Relevance of Cluster size in MMR based Summarizer: A Report

11-742: Self-paced lab in Information Retrieval
Submitted by: Madhavi K. Ganapathiraju
Advisors: Dr. Jaime Carbonell, Dr Yiming Yang
November 26, 2002

Executive Summary

Maximal Marginal Relevance Multi Document (MMR-MD) uses passage clustering to choose passages with large coverage and to aid in reducing redundancy. It is expected that the Quality of Summary (QoS) would directly depend on the cluster granularity.

The objective of this work is to study the relevance of granularity of passage clusters towards QoS, in the context of MMR-MD summarizer. This has been done, and the results on the Document Understanding Conference (DUC-2002) data set are reported.

This report also presents an overview of extractive summarization methods, features useful in selecting summary sentences, and strategies and metrics for evaluating summaries.

Observations on passage clustering by bottom-up approach, followed by results of the study on QoS versus cluster-granularity are then presented. Based on the observations from this study, a new method for extractive summarization is also proposed for future work.

1 Problem Statement

Maximal Marginal Relevance summarization presented in [6] is a cluster-based, extractive summarization method, where passages are first clustered based on similarity, prior to the selection of passages that form the extractive summary of the documents. Passage clustering forms a main component in this system that aims to extract the most relevant sentences of the documents at the same time keeping the summary non-redundant. The goal of this work is to study how the quality of the summary varies with the granularity of clusters. In this report we present the conclusions of this study, after presenting an introduction to automatic document summarization and an overview of the methods of summarization from current literature.

OVERVIEW OF SUMMARIZATION METHODS

2 Introduction

The Internet has come to be of much use primarily because of the support given by Information Retrieval (IR) tools. However with the exponential growth of the information on the Internet, a second level of abstraction of information from the results of the first round of IR becomes necessary. That is, the large number of documents returned by IR system need to be summarized. Currently this is the primary application of summarization. The many other uses of summarization are almost obvious: Information extraction, as against document-retrieval, automatic generation of comparison charts, Just-In-Time knowledge acquisition, finding answers to specific questions, a tool for information retrieval in multiple languages, biographical profiling, to name a few.

In this report we present an introduction to various methods of automatic summarization. We study one such method—MMR summarization more closely with respect to one of the parameters, namely, cluster granularity.

We first present a comprehensive report on Automatic Text Summarization, covering the following topics:

1. Types of summaries and their properties
2. Evaluation strategies and metrics
3. Approaches to automatic summarization, and their features
4. Special tools: Visualization
5. Issues that come up in various approaches

Typically evaluation metrics are discussed towards the end of a presentation. However, we feel that knowing the criteria that determine the quality of a summary helps to better-understand the approaches to summarization. Hence we present evaluation strategies and metrics of the quality of summaries soon after describing the types of summaries. The presentation will proceed in the same order as listed above.

3 Types of Summaries

The approach and the end-objective of summarization of documents explain the kind of summary that is generated. For example, it could be indicative of what a particular subject is about, or can be informative about specific details of the same. It can differ in being a "generalized summary" of a document as against a "query-specific summary". It may be a collection of sentences carefully picked from the document or can be a formed by synthesizing new sentences representing the information in the documents. Summaries may be classified by any of the following criteria [1]:

Detail: Indicative/informative
Granularity: specific events/overview
Technique: Extraction/Abstraction
Content: Generalized/Query-based
Approach: Domain/Genre specific/independent

4 Evaluation Strategies and Metrics

We first study the methods and metrics for evaluation of summarization, so that the strengths and weaknesses of the various approaches to summarization are more clearly understood. Human judgment of the quality of a summary varies from person to person. For example, in a study conducted by Goldstein, et al. [2], when a few people were asked to pick the most relevant sentences in a given document, there was very little overlap of the sentences picked by different persons. Also, human judgment usually does not find concurrence on the quality of a given summary. Hence it is difficult to quantify the quality of a summary.

However, a few indirect measures may be adopted that indicate the usefulness and completeness of a summary [1, 3-5], such as:

1. Can a user answer all the questions by reading a summary, as he would by reading the entire document from which the summary was produced?
2. What is the compression ratio between the given document and its summary?
3. If it is a summary of multiple documents with temporal dimension, does it capture the correct temporal information?
4. Redundancy—is any information repeated in the summary?

Qualities of summary that are usually difficult to measure are:

5. Intelligibility
6. Cohesiveness
7. Coherence
8. Readability (depends on cohesion/coherence/intelligibility)

A metric is said to be *intrinsic* or *extrinsic* depending on whether the metric determines the quality based on the summary alone, or based on the usefulness of the summary in completing another task [6]. For example, 1 above is an extrinsic metric. An example of intrinsic measure is the cosine similarity of the summary to the document from which it is generated. This particular measure is not of very useful, since it does not take into effect

the coverage of information or redundancy. With such a measure, a trivial way for improving the score would be to take the entire document as its summary.

A metric that is commonly employed for extractive summaries is that proposed by Edmundson [7]. Human judges hand pick sentences from the documents to create manual-extractive summaries. Automatically generated summaries are then evaluated by computing the number of sentences common to the automatic and manually generated summaries. In Information Retrieval terms, these measures are called *Precision* and *Recall*. This method is currently the most used method for evaluating extractive summaries [6, 8-10]. For an experimental study of various evaluation metrics see [10].

5 Features

Sentence extraction methods for summarization normally work by scoring each sentence as a candidate to be part of summary, and then selecting the highest scoring subset of sentences.

Some features that often increase the candidacy of a sentence for inclusion in summary are [[6, 7] and references therein]:
Keyword-occurrence: Selecting sentences with keywords that are most often used in the document usually represent theme of the document
Title-keyword: Sentences containing words that appear in the title are also indicative of the theme of the document
Location heuristic: In Newswire articles, the first sentence is often the most important sentence; in technical articles, last couple of sentences in the abstract or those from conclusions is informative of the findings in the document [11].
Indicative phrases: Sentences containing key phrases like "*this report …*"
Short-length cutoff: Short sentences are usually not included in summary.
Upper-case word feature: Sentences containing acronyms or proper names are included.

