Volume 4, Issue 1, January 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Survey of Various Summary Evaluation Techniques

Vishal Gupta

UIET, Panjab University Chandigarh,

India

Abstract— Evaluation of text summary is very essential issue of summarization. Normally, we can judge the relevance of any summary by applying extrinsic or intrinsic evaluation measures. Intrinsic evaluation techniques try to judge the relevance of any summary with evaluation by human experts. Extrinsic evaluation techniques judge quality of summary with task oriented techniques like task related to extraction of information. Usually we should must judge two attributes of any summary i) Compression Ratio (C.R.) which states up to which extent the summary is to be compressed as compared to source text and ii) Retention Ratio (RR) which states up to which extent we have to retain the information in the summary. RR is ratio of information present in summary to the information present in source text. At many places RR is also termed as omission ratio. This paper describes survey of different summary evaluation techniques including intrinsic & extrinsic evaluation techniques.

Keywords— Summary evaluation, intrinsic evaluation, extrinsic evaluation

I. INTRODUCTION

Evaluation of text summary [1] is very essential issue of summarization. Normally, we can judge the relevance of any summary by applying extrinsic or intrinsic evaluation measures. Intrinsic evaluation techniques try to judge the relevance of any summary with evaluation by human experts. Extrinsic evaluation techniques judge quality of summary with task oriented techniques like task related to extraction of information. Usually we should must judge two attributes of any summary i) Compression Ratio (C.R.) which states up to which extent the summary is to be compressed as compared to source text. C.R. is ratio of number of terms in summary to number of terms in source text. ii) Retention Ratio (RR) which states up to which extent we have to retain the information in the summary. RR is ratio of information present in summary to the information present in source text. At many places RR is also termed as omission ratio. This paper describes survey of different summary evaluation techniques including intrinsic & extrinsic evaluation techniques [2].

II. INTRINSIC TECHNIQUES OF SUMMARY EVALUATION

Intrinsic metrics of summary evaluation judge the relevance of text summary by matching it with reference summary (gold summary) created by human experts. The main stress in intrinsic summary evaluation is on the fact that up to which extent information is retained in summary and maintaining coherence among different sentences. Different factors involved in intrinsic evaluation of summary are discussed below:

A. Maintaining Coherence in Summary:

Text summaries created on the basis of text retrieval techniques i.e. cutting the text & pasting the text usually have problem that different extracted sections of summary can be out of context, leading to problem of coherence among sentences. The technique to judge it is to allow scoring and grading of summary by subjects for the purpose of coherence & after this match scores of summary lines with grades of gold summaries or with grades of input lines[1][2].

B. Maintaining Information in Summary:

Technique to judge information present in the created summary is to match it with source input text for the purpose of finding up to which extent actual information from input text is retained in the created summary. The second technique is to match created summary with gold summary already created by human experts and judging up to which extent actual information from gold summary is retained in the created summary. In single text documents we can apply recall & precision metrics to judge the quality and contents in created summary [1] [2].

C. Calculating Recall and Precision:

Recall for created summary can be calculated as number of lines present in gold summary which are also lying in created summary. In the same way we can also calculate precision. In information extraction, recall and precision are traditional metrics & these two can be clubbed together in single metric named F-measure. Difficulty in these metrics is that these are unable to find differences in various better summaries & those summaries which are very much different in content wise but can obtain identical scores using these measures[1][2].

D. Ranking of Lines

Ranking of lines is better technique than recall, precision & F-measure, in which gold summary is generated by scoring the lines present in input text considering worthiness of lines for including in final summary. Then we can apply correlation metrics for match of created summary to gold summary [1] [2].

© 2014, IJARCSSE All Rights Reserved

E. Technique of Utility

It was discussed by Radev et al. (2000) [3] and permitted gold summaries to contain retrieved units of text like: lines and passages etc. including membership of fuzzy in gold summary. In this method, gold summary consists of all lines of input text document including their values of confidence to include them in final summary. Moreover, these techniques might be broadened for permitting retrieved text units for having -ve support on each another. It is very much helpful while judging summaries of multiple documents, in which one line causing other line redundant it is able to automatically calculate scores of evaluation. For example any method which retrieves two or more than two similar lines will be penalized greater than the method which retrieves only single line among the before-said lines. This method is normally very much essential for judging retrieval oriented summaries. But much of recent research has talked about improved version of utility method called as relative utility method [4].

