SUMMARIZATION: (1) USING MMR FOR DIVERSITY- BASED RERANKING AND (2) EVALUATING SUMMARIES

Jade Goldstein and Jaime Carbonell
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213 USA
jade@cs.cmu.edu, jgc@cs.cmu.edu

ABSTRACT:

This paper[1] develops a method for combining query-relevance with information-novelty in the context of text retrieval and summarization. The Maximal Marginal Relevance (MMR) criterion strives to reduce redundancy while maintaining query relevance in re-ranking retrieved documents and in selecting appropriate passages for text summarization. Preliminary results indicate some benefits for MMR diversity ranking in ad-hoc query and in single document summarization. The latter are borne out by the trial-run (unofficial) TREC-style evaluation of summarization systems. However, the clearest advantage is demonstrated in the automated construction of large document and non-redundant multi-document summaries, where MMR results are clearly superior to non-MMR passage selection. This paper also discusses our preliminary evaluation of summarization methods for single documents.

1. INTRODUCTION

With the continuing growth of online information, it has become increasingly important to provide improved mechanisms to find information quickly. Conventional IR systems rank and assimilate documents based on maximizing relevance to the user query [1, 8, 6, 12, 13]. In cases where relevant documents are few, or cases where very-high recall is necessary, pure relevance ranking is very appropriate. But in cases where there is a vast sea of potentially relevant documents, highly redundant with each other or (in the extreme) containing partially or fully duplicative information we must utilize means beyond pure relevance for document ranking.

In order to better illustrate the need to combine relevance and anti-redundancy, consider a reporter or a student, using a newswire archive collection to research accounts of airline disasters. He composes a well-though-out query including "airline crash", "FAA investigation", "passenger deaths", "fire", "airplane accidents", and so on. The IR engine returns a ranked list of the top 100 documents (more if requested), and the user examines the top-ranked document. It's about the suspicious TWA-800 crash near Long Island. Very relevant and useful. The next document is also about "TWA-800", so is the next, and so are the following 30 documents. Relevant? Yes. Useful? Decreasingly so. Most "new" documents merely repeat information already contained in previously offered ones, and the user could have tired long before reaching the first non-TWA-800 air disaster document. Perfect precision, therefore, may prove insufficient in meeting user needs.

A better document ranking method for this user is one where each document in the ranked list is selected according to a combined criterion of query relevance and novelty of information. The latter measures the degree of dissimilarity between the document being considered and previously selected ones already in the ranked list. Of course, some users may prefer to drill down on a narrow topic, and others a panoramic sampling bearing relevance to the query. Best is a user-tunable method that focuses the search from a narrow beam to a floodlight. Maximal Marginal Relevance (MMR) provides precisely such functionality, as discussed below.

If we consider document summarization by relevant-passage extraction, we must again consider anti-redundancy as well as relevance. Both query-free summaries and query-relevant summaries need to avoid redundancy, as it defeats the purpose of summarization. For instance, scholarly articles often state their thesis in the introduction, elaborate upon it in the body, and reiterate it in the conclusion. Including all three in versions in the summary, however, leaves little room for other useful information. If we move beyond single document summarization to document cluster summarization, where the summary must pool passages

[1] This research was performed as part of Carnegie Group Inc.'s Tipster III Summarization Project under the direction of Mark Borger and Alex Kott.

from different but possibly overlapping documents, reducing redundancy becomes an even more significant problem.

Automated document summarization dates back to Luhn's work at IBM in the 1950's [12], and evolved through several efforts including Tait [24] and Paice in the 1980s [17, 18]. Much early work focused on the structure of the document to select information. In the 1990's several approaches to summarization blossomed, include trainable methods [10], linguistic approaches [8, 15] and our information-centric method [2], the first to focus on query-relevant summaries and anti-redundancy measures. As part of the TIPSTER program [25], new investigations have started into summary creation using a variety of strategies. These new efforts address query relevant as well as "generic" summaries and utilize a variety of approaches including using co-reference chains (from the University of Pennsylvania) [25], the combination of statistical and linguistic approaches (Smart and Empire) from SaBir Research, Cornell University and GE R&D Labs, topic identification and interpretation from the ISI, and template based summarization from New Mexico State University [25].

In this paper, we discuss the Maximal Marginal Relevance method (Section 2), its use for document reranking (Section 3), our approach to query-based single document summarization (Section 4), and our approach to long documents (Section 6) and multi-document summarization (Section 6). We also discuss our evaluation efforts of single document summarization (Section 7-8) and our preliminary results (Section 9).

2. MAXIMAL MARGINAL RELEVANCE

Most modern IR search engines produce a ranked list of retrieved documents ordered by declining relevance to the user's query [1, 18, 21, 26]. In contrast, we motivated the need for "'relevant novelty" as a potentially superior criterion. However, there is no known way to directly measure new-and-relevant information, especially given traditional bag-of-words methods such as the vector-space model [19, 21]. A first approximation to measuring relevant novelty is to measure relevance and novelty independently and provide a linear combination as the metric. We call the linear combination "marginal relevance" -- i.e. a document has high marginal relevance if it is both relevant to the query and contains minimal similarity to previously selected documents. We strive to maximize marginal relevance in retrieval and summarization, hence we label our method "maximal marginal relevance" (MMR).

The Maximal Marginal Relevance (MMR) metric is defined as follows:

Let C = document collection (or document stream)
Let Q = ad-hoc query (or analyst-profile or topic/category specification)
Let R = IR (C, Q, q) – i.e. the ranked list of documents retrieved by an IR system, given C and Q and a relevance threshold theta, below which it will not retrieve documents. (q can be degree of match, or number of documents).
Let S = subset of documents in R already provided to the user. (Note that in an IR system without MMR and dynamic reranking, S is typically a proper prefix of list R.) R\S is the set difference, i.e. the set of documents in R, not yet offered to the user.

$$MMR(C,Q,R,S) \overset{def}{=} \underset{D_i \in R \setminus S}{Argmax}[\lambda * Sim_1(D_i,Q) - (1-\lambda) \underset{D_j \in S}{Max}(Sim_2(D_i,D_j))]$$

Given the above definition, MMR computes incrementally the standard relevance-ranked list when the parameter λ=1, and computes a maximal diversity ranking among the documents in R when λ=0. For intermediate values of λ in the interval [0,1], a linear combination of both criteria is optimized. Users wishing to sample the information space around the query, should set λ at a smaller value, and those wishing to focus in on multiple potentially overlapping or reinforcing relevant documents, should set λ to a value closer to 1. For document retrieval, we found that a particularly effective search strategy (reinforced by the user study discussed below) is to start with a small λ (e.g. λ = .3) in order to understand the information space in the region of the query, and then to focus on the most important parts using a reformulated query (possibly via relevance feedback) and a larger value of λ (e.g. λ = .7).

Note that the similarity metric Sim_1 used in document retrieval and relevance ranking between documents and query could be the same as Sim_2 between documents (e.g., both could be cosine similarity), but this need not be the case. A more accurate, but computationally more costly metric could be used when applied only to the elements of the retrieved document set R, given that |R| << |C|, if MMR is applied for re-ranking the top portion of the ranked list produced by a standard IR system.

query : Brazil external debt figure			
		λ	
Article Title	1	0.7	0.3
BRAZIL SEEN AS VANGUARD FOR CHANGING DEBT STRATEGY	76	76	76
FUNARO REJECTS UK SUGGESTION OF IMF BRAZIL PLAN	1308	1308	1293
ECONOMIC SPOTLIGHT - BRAZIL DEBT DEADLINES LOOM	1431	1431	1308
U.S. URGED TO STRENGTHEN DEBT STRATEGY	104	**2149**	133
U.S. URGES BANKS TO DEVELOP NEW 3RD WLD FINANCE	50	104	14
FUNARO'S DEPARTURE COULD LEAD TO BRAZIL DEBT DEAL	**2149**	**1388**	**1388**
U.S. OFFICIALS SAY BRAZIL SHOULD DEAL WITH BANKS	1713	1293	1762
BRAZIL SEEKS TO REASSURE BANKS ON DEBT SUSPENSION	**1388**	1713	**2149**
BRAZIL SEEKS TO REASSURE BANKS ON DEBT SUSPENSION	**1403**	50	**69**
BRAZIL CRITICISES ADVISORY COMMITTEE STRUCTURE	1291	133	1713
LATIN DEBTORS MAKE NEW PUSH FOR DEBT RELIE	32	1291	104
BRAZIL DEBT SEEN PARTNER TO HARD SELL TACTICS	99	99	1431
BRAZIL DEBT POSES THORNY ISSUE FOR U.S. BANKS	54	14	99
U.S. URGES BANKS TO WEIGH PHILIPPINE DEBT PLAN	44	54	1291
U.K. SAYS HAS NO ROLE IN BRAZIL MORATORIUM TALKS	1293	32	54
TALKING POINT/BANK STOCKS	53	**69**	44
CANADA BANKS COULD SEE PRESSURE ON BRAZIL LOANS	1762	1762	32
TREASURY'S BAKER SAYS BRAZIL NOT IN CRISIS	133	44	50
BRAZIL'S DEBT CRISIS BECOMING POLITICAL CRISIS	14	**1403**	**1403**
BAKER AND VOLCKER SAY DEBT STRATEGY WILL WORK	**69**	53	53