While the above features increase the score of a sentence to be included in the summary, those that reduce its score are:
Pronouns: Pronouns such as "she, they, it" cannot be included in summary unless they are expanded into corresponding nouns.
Redundancy in summary: Anti-redundancy was not explicitly accounted for in earlier systems, but forms a part of most of the current summarizers. This score is computed dynamically as the sentences are included in the summary, to ensure that there is no repetitive information in the summary. The following are two examples of anti-redundancy scoring, when a new sentence is added to the summary:

- Scale down the scores of all the sentences not yet included in the summary by an amount proportional to their similarity to the summary generated so far [2, 9, 12].
- Recompute the scores of all the remaining sentences after removing the words present in the summary from the query/centroid of document [13].

6 Approaches to Summarization

Abstraction of documents by humans is complex to model as is any other information processing by humans. The abstracts differ from person to person, and usually vary in the style, language and detail. The process of abstraction is complex to be formulated mathematically or logically [14]. In the last decade some systems have been developed that generate abstractions using the latest natural language processing tools. These systems extract phrases and lexical chains from the documents and fuse them together with generative tools to produce a summary (or *abstraction*). A comparatively less complex approach is to make an *extractive* summary in which sentences from the original documents are selected and presented together as a summary.

PROBLEMS WITH EXTRACTIVE METHODS:

- Extracted sentences usually tend to be longer than average. Due to this, part of the segments that are not essential for summary also get included, consuming space.
- Important or relevant information is usually spread across sentences, and extractive summaries cannot capture this (unless the summary is long enough to hold all those sentences).
- Conflicting information may not be presented accurately.

PROBLEMS WITH ABSTRACTIVE METHODS:

- It has been shown that users prefer extractive summaries instead of glossed-over abstractive summaries [15]. This is because extractive summaries present the information as-is by the author, and would allow the users to read between-the-lines information.
- Sentence synthesis is not a well-developed field yet, and hence the machine generated automatic summaries would result in incoherence even within a sentence. In case of extractive summaries, incoherence occurs only at the border of two sentences.

The work presented in this report is relevant to extractive summaries. In the rest of this section we study some specific methods producing extractive summaries.

6.1 EXTRACTIVE METHODS

Extractive summarizers aim at picking out the most relevant sentences in the document while also maintaining a low redundancy in the summary. While anti-redundancy was not explicitly documented in older systems, most of the current systems account for it in their own novel ways.

6.1.1 CLASSICAL METHOD:

Though text summarization has drawn attention primarily after the information explosion on the Internet, the seminal work has been done as early as in the 1950's. Edmundson presents a survey of the then existing methods to automatic summarization in [16] and a

systematic approach to summarization which forms the core of the extraction methods even today in [7]. Edmundson's approach is summarized in Figure 2.

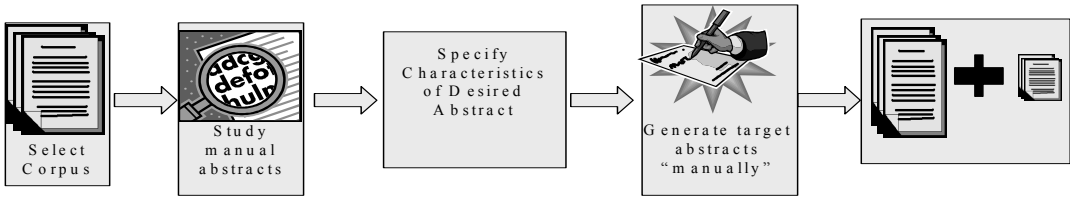

Figure 1: Step 1 in of the summarization process elicited by Edmondson: Manually generate abstracts, in the form desired to be generated by automatic summarizer

The key ideas in this approach are:

1. Study human generated abstracts, and *specify* characteristics expected in automatically generated abstracts
2. Generate such abstracts manually.
3. Design mathematical and logical formulations to score and pick out sentences from the documents to match these manually generated abstracts. A system of recent times, that automatically learns from training documents and their corresponding abstracts is described in [17].
4. Iteratively improve the sentence-scoring scheme to match the automatic abstracts to manually generated abstracts.

Computationally representable features of sentences that are useful to score sentences for potential inclusion in the summaries have been proposed in [7], and are used even in the systems of today. Stop words are removed. Sentences are then scored according to four factors:

Cue: Those containing cue words/phrases like *conclusion, according to the study, hardly* are given a higher weight than those not containing them.

Key words: Statistically significant words are given higher scores. Score of sentence is then computed as the sum of the scores of its constituent words. Edmundson [7] reports that he considered the words present in the sentences containing cue words, as significant words. Later the score of words is modified to be count of that word in the document. This is later made into a relative measure, and is modified to be the frequency of this word in the document.

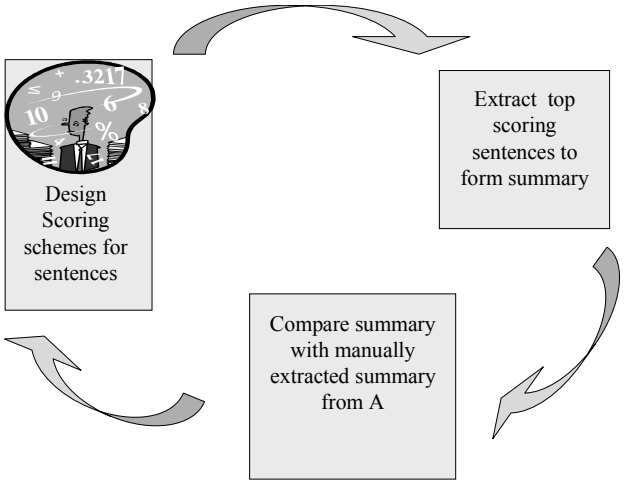

Figure 2: Step 2 in of the summarization process elicited by Edmondson: Design sentence scoring metrics such that extracted sentences are close to manually generated summaries of Step 1. Iteratively improve scoring schemes.

