Domain Specific e-Document Summarization Using Extractive Approach

Sunita R. Patil.
Research Scholar, MPSTME, NMIMS, Vile Parle

Sunita M.Mahajan.
Principal, ICS, MET, Bandra, Mumbai

ABSTRACT
With the rapid growth of online information availability, it becomes more and more important to find and describe textual information effectively from multiple related e-documents. Domain specific related e-documents contain information which is much relevant, similar or repeated and shares same background. Reading these all multiple relevant e-documents completely for accurate & brief contents is time-consuming, unnecessary and impossible. In this scenario multidocument summarization is useful to give an outline of a topic from multiple related source documents and allow users to zoom in for more details as per interest.

We discuss here a part of research work for summarizing multiple related research papers as e-documents presented by research scholars using various approaches and techniques aiming at giving an overview of the researches of this area and describing a method for automatic summarization of sets of research papers of desired subject that may be retrieved by a digital library system or search engine in response to a user query. Many digital libraries or online services provide research papers published in journals, conferences or workshops for reference purpose to the user. Research papers contain a wealth of high quality information by specifying research objectives, research methods, evaluation of research objectives, research results and concluding remarks. However, a research paper is relatively long and browsing too many of such research papers results in information overload. Therefore, it would be helpful to summarize a set of research papers to assist users in grasping the main ideas on a desired topic in a specific area.

General Terms

Data Mining and Intelligent Systems.

Keywords

Multi-document Summarization, Clustering, Extraction.

1. INTRODUCTION

Nowadays, people need much more information in work and life, especially the use of internet make information more easily gained. So, automatic summarization draws substantial interest since it provides a solution to the information overload problem people face in this digital era**.** Multi-document summarization is the process of dealing with a large amount of information present in multiple related source documents by comprises only the essential material or main ideas in a document in less space. Thus multidocument summary is useful to give an outline of a topic from multiple relevant documents and allow users to zoom in for more details as per interest which may be identifying the important concepts, entities, keywords, paragraphs, variables, relations or features in the text. Researchers are investigating summarization tools and methods that automatically extract or abstract content from a range of information sources, including multimedia. Moreover, the source may not always have text for example; a sports event on videotape or tables displaying economic data and current tools cannot summarize nontextual media.

The purpose of automatic summarization in technical literature is to facilitate quick, condenced and accurate identification of the topic from multiple related documents which are domain specific. The objective is to save a prospective reader time and effort in finding useful brief information from many more relevant e-documents in a specific area.

Typical information retrieval (IR) systems have two steps in doing this: the first is to find documents based on the user's query, and the second is to rank relevant documents and present them to the user based on their relevance to the query. Then the user has to read all of these documents. The problem is that these documents are much relevant and reading them all is time-consuming and unnecessary. In this scenario multi-document summarization is useful to give an outline of a topic from multiple related source documents and allow users to zoom in for more details as per interest.

We presents a part of research work to develop a method for automatic summarization of sets of research paper of specific area that may be retrieved by a digital library system or search engine in response to a user query. As the research done in a particular area is always published as a paper by researchers in journals and or conferences, many digital libraries or online services provide this published papers for general reference.

Research papers contain a wealth of high quality information by specifying research objectives, research methods, evaluation of research objectives, results of research and concluding remark. However, a research paper is relatively long and browsing too many of such research paper results in information overload. Therefore, we suggested a way to summarize a set of research papers to assist users in grasping the main ideas on a desired topic in a specific area.

2. BACKGROUND

2.1 Summary Categorization

Many types of summary have been identified since long by many researchers (Borko and Bernier 1975; Cremmins 1996; Sparck Jones 1999; Hovy and Lin 1999). Basically there are two approaches to categories summaries linguistic approach &

statistical approach. Under these two summaries can be further categorized as Extractive summaries, Abstractive summaries, Indicative summaries, Informative summaries, Topic-oriented or User Focused summaries, Fixed Length Summary / Variable Length Summaries, Critical and Generic summaries. At the most basic level, summaries differ according to whether they are extracts or abstracts.

• *Extractive summaries*- created by reusing portions (words, sentences, etc.) of the input text verbatim.

