ISSN 2394-739X

International Journal for Research in Science Engineering and Technology

EVALUATION ON ONTOLOGY BASED DOCUMENT SUMMARIZATION USING NOVEL METHOD

[1] Dr.A. Mekala,
[1] Department of Computer Application
[1] Sacred Heart College
[1] Tirupattur

ABSTRACT: With widespread use of Internet and the emergence of information aggregation on a large scale, a quality text summarization is essential to effectively condense the information. Automatic summarization systems condense the documents by extracting the most relevant facts. Summarization is commonly classified into two types, extractive and abstractive. Summarization of abstraction needs understanding of the original text and then generating the summary which is semantically related. Ontology is one among the approach used for getting summary for a specific domain. In this paper, we discuss about various works carried out using ontology for text summarization.

Keywords: [Summarization, Ontology, Abstractive, Extractive]

1. INTRODUCTION

Text summarization is the task of creating a document from one or more textual sources that is smaller size but retains some or most of the information contained into the original sources. What information and which other characteristics of the source documents are kept depends on the intended use of the summary. Ultimately, the goal of automatic text summarization is to create summarised that are similar to human-created abstracts. Since this is a challenging task that involves text analysis, text understanding, the use of domain knowledge and natural language generation, research in automatic text summarization has largely focussed on generating extractive summaries. Extracts the summaries that consist of textual units selected from source documents, based on their usefulness for a summary. This usefulness is often equated with salience, hence most approaches evaluate which properties of textual units determine key information that therefore should be contained in a summary. Text Summarization aimed to generate concise and compressed form of original documents. With text mining, the information to be extracted is clearly and explicitly stated in the text. Text mining summarizes salient features from a large body of text, which is a subfield of text summarization. Summarization can be classified into two main categories i.e. extractive and abstractive summarization. Both techniques are used for summarizing text either for single document or for multi-documents. Extractive summarization involves assigning saliency measure to some units (e.g. sentences, paragraphs) of the documents and extracting those with highest scores to include in the summary. Abstractive summarization usually needs information fusion, sentence compression and reformulation. It is complex because it requires deeper analysis of source documents and concept-to-text generation.

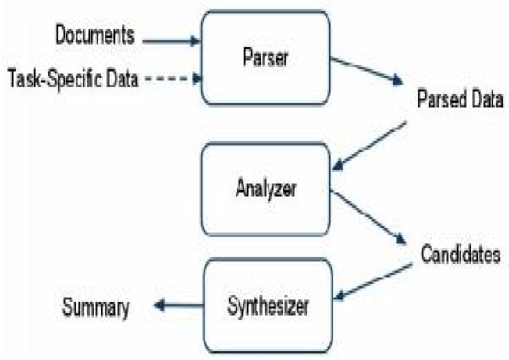


Figure 1- Summarization Systems Architecture

With Documents Summarization systems, it is possible to rephrase these steps and draw them as illustrated in Figure 1. The shown steps are shared and adapted by all of the available summarization systems. The details of the implementation of these steps are what make a system different from another. The figure illustrates that there are three main components: Parser, Analyzer and Synthesizer. The first stage may sometimes be referred to as the Pre-processing stage while the later is often called post processing.

Parser

The parser is first fed with the documents to be summarized. Task-specific data such as user's queries and compression rate may also be fed into the system. The parser then parses the fed data and prepares the documents in a suitable format acceptable by the analyzer.

Analyzer

The generated data are then fed to the analyzer where the core algorithms of the system are applied. A weight is usually attributed to each of the features detected or generated for each sentence. A score is then assigned to each sentence representing its importance. A sentence score is usually the sum of the weighted features scores. For some systems which produce abstractive summaries, sentences simplification, splitting, trimming or compression may be applied in this stage, too.

Synthesizer

The role of the synthesizer is to organize the scored candidates and present them in a form suitable to the user's needs. If a compression rate or words/sentences limit is specified, the synthesizer ensures that the output meets the given conditions. For single-document summaries, producing the summary is usually straightforward and is accomplished by choosing the highest ranked sentences according to their scores. For multi-document summaries, the process is usually more complex as it involves checking for redundancy, diversity and relevance to the user's specific needs.