Title: Sentences containing title words are considered to have scored higher. Title words are those that are present in the title of the

document, and headings and sub-headings.
Location: This is dependent on the type of the document. For example, in technical documents, sentences in the conclusion section are ranked high, while in news articles, first few sentences are ranked higher.

The score of each sentence is computed as

$$S_i = w_1 * C_i + w_2 * K_i + w_3 * T_i + w_4 * L_i \quad \text{.....(1)}$$

where S_i is the score of sentence *i*. $C_i, K_i \,\&\, T_i$ are the scores of the sentence *i* based on the number of cue words, keywords and title words it contains, respectively. L_i is the score of the sentence based on its location in the document. $w_1, w_2, w_3 \,\&\, w_4$ are the weights for linear combination of the four scores.

Comments:
Edmundson's method [7] discussed above is a very systematic approach to extractive summarization and elicits most of the characteristics that are useful for sentence selection for summarization. The cues, keywords, title words and location are what are relied upon as primary features even today. One important issue not accounted for in this approach is redundancy in the summary. Future systems such as Maximal Marginal Relevance (MMR) summarizer and MEAD make Edmundson's method more complete. MMR and MEAD are described later in the report.

6.1.2 TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY (TF-IDF) METHODS:

Bag-of-words model is built at sentence level, with the usual weighted term-frequency and inverse sentence-frequency paradigm, where sentence-frequency is the number of sentences in the document that contain that term. These sentence vectors are then scored by similarity to the query and the highest scoring sentences are picked to be part of the summary [18]. This is a direct adaptation of Information Retrieval paradigm to summarization. Summarization is query-specific, but can be adapted to be generic as described below.

To generate a generic summary, non stop-words that occur most frequently in the document(s) may be taken as the query words. Since these words represent the theme of the document, they generate generic summaries.

Comments:
Term-frequency is usually 0 or 1 for sentences—since normally the same content-word does not appear many times in a given sentence. If users create query words the way they create for information retrieval, then the query based summary generation would become generic summarization. This is because giving theme words such as "information retrieval" in a document talking about IR, the IDF would highly dominate the TF. Also relying on TF-IDF for summary generation would "insist" that the keyword be present in every sentence. This would invariably cause some of the most important and informative sentences to be excluded from the summary. It is unlikely that every sentence in the document would contain the theme words. This also leads to the summary being extremely redundant.

6.1.3 CLUSTER-BASED METHODS:

Documents are usually written such that they address different topics one after the other in an organized manner. They are normally broken up explicitly or implicitly into sections. This organization applies even to summaries of documents. It is intuitive to think that summaries should address different "themes" appearing in the documents. Some summarizers incorporate this aspect through clustering.

If the document collection for which summary is being produced is of totally different topics, document clustering becomes almost essential to generate a meaningful summary.

Systems:

Maximum Marginal Relevance Multi Document (MMR-MD) summarization is a purely extractive summarization method that is based on Maximal Marginal Relevance concept proposed for information retrieval [19].It aims at having high relevance of the summary to the query or the document topic, while keeping redundancy in the summary low [2, 6, 19]. It can accommodate a number of criteria for sentence selection such as content words, chronological order, query/topic similarity, anti-redundancy and pronoun penalty.

The core scoring algorithm from [2] is reproduced ***ad verbatim*** in Figure 3 for convenience.

$$MMR-MD \equiv Arg \max_{P_{ij} \in R \setminus S} [\lambda Sim_i(P_{ij}, Q, C_{ij}) - (1-\lambda) \max_{P_{ij} \in S} Sim_2(P_{ij}, P_{nm}, C, S)]$$

$$Sim_1(P_{ij}, Q, C_{ij}, D_i, D) = w_1 * (P_{ij}, Q) + w_2 * Coverage(P_{ij}, C_{ij}) + w_3 * Content(P_{ij}) + w_4 * time\_sequence(D_i, D)$$

$$Sim_2(P_{ij}, P_{nm}, C, S, D_i) = w_a * (P_{ij} \bullet P_{nm}) + w_b * clusters\_selected(C_{ij}, S) + w_c * documents\_selected(D_i, S)$$

$$Coverage(P_{ij}, C) = \sum_{k \in C_{ij}} wk * |k|$$

$$Content(P_{ij}) = \sum_{W \in P_{ij}} w_{type}(W)$$

$$time\_sequence(D_i, D) = \frac{timestamp(D_{\max time}) - timestamp(D_i)}{timestamp(D_{\max time}) - timestamp(D_{\min time})}$$

$$clusters\_selected(C_{ij}, S) = |C_{ij} \cap \bigcup_{v,w:P_{vw} \in S} C_{vw}|$$

$$documents\_selected = \frac{1}{|D_i|} * \sum_w [P_{iw} \in S]$$

Figure 3: MMR scoring algorithm, reproduced from original source.

where, Sim_1 is the similarity metric for relevance reranking; Sim_2 is the anti-redundancy metric; D is the document collection; P is the passages from the documents in that collection; Q is a query or user profile; R = IR(D,P,Q,θ); S is the subset of passages in R already selected; R\S is the set difference of R and S; C is the set of passage clusters for the set of documents; C_{vw} is the subset of clusters of C that contains passage P_{vw}; C_v is the subset of clusters that contain passage from document D_v; |k| is the number of passages in the individual cluster k; $|C_{vw} \cap C_{ij}|$ is the number of clusters in the

intersection of C_{vw} and C_{ij}; w_i are the weights for the terms, which can be optimized; W is a word in the passage P_{ij}; type is a particular type of word, e.g, city name; $|D_i|$ is the length of the document $_i$.
MMR uses bag-of-words model to represent individual sentences and the whole document. This model can be any of Vector Space Model (VSM) or Latent Semantic Analysis (LSA) or similar models. It then computes similarity of each sentence to the entire document or query using any of the similarity metrics such as cosine similarity. Sentences that are chosen for inclusion in summary are such that they are maximally similar to the document or query, while maintaining that they are maximally dissimilar to the sentences already included in the summary. This ensures that the sentences most representative of the document are chosen, while ensuring minimum redundancy in the summary. In addition to the maximal marginal relevance, passages may also be weighed more for special features such as having words like "in conclusion", or proper names. MMR has the advantage that a summary of any desired length (with trivial bounds) could be generated.

