International Conference on Research Advances in Integrated Navigation Systems (RAINS - 2016),
April 06-07, 2016, R. L. Jalappa Institute of Technology, Doddaballapur, Bangalore, India

Review on text summarization evaluation methods

M Indu
Mtech Student
Department of Computer Science and Engineering
SCT College of Engineering
Trivandrum, India
indu22.emp@gmail.com

Kavitha K V
Assistant Professor
Department of Computer Science and Engineering
SCT College of Engineering
Trivandrum, India
kavitha279@yahoo.co.in

Abstract— **To familiarize oneself with a subject area summaries play an important role. Text Summarization is a challenging problem these days. Summarization is very interesting and useful task that gives support to many other tasks as well as it takes advantage of techniques developed for related Natural Language Processing tasks. Evaluating summaries and automatic text summarization systems are not a straightforward process. This review paper discusses an overview of text summarization, various evaluation approaches on intrinsic and extrinsic techniques. In principle, text summarization is achieved because of the naturally occurring redundancy in text and because important (salient) information is spread irregularly in textual documents. Recognizing the redundancy is a challenge that hasn't been fully resolved yet.**

Keywords— summarization; extraction-based summarization; abstraction-based summarization; intrinsic evaluation; extrinsic evaluation.

I. INTRODUCTION

In the present age of Internet and due to rapid growth of broadcast systems, there is massive amount of information being available online. Search Engines employs the method of constant indexing in order to accumulate the growing information in World Wide Web. Once the user enters the search request, documents are retrieved. The classic problem of Information Overload comes into play as the search engine retrieves hundreds of documents as search results.

The retrieval time for the search is very less, the user has to go through the documents in order to attain at document he/she is searching for, because most of the naive users are reluctant to make cumbersome effort of going through each of the documents. With these enormous amount of information available and need for summarization not only for saving the search time but also for having a cut short understanding of information available. Summarization has been interest of study in the field of Computer Science for so long. With the growing large data sets, research on Automatic Text Summarization [1] has become the study of hour. Although the attempts to generate automatic summaries began 50 years ago [2], in recent automatic Text Summarization has experienced an exponential growth [3], [4], [5] due to these new technologies.

Automatic text summarization is a technique that gets a source text and presents the most relevant content in a condensed form as the user or task needs. Technologies that can make a logical summary are taken such as length, writing style and syntax. Traditionally, the process of automatic text summarization has been decomposed into three main stages [6], [2], [7]. The Spark Jones [7] approach, which is: the source text is interpreted to obtain a text representation, then transforming the text representation into a summary representation, and from summary representation summary of text is generated.

Effective summarizing requires an explicit and detailed analysis of context factors. In [7] three classes of context factors are distinguished: input, purpose and output factors. Input factors define the features of the text to be summarized crucially determine the way a summary can be obtained. This fall into three groups: text form; subject type and unit. Purpose factors are the most important factor which fall under three categories: refers the context within the summary to be used; audience and use. Output factors group: material (i.e. content) ; format and style.

II. OBJECTIVES OF SUMMARIZATION

Today summarization technologies are used in large number of sectors, for example in search engines (Google), document summarization, image collection summarization and video summarization. By finding the most informative sentences document summarization automatically create a representative summary or abstract of the entire document. Similarly, in image summarization the system finds the most representative and important images. Likewise, in consumer videos one would want to remove the boring or repetitive scenes, and extract out a much shorter and abstract version of the video. In surveillance videos, extraction of important events in the recorded video is only considered, since most part of the video may be uninteresting with nothing going on. The problem of information overload grows, and the amount of data increases, the interest in automatic summarization is also increasing.

III. TYPES OF SUMMARIZATION

There are two approaches to automatic summarization: extraction and abstraction. To form the summary extractive methods [8],[9] work by selecting a subset of existing words, phrases, or sentences in the original text. Abstractive methods [10],[11] build an internal semantic representation and natural language generation techniques is used to create a summary

978-1-4673-8819-8/16/$31.00 ©2016 IEEE

that is closer to what a human may generate. Such a summary could contain words not explicitly present in the original.

A. *Extraction-based summarization*

Here, without modifying the objects themselves the automatic system extracts objects from the entire collection. Examples include key phrase extraction, where the aim is to select individual words or phrases to "tag" a document. The goal for document summarization is to select whole sentences (without modifying them) for creating a short paragraph summary. Similarly, from the collection system extracts images without modifying itself, is image collection summarization.