• *Abstractive summaries*- are created by regenerating the extracted content. *Extraction* is the process of identifying important material in the text, *abstraction* the process of reformulating it in novel terms, *fusion* the process of combining extracted portions, and *compression* the process of squeezing out unimportant material. The need to maintain some degree of grammaticality and coherence plays a role in all four processes.
Within and across these two categories, summaries differ according to function and target reader as indicative, informative, or critical:

• *Indicative summaries* follow the classical information retrieval approach: They provide enough content to alert users to relevant sources, which users can then read in more depth, provide an idea of what the text is about without conveying specific content

• *Informative summaries* act as substitutes for the source, mainly by assembling relevant or novel factual information in a concise structure.

• *Critical summaries* (or *reviews*), besides containing an informative gist, incorporate opinion statements on content. They add value by bringing expertise to bear that is not available from the source alone.

A summary can also be generic or user-focused:

• *Generic summaries* address a broad community; there is no focus on special needs because the summarizer is not targeting any particular group, it reflects the author's point of view.

• *Topic-oriented or User-focused summaries*, in contrast, are tailored to the specific needs or interest of an individual or a particular group.

2.2 Summarization Approaches & Methods

The main approaches used for multi-document summarization include sentence extraction, template-based information extraction, identifying important concepts, entities, keywords, paragraphs, variables, relations or features in the text, question answer based query systems, topic focused extraction, event indexing, time stamping and identification of similarities and differences between documents. Work on multi-document summarization by (Carbonell & Goldstein,1998; Mani & Bloedorn, 2000; McKeown, Klavans, Hatzivassiloglou, Barzilay, & Eskin, 1999; Radev & McKeown, 1998), uses techniques such as graph matching, maximal marginal relevance, or language generation.

Until recently, generic summaries were more popular, but with the prevalence of full-text searching and personalized information filtering, user-focused summaries are gaining importance. Many approaches discussed above support both user-focused and generic summarization.
Sentence extraction –extractive approach uses identification of frequent keywords, title keywords, cue phrases, sentence positioning, sentence length, and cohesive links such as lexical chains, co-reference, word co-occurrences, for internally linked sentences. Extractive techniques first segment source text into smaller segments (sentences, paragraphs, etc.), which are then scored according a variety of features, e.g., position in the text [16], term and phrase frequencies [20], lexical chains (degree of lexical-connectedness between various segments) [13], topics present in the text [23], or discourse prominence [22]. A widely-adopted approach is to use machine learning techniques to determine the relative importance of various features. These systems should explicitly model similarities and differences in text to address redundancy, paraphrase, entailment (consequence), contradiction, and related linguistic issues. One general approach involves clustering, as exemplified by the MEAD frame-work [16]. Documents are first clustered to find topics present in the sources. Clusters are represented by their centroids, which are used to rank extracts (along with other features). Maximal Marginal Relevance (MMR) [17] is another effective algorithm, specifically designed for query-focused summaries (i.e., summaries that address an information need). It iteratively selects candidate segments to include in the final summary, balancing relevance and redundancy at each iteration. Redundancy is computed by content similarity between each candidate and the current summary state (using cosine similarity)—thus, candidates containing words already in the summary are penalized. Note that neither MEAD nor MMR explicitly deals with linguistic relationships such as paraphrase, but that issue has been specifically addressed in other work [18].
After scoring and selecting segments from source documents, extractive systems must decide on an ordering in the final system output. Ideally, the output should constitute a coherent piece of text. Simple baselines for ordering segments include extraction order (i.e., by score), temporal order (based on metadata or temporal expressions), and order in source document (preserving source structure). While simple to implement, these techniques frequently yield dissent summaries. Coherence can be improved by applying computational models of content and discourse [14]. Nevertheless, text structuring is a relatively under-explore area of summarization, particularly due to difficulty in evaluation

Template-based information extraction – abstractive approach creates topic specific templates or patterns, connects templates to relations such as identity, elaboration, contradiction, equivalence, continuation, stability, extract salient (most important) information, combines instantiated slots in different templates.