All the sentences in the given document or documents may initially be clustered by similarity. And a sentence closest to the centroid of the cluster may be chosen to be included in the summary. The MMR approach has a tendency to include longest sentences into the summary initially.

Comments:
The work reported in this paper is primarily on MMR summarization. See later sections for further description of the algorithm.

MEAD is a sentence level extractive summarizer that takes document clusters as input [8, 9, 12]. Documents are represented using term frequency-inverse document frequency (TF-IDF) of scores of words. Term frequency used in this context is the average number of occurrences (*per document*) over the cluster. IDF value is computed based on the entire corpus. The summarizer takes already clustered documents as input. Each cluster is considered a theme. The theme is represented by words with top ranking term frequency, inverse document frequency (TF-IDF) scores in that cluster.

Sentence selection is based on similarity of the sentences to the theme of the cluster (C_i). The next factor that is considered for sentence selection is the location of the sentence in the document (L_i). In the context of newswire articles, the closer to the beginning a sentence appears, the higher its weightage for inclusion in summary. The last factor that increases the score of a sentence is its similarity to the first sentence in the document to which it belongs (F_i).

The overall score (S_i) of a sentence *i* is a weighted sum of the above three factors:

$$S_i = w_1 * C_i + w_2 * F_i + w_3 * L_i \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2)$$

where S_i is the score of sentence *i*. C_i & F_i are the scores of the sentence *i* based on the similarity to theme of cluster and first sentence of the document it belongs to, respectively. L_i is the score of the sentence based on its location in the document.

w_1, w_2 & w_3 are the weights for linear combination of the three scores. Note the similarity between the sentence score in equations (1) and (2). The role of *F* in (2) is similar to that of *T* in (1). The difference however, is that, S_i in (2) is further re-scored using a redundancy factor. For further details on this redundancy re-ranking, see [8].

Comments:

- Once the documents are clustered, sentence selection from within the cluster to form its summary is local to the documents in the cluster. The IDF value based on the corpus statistics seems counter-intuitive. A better choice may be to take the Average-TF alone to determine the theme of the cluster, and then rely on the "anti-redundancy" factor to cover the important 'themes' *within* the cluster.
- Both the position factor (P_i) and the first-sentence similarity factor (F_i) heavily weight the first few sentences of the documents in the cluster. Thus this metric is genre-specific and applies primarily to newswire articles. For other articles such as technical papers, the scoring would have to be re-designed.

6.1.4 GRAPH THEORETIC APPROACHES:

As seen in the previous methods, the first step involved in the process of summarizing one or more documents is identifying the issues or topics addressed in the document.

Graph theoretic representation of passages provides a method of identification of these themes. After the common preprocessing steps, namely, stop word removal and stemming, sentences in the documents are represented as nodes in an undirected graph. There is a node for every sentence. Two sentences are connected with an edge if the two sentences share some common words, or in other words, their (cosine, *or such)* similarity is above some threshold.

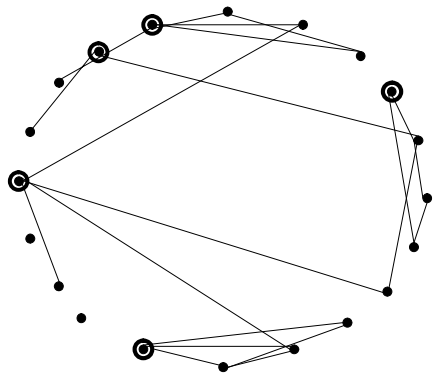

Figure 4: Graph Theoretic approach: Each Node is a sentence; an edge exists between two nodes if their similarity is above a threshold. Highlighted nodes can be seen to be "important" sentences in the document.

This representation yields two results: The partitions contained in the graph (that is those sub-graphs that are unconnected to the other sub-graphs), form *distinct topics* covered in the documents. This allows a choice of coverage in the summary. For query-specific summaries, sentences may be selected only from the pertinent sub-graph, while for generic summaries, representative sentences may be chosen from each of the sub-graphs. The second result yielded by the graph-theoretic method is the identification of the important sentences in the document. The *nodes with high cardinality* (number of edges connected to that node), are the important sentences in the partition, and hence carry higher preference to be included in the summary. Figure 3 shows an example graph for a document. It can be seen that there are about 3-4 topics in the document; the

nodes that are encircled can be seen to be informative sentences in the document, since they share information with many other sentences in the document.

The graph theoretic method may also be adapted easily for visualization of inter- and intra-document similarity.

6.1.5 MACHINE LEARNING TECHNIQUES:

Given a set of training document and their extractive summaries, the summarization process is modeled as a classification problem: sentences are classified as summary-sentences and non-summary sentences based on the features that they possess [Source unavailable]. Features that are used to distinguish summary sentences are those listed in Section 5. The classification probabilities are learnt statistically from the training data, using Bayes' rule:

$$P(s \in S \mid F1, F2, ..., FN) = P(F1, F2, \ldots, FN \mid s \in S) * P(s \in S) / P(F1, F2, \ldots, FN)$$

where s is a sentence from the document collection, F1, F2…FN are features used in classification, S is the to be generated, and $P(s \in S \mid F1, F2, ..., FN)$ is the probability that sentence s will be chosen to form the summary given that it possesses features F1, F2…FN.