Table 1: Initial Relevance Ranking ($\lambda = 1$) vs. MMR reranking ($\lambda = .7$ & $\lambda = .3$)

3. DOCUMENT REORDERING

We implemented MMR in two retrieval engines, PURSUIT (an upgraded version of the original retrieval engine inside the Lycos^{TM} search engine), [9] and SMART (the publicly available version of the Cornell IR engine) [1]. Using the scoring functions available in each system for both Sim_1 and Sim_2, we obtained consistent and expected results in the behavior of the two systems.

The results of MMR reranking are shown in Table 1. In this Reuters document collection, article 1403 is a duplicate of 1388. MMR reranking performs as expected, for decreasing values of l, the ranking of 1403 drops. Also as predicted, novel but still relevant information as evidenced by document 69 starts to increase in ranking. Relevant, but similar to the highest ranked documents, such as document 1713 drop in ranked ordering. Document 2149 's position varies depending on its similarity to previously seen information.

We also performed a pilot experiment with five users who were undergraduates from various disciplines. The purpose of the study was to find out if they could tell what was the difference between the standard ranked document order retrieved by SMART and a MMR reranked order with $\lambda = ..5$. They were asked to perform nine different search tasks to find information and asked various questions about the tasks. They used two methods to retrieve documents, known only as R and S. Parallel tasks were constructed so that one set of users would perform method R on one task and method S on a similar task. Users were not told how the documents were presented only that either "method R" or "method S" were used and that they needed to be try to distinguish the differences between methods. After each task we asked them to record the information found. We also asked them to look at the ranking for method R and method S and see if they could tell any difference between the two. The majority of people said they preferred the method which gave in their opinion the most broad and interesting topics. In the final section they were asked to select a search method and use it for a search task. 80% (4 out of 5) chose the method MMR to use. The person who chose Smart stated it was because "it tends to group more like stories together." The users indicated a differential preference for MMR in navigation and for locating the relevant candidate documents more quickly, and pure-relevance ranking when looking at related documents within that band. Three of the five users clearly discovered the differential utility of diversity search and relevance-only search. One user explicitly stated his strategy:

> *"Method R [relevance only] groups items together based on similarity and Method S [MMR re-ranking] gives a wider array. I would*

use Method S [MMR re-ranking] to find a topic ... and then use Method R [relevance-only] with a specific search from Method S [MMR re-ranking] to yield a lot of closely related items."

The initial study was too small to yield statistically significant trends with respect to speed of known-item retrieval, or recall improvements for broader query tasks. However, based on our own experience and questionnaire responses from the five users, we expect that task demands play a large role with respect to which method yields better performance.

4. SINGLE DOCUMENT SUMMARIES

Human summarization of documents, sometimes called "abstraction" is a fixed-length *generic* summary, reflecting the key points that the abstractor -- rather than the user -- deems important. Consider a physician evaluating a particular chemotherapy regimen who wants to know about its adverse effects to elderly female patients. The retrieval engine produces several lengthy reports (e.g. a 300-page clinical study), whose abstracts do not contain any hint of whether there is information regarding effects on elderly patients. A useful summary for this physician would contain *query-relevant* passages (e.g. differential adverse effects on elderly males and females, buried in page 211-212 of the clinical study) assembled into a summary. A different user with different information needs may require a totally different summary of the same document.

We developed a minimal-redundancy query-relevant summarizer-by-extraction method, which differs from previous work in summarization [10, 12, 15, 18, 24] in several dimensions.

- Optional query relevance: as discussed above a query or a user interest profile (for the vector sum of both, appropriately weighted) is used to select relevant passages. If a generic query-free summary is desired, the centroid vector of the document is calculated and passages are selected with the principal components of the centroid as the query.

- Variable granularity summarization: The length of the summary is under user control. Brief summaries are useful for indicative purposes (e.g. whether to read further), and longer ones for drilling and extracting detailed information.

- Non-redundancy: Information density is enhanced by ensuring a degree of dissimilarity between passages contained in the summary. The degree of query-focus vs. diversity sampling is under user control (the λ parameter in the MMR formula).

Our process for creating single document summaries is as follows:

1. Segment a document into passages and index the passages using the inverted indexing method used by the IR engine for full documents. Passages may be phrases, sentences, n-sentence chunks, or paragraphs. For the TIPSTER III evaluation, we used sentences as passages.
2. Within a document, identify the passages relevant to the query. Use a threshold below which the passages are discarded. We used a similarity metric based on cosine similarity using the traditional TF-IDF weights.
3. Apply the MMR metric as defined in Section 2 to the passages (rather than full documents). Depending on the desired length of the summary, select a few or larger number. If the parameter λ is not very close to 1, redundant query relevant passages will tend to be eliminated and other different, slightly less query relevant passages will be included. We allow the user to select the number of passages or the percentage of the document size (also known as the "compression ratio").
4. Reassemble the selected passages into a summary document using one of the following summary-cohesion criteria:

- Document appearance order: Present the segments according to their order of presentation in the original document. If the first sentence is longer than a threshold, we automatically include this sentence in the summary as it tends to set the context for the article. If the user only wants to view a few segments, the first sentence must also meet a threshold for sentence rank to be included.
- News-story principle: Present the information in MMR-ranked order, i.e., the most relevant and most diverse information first. In this manner, the reader gets the maximal information even if they stop reading the summary. This allows the diversity of relevant information to be presented earlier and topic introduced may be revisited after other relevant topics have been introduced.
- Topic-cohesion principle: First group together the document segments by topic clustering (using sub-document similarity criteria). Then rank the centroids of each cluster by MMR (most important first) and present the information, a

topic-coherent cluster at a time, starting with the cluster whose centroid ranks highest.

We implemented query-relevant document-appearance-based sequencing of information. Our method of summarization does not require the more elaborate language-regeneration needed by Kathy McKeown and her group at Columbia in their summarization work [15]. As such our method is simpler, faster and more widely applicable, but yields potentially less cohesive summaries. All summary results in this paper use the SMART search engine with stopwords eliminated from the indexed data and stemming.

Query: Delaunay refinement mesh generation finite element method foundations three dimension analysis; $\lambda = .3$

[1] Delaunay refinement is a technique for generating unstructured meshes of triangles or tetrahedra suitable for use in the finite element method or other numerical methods for solving partial differential equations.
[5] The purpose of this thesis is to further this progress by cementing the foundations of two-dimensional Delaunay refinement, and by extending the technique and its analysis to three dimensions.
[15] Nevertheless, Delaunay refinement methods for tetrahedral mesh generation have the rare distinction that they offer strong theoretical bounds and frequently perform well in practice.
[39] If one can generate meshes that are completely satisfying for numerical techniques like the finite element method, the other applications fall easily in line.
[131] Our understanding of the relative merit of different metrics for measuring element quality, or the effects of small numbers of poor quality elements on numerical solutions, is based as much on engineering experience and rumor as it is on mathematical foundations.
[158] Delaunay refinement methods are based upon a well-known geometric construction called the Delaunay triangulation, which is discussed extensively in the mesh generation chapter.
[201] I first extend Ruppert's algorithm to three dimensions, and show that the extension generates nicely graded tetrahedral meshes whose circumradius-to-shortest edge ratios are nearly bounded below two.
[2250] Refinement Algorithms for Quality Mesh Generation: Delaunay refinement algorithms for mesh generation operate by maintaining a Delaunay or constrained Delaunay triangulation, which is refined by inserting carefully placed vertices until the mesh meets constraints on element quality and size.
[3648] I do not know to what difference between the algorithms one should attribute the slightly better bound for Delaunay refinement, nor whether it marks a real difference between the algorithms or is an artifact of the different methods of analysis.