B. *Abstraction-based summarization*

Extraction techniques copy the information most important by the system to the summary (ex: key clauses, sentences or paragraphs), while abstraction includes paraphrasing sections of source document. In general, abstraction can shorten a text more strongly than extraction, but the programs that can do is harder to develop as they require the usage of natural language generation technology, which itself is a growing field.

In abstractive summarization an abstract synopsis like that of a human is done, while majority of summarization systems are extractive where selection of subset of sentences to place in a summary.

IV. Review Of Various Intrinsic And Extrinsic Evaluation Technique

A common way to evaluate the informativeness of automatic summaries is to compare them with human-made model summaries. First broad division for evaluating automatic text summarization systems is intrinsic and extrinsic evaluation methods [12]. An intrinsic evaluation tests the summarization system itself while an extrinsic evaluation tests the summarization based on how completion of some other task is affected. Intrinsic evaluations have assessed mainly the coherence and informativeness of summaries. Extrinsic evaluations, tested the impact of summarization on tasks like relevance assessment, reading comprehension, etc.

A. Intrinsic Evaluation

It measures the system in of itself, which is done by comparing to some old standard, (made by a reference summarization system or man-made using informants). Intrinsic evaluation mainly focuses on coherence and informativeness of summaries.

- Utility Method

The utility method (UM) [13] allows reference summaries which consist of extraction units (sentences, paragraphs etc.) along with fuzzy membership in reference summary. The reference summary contains all the sentences of the source document(s) with confidence values for the inclusion in the summary. Furthermore, this method can be expanded to allow extraction units for exerting negative support on one another. This is predominantly useful when evaluating multi-document summaries. In case of sentence making another redundant data can automatically penalize the evaluation score. A system extracting two or more “equivalent” sentences gets penalized more than a system that extracting only one of the fore mentioned sentences. This method makes similarities to the Majority Vote method [14] in that it, in contrast to P&R and Percent Agreement, allowing summaries to be evaluated at different compression rates. For extraction based summaries UM is most useful. Recent evaluation experiments led to the development of the Relative Utility metric [15].

- Content Similarity

Here both extraction based summaries and true abstracts [16] can be applied to evaluate the semantic. One such is the Vocabulary Test (VT) where standard Information Retrieval methods [17] are used to compare term frequency vectors calculated over stemmed or lemmatized summaries and reference summaries of some type. Controlled thesauri and “synonym sets” created with Latent Semantic Analysis [18] or Random Indexing [19], [20] can be used to reduce the terms in the vectors by combining the frequencies of terms synonymous, thus allowing for greater variation among summaries. This is especially useful when evaluating abstracts. Disadvantage of these methods are; quite sensitive to negation and word order differences. With LSA5 or RI6 one must also be aware of the fact that these methods do not necessarily make true synonym sets, these sets typically include antonyms, hyponyms and other terms that occur in similar semantic contexts. These methods are useful for extraction based summaries where little rewriting of the source fragments is done along with comparing fragmentary summaries, such as key phrase summaries.

- BLEU Scores

The idea here is that, as well as there may be many “perfect” translations of a given source sentence, there may be several equally good summaries for a single source document. These summaries may vary in word or sentence choice, or in word or sentence order even when they use the same words/sentences. Still humans can clearly discriminate a good summary from a bad one. The recent adoption of BLEU/NIST7 scores [21], [22] by MT community for automatic evaluation of Machine Translation, [23] have applied the same idea to the evaluation of summaries. There an automatically computed accumulative n-gram matching scores (NAMS) between ideal summaries and system summaries as a performance indicator. Only content words were used in forming n-grams and n-gram matches between the summaries being compared where treated as position independent. For comparison, IBM’s BLEU evaluation script was also applied to the same summary set. However, this showed that direct application of the BLEU evaluation procedure does not always give good results.

B. Extrinsic Evaluation

Extrinsic evaluation on the other hand measures the efficiency and acceptability of the generated summaries in some task. If the summary contains some type of instructions, it is possible to measure at what extent it is possible to follow the instructions and the result. Other measurable tasks are information gathering in a large document collection. The effort and time required to post-edit machine generated summary for some specific purpose, or the summarization system's impact on a system of which it is part. Example: relevance feedback in a search engine or a question answering system. Proposed several game like scenarios at surface methods for summarization evaluation inspired by different disciplines. Among this include The Shannon Game (information theory), The Question Game (task performance), The Classification/ Categorization Game and Keyword Association (information retrieval).