Although open-domain abstractive summarization using deep semantic representations is beyond the current state of the art, a variety of successful abstractive techniques operating on syntactic structures have been developed. Most of these techniques involve parsing source documents and manipulating the resulting parse trees. One popular approach involves "trimming", or removing inessential structures from the parse tree [19, 24]—for example, removing adjunct clauses that do not contribute much information. Other successful techniques include "splicing" fragments from multiple sentences (sometimes across multiple documents)—for example, embedding a simple sentence as a relative clause inside another [21, 15]. Of course, these operations are not mutually exclusive. Syntactic manipulations are particularly helpful in

multi-document summarization since sentences from different sources might partially overlap, e.g., a sentence contains both redundant and new information. In this case, syntactic operations can potentially deliver the best of both worlds, by eliminating redundant information and preserving new information.

Shiyan Ou, Christopher S.G. Khoo [2] reports the similar type of work for summarizing sociology dissertation abstracts using semantic-level research variables, their relationships, taxonomy construction and variable based approach.

1. OUTLINE OF SELECTED TECHNIQUE

Most research in summarization over the past two decades has been on written news, due to relatively easy access to corpora. Today domain specific summarization systems are the attraction of researchers.

We are using sentence extraction and clustering approach in our study. [1]With sentence extraction approach sentences across all the research paper subtopics are clustered, following which, a small number of most related sentences are selected from each cluster of the particular category to form a summary. The sentence extraction strategy ranks and extracts representative sentences from multiple research papers. Radev et al. [3] described an extractive multi-document summarizer, which extracts a summary from multiple documents based on the document cluster centroids. Sentences extracted from the documents can describe part contents in a certain extent. Extraction-based summarization is a promising solution especially when the speed is concerned. In the sentence extraction strategy, clustering is frequently used to eliminate the redundant information resulted from the multiplicity of the original documents [4]. Some issues related to multi-document summarization based on clustering and sentence extraction strategy are:

i) Representation of a sentence: Vector space model (VSM) [5] handle massive real topics representing a sentence as a vector, in which words are used as features. Commonly, the segmentation tools are employed to break a sentence into a list of "words". In this paper, we use term as the feature of sentence vector. Here a term may refer to a word or a phrase, which has a relative complete meaning.

ii) Number of clusters appropriate for document collection:

To form the appropriate number or similarity threshold of the clusters in advance, we used the strategy to automatically assume the cluster number by using summary length fixed by the user.

iii) Select representative sentences from the clusters: Here, the sentence selection is a key problem. [9]In general, there are two kinds of search strategies. The local strategy tries to find a representative sentence for each cluster based on the information configuration of the cluster itself, while the global strategy tries to find the representative sentence based on the overall performance of the whole summary.

iv) Summary Evaluation Quality: There are two kinds of evaluating approaches: intrinsic and extrinsic [7].Intrinsic compares the final summaries generated by computer with those by hand [6, 8, 9]. Extrinsic is task-oriented approach [10]. The task-oriented method is more objective and therefore, an extrinsic summary evaluation method based on classification task is adopted in our research.

2. TERM IDENTIFICATION & EXTRACTION

For identifying key terms or words from the texts seeding-and-expansion mechanism is used. Term extraction using this mechanism consists of two phases, seed positioning and term determination. Where seed for a candidate term is an individual word, seed positioning is to locate the rough position of a term or word in the text, while term determination is to figure out which string covering the seed in the position forms a term. To determine a seed needs to consider a word reflecting their significance in the text. The relative probabilities of words occurring in reference documents and text document are calculated.

Any research paper contains the categories such as research objectives, research methods, evaluation of research objectives, research results and concluding remarks, which are considered as term. A key term can be defined with following assumption to extract it from document collection: [11]

i) A term contains at least one seed.

ii) A term occurs at least *T* (assume *T*=3) times in the document.

iii) A maximal word string satisfying i) and ii) is a term.

iv) For a term, a maximal substring satisfying i) and ii) without considering their occurrence in all those terms containing the sub-string is also a term.

Here a maximal word string satisfying i) and ii) refers to a word string satisfying i) and ii) while no other longer word strings containing it meet i) and ii). A maximal substring satisfying i) and ii) refer to a substring satisfying i) and ii) while no other longer substrings containing it meet i) and ii). For example, if the word string "Institute of social science and studies" is a term, "social" is a seed, then the "social science and studies" is a term too. The above assumptions tell us a term is an independent maximal string that must contain a seed and occur at least 3 times in an abstract document collection.