Figure 1: Generic MMR- generated summary of dissertation.

Query: sliver mesh boundary removal small angles; $\lambda = .7$

[1] Delaunay refinement is a technique for generating unstructured meshes of triangles or tetrahedra suitable for use in the finite element method or other numerical methods for solving partial differential equations.
[129] Hence, many mesh generation algorithms take the approach of attempting to bound the smallest angle.
[2621] Because s is locked, inserting a vertex at c will not remove t from the mesh.
[2860] Of course, one must respect the PSLG; small input angles cannot be removed.
[3046] The worst slivers can often be removed by Delaunay refinement, even if there is no theoretical guarantee.
[3047] Meshes with bounds on the circumradius-to-shortest edge ratios of their tetrahedra are an excellent starting point for mesh smoothing and optimization methods designed to remove slivers and improve the quality of an existing mesh (see smoothing section).
[3686] If one inserts a vertex at the circumcenter of each sliver tetrahedron, will the algorithm fail to terminate?
[3702] A sliver can always be eliminated by splitting it, but how can one avoid creating new slivers in the process?
[3723] Unfortunately, my practical success in removing slivers is probably due in part to the severe restrictions on input angle I have imposed upon Delaunay refinement.
[3724] Practitioners report that they have the most difficulty removing slivers at the boundary of a mesh, especially near small angles.

Figure 2: Focused-query MMR-generated summary of dissertation.

5. SUMMARIZING LONGER DOCUMENTS

The MMR-passage selection method for summarization works better for longer documents (which typically contain more inherent passage redundancy across document sections such as abstract, introduction, conclusion, results, etc.). To demonstrate the quality of summaries that can be obtained for long documents, we summarized an entire dissertation containing 3,772 sentences with a generic topic query constructed by expanding the thesis title (Figure 1). In contrast, Figure 2 shows the results of a more specialized query with a larger λ value to focus summarization less on diversity and more on topic.

The above example demonstrates the utility of query relevance in summarization and the incremental utility of controlling summary focus via the lambda parameter. It also highlights a shortcoming of summarization by extraction, namely coping with antecedent references. Sentence [2621] refers to coefficients "s", "c", and "t," which do not make sense outside the framework that defines them. Such referential problems are ameliorated with increased passage length, for instance using paragraphs rather than sentences. However, longer-passage selection also implies longer summaries. Another solution is co-reference resolution [25].

6. MULTI-DOCUMENT SUMMARIES

As discussed earlier, MMR passage selection works equally well for summarizing single documents or clusters of topically related documents. Our method for multi-document summarization follows the same basic procedure as that of single document summarization (see section 4). In step 2 (Section 4), we identify the N most relevant passages from each of the documents in the collection and use them to form the passage set to be MMR re-ranked. N is dependent on the desired resultant length of the summary. We used N relevant passages from each document collection rather than the top relevant passages in the entire collection so that each article had a chance to provide a query-relevant contribution. In the future we intend to compare this to using MMR ranking where the entire document set is treated as a single document. Steps 2, 3 and 4 are primarily the same.

The TIPSTER evaluation corpus provided several sets of topical clusters to which we applied MMR summarization. In one such example on a cluster of apartheid-related documents, we used the topic description as the query (see Figure 3) and N was set to 4 (4 sentences per article were reranked). The top 10 sentences for $\lambda = 1$ (effectively query relevance, but no MMR) and $\lambda = .3$ (both query relevance and MMR anti-redundancy) are shown in Figures 4 and 5 respectively.

The summaries clearly demonstrate the need for MMR in passage selection. The $\lambda = 1$ case exhibits considerable redundancy, ranging from near-replication in passages [4] and [5] to redundant content in passages [7] and [9]. Whereas the $\lambda = .3$ case exhibits no such redundancy. Counting clearly distinct propositions in both cases yields a 20% greater information content for the MMR case, though both summaries are equivalent in length.

Topic:

<head> Tipster Topic Description
<num> Number: 110
<dom> Domain: International Politics
<title> Topic: Black Resistance Against the South African Government
<desc> **Description:**
Document will discuss efforts by the black majority in South Africa to overthrow domination by the white minority government.
<smry> **Summary:**
Document will discuss efforts by the black majority in South Africa to overthrow domination by the white minority government.
<narr> **Narrative:**
A relevant document will discuss any effort by blacks to force political change in South Africa. The reported black challenge to apartheid may take any form -- military, political, or economic -- but of greatest interest would be information on reported activities by armed personnel linked to the African National Congress (ANC), either in South Africa or in bordering states.
<con> **Concept(s):**
1. African National Congress, ANC, Nelson Mandela, Oliver Tambo
2. Chief Buthelezi, Inkatha, Zulu
3. terrorist, detainee, subversive, communist
4. Limpopo River, Angola, Botswana, Mozambique, Zambia
5. apartheid, black township, homelands, group areas act, emergency regulations

Query:

Black Resistance Against South African Government black majority South Africa overthrow domination white minority government blacks force political change South Africa black challenge apartheid military political economic activities armed personnel African National Congress (ANC) South Africa bordering states African National Congress ANC Nelson Mandela Oliver Tambo Chief Buthelezi Inkatha Zulu terrorist detainee subversive communist Limpopo River Angola Botswana Mozambique Zambia apartheid black township homelands group areas act emergency regulations

Query (short version - no narrative or concepts):

Black Resistance South African Government black majority South Africa overthrow domination white minority government

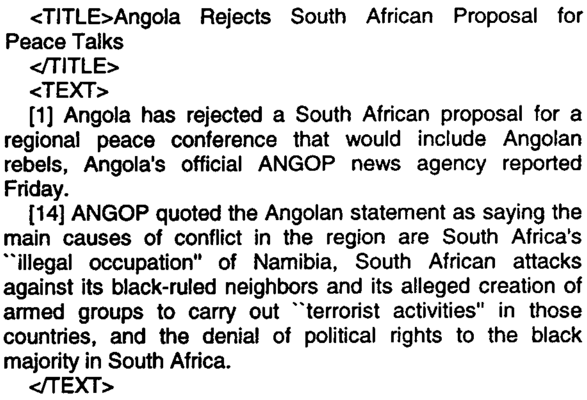
Figure 3: Topic and Query for Tipster Topic 110

[1] [761] AP880212-0060 [15] ANGOP quoted the Angolan statement as saying the main causes of conflict in the region are South Africa's ``illegal occupation" of Namibia, South African attacks against its black-ruled neighbors and its alleged creation of armed groups to carry out ``terrorist activities" in those countries, and the denial of political rights to the black majority in South Africa.
[2] [758] AP880803-0080 [25] Three Canadian anti-apartheid groups issued a statement urging the government to sever diplomatic and economic links with South Africa and aid the African National Congress, the banned group fighting the white-dominated government in South Africa.
[3] [756] AP880803-0082 [25] Three Canadian anti-apartheid groups issued a statement urging the government to sever diplomatic and economic links with South Africa and aid the African National Congress, the banned group fighting the white-dominated government in South Africa.
[4] [790] AP880802-0165 [27] South Africa says the ANC, the main black group fighting to overthrow South Africa's white government, has seven major military bases in Angola, and the Pretoria government wants those bases closed down.
[5] [654] AP880803-0158 [27] South Africa says the ANC, the main black group fighting to overthrow South Africa's white-led government, has seven major military bases in Angola, and it wants those bases closed down.
[6] [92] WSJ910204-0176 [2] de Klerk's proposal to repeal the major pillars of apartheid drew a generally positive response from black leaders, but African National Congress leader Nelson Mandela called on the international community to continue economic sanctions against South Africa until the government takes further steps.
[7] [781] AP880823-0069 [18] The ANC is the main guerrilla group fighting to overthrow the South African government and end apartheid, the system of racial segregation in which South Africa's black majority has no vote in national affairs.
[8] [375] WSJ890908-0159 [24] For everywhere he turns, he hears the same mantra of demands -- release, lift bans, dismantle, negotiate -- be it from local anti-apartheid activists or from foreign governments: release political prisoners, like African National Congress leader Nelson Mandela; lift bans on all political organizations, such as the ANC, the Pan Africanist Congress and the United Democratic Front; dismantle all apartheid legislation; and finally, begin negotiations with leaders of all races.
[9] [762] AP880212-0060 [14] The African National Congress is the main rebel movement fighting South Africa's white-led government and SWAPO is a black guerrilla group fighting for independence for Namibia, which is administered by South Africa.
[10] [91] WSJ910404-0007 [8] Under an agreement between the South African government and the African National Congress, the major anti-apartheid organization, South Africa's remaining political prisoners are scheduled for release by April 30.

