REVIEW ON TEXT SUMMARIZATION EVALUATION METHODS

Saiyed Saziyabegum, Ph.D. Scholar,
Sardar Patel University,
Vallabh vidhyanagar, Gujarat, India
saziya.saiyed2013@gmail.com

Dr. Priti S. Sajja, Professor,
G.H.Patel PG Department of computer Science, Sardar Patel University,
Vallabh vidhyanagar, Gujarat, India
priti@pritisajja.info
http://www.pritisajja.info/

Abstract There is ample amount of information available on internet. Important information can be gained by creating summary from available information. Manual creation of summary is difficult task. Hence research community is developing new approaches for automatic text summarization that creates summary automatically. Summary is shorter text that covers important information from original text. This paper discusses basics of automatic text summarization. To evaluate automatic summaries are also challenging task. The challenges in evaluating summaries are also described. Methods for evaluation of summary- Both intrinsic and extrinsic are described in detail. The paper concludes with some suggestions for future directions for summary evaluation.

***Keywords*:** Text Summarization; Summary Evaluation; Intrinsic evaluation; extrinsic evaluation;

1. Introduction to Automatic text summarization and summary Evaluation

In today's fast rising world of information, text summarization [Jezek and Steinberger (2008)] is very important tool for understanding text information. There is lot of text materials accessible on the internet which provides information beyond requirements and creates the situation called "infobesity". To select information from huge amount of information from different sources is difficult for human beings. Due to the volume of information manual summarization of information is really challenging, complicated and difficult task. The aim of automatic text summarization is to reduce the source text into a compact version which will preserve contents and general meaning. Advantage of Summary is that it minimizes reading time and efforts. Generating automatic text summaries is not enough; evaluation of those summaries is also an important task.

There are number of severe challenges in evaluating summaries [Mani (2001)], which make summarization evaluation a very interesting problem:

1. There is the chance of a system generating a good summary that is moderately different from summary generated by human.
2. Human evaluated the summary is expensive way of summarization. Instead of human evaluation, scoring program is preferable, that can be repeatable.
3. The scale and complexity of evaluation increases due to different compression rate of summary.

Text summarization Evaluation methods can be broadly classified into two categories [Lin (2004)] extrinsic evaluation and intrinsic evaluation.

2. Text Summarization Evaluation methods

2.1. *Extrinsic evaluation*

It checks the summarization based on how it influences the completion of some other task such as text classification, information retrieval, answering of question etc. It evaluates the impact of summarization on tasks like reading comprehension, relevance assessment etc. Therefore a summary is considered as good if it is helpful to some other tasks.

(1) Reading comprehension: This method determines whether it is capable to answer multiple choice test after comprehension of the summary.

(2) Relevance assessment: Variety of methods are used for evaluating a relevance of subject present in the summary or the original document.

2.2. *Intrinsic evaluation*

It checks the summarization system in of itself. It determines the summary quality on the basis of comparison between the automatically generated summary and the human made manual summary. Summary is evaluated on

basis of two aspects Quality or informativeness. The informativeness of a summary is evaluated by comparing it with a human-made summary, i.e., reference/ ideal summary. There is another model called fidelity to the source which checks whether the summary consists of the same or similar content as present in the original document.

(1) Informativeness evaluation

i. Some of the methods for informativeness evaluation are as follows.