3. SENTENCE CLUSTERING

The terms extracted from the research paper collection are used to represent the features of vector in VSM. According to this, we set up the sentence VSM, where each sentence si in research paper collection is represented as the weights of terms, Vi. $Vi = (v_{i1}, v_{i2}, \ldots, v_{iN})$, i=1,2,…..M, where *M* is the number of sentences and *N* is the number of total terms in the document collection, vij denotes the weight of the *j*th term in the *i*th sentence. In this paper, we adopt the normalized term frequency in sentence as the term weight: [11]

$$Tij = TF(t_{ij}) / \sqrt{\sum_{t=1}^{N} TF^2(t_{ij})} \quad (1)$$

Where $TF(t_{ij})$ denotes the occurrence number of the *j*th term in the *i*th sentence.

For sentence clustering, cosine similarity and k-means clustering methods are used.

$$Sim(Si,Sj)=Cos(Vi,Vj)=(\sum^{N}_{t=1} v_{it}*v_{jt}) / (\sqrt{\sum^{N}_{t=1} v^2_{it}}*\sqrt{\sum^{N}_{t=1} v^2_{jt}}) \quad (2)$$

We fixed the number of cluster length because the clusters contains similar sentences of categories such as research objectives, research methods, evaluation of research objectives, research results and concluding remarks which forms the research paper abstract. On the other hand, to generate an anti-redundant summary, summarizer usually extracts only one sentence from each cluster. So, the number of sentences in fixed-length-summary is an acceptable value for the number of clusters. The most probable number of sentences in a fixed-length-summary is the length of summary fixed by user divided by the average length of sentences in abstract document collection. Thus, we determine the approximate number of clusters as:

$$K=LSM/avg(LS) \quad (3)$$

Where LSM denotes the summary length fixed by the user, avg (LS) denotes the average length of sentences in the abstract document collection.

After the number of optimal clusters has been chosen, we adopted the k-means algorithm for the clustering phase. Each of the output sentence clusters is supposed to denote one feature in the document collection.

4. SUMMARY SENTENCE SELECTION

For each sentence cluster, we need to select one sentence to represent the category denoted by the cluster. Now that the terms extracted from the texts (sentence cluster or the whole document collection) are supposed to denote the main concepts in the texts, we weight the sentence based on the terms included in the sentences.

6.1 Local Search Strategy

For local search strategies, we select the representative sentences based on the clusters themselves. We try 3 methods to select the representative sentence: centroid sentence, TF*IDF, TF. We use the terms extracted from the clusters, rather than those from the abstract document collections, because we are supposed to determine the focus or category of each cluster.

i) Centroid Sentence- Centroid sentence is selected by two steps. First, the centroid vector of the cluster is calculated. Second, the sentence, which has the smallest distance with the centroid vector, is selected. We use cosine distance as well.

*ii)TF*IDF-* TF*IDF value of a sentence is based on TF*IDF value of the terms occurred in the sentence.

Among a cluster, the sentence with the highest score is selected as the representative sentence.

iii) Term Frequency- Term Frequency value of a sentence is similar to TF*IDF, except that cluster frequency is not considered.

6.2 Global Search Strategy

For global search strategy, we select a sentence according to its contribution to the performance of the whole summary. To do that, we need a global criterion to measure the summary. The criterion is defined as follows:

$$w_{summary}=(\sum_{t\ \epsilon\ summary} \log(1+f_t^D)*\log(1+l_t))/(\log(1+l_{summary})) \quad (4)$$

Where t, is the term in the summary, f_t^D is term frequency in document collection, l_t is the term length. Intuitively, the criterion reflects the global term density of a summary. In general, we expect the summary to contain more terms, more longer terms, and as short as possible in each selecting step. The search process contains two phases: Firstly, the clusters are ordered by their size (the number of sentences). Secondly, for the first cluster, we choose a sentence which maximizes the criterion ($w_{summary}$), and then for the each remainder cluster, we choose a sentence, which, together with the sentences have been selected, maximizes the criterion.