2. LITERATURE SURVEY

In recent days, there has been an explosive growth in the volume of textual information available. Hence it is very important to present the data to the user in an abstract version. Summarization will make this process easy. Ontology based summarization methods involve reduction of sentences by compressing and reformulation.

Thanh Tran and Philipp Cimiano presented an approach for interpreting keyword queries using background knowledge available in ontology's. Based on a few assumptions about how people describe their information needs, an approach was presented which translates a keyword query into a DL conjunctive query which can be evaluated with respect to an underlying knowledge base (KB). One major problem the approach suffers from is the fact that it does not consider that keywords can be ambiguous with respect to labels in the ontology and simply considers the first matching ontology element to start the exploration.

Peroni, S., Motta, E., d'Aquin, M. address the issue of identifying the concepts in an ontology, which best summarize what the ontology is about. A number of criteria were jointly considered, and correspondingly a number of algorithms were developed and linearly combined, to identify key concepts of ontology. The criteria include: name simplicity which favours concepts that are labelled with simple names while penalizing compounds; basic level which measures how "central" a concept is in the taxonomy of the ontology; density highlights concepts which are richly characterized with properties and taxonomic relationships; coverage aims to ensure that no important part of the ontology is neglected; and popularity identifies concepts that are commonly used. The summarization results, i.e. key concepts, were evaluated against human assessors' summaries, referred to as ground truth.

Nesrine Ben Mustapha introduced a comprehensive framework for building a domain-specific ontology. Two methods for ontology acquisition were applied in order to create the domain ontology. The first was to create small domain-specific core ontology from scratch and then apply a focused web crawler to this ontology in order to retrieve domain related web pages and interesting domain terms for extending the knowledge base. The second acquisition approach takes a well-established thesaurus as a basic vocabulary reference set and converts it to an ontology representation. Then a domain specific and a general corpus of texts were used in order to remove concepts that are not descriptive for the domain.

Xiang Zhang proposed a novel approach to automatic ontology summarization based on RDF Sentence Graph. Summaries are customizable: users can specify the length of summaries and their navigational preferences. The authors compared five different centrality measurements in assessing the salience of RDF sentence and defined a reward-penalty re-ranking algorithm to make the summaries comprehensive. The evaluation showed that weighted in-degree centrality measures and several eigenvector centralities all have good performance in producing qualified summaries after re-ranking. Shown by the experiments, the approach of ontology summarization was feasible and promising.

Mithun and Munirathnam presented a semi-automatic development of an ontology library for the topics defined in the National Intelligence Priorities Framework (NIPF). They use Jaguar-KAT, a state-of-the-art tool for knowledge acquisition and domain understanding, with minimized manual intervention to create NIPF ontology's loaded with rich semantic content. Jaguar automatically builds domain-specific ontology's from text. The text input to Jaguar can come from a variety of document sources, including Text, MS Word, PDF and HTML web pages, etc. The ontology/knowledge base created by Jaguar includes ontological concepts, hierarchy and contextual knowledge base.

3. ONTOLOGY: AN OVERVIEW

Ontology is defined as a formal and explicit specification of a shared conceptualization. Generally, ontology's are defined for particular domains. Since information extraction is essentially concerned with the task of retrieving information for a particular domain, formally and explicitly specifying the concepts of that domain through ontology can be helpful to this process. Ontology together with a set of individual instances of classes constitutes a knowledge base. Classes are the focus of most ontology's. Classes describe concepts in the domain. For example, a class of wines represents all wines. Specific wines are instances of this class. A class can have subclasses that represent concepts that are more specific than the super class. For example, we can divide the class of all wines into red, white, and rose wines. A concept can be referenced by several terms and a term can reference several concepts. The roles of linguistic ontology's are two field: The first one is to present and define the vocabulary used. This is achieved by a dictionary which list all the terms actually used in language. Secondly, linguistic ontology is the result of a terminology agreement between users' community. This agreement defines which term is used to represent a concept in order to avoid ambiguity. This process is called vocabulary normalization. When a concept could be described by two synonym terms, the normalization process selects one of those to be the preferred label of the concept.

4. REASONS FOR DEVELOPING ONTOLOGY

1. Sharing regular comprehension of the structure of data among individuals or programming operators is one of the objectives in creating ontology's. For instance, assume a few diverse Web destinations contain restorative data or give medicinal online business administrations.