- **Relative utility:** In this method [Radev and Tam (2003)] score is assigned by adjudicators whose range is between 0 and 10 to each sentence in the input document according to its importance. The sentences with high score are considered suitable for the summary.
- **Text grammars:** This method [Branny (2007)] helps to evaluate text summaries. It addresses a text as a complex structure, and elements of the structure are interconnected both on the level of form and meaning, and the well-form aspect of which should be described on both of these levels.
- **Factoid Score :** In this method [Teufel and Van Halteren (2004)] evaluation of automatic summaries is done based on factoids. Factoids are atomic units of information that can be used to convey the meaning of a sentence. Different reference summaries are used as gold standards and common information is measured between them.
- **Basic Elements:** In this method [Hovy et al (2005)] sentence is divided into small unit of content, which is known as Basic Element, those are expressed as words' triplets. Each triplet containing a head, modifier or argument along with the relationship of modifier to the head. Different similar expressions are matched with more flexibility using this method.
- **Pyramid Method:** it looks for information with same meaning across different human-made summaries, which are known as Summary Content Units (SCU). A weight is assigned to each SCU related to the number of human assessments which recognize the same content. These weights have a distribution such that it distinguishes more related information from less related information.
- **AutoSummENG (Automatic Summary Evaluation based on N-gram Graphs):** This is language independent automatic method [Giannakopoulos (2008)] which has high relationship with human decision. This method differs from the others in three main aspects: (1) the type of statistical information extracted; (2) the representation chosen for this extracted information, and (3) the method used to calculate the similarity between summaries. For comparing the summaries, First n-gram character graphs are built and then their representations are compared to find various type of similarity among the graphs.
- **QARLA**: This is summary evaluation framework [Amigó et al (2005)]. On passing some automatic and reference summaries and some similarity metrics, this approach provides some measures like QUEEN, KING and JACK. QUEEN calculates the quality of a machine-generated summary. KING calculates a similarity metric's quality and JACK is used for estimating the reliability of machine-generated summaries. This framework uses different similarity metrics like precision, recall, frequency and sentence length and metrics for grammatical distribution.
- **ParaEval**: This method [Zhou et al (2006)] is used for identifying paraphrase matching. Process is as follows. 1. First paraphrases consisting of multiple words are searched between phrases in the reference and automatic summaries. 2. This method tries to look for synonyms between single words for those unmatched fragments. 3. Finally, if no synonym is found between single words, then simple lexical matching is done.
- **DEPEVAL (summ):** It is a dependency-based metric [Owczarzak (2009)]. It has a concept similar to Basic Elements (BE) but parsers are used in this method. Dependency triples are selected from automatic summaries and reference summaries and then they compared with one another.
- **GEMS (Generative Modelling for Evaluation of Summaries)**: This method [Katragadda (2010)] suggests the use of signature for analyzing that how they are captured in automatic summaries. The signature terms are calculated on the basis of part-of-speech tags, such as nouns or verbs; query terms and terms of reference summaries. The distribution of the signature terms is calculated in the source document and then the possibility of a summary being biased towards such signature-terms is gained.

ii. Some of the metrics for informativeness evaluation are as follows.

- **ROUGE**: A set of metrics called recall oriented understudy of gisting Evaluation (ROUGE) was introduced [Lin (2004)] gives a score based on the similarity in the sequences of words between a human-written model summary and the machine summary. Thus, it helps to automatically evaluate the summary. ROUGE includes five measures like ROUGE-N, ROUGE-L, ROUGE-W, ROUGE-S and ROUGE-SU.

i. ROUGE-N: measures the N-gram units common between a particular summary and a collection of reference summaries where N determines the N-gram's length. E.g., ROUGE-1for unigrams and ROUGE-2 for bi-grams.

ii. ROUGE-L: It computes Longest Common Subsequence (LCS) metric. LCS is the maximum size of common subsequence for two given sequences. It calculates ratio between size of two summaries' LCS and size of reference summary.

iii. ROUGE-W: It is the weighted longest common subsequence metric. It's the improvement over the simple LCS approach. ROUGE-W prefers LCS with consecutive common units. It can be computed efficiently using dynamic programming.

iv. ROUGE-S (Skip-Bigram co-occurrence statistics) :It evaluates the amount of skip bigrams common between a particular summary and a collection of reference summaries. Skip bigrams are any word pair in the sentence orders with random.

v. ROUGE-SU: it is the weighted average between ROUGE-S and ROUGE-1 and it extends ROUGE-S with counting unit as unigram. Actually this is an improvement over ROUGE-S.

- **Other popular metrics**: For intrinsically evaluating the summary, other popular metrics are precision, recall and F-measure [Steinberger and Ježek (2012)]. They are required to predict coverage between human-made ideal summary and automatically generated machine-made summaries. With the help of above metrics, it is also feasible that two summaries generate different evaluation results even being identically good. These metrics are explained below:

i. Precision: It determines what fraction of the sentences chosen by the humans and selected by the system are correct. Precision is the number of sentences found in both system and ideal summaries divided by the number of sentences in the system summary.

ii. Recall: It determines what proportion of the sentences chosen by humans is even recognized by the machine. Recall is the number of sentences found in both system and ideal summaries divided by the number of sentences in the ideal summary.

iii. F-measure: It is computed by combining recall and precision.

(2) Quality evaluation

In Quality evaluation linguistic aspects of the summary are considered. In the conferences of DUC and TAC, following factors related to linguistic quality are used for evaluating summaries.

i. Redundancy :

The text should not contain redundant information.

ii. Grammaticality:

The text should not contain non-textual items (i.e., markers) or punctuation errors or incorrect words.

iii. Referential clarity:

The nouns and pronouns should be clearly referred to in the summary. For example, the pronoun 'she' has to mean that it is referring somebody in the context of the summary.

iv. Structure and Coherence:

The summary should have good structure and the sentences should be coherent.