6. EVALUATION

From the above discussion we have reached to the fact that multi-document summarization have two core tasks:
1. Determine what is salient (or relevant or important) in the source being summarized and
2. Decide how to reduce (or condense or abbreviate) it's content.
But within and across these two categories, summaries differ according to function and target user.

We adopt the extrinsic method to evaluate the quality of summarization by evaluating the results of classifying task. The training and testing data set are the research document collection and their summaries produced by the summarizer to be evaluated. For this classifying task, the abstract document collection *D* and their summaries set *S* are divided into two equal parts *D*1, *D*2, *S*1 and *S*2 respectively. The effectiveness of summarization can be evaluated through comparing the effectiveness of following 4 subtasks:[1]

i) S1D2: classify *S*1 using the classifier trained with *D*2;

ii) S1S2: classify *S*1 using the classifier trained with *S*2;

iii) D1S2: classify *D*1 using the classifier trained with *S*2;

iv) D1D2: classify *D*1 using the classifier trained with *D*2;

In this paper, we employ the naive Bayes method as the classifier [12] and F1-measure as the evaluation criterion. Suppose that α denotes the number of documents correctly assigned to this category, β denotes the number of documents incorrectly assigned to this category, and γ denotes the number of documents incorrectly rejected from this category. F1-measure defined as follows:

$$Recall = \alpha/(\alpha+\gamma)$$

$$Precision = \alpha/(\alpha+\beta)$$

$$F1\ measure = (2*precision*recall)/(precision+recall) \quad (5)$$

For evaluating performance average across categories, first overall performance scores are determined by computing the performance measures per category and then averaging these to compute the global means. Secondly in particular performance scores are determined by first computing the totals of α, β and γ for all categories and then use these totals to compute the performance measures.

7. RESULTS

Evaluation gives the measure of will the summaries generated various approaches actually help end-users to make better use? Multi document summaries should enable users to more efficiently find the information they need.

While evaluating above approach, we were interested for whether this will be an effective tool for assisting the processing of large volumes of research document contents to answer the following questions:

• Do our summary help the user find information needed to perform a task of interest?
• Do users use information from the summary in gathering their facts?
• Do our summary increase user satisfaction with the online information services?
• Do users create better fact sets with an online information services which includes multi-document summarization than one that not?
• In the context of every summary, what is the comparison of information quality in this task, and user satisfaction, when users have access to various summarization approaches versus minimal or human summaries?

The comparison of scores of classification tasks S1D2, S1S2, D1S2, D1D2 and different sentence selection strategies as Centroid Sentences, TF*IDF, Term Frequency, Global Search are measured for overall performance and in particular performance. Summaries generated by the summarizer using word as the VSM feature and the summarizer using term show that the term-based summarizer outperforms the word-based, which confirm that term or feature is more than the word.

8. CONCLUSION

In this paper we proposed a method for domain specific multiple related research papers summarization which are published at various conferences or journals. It mainly consists of two steps: sentence clustering and sentence selection for summary. For sentence clustering, we proposed the strategy to determine the number of clusters automatically as use of the summary length fixed by the user. For sentence selection, we present a global search method and compare this method with other local methods. To evaluation the summarization, we proposed an extrinsic evaluation method based on a classification task. Experimental results show that our summarization strategy is effective and efficient for classification tasks.

9. REFERENCES

[1] Sunita R. Patil.and Sunita M.Mahajan, A Novel Approach For Research Paper Abstracts Summarization Using Cluster Based Sentence Extraction, ACM-ICWET2011,TCET,Mumbai

[2] Shiyan Ou, Christopher S.G. Khoo and Dion H. Goh,Design and development of a concept-based multidocument summarization system for research abstracts. Published in Journal of Information Science OnlineFirst, on December 3, 2007 as doi:10.1177/0165551507084630.

[3] Dragomir R. Radev, Hongyan Jing, and Malgorzata Budzikowska, Centroid-Based Summarization of Multiple Documents: Sentence Extraction, Utility-Based Evaluation and User Studies. Information Processing and Management, Vol. 40, pp. 919–938, 2004.