On the off chance that these Web locales share and distribute the same basic ontology of the terms they all utilization, at that point PC operators can concentrate and total data from these diverse destinations. The specialists can utilize this totalled data to answer client questions or as info information to different applications.

2. Empowering reuse of area information was one of the main thrusts behind late surge in ontology look into. For instance, models for a wide range of spaces need to speak to the thought of time. This portrayal incorporates the thoughts of time interims, focuses in time, relative measures of time, et cetera. On the off chance that one gathering of analysts grows such ontology in detail, others can essentially reuse it for their areas. Furthermore, on the off chance that we have to construct a huge ontology, we can coordinate a few existing ontology's depicting segments of the huge space.

3. Making unequivocal area suppositions fundamental a usage rolls out it conceivable to improvement these presumptions effectively if our insight about the space changes. Hard-coding suspicions about the world in programming-dialect code make these presumptions elusive and comprehend as well as difficult to change, specifically for somebody without programming ability. Likewise, unequivocal determinations of space information are helpful for new clients who must realize what terms in the area mean.

4. Isolating the space learning from the operational information is another regular utilization of ontology's. We can portray an undertaking of designing an item from its parts as indicated by a required determination and execute a program that does this setup autonomous of the items and segments themselves. We would then be able to build up ontology of PC-segments and qualities and apply the calculation to arrange made-to-arrange PCs.

5. Dissecting space learning is conceivable once an explanatory detail of the terms is accessible. Formal examination of terms is to a great degree important when both endeavouring to reuse existing ontology's and broadening them.

5. SUMMARIZATION TECHNIQUES

1. Abstractive Summarization Techniques

Abstractive summarization methods comprise of understanding the first text and re-letting it know in less word. It utilizes etymological methods to inspect and translate the text and afterward to locate the new concepts and articulations to best depict it by creating another shorter text that passes on the most essential information from the first text document. Abstractive summarization is grouped into two classes organized based (Rule based method, tree based method, ontology method and so forth.) and semantic based (Multimodal semantic model, information item based method, semantic graph based method and so forth.) methods.

2. Extractive Summarization Techniques

Extractive summarizers discover the most significant sentences in the document. It likewise stays away from the excess information. It is less demanding than abstractive summarizer to draw out the rundown. The regular methods for extractive are Term Frequency/Inverse Document Frequency (TF/IDF) method, cluster based method, graph theoretic approach, machine learning approach, LSA Latent Semantic Analysis (LSA) method, artificial neural networks, fuzzy logic, query based, concept-obtained text summarization, utilizing relapse for evaluating highlight weights, multilingual, subject driven summarization, Maximal Marginal Relevance (MMR), centroid-based summarization and so on. A general methodology for extractive methods includes three stages to be performed which are examined beneath.

Step 1: First step makes a portrayal of the document. Some pre-handling, for example, tokenization, stop word removal, noise removal, stemming, sentence splitting, frequency computation and so forth is connected here.

Step 2: In this progression, sentence scoring are performed. All in all, three approaches are tailed: (i) Word scoring–assigning scores to the most vital words. (ii) Sentence scoring–verifying sentences components, for example, its position in the document, similitude to the title and so on (iii) Graph scoring–analyzing the connection between sentences. The general methods for figuring the score of any word will be word frequency, TF/IDF, capitalized, formal person, place or thing, word co-event, lexical similitude, etc.The basic marvels utilized for scoring any sentences are Cue-phrases ("in outline", "in conclusion", "our examination", "the paper depicts and underscores, for example, "the best", "the most essential", "as per the investigation", "altogether", "critical", "specifically", "barely", "impossible‟), sentence incorporation of numerical information, sentence length, sentence centrality, sentence similarity to the title, and so on. Additionally the well known graph scoring methods are text rank, rugged way of the hub, total likeness and so forth.

Step 3: In this progression, high score sentences utilizing a particular arranging request for separating the substance are chosen and afterward the last synopsis is created on the off chance that it is a solitary document summarization. For multi-document summarization, the procedure needs to expand. Each document produces one synopsis and afterward any clustering calculation is connected to cluster the applicable sentences of every rundown to create the last outline. Table 1 presents a comparison of summarization methods based on type of summary.