Fig 4: $\lambda = 1.0$ Multi Document Summarization
[Rank] Document ID [Sentence Number] Sentence

[1] [1] [761] AP880212-0060 [15] ANGOP quoted the Angolan statement as saying the main causes of conflict in the region are South Africa's ``illegal occupation" of Namibia, South African attacks against its black-ruled neighbors and its alleged creation of armed groups to carry out ``terrorist activities" in those countries, and the denial of political rights to the black majority in South Africa.
[2] [2] [758] AP880803-0080 [25] Three Canadian anti-apartheid groups issued a statement urging the government to sever diplomatic and economic links with South Africa and aid the African National Congress, the banned group fighting the white-dominated government in South Africa.
[3] [6] [92] WSJ910204-0176 [2] de Klerk's proposal to repeal the major pillars of apartheid drew a generally positive response from black leaders, but African National Congress leader Nelson Mandela called on the international community to continue economic sanctions against South Africa until the government takes further steps.
[4] [8] [375] WSJ890908-0159 [24] For everywhere he turns, he hears the same mantra of demands -- release, lift bans, dismantle, negotiate -- be it from local anti-apartheid activists or from foreign governments: release political prisoners, like African National Congress leader Nelson Mandela; lift bans on all political organizations, such as the ANC, the Pan Africanist Congress and the United Democratic Front; dismantle all apartheid legislation; and finally, begin negotiations with leaders of all races.
[5] [4] [790] AP880802-0165 [27] South Africa says the ANC, the main black group fighting to overthrow South Africa's white government, has seven major military bases in Angola, and the Pretoria government wants those bases closed down.
[6] [11] [334] AP890703-0114 [14] The white delegation chief, Mike Olivier, said the ANC members, including President Oliver Tambo and South African Communist Party leader Joe Slovo, said some white anti-apartheid members of Parliament could make a difference, although the organization believes Parliament as a whole is not representative of South Africans.
[7] [14] [788] WSJ880323-0129 [11] These included a picture of Oliver Tambo, the exiled leader of the banned African National Congress; a story about 250 women attending an ANC conference in southern Africa; a report on the crisis in black education; and an advertisement sponsored by a Catholic group in West Germany that quoted a Psalm and called for the abolition of torture in South Africa.
[8] [12] [303] AP880621-0089 [8] There was no immediate comment from South Africa, which in the past has staged cross-border raids on Botswana and other neighboring countries to attack suspected facilities of the African National Congress, which seeks to overthrow South Africa's white-led government.
[9] [24] [502] WSJ900510-0088 [24] While the membership of Inkatha, the religiously and politically conservative group that is the ANC's chief rival for power in black South Africa, is overwhelmingly Zulu, Inkatha's leader, Mangosutho Buthelezi, has very seldom appealed to sectional tribal loyalties.
[10] [16] [593] AP890821-0092 [11] Besides ending the emergency and lifting bans on anti-apartheid groups and individual activists, the Harare summit's conditions included the removal of all troops from South Africa's black townships, releasing all political prisoners and ending political trials and executions, and a government commitment to free political discussion.

Fig 5: $\lambda = .3$ Multi Document Summarization.
[Rank] [Previous Rank in $\lambda = 1.0$ Version] Document ID [Sentence Number] Sentence



<TITLE>Angola Rejects South African Proposal for Peace Talks
</TITLE>
<TEXT>
[1] Angola has rejected a South African proposal for a regional peace conference that would include Angolan rebels, Angola's official ANGOP news agency reported Friday.
[14] ANGOP quoted the Angolan statement as saying the main causes of conflict in the region are South Africa's ``illegal occupation" of Namibia, South African attacks against its black-ruled neighbors and its alleged creation of armed groups to carry out ``terrorist activities" in those countries, and the denial of political rights to the black majority in South Africa.
</TEXT>

Figure 6: Single Document Summary AP880212-0060, 10% of document length.

As can be seen from the above summaries, multi-document synthetic summaries require support in the user interface. In particular, the following issues need to be addressed:

- Attributability: The user needs to be able to access easily the source of a given passage. This could be the single document summary (see Figure 6).
- Contextually: The user needs to be able to zoom in on the context surrounding the chosen passages.
- Redirection: The user should be able to highlight certain parts of the synthetic summary and give a command to the system indicating that these parts are to be weighted heavily and that other parts are to be given a lesser weight.

7. EVALUATION OF SINGLE DOCUMENT SUMMARIZATION

An ideal text summary contains the relevant information for which the user is looking, excludes extraneous information, provides background to suit the user's profile, eliminates redundant information and filters out relevant information that the user knows or has seen. The first step in building such summaries is extracting the relevant pieces of articles to a user query. We performed a pilot evaluation in which we used a database of assessor marked relevant sentences to examine how well a summarization system could extract the relevant sections of documents.

Automatically generating text extraction summaries based on a query or high frequency words from the text can produce a reasonable looking summary, yet this summary can be far from the optimal goal of quality summaries: readable, useful, intelligible, appropriate length summaries from which the information that the user is seeking can be extracted. Jones & Galliers define this type of evaluation as intrinsic (measuring a system's quality) compared to extrinsic (measuring a system's performance in a given task) [7].

In the past year, there has been a focus in TIPSTER on both the intrinsic and extrinsic aspects of summarization evaluation [4]. The evaluation consisted of three tasks (1) determining document relevance to a topic for query-relevant summaries (an indicative summary), (2) determining categorization for generic summaries (an indicative summary), (3) establishing whether summaries can answer a specified set of questions (an informative summary) by comparison to an ideal summary. In each task, the summaries are rated in terms of confidence in decision, intelligibility and length. Jing, Barzilay, McKeown and Elhadad [6] performed a pilot experiment (40 sentences) in which they examined the performance (precision-recall) of three summarization systems (one using notion of number of sentences, the other two using numbers of words or number of clauses). They compared the performance of these systems against human ideal summaries and found that different systems achieved their best performances at different lengths (compression ratios). They also found the same results for determining document relevance to a topic (one of the TIPSTER tasks) for query-relevant summaries.

Our approach to summarization is different from Columbia and TIPSTER in that the focus is not on an "ideal human summary" of any particular document cutoff size. An ideal summarization system must first be able to recognize the relevant sentences (or parts of a document) for a topic or query and then be able to create a summary from these relevant segments. Although a list of words, an index or table of contents, is an appropriate label summary and can indicate relevance, informative summaries need at least noun-verb phrases. We choose to use the sentence as our underlying unit and evaluated summarization systems for the first stage of summary creation - coverage of relevant sentences. Other systems [16, 23] use the paragraph as a summary unit. Since the paragraph consists of more than one sentence and often more than one information unit, it is not as suitable for this type of evaluation, although it may be more suitable for a construction unit in summaries due to the additional context that it provides. For example., paragraphs will often solve co-reference issues, yet provide additional non-relevant information. One of the issues in summarization evaluation is how to score (penalize) extraneous non-useful information contained in a summary.

Unlike document information retrieval, text summarization evaluation has not extensively addressed the performance of different methodologies by evaluating the effects of different components. Most summarization systems use linguistic knowledge as well as a statistical component [3, 5, 16, 23]. We applied the monolingual information retrieval method of query expansion [20, 27, 28] to summarization, using parts of the document to expand our queries. We also performed compression experiments. We used a modified version of the 11-pt average recall/precision (Section 9.2) to evaluate our results.

8. EXPERIMENT DESIGN

For our pilot experiment, we created two data sets, one based on relevant sentence judgments, the other based on model summaries (Section 8.1). We then defined a modified version of the 11-point average recall precision (Section 8.2) to use as our evaluation measure. We then performed experiments as described in Section 9 to evaluate the effects of MMR, query expansion, and compression.

