarization: MIT Press; 1999.

[2] Mani I. Automatic summarization: John Benjamins Publishing; 2001.

[3] Jones KS, Galliers JR. Evaluating natural language processing systems: An analysis and review: Springer; 1996.

[4] Donaway RL, Drummey KW, Mather LA. A comparison of rankings produced by summarization evaluation measures. Proceedings of the 2000 NAACL-ANLPWorkshop on Automatic summarization-Volume 4: Association for Computational Linguistics; 2000. p. 69-78.

[5] Radev DR, Hovy E, McKeown K. Introduction to the special issue on summarization. Computational linguistics 2002;28:399-408.

[6] Radev DR, Jing H, Styś M, Tam D. Centroid-based summarization of multiple documents. Information Processing & Management 2004;40:919-38.

[7] Saggion H, Lapalme G. Generating indicative-informative summaries with sumUM. Computational linguistics 2002;28:497-526.

[8] Papineni K, Roukos S, Ward T, Zhu W-J. BLEU: a method for automatic evaluation of machine translation. Proceedings of the 40th annual meeting on association for computational linguistics: Association for Computational Linguistics; 2002. p. 311-8.

[9] Pastra K, Saggion H. Colouring summaries BLEU. Proceedings of the EACL 2003 Workshop on Evaluation Initiatives in Natural Language Processing: are evaluation methods, metrics and resources reusable?: Association for Computational Linguistics; 2003. p. 35-42.

[10] Zipitria I, Elorriaga JA, Arruarte A, de Ilarraza AD. From human to automatic summary evaluation. Intelligent Tutoring Systems: Springer; 2004. p. 432-42.

[11] Landauer TK, Laham D, Rehder B, Schreiner ME. How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans. Proceedings of the 19th annual meeting of the Cognitive Science Society1997. p. 412-7.

12] Landauer TK, Foltz PW, Laham D. An introduction to latent semantic analysis. Discourse processes 1998;25:259-84.

[13] Franzke M, Streeter LA. 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.

[14] Lin C-Y. Rouge: A package for automatic evaluation of summaries. Text Summarization Branches Out: Proceedings of the ACL-04 Workshop2004. p. 74-81.

[15] Louis A, Nenkova A. Automatically evaluating content selection in summarization without human models. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1: Association for Computational Linguistics; 2009. p. 306-14.

[16] Lin J. Divergence measures based on the Shannon entropy. Information Theory, IEEE Transactions on 1991;37:145-51.

[17] He Y, Hui SC, Quan TT. Automatic summary assessment for intelligent tutoring systems. Computers & Education 2009;53:890-9.

[18] Pérez D, Alfonseca E, Rodrıguez P. Upper bounds of the BLEU algorithm applied to assessing student essays. Proceedings of the 30th international association for educational assessment (IAEA) conference2004.

[19] Lin C-Y, Hovy E. Automatic evaluation of summaries using n-gram co-occurrence statistics. Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1: Association for Computational Linguistics; 2003. p. 71-8.

[20] Landauer TK. On the computational basis of learning and cognition: Arguments from LSA. Psychology of learning and motivation 2002;41:43-84.

[21] Landauer TK, Dumais ST. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological review 1997;104:211.

[22] Choi FY, Wiemer-Hastings P, Moore J. Latent semantic analysis for text segmentation. In Proceedings of EMNLP: Citeseer; 2001.

[23] Lin C-Y, Cao G, Gao J, Nie J-Y. An information-theoretic approach to automatic evaluation of summaries. Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics: Association for Computational Linguistics; 2006. p. 463-70.

[24] Kullback S, Leibler RA. On information and sufficiency. The Annals of Mathematical Statistics 1951:79-86.

[25] Nenkova A. Summarization evaluation for text and speech: issues and approaches. INTERSPEECH2006.

[26] Li Y, McLean D, Bandar ZA, O'shea JD, Crockett K. Sentence similarity based on semantic nets and corpus statistics. Knowledge and Data Engineering, IEEE Transactions on 2006;18:1138-50.