- Shannon Game

A variant of Shannon's measures in Information Theory is attempting to quantify information content by guessing the next token, e.g. letter or word, recreation of original text. The idea has been retrieved from Shannon's measures in Information Theory where three groups of informants to reconstruct important passages from the source article having seen either the full text, a generated summary, or no text at all. The information retention is measured in a number of keystrokes it takes to recreate the original passage. Hovy [24] has shown that there is a magnitude of difference across the three levels (about factor 10 between each group). The problem in Shannon's work is relative to the person doing the guessing and it is therefore implicitly conditioned on the reader's knowledge. Thus information measure will infallibly change with more knowledge of the language, the domain, etc.

- The Question Game

The purpose here is to test the readers' understanding of the summary and an ability to convey key facts of the source article. This evaluation task is carried out in two steps. First the source articles are read by testers, marking central passages as they identify them. The testers then create questions corresponding to certain factual statements in the central passages. Next, assessors answer the questions 3 times: without seeing any document (baseline 1), after seeing a system generated summary, and after seeing the original document (baseline 2). A summary successfully convey the key facts of the source article. It should be able to answer most questions, i.e. being closer to baseline 2 than baseline 1. This evaluation scheme has been used in the TIPSTER SUMMAC text summarization evaluation Q&A8 task, where [25] found an informativeness ratio of accuracy to compression of about 1.5.

- Keyword Association

An inexpensive, shallower approach. It relies on keywords associated to the documents being summarized. For example [26] presented a human judges with summaries generated by their summarization system together with five lists of keywords taken from the source article as presented in the publication journal. The judges were given the task to associate each summary with the correct list of keywords. If successful, the summary is said to cover the central aspects of the article since the keywords associated to the article by the publisher were content indicative. Major advantage is that it requires no cumbersome manual annotation.

V. Comparison Of Various Evaluation Techniques

A study of various evaluation techniques in the field of text summarization is done and a comparison table is illustrated regarding the review.

TABLE I. Comparison of intrinsic evaluation techniques

Method	Methodology	Advantage
Utility Method	Fuzzy membership in the reference summary	Includes all sentences of source document(s)
Content Similarity	Evaluate the semantic content in both extraction based summaries and true abstracts.	Useful for evaluating abstracts
BLEU Scores	MT community for automatic evaluation of Machine Translation	Automatically computed accumulative n-gram matching scores (NAMS)

TABLE II. Comparison of extrinsic evaluation techniques

Method	Methodology	Advantage
Shannon Game	Quantify information content by guessing the next token	Reconstruct important passages from the source article
The Question Game	Test the readers' understanding of the summary and its ability	Using power nodes and LRW strategy
Keyword Association	Inexpensive, shallower approach	Cover the central aspects of the article

VI. Conclusion

A general overview of automatic text summarization has been reviewed in this paper while focusing on various intrinsic and extrinsic evaluation techniques. The status, and state, of automatic summarizing has radically changed through the years. The biggest challenge for text summarization is to summarize content from a number of textual and semi structured sources, including databases and web pages, in the right way (language, format, size, time) for a specific user. Text summarization software should produce an effective summary in less time and with least redundancy. Summaries can be evaluated using intrinsic or extrinsic measures. While intrinsic methods attempt to measure summary quality using human evaluation, extrinsic methods measure the same through a task based performance measure such as information retrieval oriented task. Research on this field will continue due to the fact that text summarization task has not been finished yet and there is still much effort to do, to investigate and to improve.

References

[1] Karel Jezek and Josef Steinberger, "Automatic Text summarization", Vaclav Snasel (Ed.): Znalosti 2008, pp.1-12, ISBN 978-80-227-2827-0, FIIT STU Brarislava, Ustav Informatiky a softveroveho inzinierstva, 2008.

[2] Mani, I., Maybury, M. T., Ed. Advances in Automatic Text Summarization. The MIT Press, 1999.

[3] Hovy, E., Lin, C. Y., Zhou, L., et al. Automated Summarization Evaluation with Basic Elements. In Proceedings of the 5th International Confer- ence on Language Resources and Evaluation (LREC). Genoa, Italy, 2006.

[4] Jackson, P., Moulinier, I. Natural language processing for online applications. John Benjamins Publishing Company, 2002.

[5] Padro Cirera, L., Fuentes, M.J.,Alonso, L., et al. Approaches to Text Summarization: Questions and Answers. Revista Iberoamericana de Inteligencia Arti_cial, ISSN 1137-3601,(22):79{102, 2004.