8.1 Data Sets

We created two data sets for our pilot experiments. For the first {110 Set} we took 50 documents from the TIPSTER evaluation provided set of 200 news articles spanning 1988-1991. All these documents were on the same topic (see Figure 3). Three evaluators ranked each of the sentences in the document as relevant, somewhat relevant and not relevant. For the purpose of this experiment, somewhat relevant was treated as relevant and the final score for the sentence was determined by a majority vote. The sentences that received this majority vote were tabulated as a relevant sentence (to the topic). The document was ranked as relevant or not relevant. All three assessors had 68% agreement in their relevance judgments. The query was extracted from the topic (see Figure 3).

The second data set {Model Summs} was provided as a training set for the Question and Answer portion of the TIPSTER evaluation. It consisted of "model summaries" which contained sentences of an article that answered a list of questions. These model sentences were used to score the summarizer. The query was extracted from the questions.

8.2 Evaluation Code

We modified the 11-pt recall-precision curves [21] commonly used for document information retrieval. Since many documents only have a few relevant sentences, corresponding curves for summarization have a lot of intervals with missing data items. To remedy this situation, we implemented a step function for the precision values. This allowed the recall intervals that would not naturally be filled to be assigned an actual precision value. For example, in the case of two relevant sentences in the document, points 0-5 (the first five intervals) would all have the first precision value (naturally occurring at point 5) and points 6-10 (the second value), the second value (naturally occurring at point 10). We interpolated the results of each query for the composite graph to form modified interpolated recall-precision curves.

In order to account for the fact that a compressed summary does not have the opportunity to return the full set of relevant sentences, we use a normalized version of recall and a normalized version of F1 as defined below.

Given:

Rel = Number of Relevant Sentences in Document
RelSum = Number of Relevant Sentences in Summary
SentSum = Number of Sentences in Summary

Definitions:

Precision P = RelSum / SentSum
Recall R = RelSum / Rel
F1 = 2P*R / (P + R)

NorR = RelSum / min (Rel, SentSum)
NorF1 = 2P*NorR/(P+NorR)

9. EXPERIMENTS AND RESULTS

In this section we describe the experiments we performed and results obtained in evaluating the diversity gain - MMR (Section 9.1), query expansion (Section 9.2) and compression (Section 9.3).

9.1 MMR (Diversity Gain)

In order to evaluate what the relevance loss for the MMR diversity gain in single document summarization, we created summaries for two document length percentages (measured by number of sentences) and determined how many relevant sentences the summaries contained.

Sentence Precision			
Percentage of Document Length	λ	TREC and CMU Relevant	CMU Relevant
10%	1.0	0.78	0.83
10%	0.7	0.76	0.83
10%	0.3	0.74	0.79
10%	Baseline	0.74	0.83
25%	1.0	0.74	0.76
25%	0.7	0.73	0.74
25%	0.3	0.74	0.76
25%	Baseline	0.60	0.65

Table 2: Precision Scores

	Model Summaries	110 Set
task	Q & A	indicative summaries
number of documents	48	50
source	provided by Tipster	3 people marked each sentence
relevant documents	all	15
average sentences per document	22.6	25.1
median sentences per document	19	23
maximum sentences per document	51	50
minimum sentences per document	11	11
query formation	provided questions	topic
statistics	all documents	40 documents
percent of document length	19.4%	24.9%
summary includes first sentence	72%	47%, 73% (only relevant docs)
average summary size (sentences)	4.3	6.1
median summary size (sentences)	4	5

Table 3: Data Set Comparison

110 Set Comparison	Relevant Documents	Non-Relevant Documents
number of documents	15	25
average sentences per document	27.5	23.8
median sentences per document	23	23
maximum sentences per document	51	44
minimum sentences per document	15	11
percent of document length	36.2%	17.7%
summary includes first sentence	73%	32%
average summary size (sentences)	10.1	3.7
median summary size (sentences)	7	4

Table 4: 110 Set - Relevant vs. Non-Relevant Documents with relevant sentences.

The results are given in Table 2 for document percentages 0.25 and 0.1. Two precision scores were calculated, (1) that of TREC relevance plus at least one CMU assessor marking the document as relevant (yielding 23 documents) and (2) at least two of the three CMU assessor marking the document as relevant (yielding 15 documents). From these scores we can see there is no significant statistical difference between the $\lambda=1$, $\lambda=.7$, and $\lambda=.3$ scores. This is often explained by cases where the $\lambda=1$ article failed to pick up a piece of relevant information and the reranking of $\lambda=.7$ or .3 might or vice versa. The baseline (baseln) contains the first N sentences of the document, where N is the number of sentences in the summary.

9.2 Query Expansion

We expanded the original queries by: (1) adding the highest ranked sentence of the document (a form of pseudo-relevance feedback), (2) adding the title, and (3) adding the title and the highest ranked sentence.

The most significant effects were shown for short queries (see Figures 7, 9). For the longer queries, the effect was less (see Figures 8, 10). For 20% document length (characters rounded up to the sentence boundary) adding the highest ranked sentence (prf) and title to the query helps performance for the 110 set relevant summary judgments (Figures 7, 8). For 10% document length,

Figure 7: Query Expansion Effects - 110 Set, SMART weighting lnn, 20% document length, relevant documents only