Type of summarization methods	**Sub type**	**Concept**	**Advantages**	**Dis-Advantages**
Approaches	Abstractive	It is the process of reducing a text document in order to create a summary that is semantically related	Good compression ratio. More reduced text and semantically related summary	Difficult to compute
	Extractive	It consists of selecting important sentences from original document based on statistical features	Easy to compute because it does not deal with the semantics and more successful	Suffers from inconsistencies, lack of balance, results in lengthy summary

Table 1- Comparison of Summarization Methods

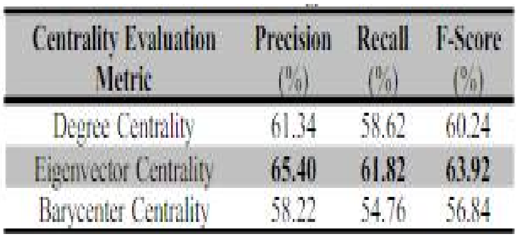
6. ASSESSMENT EXPERIMENTS AND RESULT

Assessment Condition

The assessment strategies for the programmed rundown frameworks by and large are isolated into two fundamental segments: extraneous and characteristic technique. In outward assessment strategies the nature of the rundowns in playing out specific undertakings is assessed, while in characteristic techniques the synopses autonomously and in view of the examinations from the outlines are assessed. The base of examination for the programmed synopsis frameworks is the rundowns which have been delivered by people and are called brilliant or reference outline. The brilliant outlines for an arrangement of records are being delivered by human here. We have utilized inherent assessment measures including exactness, review and furthermore F-score measure to assess the got after effects of programmed synopsis. The exactness is the portion of recovered occasions that are relevant and the review is the part of significant cases that are recovered. As to the way that we manage sentences as craved content units, we can express that the exactness breaks even with the quantity of basic sentences between the brilliant rundown and framework synopsis, partitioned by the quantity of the sentences of the framework outline. The review breaks even with the quantity of the basic sentences between the brilliant rundown and the framework outline, partitioned by the quantity of the sentences of the brilliant synopsis. The F-score measure rises to the consonant normal of the accuracy and review measures and equivalents (2×P×R)/(P+R).

Assessment Results

As said some time recently, in this investigation there are three approaches to assess the centrality of the charts vertices and therefore the centrality of sentences, including the measures of degree centrality, eigenvector centrality and bary focus centrality. In this manner there are three conceivable strategies to deliver programmed rundown. Table 2 incorporates the assessment after effects of the rundowns delivered by utilizing each of the three strategies for accuracy, review and F-score measures. It ought to be specified that esteems displayed in this table are the after effects of normal esteems, acquired from assessment of an arrangement of reports.



Centrality Evaluation Metric	Precision (%)	Recall (%)	F-Score (%)
Degree Centrality	61.34	58.62	60.24
Eigenvector Centrality	**65.40**	**61.82**	**63.92**
Barycenter Centrality	58.22	54.76	56.84

Table 2- Evaluation comes about for cosmology based synopsis

Centrality has the best outcomes among these three measures. At the end of the day, among the three measures of centrality the measure which considers the level of every vertex notwithstanding the vertices identified with it, will have a superior capacity in the valuation the significance of various literary writings. The rundowns created by utilizing degree centrality measure accomplished higher esteems in contrast with barycentre centrality measure for the accuracy, review and F-score. It can be derived that in assessing the centrality of the vertices of the chart, the measures which depend on the degrees of vertices (in the main rank is the one which contains the level of the present vertex not with standing the neighbouring vertices and the second rank is the measure which just contains the level of current vertex), will have a superior capacity in contrast with those which depend on separation of vertices from each other. Regardless of that they acquired outcomes demonstrate a worthy quality in created rundowns, we should focus on the way that the summarization has been the base for basic leadership for sentence determination and outline generation, so it is normal that utilizing the future and more entire renditions of that, which contains more substances and more semantic relations, prompts better outcomes in programmed synopsis.