These do not need to be compared against the ideal summary. Expert human evaluator evaluates the summary manually by assigning a score to the summary corresponding to five-point scale on the basis of its quality.

Text quality of summary can also be assessed by analyzing different factors for readability[Pitler and Nenkova (2008)]. Text quality is analyzed through different criterions like vocabulary, syntax or discourses so that correlation can be estimated between these factors and already obtained human readability ratings. Vocabulary is expressed by unigrams and syntax by features like average number of verb-phrases or noun-phrases. Other text quality evaluation paradigms are local coherence [Barzilay and Lapata (2008)], centering theory [Grosz (1995)] and syntactic and semantic models and grammaticality of a grammar [Vadlapudi and Katragadda (2010)].

3. Conclusion

Summary evaluation process is a big challenge. This paper discussed both the types of evaluation methods, intrinsic and extrinsic. Most of the evaluation is intrinsic in nature which is further classified into informativeness and quality evaluation for which recent methods and tools are used. Majority of the recent tools assess the information present in the summary and only some methods try to evaluate the summary quality. Research can be carried out in intrinsic evaluation, thus develop new methods to evaluate the summary on the basis of information it contains and presentation of that information. Some good criteria should be defined for evaluation of summaries.

References

[1] Amigó, E., Gonzalo, J., Penas, A., & Verdejo, F. (2005, June). QARLA: a framework for the evaluation of text summarization systems. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (pp. 280-289). Association for Computational Linguistics

[2] Barzilay, R., & Lapata, M. (2008). Modeling local coherence: An entity-based approach. Computational Linguistics, 34(1), 1-34.

[3] Branny, E. (2007). Automatic summary evaluation based on text grammars. Journal of Digital Information, 8(3).

[4] Giannakopoulos, G., Karkaletsis, V., Vouros, G., & Stamatopoulos, P. (2008). Summarization system evaluation revisited: N-gram graphs. ACM Transactions on Speech and Language Processing (TSLP), 5(3), 5

[5] Grosz, B. J., Weinstein, S., & Joshi, A. K. (1995). Centering: A framework for modeling the local coherence of discourse. Computational linguistics, 21(2), 203-225

[6] Hovy, E., Lin, C. Y., & Zhou, L. (2005, October). Evaluating duc 2005 using basic elements. In Proceedings of DUC (Vol. 2005)

[7] Jezek, K., & Steinberger, J. (2008). Automatic text summarization. InZnalosti (pp. 1-12).

[8] Katragadda, R. (2010, March). GEMS: generative modeling for evaluation of summaries. In International Conference on Intelligent Text Processing and Computational Linguistics (pp. 724-735). Springer Berlin Heidelberg.

[9] Lin, C. Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In Text summarization branches out: Proceedings of the ACL-04 workshop (Vol. 8).

[10] Mani, I. (2001). Summarization evaluation: An overview.

[11] Nenkova, A., & Passonneau, R. J. (2004, May). Evaluating Content Selection in Summarization: The Pyramid Method. In HLT-NAACL (Vol. 4, pp. 145-152).

[12] Owczarzak, K. (2009, August). Depeval (summ): dependency-based evaluation for automatic summaries. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1 (pp. 190-198). Association for Computational Linguistics.

[13] Pitler, E., & Nenkova, A. (2008, October). Revisiting readability: A unified framework for predicting text quality. In Proceedings of the conference on empirical methods in natural language processing (pp. 186-195). Association for Computational Linguistics.

[14] Radev, D. R., & Tam, D. (2003, November). Summarization evaluation using relative utility. In Proceedings of the twelfth international conference on Information and knowledge management (pp. 508-511). ACM.

[15] Steinberger, J., & Ježek, K. (2012). Evaluation measures for text summarization. Computing and Informatics, 28(2), 251-75.

[16] Teufel, S., & Van Halteren, H. (2004). Evaluating Information Content by Factoid Analysis: Human annotation and stability. In EMNLP (pp. 419-426).

[17] Vadlapudi, R., & Katragadda, R. (2010, June). On automated e14valuation of readability of summaries: Capturing grammaticality, focus, structure and coherence. In Proceedings of the NAACL HLT 2010 student research workshop (pp. 7-12). Association for Computational Linguistics.

[18] Zhou, L., Lin, C. Y., Munteanu, D. S., & Hovy, E. (2006, June). Paraeval: Using paraphrases to evaluate summaries automatically. In Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (pp. 447-454). Association for Computational Linguistics.