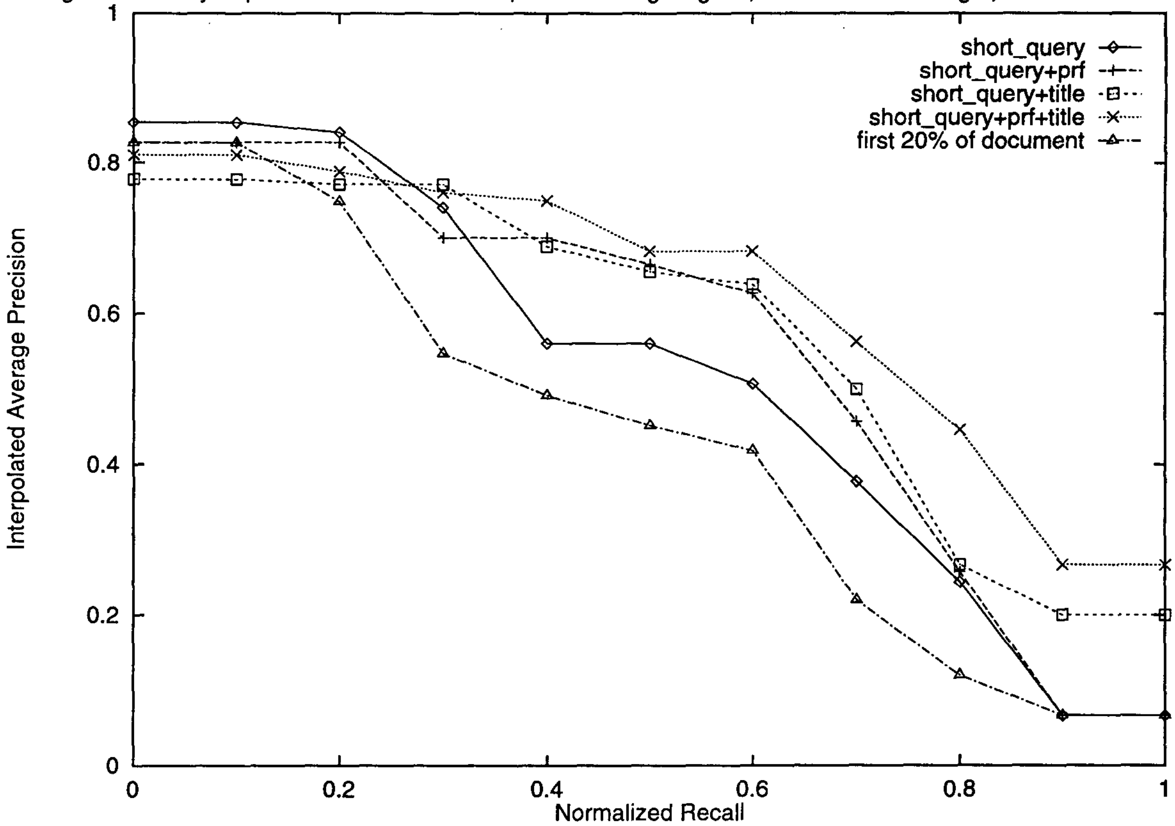


Figure 8: Query Expansion Effects - SMART weighting lnn, 20% document length

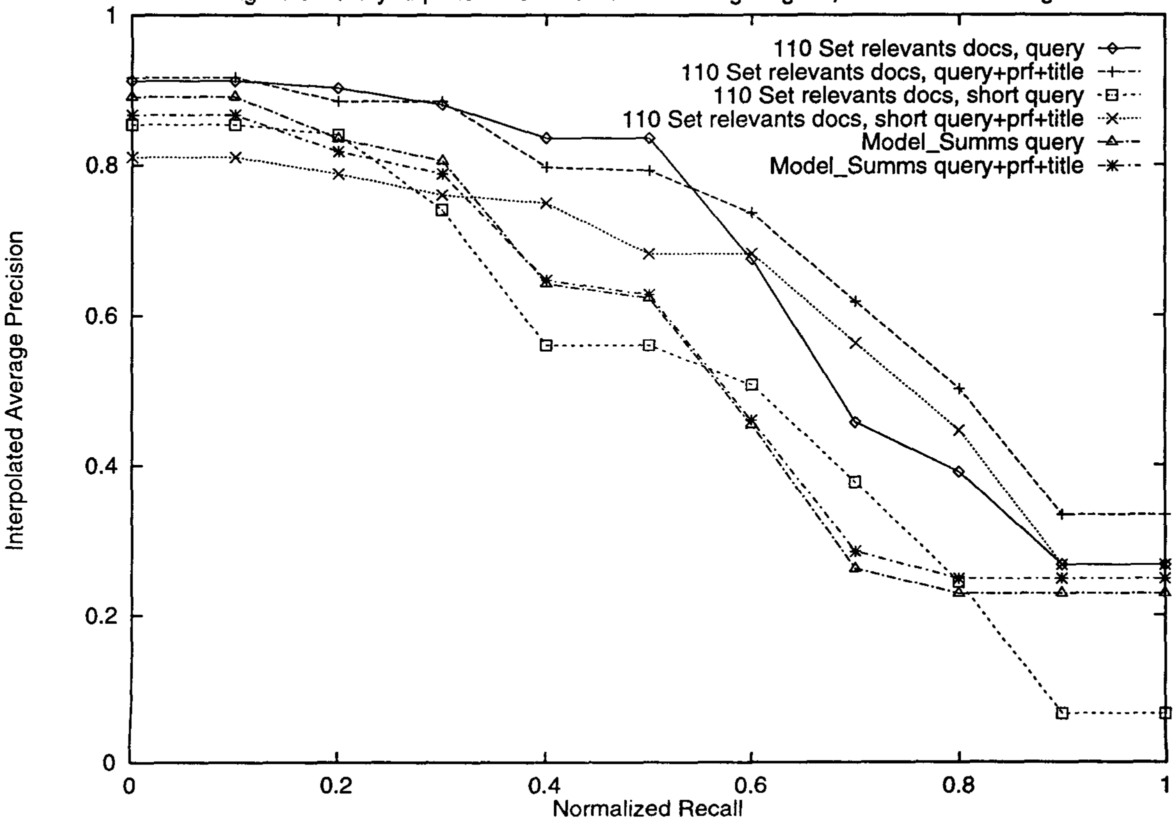


Figure 9: Query Expansion Effects - 110 Set, SMART weighting lnn, 10% document length, relevant documents only

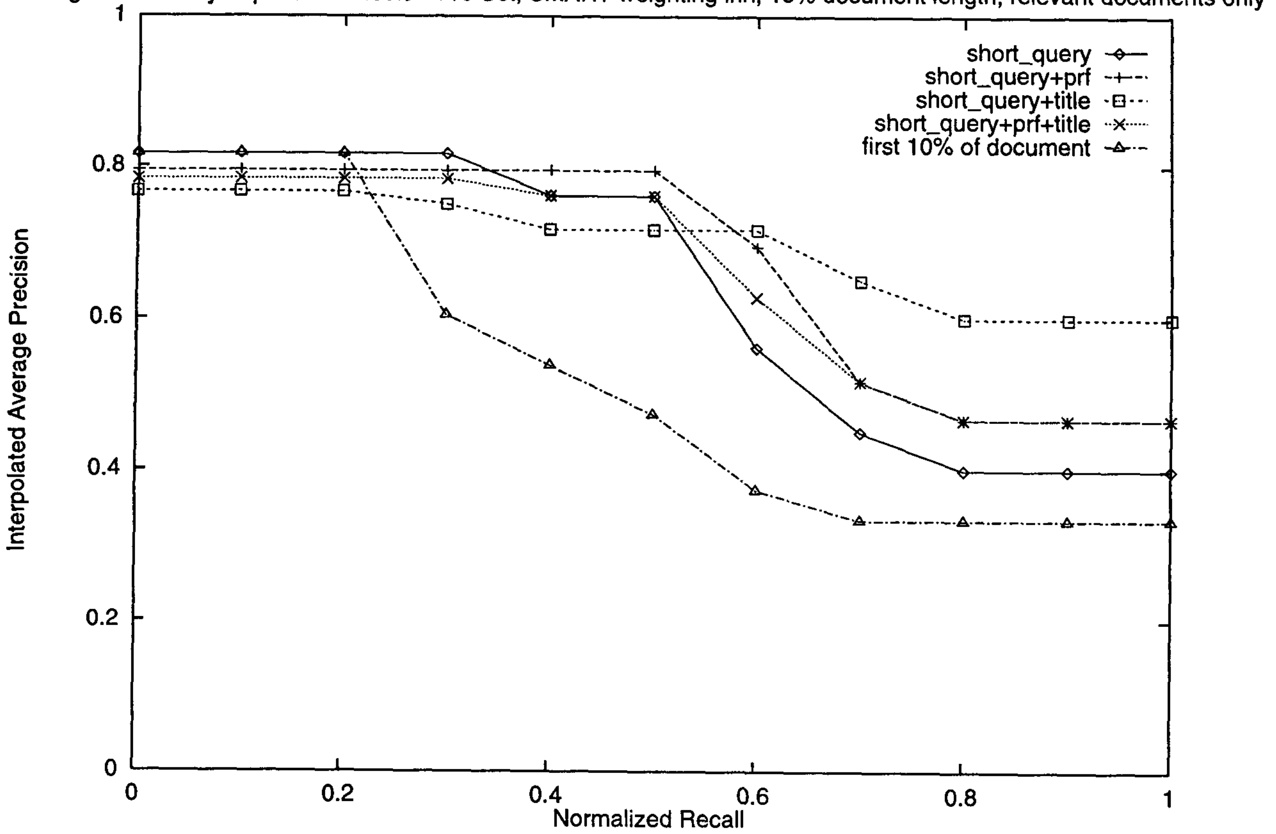


Figure 10: Query Expansion Effects - SMART weighting lnn, 10% document length, relevant documents only