6.1.6 LSA METHODS:

Singular Value Decomposition (SVD) is a very powerful mathematical tool that can find principal orthogonal dimensions of multidimensional data. It has applications in many areas and is known by different names: Karhunen-Loeve Transform in image processing, Principal Component Analysis (PCA) in signal processing and Latent Semantic Analysis (LSA) in text processing. It gets this name LSA because SVD applied to document-word matrices, groups documents that are semantically related to each other, even when they do not share common words [20]. For a very good and detailed tutorial in LSA see [21]. Words that usually occur in related contexts are also related in the same singular space. This method can be applied to extract the topic-words and content-sentences from documents. The advantage of using LSA vectors for summarization rather than the word-vectors is that conceptual (or semantic) relations *as represented in human brain* are automatically captured in the LSA [22], while using word vectors without the LSA transformation requires design of explicit methods to derive conceptual relations. Since SVD finds *principal* and *mutually orthogonal* dimensions of the sentence vectors, picking out a representative sentence from each of the dimensions ensures relevance to the document, and orthogonality ensures non-redundancy [23]. It is to be noted that this property applies only to data that *has principal dimensions inherently*—however, LSA would probably work since most of the text data has such principal dimensions owing to the variety of topics it addresses.

LSA is also used for visualization of topics, summaries, and distribution of documents/sentences on each of the topics—One such comprehensive tool is described in [24].

7 Issues in Summarization

- Temporal ordering: Information has to be presented in a chronological order. Conflicts in information have to be resolved. Temporal normalization is required. For example in newswire articles, the word *today* means different dates in different articles. Also words like next week, Monday etc need to be resolved.
- Algorithms that make use of clustering methods should be designed carefully around the problem of data sparseness. Due to the high dimensionality and data sparseness, the clustering algorithms often fail in forming meaningful clusters of passages/documents
- Evaluation methodologies are not yet standardized.

PASSAGE CLUSTERING IN THE CONTEXT OF MMR

8 Study of Quality-of-Summary in MMR with respect to Granularity of Clusters

Maximum Marginal Relevance Multi-Document (MMR-MD) summarizer is the most systematic approach to sentence-extractive summarization available to date. It accounts for most of the features mentioned in literature, and allows for flexibility to include or exclude any of these features. It also does not impose restrictions on the choice of sentence representations. For example, passages may be represented as vectors obtained by Latent Semantic Analysis or Generalized Vector Space Model.

MMR-MD summarizer defines criteria for selecting sentences to compose a summary. Its aim is to increase the relevance of the sentences to the query, while maintaining minimum redundancy in the summary.

The core algorithm for sentence scoring and selection for summarization, has been reproduced in Figure 3 from [2]. In the following discussion, a passage can be a sentence or a paragraph or any other such appropriately chosen text unit. As shown in [2], the following features of a passage improve the candidature of the passage to belong to the summary:

(+) 1. Similarity to query.
(+) 2. Coverage of the passage w.r.t the cluster(s) that it belongs to.
(+) 3. Content in the passage, eg., proper nouns, dates, etc.
(+) 4. Time Sequence: Passages that are more recent are considered more important (or more accurate with information).

Features that reduce the candidature of passages are:

(-) 5. Similarity to passages already included in the summary
(-) 6. Belonging to the cluster that has already contributed a passage to the summary
(-) 7. Belonging to a document that has already contributed a passage to the summary

Passages that have negative (-) features can easily be seen to be redundant to the summary generated thus far.

To restate the objective of the work presented in this report, in the light of the above discussion: the objective is to study how the quality of the summary varies with a change in the granularity of the passage clusters described in MMR summarizer. No systematic effort has been reported in the literature so far and the work described below is an attempt to fill this gap.

8.1 SYSTEM DESCRIPTION

The MMR-MD system that was provided at the beginning of the project was a partial implementation of the MMR-MD summarizer. It included only the following functionality out of those described in the reference [2].
Segmentation of documents into passages: Passages are taken to be sentences. Segmentation is performed by looking for sentence terminating characters-., ! or ?, and avoiding segmentation at 'Ms.', 'Jr.' and so on.
Relevance to query**:** Query relevance is computed by cosine similarity. If no query is given, the first sentence of the document is taken as auto-query.
Content--Pronoun penalty**:** Passages containing pronouns are penalized by a factor defined by the user.

Thus, the previous version of summarizer takes into account only points 1, 3 and 5 of those mentioned in Section 8. The features that are not taken into account are time-sequence, clustering of passages and clustering of documents.

8.1.1 COMPONENTS ADDED TO THE SYSTEM AS A PART OF THIS PROJECT:

Module to read DUC format data:
Passage Clustering: A bottom-up clustering module has been added to cluster passages based on similarity.
Evaluation module: An evaluation module to compare the automatically generated summary to human generated summary has been added to the system. Details of the evaluation strategy are described later in the report.

8.2 PASSAGE CLUSTERING

Passages are represented as word vectors like in general information retrieval sense. Similarity metric used is Cosine similarity.

The following two points are to be noted regarding clustering at the sentence level:
- Sentence-vectors are very sparse, that is, its value in most of the dimensions is zero.
- Similarity measure of sentence-vectors typically employed, such as cosine similarity or product of vectors, is more "hamming" like. Thus only those dimensions that are non-zero in both the vectors under consideration, contribute to the similarity mass.

Because of these two points above, traditional clustering algorithms do not always lead to meaningful clusters.

Even when the cosine-similarity of sentences is high, owing to more than a couple of common words between the sentences, the two sentences may be semantically much different—thus, they might look much "dissimilar" to a human reader, even though they score high on similarity measure. This aberration becomes much more unacceptable, in the process of clustering. This is because, in this process, a sentence is merged with a cluster when its similarity to the centroid is higher than a threshold. Whereas, a high-similarity to the centroid can be achieved even if the sentence shares a word each with 2-3 different sentences in the cluster. This means, the sentence is not similar to any one of the sentences in the cluster, but gets tagged to that cluster as a member. ***Thus the clustering looks very arbitrary. The clustering works pretty well in bringing together "synonymous" sentences.*** However, clustering of synonymous sentences is not very useful in case of MMR-MD summarization, as discussed further below.