F. Similarity of Contents

Technique related to similarity of contents [5] are used for evaluation of contents which are semantic in nature for abstractive summary and retrieval oriented summaries. Technique of this type of summary evaluation is test of vocabulary in which information extraction oriented techniques are applied for matching of frequency of words with vectors of frequency obtained for extraction oriented or abstractive summaries & gold summaries in a particular manner. Sets related to synonyms generated with LSA (i.e. Latent Semantic Analysis) [6] or RI (i.e. Random Indexing) [7] may be applied to shorten number of words in term-vectors using mixing of word-frequencies and hence permitting for more difference in text summaries. It is very much helpful for evaluation of abstractive summaries. Shortcomings of techniques is that these are very much sensitive for term order dissimilarities & negation. With random indexing and Latent semantic we should be familiar with fact that these techniques usually will not create synonym sets which will be true and sets may also have hyponyms, antonyms and words which will occur in same contexts. These techniques are very much useful in retrieval oriented text summaries in which vary less rearrangement among input sections of text is done.

G. Similarity Measure of Summary using Cosine similarity

Text documents in cosine similarity [8] are indicated in terms of word vectors. Similarity among any two text documents relates to correlation among these vectors. Suppose we are given with two text documents with vectors *D* and *E* are word frequency vectors of text documents on word set S= $\{s_1,\ldots\ldots, s_k\}$ The cosine similarity among any two vectors is determined as:

$$\text{VALUE\_OF\_COSINE\_SIMILARITY } (D, E) = (D \cdot E) / (|D| |E|)$$
$$= \sum D_c \times E_c / \sqrt{\sum (D_c)^2} \times \sqrt{\sum (E_c)^2} \text{ where } c= 1 \text{ to } n$$

Here every dimension denotes a word along with frequency of that word in text document and it is always non negative. Cosine similarity is also positive & lying from 0 to 01. If value of cosine similarity is near to 01, then it means that two text documents are almost identical with each other. Those documents which are not identical with each other the value of cosine similarity will be closer to 0 for them.

H. Similarity Measure of Summary using Jaccard Coefficient

Jaccard coefficient [8] finds similarity by taking ratio of intersection of two objects to union of them. Suppose we are given with two text documents with word frequency vectors *D* and *E* for these text documents on word set S= $\{s_1,\ldots\ldots, s_k\}$ then we can find value of Jaccard Coefficient as:

$$\text{Value \_of\_Jaccard Coefficient} = \text{SIMILARITY-VALUE } (D, E) = (D \cdot E)/ (|D|^2 + |E|^2 - D \cdot E)$$
$$= (D \cdot E) / (\sqrt{\sum (D_c)^2} \times \sqrt{\sum (D_c)^2} + \sqrt{\sum (E_c)^2} \times \sqrt{\sum (E_c)^2} - D \cdot E) \quad \text{Where } c= 1 \text{ to } n$$

Here every dimension denotes aword along with frequency of that word in input document. Value of Jaccard Coefficient lies from 0 to 1. If its value is near to 01 then two input documents are very similar with each other. If its is moving near 0 then we can say that these two text documents are not similar with each other.

I. Similarity Measure of Summary using Euclidean Distance

Euclidean distance [8] is normally used as measure in case of geometrical field. Here it is used for calculating distance among system generated summary with gold summary produced by human experts. Suppose we are given with 02 text documents along their frequency vectors of key words E_{id} and E_{jd} respectively, where d= 1 to n keywords.

Euclidean distance between 02 text documents is calculated as:

$$\text{Value\_of\_Euclidean distance}(E_{id}, E_{jd}) = (\sum (E_{id} - E_{jd})^2)^{1/2} \quad \text{for d=1 to n keywords.}$$

J. Similarity Evaluation by ROUGE

ROUGE is one of metric used in judging quality in any summary. Meaning of ROUGE [12] is recall-oriented understudy for gisting evaluation. It uses different type of measures for automatically determining quality & correctness in summary by matching sentences in summary with sentences of reference summary created by human experts. It calculates overlapping text units frequency i.e. n-grams, different sequences of words & combination of words present in system developed summary & reference summaries created by human experts.

© 2014, IJARCSSE All Rights Reserved

III. EXTRINSIC TECHNIQUES OF SUMMARY EVALUATION

Evaluating text summaries using extrinsic evaluation techniques judge acceptability & accuracy of created summaries by performing tasks on the summary like: judging relevance or comprehension reading. Moreover, if there are some type of instructions in summary, it is feasible to judge up to which level it to adopt those instructions. & result also. Another type of feasible tasks are collection of information from a bulk set of documents, the time and labour actually needed for changing system created summary for a particular motive, or effect of summarizer on whole system of whom this summarizer is part. For example query expansion in a question answering system or in a search engine. particular game type scenes are suggested as base techniques for evaluation of summarization motivated by diverse fields, among them are: game proposed by Shannon in information theory, game of question, game of classification & clustering and association of key terms in information retrieval.

A. Game proposed by Shannon:

Game proposed by Shannon is different form of metric proposed by Shannon in information theory [9]. It is an effort for information contents quantification by predicting next word, i.e. term or letter and hence regenerating source input text. This concept is taken from metrics of Shannon of information theory in which we can tell 03 groups for regenerating relevant paragraphs of input article. We can calculate retention of information in keystrokes frequency it uses for regenerating source text. Work of Shannon is useful for the human being which is involved in predicting & so conditioned implicitly on knowledge of reader.