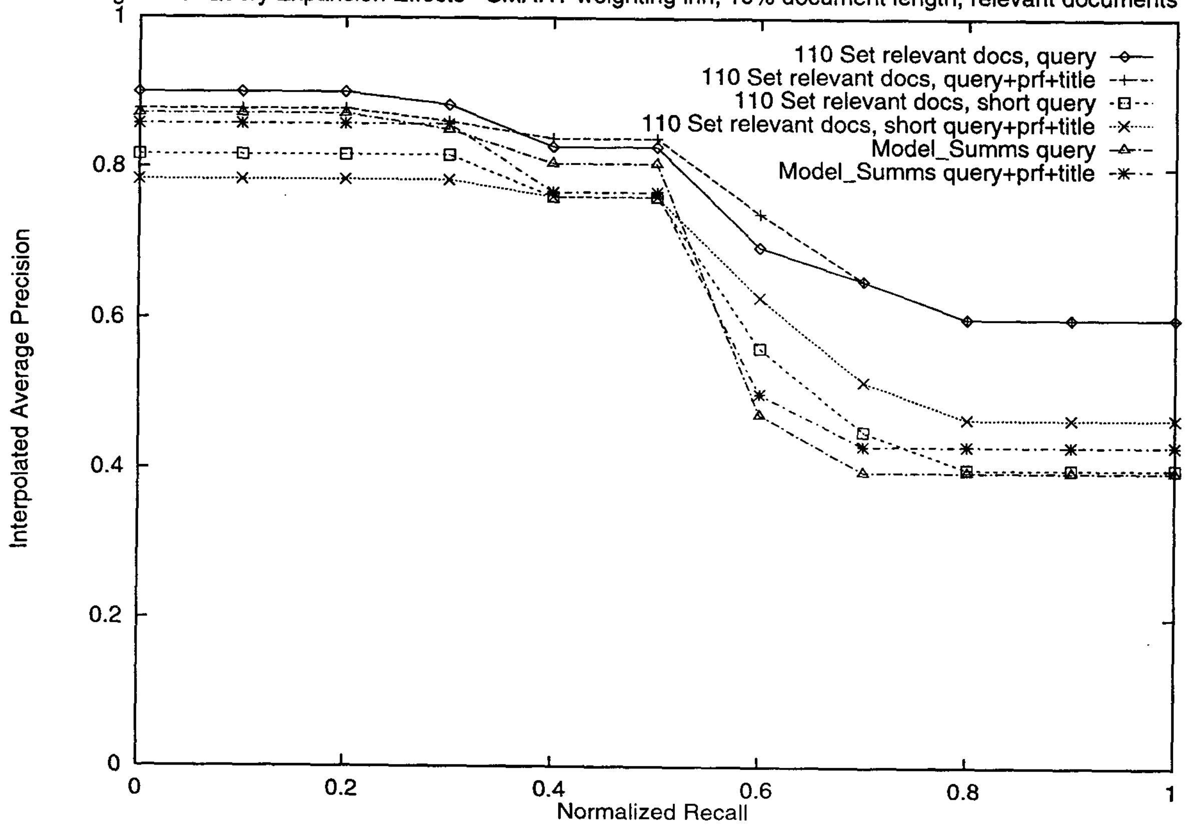


Figure 11: Compression Effects - SMART weighting lnn, 110 Set relevant documents only

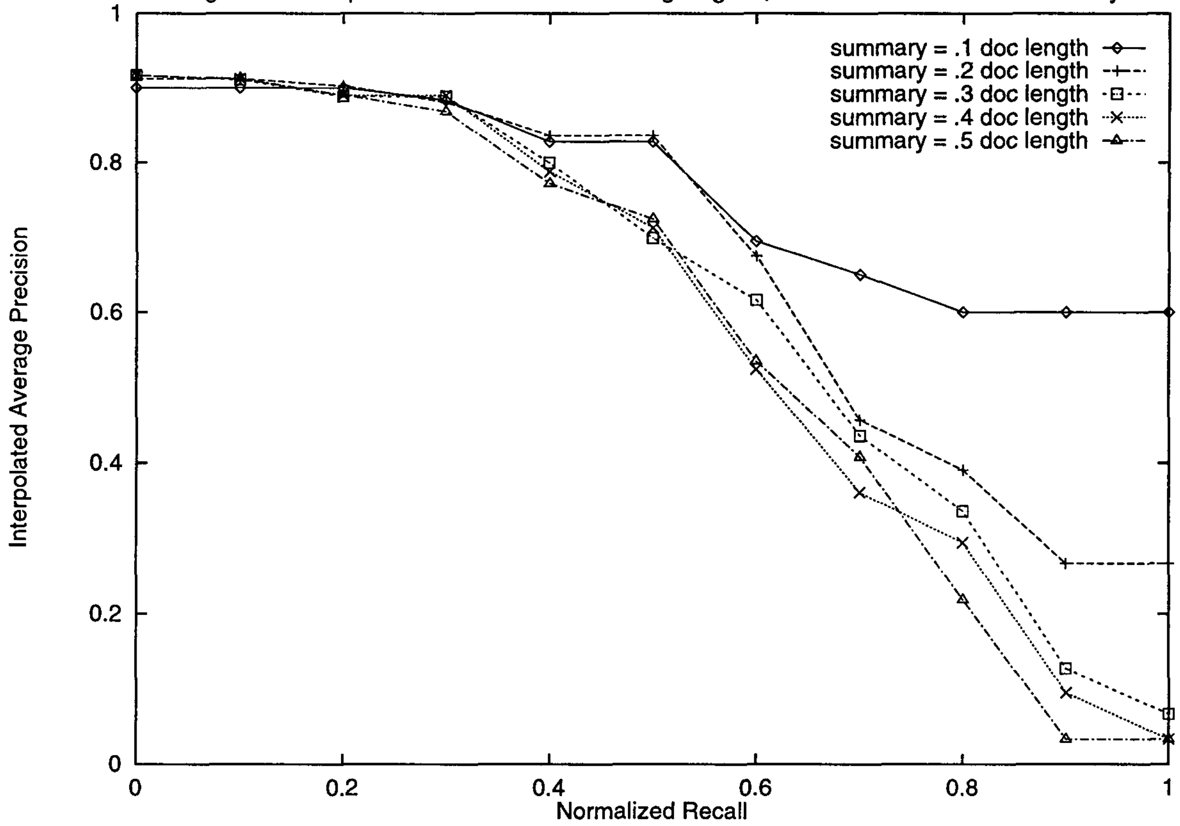
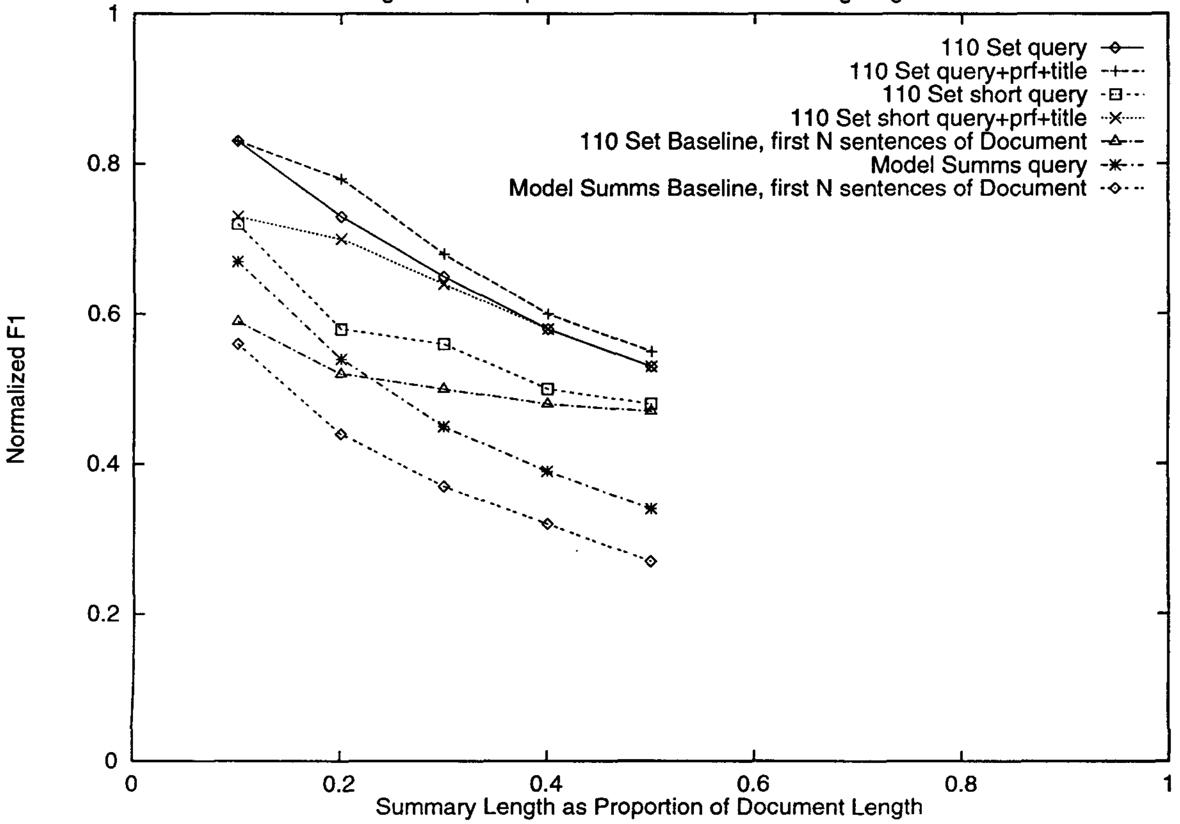


Figure 12: Compression Effects - SMART weighting lnn



for short queries just adding the title performed better than prf and the title (Figures 9,10). We will determine if these results hold over more extensive data.