Consider the following two sentences

1. President Fidel V. Ramos yesterday lashed at his critics who accused him of being ``preoccupied" with campaigning for the Lakas-NUCD national candidates and not with fixing the economy.

2. At least 100,000 candidates are expected to run for 17,340 national and local posts on May 11, including more than 80 people seeking to succeed President Fidel Ramos, whose single six-year term ends in June.

These two sentences share five words in common (president, fidel, ramos, national, candidates), and hence the cosine similarity ~0.3. However, in the context of all the sentences that belong to the documents describing this particular presidential election, these two would be called “dissimilar” by human judges.

The above observations are also supported by passage clustering by k-means algorithm. The clusters were totally arbitrary and suffered from *“large cluster gets larger”* problem. This is a direct consequence of the properties discussed above.

Given all the above observations, agglomerative clustering appeared to be a better choice for passage clustering, and has been used in this study. The algorithm used is described below:

Let $P_{1,}P_{2,}...,P_N$ be the collection of passages from all the documents.

Initialization:

Define clusters $C_{1,}C_{2,}...,C_N$ such that $P_1 \in C_1, P_2 \in C_2,..., P_N \in C_N$, and so on.

Merge Clusters: Let $Sim(C_i, C_j)$ be the cosine similarity between centroids of clusters C_i and C_j. Merge clusters C_i and C_j where,

$$Sim(C_i, C_j) = \max_{m,n}(Sim(C_m, C_n))$$

and

$$Sim(C_m, C_n) > threshold$$

Termination:

Stop merging clusters when

$$\max_{m,n}(Sim(C_m, C_n)) < threshold$$

A threshold on similarity between clusters is used to determine whether or not to merge two clusters. This threshold controls granularity.

- If the threshold is too high only *almost identical* sentences group together, none other.
- If the threshold is too low, clustering is not meaningful. Sentences belonging to a cluster share not more than a word or two between one another.

Thus, in this study of relevance of cluster size, only a small range of similarity threshold, between 0.3-0.5 was found meaningful.

As a next step in the study, the passage that has maximum coverage to the cluster is chosen as a representative passage. The maximum coverage passage is that which has maximum similarity to the centroid of the document. It is possible that each cluster requires more than one sentence to represent its content.

However, owing to the points discussed above, namely,

- meaningful clustering happens only for a limited threshold, and
- sentences in these clusters are "very similar"
- going out of this range of thresholds makes clustering arbitrary

suggests that *not more than one passage would be required from each cluster*.

Hence, the representative passages from all the clusters are collected, and an MMR summarization is performed on this collection similar to what was done originally without clustering.

8.3 DATA SET

Data provided by Document Understanding Conference (DUC 2002) has been used to study the Quality of Summary. The data consists of newswire articles of about 38 events. Each event is reported by four or five articles. Two human judges individually summarize the newswire articles pertaining to each event. One summary is created per event per judge. The summary consists of sentences chosen from the original texts. The corpus provides human generated summaries of length 200 words and 400 words each. In this study the 200 word summaries have been used. Length of the MMR summary is fixed at 10 sentences.

8.4 EVALUATION METRIC

Quality of Summary (QoS) is studied in comparison to the human-judged summaries. QoS is computed for each human-judged summary separately and is then averaged across all of them. This process is repeated for each of the events, and an overall average across events is computed. Thus, in the Figure 5 below, average similarity refers to "average similarity of MMR summarization across human judges across all events".

Similarity between MMR summary and Human-Judged (HJ) summary is computed as follows:

1. Find one-to-one similarity between all sentences in MMR and those in HJ.
2. Set Cumulative similarity to zero.
3. Pick the two sentences, one from MMR summary and one from HJ summary, with maximum similarity.
4. Add this similarity to cumulative similarity
5. Delete these two sentences from the summaries.
6. Repeat steps 3 and 4 till there some more sentences left in either of the two summaries.
7. Divide cumulative similarity by number of paired sentences to get average similarity between MMR and HJ.

8.5 RESULTS

Conclusion-1: ***It is difficult to achieve meaningful clustering of sentences with traditional distance-based clustering methods.***

Table 1 shows the granularity and tightness of clusters achieved at various thresholds. A threshold of 1.0 in the clustering procedure above in effect means no two passages will be merged to form a cluster since no two can satisfy the condition $Sim(C_m, C_n) > threshold$; this means, that every passage forms a cluster of its own.

Tightness of clusters is represented by average intra-cluster similarity, which in turn is the average similarity between passages within the cluster. Average intra-cluster similarity is computed as

$$AvgSim = \frac{\sum_i AvgSim_i(C_i) * |C_i|}{\sum_i |C_i|}$$

$$= \frac{\sum_i \sum_{s_j \in C_i} Sim(s_j, centroid(C_i))}{\sum_i |C_i|}$$

where, *Sim* is the cosine similarity measure and $|C_i|$ is the number of passages in cluster C_i. Centroid(C_i) is the sum of all passages vectors in the cluster C_i, and represents its centroid.