B. Game of Questions:

Motive of game of questions is to judge understanding of readers for summary & ability of summary for conveying key points of input text. It is done in two sub phases: i) Testers read input articles and creating main paragraphs as they recognize them. Then testers generate questions which belong to particular statements which are factual in main paragraphs. After this, experts give answer of queries three number of times without looking any type of text after looking at system created summary, & after looking at input source text. That summary which will successfully convey key points belonging to input text should give answers of most of queries.

C. Game of Classification:

In game of classification we can try to match classifiability by telling experts to categorize either input text documents or text summaries into any one of n classes. Then equal-valance of categorization of text summaries to inputs is judged. Summary should be categorized to same class as of its input text document. SUMMAC has applied two types of this test [10].

D. Association of Keywords

Association of Keywords is very inexpensive. This approach fully depends on associated keywords automatically or manually to text documents which have been summarized. Saggion et al. (2000) [11] discussed judges having text summaries created by their corresponding summarizers along with 05 keywords lists obtained from input source text. Then judges were assigned job for associating each type of summary with proper list of key terms. If this task is successful then we can say that summary is covering all central themes of input text because associated key terms with input text are content indicative. The main benefit of this approach is that it does not demands bulky annotation manually.

IV. CONCLUSIONS

Thus we can conclude that there are two types of measures for automatically evaluating the summary generated by any summarization system which are intrinsic and extrinsic evaluation metrics. Intrinsic metrics try to judge the quality of summary by evaluation of summary by human judges. Extrinsic measures try to evaluate the quality of summary by performing some of tasks on the summary. We should apply both types of measures (intrinsic and extrinsic) to judge the overall quality of summary.

REFERENCES

[1] M. Hassel, "Evaluation of Automatic Text Summarization," *Licentiate Thesis*, Stockholm, Sweden, 1-75, 2004.

[2] K. Spark-Jones and J. R. Galliers, "Evaluating Natural Language Processing Systems: An Analysis and Review," *Lecture Notes in Artificial Intelligence*, Springer, Berlin, Germany, 1995.

[3] D. R. Radev, H. Jing, and M. Budzikowska, "Centroid-Based Summarization of Multiple Documents: Sentence Extraction, Utility-Based Evaluation, and User Studies," *In Proceedings of the Workshop on Automatic Summarization at the 6th Applied Natural Language Processing Conference and the 1st Conference of the North American Chapter f the Association for Computational Linguistics*, Seattle, WA, 2000.

[4] D.R. Radev, D. Tam and Gunes Erkan, "Single-Document and Multi-Document Summary Evaluation via Relative Utility," *Proceedings of the ACM CIKM Conference*, New Orleans, LA., 2003.

[5] R. L. Donaway, K. W. Drummey and L. A. Mather, "A Comparison of Rankings Produced by Summarization Evaluation Measures," *Workshop on Automatic Summarization at the 6th Applied Natural Language Processing Conference and the 1st Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 69-78, 2000.

[6] T. K. Landauer, P. W. Foltz, and D. Laham, "Introduction to Latent Semantic Analysis," *Discourse Processes*, vol. 25, pp. 259-284, 1998.

[7] P. Kanerva, J. Kristoferson, and A. Holst, "Random Indexing of text samples for Latent Semantic Analysis," *22nd Annual Conference of the Cognitive Science Society*, Pennsylvania, 2000.

© 2014, IJARCSSE All Rights Reserved

[8] A. Huang, "Similarity Measures for Text Document Clustering," In the Proceedings of New Zealand Computer Science Research Conference, Christchurch New Zealand, pp. 49-56, 2008.
[9] C.E. Shannon, "A mathematical theory of communication," The Bell System Technical Journal, pp. 623-656, 1948.
[10] I. Mani, D. House, G. Klein, L. Hirshman, L. Orbst, T. Firmin, M. Chrzanowski, and B. Sundheim, "SUMMAC: A Text Summarization Evaluation," International Journal of Natural Language Engineering, Cambridge University Press, vol. 8, pp. 43-68, 2002.
[11] H. Saggion and G. Lapalme, "Concept Identification and Presentation in the Context of Technical Text Summarization," *Proceedings of the Workshop on Automatic Summarization at the 6th Applied Natural Language Processing Conference and the 1st Conference of the North American Chapter of the Association for Computational Linguistics*, Seattle, WA, USA, 2000.
[12] C.W. Lin, "ROUGE: Package for Automatic Evaluation of Summaries," *Workshop on Text Summarization Branches Out WAS* , Spain, 2004.

© 2014, IJARCSSE All Rights Reserved