These results are similar to those obtained for document information retrieval [27]. Since 72% of the first sentences were marked relevant (Table 3), one area we plan to explore is results using the first sentence in the summary and/or query under specified circumstances, such as our first sentence heuristics (Section 4).

9.3 Compression

An important evaluation criteria for summarization is what is the ideal summary output length (compression of the document) and how does it affects the user's task. To begin looking at this issue, we evaluated the performance of our system at different summary lengths as a percentage of the document length.

We used a document compression factor based on the number of characters in the document. If this cutoff fell in the middle of a sentence the rest of the sentence was allowed, thus the output summary ends up being slightly longer than the actually compression factor.

The data set statistics are shown in Tables 3 and 4. Note that non-relevant documents (Table 4) still have a high percentage of relevant sentences. Ten documents in the 110 set were non-relevant and had no relevant sentences. We also see that the summary length or number of relevant sentences chosen per document varies significantly.

Summaries were compared using the modified interpolated normalized recall-precision curve as previously described (Section 8.2).

In Figure 11, we examine the effect of compression on normalized recall and precision and in Figure 12, we show a plot of normalized F1. This F1 graph indicates that the normalized F1 score is helped by having the pseudo-relevance feedback and title in the query thereby extracting relevant sentences that would otherwise be missed. As the number of sentences that are allowed in the summary grows, the difficulty of finding relevant sentences grows and thus the added prf sentence and title to the query help to find relevant sentences for their particular document. We need to do more studying on the effects of query expansion and compression on summarization, as well as see how our preliminary results hold for additional data sets.

If we calculate the normalized F1 score for the first sentence retrieved in the summary, we obtain a score of .80 for 110 Set standard query, .67 for 110 Set short query and .79 for the Model Summaries. This indicates that even for the short query we obtain a relevant sentence two thirds of the time. However, ideally this first sentence retrieval score would be 1.0 and we will explore methods to increase this score as well as select a "highly relevant" first retrieved sentence for the document.

10. CONCLUSION

We have shown that MMR ranking provides a useful and beneficial manner of providing information to the user by allowing the user to minimize redundancy. This is especially true in the case of query-relevant multi-document summarization in this one data collection. We are currently performing studies on how this extends to additional document collections. In the future we will also be investigating how to handle co-reference in our system as well as analyzing the most suitable λ parameters and clustering the output results.

Text Summarization is still in the infant stage in terms of evaluation. Many monolingual document information retrieval results can be applied to text summarization, but as of yet, there has been little evaluation of these techniques. This pilot experiment showed many areas that need to be examined in further detail, including whether the summary selects the most relevant sentences in the document and whether these results generalize to more data sets and other document genres. We also plan to explore further the effects of query expansion using WordNet, as well as the use the first sentence (for news stories) in the query and/or summary. We also plan to run experiments fixing the number of sentences for each document as the number of relevant sentences chosen by the assessors as well as a small number, such as three. We are currently in the process of building a more extensive sentence relevance database for further evaluation. In this database, we are collecting data on the user selected most relevant sentence(s) for each document. We also plan to explore how to join the relevant sections to provide a "good", understandable, readable, relevant, non-redundant summary.

REFERENCES

[1]. C. Buckley, Implementation of the SMART Information Retrieval System. *Department of Computer Science Technical Report* Cornell University, TR 85-686.

[2]. J.G. Carbonell, and J. Goldstein, The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries, In *Proceedings of SIGIR 98*, Melbourne, Australia, 24-28 August 1998, p. 335-336.

[3] J. Cowie, K. Mahes, S. Nirenbug, R. Zajac, MINDS -- Multi-lingual Interactive Document Summarization, *AAAI Intelligent Text Summarization Workshop*, p. 131-1328, Stanford, CA March 1998

[4] T.F. Hand, A Proposal for Task-Based Evaluation of Text Summarization Systems In *ACL/EACL-97 Summarization Workshop}*, 31-36, Madrid, Spain., July 1997.

[5] E. Hovy and C.Y. Lin, Automated Text Summarization in SUMMARIST, In *ACL/EACL-97 Summarization Workshop*, 18-24, Madrid, Spain July 1997

[6] H. Jing, R. Barzilay, K. McKeown, M. Elhadad, Summarization Evaluation Methods Experiments and Analysis, AAAI Intelligent Text Summarization Workshop, p. 60-68, Stanford, CA March 1998

[7]. K.S. Jones and J.R. Galliers, *Evaluation Natural Language Processing Systems: an Analysis and Review.* New York: Springer 1996

[8]. J.L Klavans and J. Shaw, Lexical Semantics in Summarization, In *Proceedings of the First Annual Workshop of the IFIP Working Group FOR NLP and KR*, Nantes, France, April 1995.

[9]. G. Kowalski, *Information Retrieval Systems: Theory and Implementation*, Kluwer Academic Publishers, 1997.

[10]. J.M. Kupiec, J. Pedersen, J. and F. Chen, A Trainable Document Summarizer, In *Proceedings of the 18th Annual Int. ACM/SIGIR Conference on Research and Development in IR*, Seattle, WA, July 1995, pp. 68-73.

[11] D.D. Lewis, B. Croft, B., and N. Bhandaru, "Language-Oriented Information Retrieval," *International Journal of Intelligent Systems*, Vol 4 (3), Fall 1989.

[12] H.P. Luhn, Automatic Creation of Literature Abstracts, *IBM Journal*, 1958, pp. 159-165.

[13] M.L Mauldin, Retrieval Performance in FERRET: A Conceptual Conference on Research and Development in Information Retrieval, *Proceedings of the 14th International Conference on Research and Development in Information Retrieval*, October 1991.

[14]. M.L. Mauldin and J.R. Leavitt, Web Agent Related Research at the Center for Machine Translation. In *Proceedings of SIGNIDR V*, McLean Virginia, August 1994.

[15]. K. McKeown, J. Robin, and K. Kukich, Empirically Designing and Evaluating a New Revision-based Model for Summary Generation. In *Information Processing and Management*, 31 (5) 1995.

[16] M. Mitra, A. Singhal and C. Buckley, Automatic Text Summarization by Paragraph Extraction, In *ACL/EACL-97 Summarization Workshop*, 39-46, Madrid, Spain July 1997.

[17] C.D. Paice, Automatic Generation of Literature Abstracts - An Approach Based on the Indification of Self-Indicated Phrases, in *Information Retrieval Research*, R.N. Oddy, S.E. Robertson, C.J. van Rijsbergen and P.W. Williams, editors, Butterworths, London, 1981, 172-191.

[18]. C.D. Paice, Constructing Literature Abstracts by Computer: Techniques and Prospects, In *Information Processing and Management*, Vol. 26, 1990, pp.171-186.

[19] G. Salton G and C. Buckley Improving retrieval performance by relevance feedback. *Journal of American Society for Information Sciences*, 41:288-297, 1990.

[20]. G. Salton *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*, Addison-Wesley 1989.

[21]. G. Salton and M.J. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill, New York, McGraw-Hill Computer Science Series, 1983.

[22] G. Salton, A. Singhal, M. Mitra,. and C. Buckley, Automatic Text Structuring and Summarization, *Information Processing and Management*, 33(2), 193-208, 1997.

[23] T. Strzalkowski, J. Wang, and B. Wise, A Robust Practical Text Summarization, *AAAI Intelligent Text Summarization Workshop*, p. 26-3, Stanford, CA March 1998.

[24] J.I. Tait, *Automatic Summarizing of English Texts*, PhD dissertation, University of Cambridge, 1983.

[25] TIPSTER Text Phase III 18-Month Workshop, Fairfax, VA 4-6 May, 1988,

[26] C.J. van Riesburg, *Information Retrieval*, London Butterworths 1979.

[27] J. Xu and B. Croft. Query expansion using local and global document analysis in 19th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '96), pages 4-11, 1996.

[28] E.M. Vorhees. Using Wordnet to disambiguate words senses for text retrieval. In *Proceedings of ACM SIGIR Conference (SIGIR '93)* pages, 171-180, 1993.