Threshold	Number of Clusters	Average Cluster Similarity	Number of Clusters of different sizes **x(i) :=** x number of clusters of size i ***segments***
0.3	54	0.51	1(**84**), 2(**6**), 1(**5**), 2(**4**), 7(**3**), 15(**2**), 26(**1**)
0.4	90	0.81	3(**13**), 1(**11**), 1(**5**), 4(**4**), 6(**3**), 25(**2**), 50(**1**)
0.5	121	0.89	1(**11**), 1(**10**), 1(**5**), 3(**4**), 6(**3**), 21(**2**), 78(**1**)
0.6	153	0.96	1(**6**), 5(**3**), 18(**2**), 129(**1**)
0.7	165	0.98	2(**3**), 17(**2**), 146(**1**)
0.8	177	0.99*	9(**2**), 169(**1**)
0.9	180	0.99*	6(**2**), 174(**1**)
1.0	182	1.00*	182(**1**)
**Note here, that these thresholds create almost no clusters. Hence the average similarity is that of every sentence with itself, which is 1.*			

Table 1: Granularity and tightness of clusters at various thresholds

Conclusion-2: ***Prior clustering of passages does not have any additional influence on quality of summary—MMR already has the effect of passage clustering in an implicit way. The very minor effect it has, is only to reduce the quality of the summary.***

The primary result of this study is shown in Figure 5. "Average Similarity" in this figure stands for "average similarity of MMR summarization across human judges across all events". The following two observations can be made from this figure:

1. The variation in QoS with clustering is very minor—it varies only between 0.4-0.5.
2. Even that minor variation favors absence of clustering. Note again, from Table 1 that for thresholds 0.6-1.0, almost no clustering took place between sentences. Hence when the sentences were grouped together based on similarity, the QoS only dropped. Hence, it can be concluded that MMR-MD works very well even without Passage Clustering.

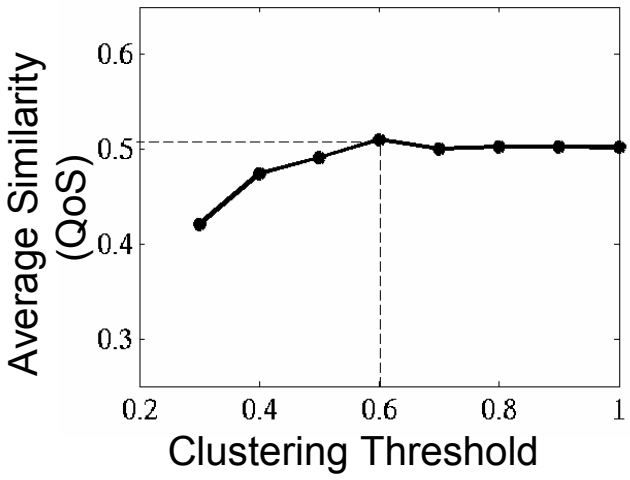

Figure 5. Average Similarity of Machine-Generated Summary to Human-Judged Extractive Summary

For further analysis, we compare one sample MMR-MD generated summary, at various thresholds of clustering. We take the summary with the threshold 1.0 as the base case and study how decreasing the threshold to allow for larger clusters affects the selected sentences. This comparison is shown in Table 2. It can be seen that though almost all the chosen sentences in the summary are different from a clustered case (threshold 0.3 or 0.4) to the unclustered case (1.0), the QoS has had very little affect (Figure 5). This indicates that the sentences in clustering must have been very good, and that the "representative sentence that has maximum coverage" may not be what would be chosen for a good summary from a given cluster. Since the additional computing time and effort of clustering sentences did not contribute any additional advantage, it may be removed.

Threshold	Number of passages **different** between this summary and summary with threshold 1.0
0.3	9
0.4	7
0.5	5
0.6	1
0.7	2
0.8	1
0.9	1
1.0	0

Table 2: Commonality between summaries at various thresholds and that with threshold 1.0

8.6 TOPICAL MMR: ALTERNATIVE APPROACH TO CLUSTERING FOR IMPROVING QUALITY OF SUMMARY

A method for improving the quality of the summary would be to cluster sentences based on *topics* rather than on sentence similarity. If the primary topics discussed in the document collection can be identified either explicitly by users or by word co-occurrence methods, sentences can be clustered into those topics.

One difference in the method of summarization with topic-based clustering to that based on sentence similarity is: Clusters formed this way are likely to contain more than one sentence that is important to the summary, and the number of sentences each cluster contributes is dependant on the relevance of that cluster to the query. Hence, the following steps may be adopted to generate a meaningful summary:

- Identify *topics* or *issues* described in the collection of documents.
- Form clusters of passages based on these topics.
- For each cluster, generate an MMR summary independently. The ratio of the length of the summary generated from a cluster to the length of the final summary desired is proportional to the relevance of the cluster to the query. In case of generic summary, where a query absent, an auto query can be used, as in MMR [2].

$$\frac{N_{C_i}}{N} \propto Sim(C_i, Q)$$

where, N_{C_i} is the number of sentences in the summary generated from cluster *i, N* is the number of sentences in the total summary and $Sim(C_i, Q)$ is the similarity of cluster i to query Q.

- Concatenate summaries from each of the clusters and use another round of MMR to ensure there are no redundant sentences coming from different clusters.

Generating Topics

Given a set of related documents, it is most likely that the documents address a few closely related but distinct issues. Each of these different issues, or topics, may be described by a set of keywords either by human or be generated automatically.

In situations where human intervention is not possible, the topics may be identified by using Graph Theoretic approach described in Section 6.1.4.

Let w_1, w_2, ..., w_M be the most frequent M words in the collection of documents. Cluster these words in the following way.

Initialization:

For i = 1,2…M, define clusters $C_{0,i}$, such that $C_{0,i}$, contains all the passages containing the word w_i.

Iteration:

For all i,j such that $Cc(C_{n,i}, C_{n,j}) > threshold$

$merge(C_{n,i}, C_{n,j})$

where, n is the number of iteration, $Cc(C_{n,i}, C_{n,j})$ is the *Co-location Count* defined as the number of passages containing atleast one word from each of the clusters $C_{n,i}$ and $C_{n,j}$. $Merge(C_{n,i}, C_{n,j})$ by creating a new cluster with all the passages in clusters $C_{n,i}$ and $C_{n,j}$. Mark clusters $C_{n,i}$ and $C_{n,j}$ for deletion in iteration n.

Delete all $C_{n,i}$ that are marked for deletion in this iteration before going to next iteration. Rename remaining $C_{n,i}$ as $C_{n+1,i}$

Termination:

Stop merging clusters when

$max(Cc(C_{n,i},C_{n,j})) < threshold$

This produces clusters of passages, such that each cluster contains passages addressing a particular topic.