[6] Hovy, E. H. Automated Text Summarization. In R. Mitkov (ed), The Oxford Handbook of Computational Linguistics, chapter 32, pages 583{598.Oxford University Press, 2005.

[7] Spark Jones, K. Automatic summarizing: factors and directions. In Inderjeet Mani and Mark Marbury, editors, Advances in Automatic Text Summarization. MIT Press, 1999.

[8] Farshad Kyoomarsi, Hamid Khosravi, Esfandiar Eslami and Pooya Khosravyan Dehkordy, “Optimizing Text Summarization Based on Fuzzy Logic”, In proceedings of Seventh IEEE/ACIS International Conference on Computer and Information Science, IEEE, University of Shahid Bahonar Kerman, UK, 347-352, 2008.

[9] Vishal Gupta, G.Sl Lehal, “A Survey of Text Mining Techniques and Applications”, Journal of Emerging Technologies in Web Intelligence, VOL. 1, NO. 1, 60-76, AUGUST 2009.

[10] G Erkan and Dragomir R. Radev, “LexRank: Graph-based Centrality as Salience in Text Summarization”, Journal of Artificial Intelligence Research, Re-search, Vol. 22, pp. 457-479 2004.

[11] Udo Hahn and Martin Romacker, “The SYNDIKATE text Knowledge base generator”, Proceedings of the first International conference on Human language technology research, Association for Computational Linguistics , ACM, Morristown, NJ, USA , 2001.

[12] Spark-Jones, K. and J. R. Galliers (1995). Evaluating Natural Language Processing Systems: An Analysis and Review. Number 1083 in Lecture Notes in Artificial Intelligence. Springer.

[13] Radev, D. R., H. Jing, and M. Budzikowska (2000). Centroid-Based Summarization of Multiple Documents: Sentence Extraction, Utility-Based Evaluation, and User Studies, Proceedings of the Workshop on Automatic Summarization at the 6th Applied Natural Language Processing Conference

[14] Hassel, M. (2003). Exploitation of Named Entities in Automatic Text Summarization for Swedish. In Proceedings of NODALIDA’03 - 14th Nordic Conference on Computational Linguistics, Reykjavik, Iceland.

[15] Radev, D. R. and D. Tam (2003). Single-Document and Multi-Document Summary Evaluation via Relative Utility. In Poster Session, Proceedings of the ACM CIKM Conference, New Orleans, LA.

[16] Donaway, R. L., K. W. Drummey, and L. A. Mather (2000). A Comparison of Rankings Produced by Summarization Evaluation Measures, Proceedings of the Workshop on Automatic Summarization at the 6th Applied Natural Language Processing Conference.

[17] Salton, G. and M. J. McGill (1983). Introduction to Modern Information Retrieval. McGraw-Hill Book Company.

[18] Landauer, T. K., P. W. Foltz, and D. Laham (1998), Introduction to Latent Semantic Analysis. Discourse Processes, 25:259–284.

[19] Kanerva, P., J. Kristoferson, and A. Holst (2000). Random Indexing of text samples for Latent Semantic Analysis. In Gleitman, L. and A. Josh (editors), Proceedings 22nd Annual Conference of the Cognitive Science Society, Pennsylvania.

[20] Karlgren, J. and M. Sahlgren (2001). Vector-based Semantic Analysis using Random Indexing and Morphological Analysis for Cross-Lingual Information Retrieval.Technical report, SICS, Sweden.

[21] Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu (2001). BLEU: A Method for Automatic Evaluation of Machine Translation. Research Report RC22176, IBM.

[22] NIST (2002). Automatic Evaluation of Machine Translation Quality using N-gram Co-Occurrence Statistics. http://www.nist.gov/speech/tests/mt/doc/ngramstudy. pdf.

[23] Lin, C.-Y. and E. Hovy (2003). Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In Proceedings of 2003 Language Technology Conference (HLT-NAACL 2003), Edmonton, Canada.

[24] Hovy, E. and D. Marcu (1998). Automated Text Summarization Tutorial at COLING/ ACL’98. http://www.isi.edu/˜marcu/acl-tutorial.ppt.

[25] Mani, I. and M. T. Maybury (editors) (1999). Advances in Automatic Text Summarization, MIT Press, Cambridge, MA.

[26] Saggion, H. and G. Lapalme (2000). 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.