[4] Endre Boros, Paul Kantor, and David J. Neu, A Clustering Based Approach to Creating Multi-Document Summaries, http://www.nlpir.nist.gov/projects/duc/pubs/2001papers/rutgers_final.pdf

[5] Patrick Pantel, and Dekang Lin, Document Clustering with Committees, Proceedings of ACM, SIGIR'02, New York: ACM, pp. 199–206, 2002.

[6] Po Hu, Tingting He, Donghong Ji, and Meng Wang, A Study of Chinese Text Summarization Using Adaptive Clustering of Paragraphs, Proceeding of the Fourth International Conference on Computer and Information Technology, Wuhan, pp. 1159–1164, 2004.

[7] Inderjeet Mani, Summarization Evaluation: An Overview, Proceedings of the NTCIR Workshop 2 Meeting on Evaluation of Chinese and Japanese Text Retrieval and Text Summarization, 2001.

[8] Daniel Marcu, From Discourse Structures to Text Summaries, ACL/EACL-97 Workshop on Intelligent Scalable Text Summarization, pp. 82–88, 1997.

[9] ChinYew Lin, and Eduard Hovy, Automatic Evaluation of Summaries Using N-gram Co-Occurrence Statistics, Proceedings of the Human Technology Conference (HLTNAACL-2003), Edmonton, Canada, 2003

[10] T. F. Hand, A proposal for task-based evaluation of text summarization systems, ACLEACL-97 summarization workshop, pp. 31–36, 1997.

[11] De-Xi Liu, Yan-Xiang He, Dong Hong Ji, Hua Yang, A Novel Chinese Multi-Document Summarization Using Clustering Based Sentence Extraction. Proceedings of the Fifth International Conference on Machine Learning and Cybernetics, Dalian, 1-4244-0060-0/06/$20.00 ©2006 IEEE

[12] T. Joachims, A probabilistic analysis of the rocchio algorithm with TFIDF for text categorization, Int.Conf. Machine Learning. 1997.

[13] Regina Barzilay and Michael Elhadad. Using lexical chains for text summarization. In *Proceedings of the ACL/EACL Workshop on Intelligent Scalable Text Summarization*, Madrid, Spain, 1997.

[14] Regina Barzilay and Lillian Lee. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *Proceedings of the 2004 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL 2004)*, pages 113–120, Boston, Massachusetts, 2004.

[15] Regina Barzilay and Kathleen R. McKeown. Sentence fusion for multidocument news summarization. *Computational Linguistics*, 31(3):297–327, 2005.

[16] Harold P. Edmundson. New methods in automatic extracting. *Journal of the ACM* , 16(2):264–285, 1969.

[17] Jade Goldstein, Vibhu Mittal, Jaime Carbonell, and Jamie Callan. Creating and evaluating multi-document sentence

extract summaries. In *Proceedings of the Ninth International Conference on Information and Knowledge Management (CIKM 2000)*, pages 165–172, McLean, Virginia, 2000.

[18] Vasileios Hatzivassiloglou, Judith L. Klavans, and Eleazar Eskin. Detecting text similarity over short passages: Exploring linguistic feature combinations via machine learning. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-1999)* , 1999.

[19] Kevin Knight and Daniel Marcu. Statistics-based summarization—step one: Sentence compression. In *Proceedingsof the Seventeenth National Conference on Artificial Intelligence (AAAI-2000)* , pages 703–710, Austin, Texas, 2000.

[20] Hans Peter Luhn. The automatic creation of literature abstracts.*IBM Journal of Research Development*, 2(2):159–165, 1958.

[21] Inderjeet Mani, Barbara Gates, and Eric Bloedorn. Improving summaries by revising them. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL 1999)* , pages 558–565, College Park, Maryland, 1999.

[22] Daniel Marcu. *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts* . PhD thesis, University of Toronto, 1997.

[23] Dragomir R. Radev, Sasha Blair-Goldensohn, and Zhu Zhang. Experiments in single and multi-document summarization using MEAD. In *Proceedings of the 2001 Document Understanding Conference (DUC 2001)* , 2001.

[24] David Zajic, Bonnie Dorr, Jimmy Lin, and Richard Schwartz. Multi-Candidate Reduction: Sentence compression as a tool for document summarization tasks. *Information Processing and Management* , 43(6):1549–1570, 2007.