CONCLUSION

Summarization methods create profoundly steady information and less repetitive rundown. By and large, a kind of summarization is a testing zone in light of the intricacy of common dialect handling. Many works are being completed in the field of summarization particularly by making utilization of ontology in different spaces. Presently we ought to think about how possible it is of pleasing all these express spaces in a solitary platform to manufacture a hearty, extensible summarization framework which will empower us to get outline from various areas. This examination analyzes an audit on ontology based summarization methods and its significance in various areas. A portion of the methods for assessing ontology are likewise determined. Positively, this examination can be adjusted in a way that new specialists to the region of text summarization can show signs of improvement understanding on ontology based approaches.

REFERENCES

[1] T. J. Siddiki and V. K. Gupta, Multi-document Summarization using Sentence Clustering, IEEE Proceedings of 4th International Conference on Intelligent Human Computer Interaction, India, 2012.

[2]. A. R. Deshpande, and Lobo L. M. R. J, Text Summarization using Clustering Technique, International Journal of Engineering Trends and Technology (IJETT), 4(8), 2013.

[3]. A. Agrawal and U. Gupta, Extraction based approach for text summarization using

k-means clustering, International Journal of Scientific and Research Publications, 4(11), 2014. [4]. M. A. Uddin, K. Z. Sultana and M. A. Alom, A Multi-Document Text Summarization for Bengali Text, IEEE International Forum on Strategic Technology (IFOST), Bangladesh, 2014.

[5]. M. I. A. Efat, M. Ibrahim , H. Kayesh, Automated Bangla Text Summarization by Sentence Scoring and Ranking, IEEE International Conference on Informatics, Electronics & Vision (ICIEV), Bangladesh, 2013.

[6]. L. C. Reddy and Venkatadri. M, A Review on Data mining from Past to the Future, International Journal of Computer Applications, 15(7), 2011.

[7]. H. Dave and S. Jaswal, Multiple Text Document Summarization System using Hybrid Summarization Technique,1st International Conference on Next Generation Computing Technologies (NGCT-2015), India, 2015, 4-5.

[8]. J. A. Kwak and H.-S. Yong, Ontology Matching Based On Hypernym, Hyponym, Holonym, and Meronym Sets in Wordnet, International Journal of Web Semantic Technology (IJWesT),1(2), 2010.

[9]. F. E. Gunawan, A. V. Juandi and B. Soewito, An Automatic Text Summarization using Text Features and Singular Value Decomposition for Popular Articles in Indonesia Language, IEEE International Seminar on Intelligent Technology and Its Applications, 2015.

[10] Thanh Tran, Philipp Cimiano, Sebastian Rudolph and Rudi Studer," Ontology-based Interpretation of Keywords for Semantic Search" Institute AIFB, Universität Karlsruhe, Germany

[11] Peroni, S., Motta, E., d'Aquin, M.: "Identifying Key Concepts in an Ontology Through the Integration of Cognitive Principles with Statistical and Topological Measures". In: 3rd AsianSemantic Web Conference, Bangkok, Thailand (2008)

[12] Nesrine Ben Mustapha, Hajer Baazaoui Zghal, Marie-Aude Aufaure, and Henda Ben Ghezala," Combining Semantic Search and Ontology Learning for Incremental Web Ontology Engineering" National School of Computer Sciences, University of Manouba, 2010 la Manouba, Tunisia

[13] Zhang, X., Cheng, G., Qu, Y.: "Ontology Summarization Based on RDF Sentence Graph". In: 16th Inter. World Wide Web Conference Banff, Alberta, Canada, May 8-12 (2007)

[14] C.-S. Lee, et al., "A fuzzy ontology and its application to news summarization, "Systems, Man, and Cybernetics, Part B:Cybernetics, IEEE Transactions on, vol.35, pp. 859-880, 2005.

[15] Mithun Balakrishna, Munirathnam Srikanth," Automatic Ontology Creation from Text for National Intelligence Priorities Framework (NIPF)", Lymba Corporation Richardson, TX, 75080, USA

[16] An Empirical Study of Ontology-Based MultiDocument Summarization in Disaster Management Lei Li and Tao Li-IEEE transactions on systems, man, and cybernetics: systems, vol. 44, no. 2, february 2014.

[17] J. Jayabharathy, S. Kanmani and A. Ayeshaa Parveen, "Document Clustering and Topic Discovery based on Semantic Similarity in Scientific Literature" 2nd international conference on Data Storage and Data Engineering, DSDE2011, 13th-15th May 2011, china.