To produce a summary that covers all the important issues discussed in the document, it is necessary to include information from each of the topics identified. This may be achieved by following these steps:

1. Generate MMR-WC summaries for each of these categories/clusters
2. Concatenate summaries from all the categories
3. Re-run MMR-WC summarizer on the summary to remove redundant passages if any

The method proposed above derives its strength by combining the novel features of – MMR-MD summarization and Graph Theoretic method of partitioning the documents. It is fairly straightforward and intuitive to see that the proposed method would perform significantly better than the traditional MMR techniques with and without granularity control.

9 Key Observation and Conclusion

Owing to the very sparse sentence vectors and also due to the nature of the hamming-like distance measures employed for sentence similarities, clustering of sentences based on similarity metric, is meaningful only for a small range of similarity thresholds: ~0.3 to 0.5 (Table 1). It can be seen that for larger thresholds the clusters contain just one sentence each (which essentially means that *clustering does not take place*). Lower thresholds of 0 to 0.2 are meaningless since they join together *any* and *all* sentences together into one or two clusters—which is equivalent to no-clustering.

The second term in MMR algorithm (1st line in Figure 3), namely $-(1-\lambda)$ *max* (Sim_2) already ensures that a sentence similar to previously selected sentences is not chosen, which in other words means a sentence from same cluster as previous sentences is not selected. This is reflected in Figure 5 QoS lies between 0.4 and 0.5 with very little variation for the entire range of cluster granularity, including the case where the passages are not clustered. Thus, there seems to be no additional advantage of pre-computing of the passage clusters.

The primary conclusion of this study is that passage clustering has little affect on the quality of summary. Whatever little effect it has is only to "reduce" the quality of summary.

Based on a careful study of results of passage clustering, the MMR-MD algorithm, and the literature surveyed, a new method has been proposed, which has all the features and strengths of MMR-MD summarization. It would have an additional advantage that it identifies different topics covered in the document collection, and addresses them individually. It is beyond the scope of this project to implement the new approach and study the results. It is left for implementation in some future project. However, the observations that led to the formulation of this novel technique, Topical MMR, indicate that the proposed technique is likely to perform better for documents with inherent "topical" nature, such as news reports.

10 References

1. Mani, I., *Automatic Summarization*. 2001: John Benjamin's Publishing Company.

2. Goldstein, J., V. Mittal, and J. Carbonell. *Creating and Evaluating Multi-Document Sentence Extract Summaries*. in *CIKM'00: Ninth International Conference on Information Knowledge Management*. 2000.

3. Hahn, U. and I. Mani, *The Challenges of Automatic Summarization.* Computer, 2000: p. 29-36.

4. Mittal, V. and e. al, *Selecting Text Spans for Document Summaries: Heuristics and Metrics.*

5. Goldstein, J., et al., *Sentence Selection and Evaluation Metrics*.

6. Goldstein, J., et al. *Summarizing Text Documents: Sentence Selection and Evaluation Metrics*. in *ACM-SIGIR'99*. 1999. Berkeley, CA, USA.

7. Edmundson, H.P., *New Methods in Automatic Extracting.* Journal of the ACM, 1969. **16**(2): p. 264-285.

8. Radev, D.R., H. Jing, and M. Budzikowska. *Centroid-Based Summarization of Multiple Documents: Sentence Extraction, Utility-Based Evaluation and User Studies*. in *ANLP/NAACL 2000 Workshop*. 2000.

9. Radev, D.R., W. Fan, and Z. Zhang. *Webinessence: A Personalizedweb-Based Multi-Document Summarization and Recommendation System*. in *NAACL'01 Workshop on Automatic Summarization*. 2001. Pittsburgh, PA, USA.

10. Jing, H., et al., *Summarization Evaluation Methods: Experiments and Analysis.*

11. Goldstein, J., *The Use of Genre in Summarization.*

12. Radev, D.R. and e. al, *Experiments in Single and Multi-Document Summarization Using Mead.*

13. Gong, Y. and X. Liu. *Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis*. in *SIGIR-2001*. 2001.

14. Jing, H. and K.R. McKeown, *The Decomposition of Human-Written Summary Sentences.*

15. Endres-Niggemeyer, B., *Human-Style Www Summarization.* 2000.

16. Edmundson, H.P. and R.E. Wyllys, *Automatic Abstracting and Indexing-Survey and Recommendations.* Communications of the ACM, 1961. **4**(5): p. 226-234.

17. Witbrock, M.J. and V. Mittal. *Ultra-Summarization: A Statistical Approach to Generating Highly Condensed Non-Extractive Summaries*. in *SIGIR'99*. 1999.

18. Neto, J.L. and A.D. Santos. *Document Clustering and Summarization*. in *PADD'00: 4th Int. Conference on Practical Applications of Knowledge Discovery and Data Mining*. 2000. London.

19. Carbonell, J. and J. Goldstein. *The Use of Mmr, Diversity-Based Reranking for Reordering Documents and Producing Summaries*. in *SIGIR'98*. 1998. Melbourne, Australia.

20. Wolfe, M., et al., *Learning from Text: Matching Readers and Texts by Latent Semantic Analysis*. 1998.

21. Bellegarda, J.R., *Exploiting Latent Semantic Information in Statistical Language Model.* Proceedings of the IEEE, 2000. **88**(8).

22. Landauer, T.K. and S.T. Dumais, *A Solution to Plato's Problem: Thelatent Semantic Analysis Theory of Acquisition, Induction Andrepresentation of Knowledge.* Psychological Review, 1997. **104**: p. 211-240.

23. Gong, Y. and X. Liu. *Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis*. in *SIGIR'01*. 2001.

24. Ando, R.K., et al. *Multi-Document Summarization by Visualizing Topical Content*. in *NAACL-2000.* 2000.