The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries

Jaime Carbonell
Language Technologies Institute
Carnegie Mellon University
jgc@cs.cmu.edu

Jade Goldstein
Language Technologies Institute
Carnegie Mellon University
jade@cs.cmu.edu

Abstract This paper presents 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 document retrieval and in single document summarization. The latter are borne out by the recent results of the SUMMAC conference in the evaluation of summarization systems. However, the clearest advantage is demonstrated in constructing non-redundant multi-document summaries, where MMR results are clearly superior to non-MMR passage selection.

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, 5]. 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.

A new document ranking method 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; Maximal Marginal Relevance (MMR) provides precisely such functionality, as discussed below.

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. In contrast, we motivated the need for "relevant novelty" as a potentially superior criterion. 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).

$$MMR \stackrel{\text{def}}{=} Arg \max_{D_i \in R \backslash S} \left[\lambda(Sim_1(D_i, Q) - (1-\lambda) \max_{D_j \in S} Sim_2(D_i, D_j)) \right]$$

Where C is a document collection (or document stream); Q is a query or user profile; R = IR(C, Q, θ), i.e., the ranked list of documents retrieved by an IR system, given C and Q and a relevance threshold θ, below which it will not retrieve documents (θ can be degree of match or number of documents); S is the subset of documents in R already selected; $R \backslash S$ is the set difference, i.e, the set of as yet unselected documents in R; Sim_1 is the similarity metric used in document retrieval and relevance ranking between documents (passages) and a query; and Sim_2 can be the same as Sim_1 or a different metric.

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 λ. We found that a particularly effective search strategy (reinforced by the user study discussed below) is to start with a small λ (e.g. $\lambda = .3$) 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. $\lambda = .7$).

3 Document Reordering

We 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 a standard ranking method and

MMR. The users were asked to find information from documents and were not told how the order in which documents were presented - only that either "method R" or "method S" were used. The majority of people said they preferred the method which gave in their opinion the most broad and interesting topics (MMR). In the final section they were asked to select a search method and use it for a search task. 80% chose the method MMR. 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.

4 Summarization

If we consider document summarization by relevant-passage extraction, we must again consider relevance as well as anti-redundancy. Summaries need to avoid redundancy, as it defeats the purpose of summarization. If we move beyond single document summarization to document cluster summarization, where the summary must pool passages 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 [4], and evolved through several efforts to the recent TIPSTER effort which includes trainable methods [3], linguistic approaches [6] and our information-centric method [2], the first to focus on anti-redundancy measures.

Human summarization of documents, sometimes called "abstraction" is a fixed- length summary, reflecting the key points that the abstractor – rather than the user – deems important. A different user with different information needs may require a totally different summary of the same document. We created single document summaries by segmenting the document into passages (sentences in our case) and using MMR with a cosine similarity metric to rerank the passages in response to a user generated or system generated query. The top ranking passages were presented in the original document order.

In the May 1998 SUMMAC conference [6], featuring a government-run evaluation of 15 summarization systems, our MMR-based summarizer produced the highest-utility query-relevant summaries with an F-score of .73 – derived from precision and recall by assessors making topic-relevance judgements from summaries. Our system also scored highest (70% accuracy) on informative summaries, where the assessor judged whether the summary contained the information required to answer a set of key questions. It should be noted that some parameters, such as summary length, varied among systems and therefore the evaluation results are indicative but not definitive measures of comparative performance.

In order to evaluate what the relevance loss for a diversity gain in single document summarization, three assessors went through 50 articles from 200 articles of a TIPSTER topic and marked each sentence as relevant, somewhat relevant and irrelevant. The article was also marked as relevant or irrelevant. The assessor scores were compared against the TREC relevance judgments provided for the topic.

The sentence precision results are given in Table 1 for compression factors .25 and .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 assessors marking the document as relevant (yielding 18 documents). From these scores we can see there is no significant statistical difference between the λ=1, λ=.7, and λ=.3 scores. This is often explained by cases where the λ=1 summary failed to pick up a piece of relevant information and the reranking with λ=.7 or .3 might.

Sentence Precision			
Document Percentage	λ	TREC and CMU Relevant	CMU Relevant
10	1	.78	.83
10	.7	.76	.83
10	.3	.74	.79
10	Lead Sentences	.74	.83
25	1	.74	.76
25	.7	.73	.74
25	.3	.74	.76
25	Lead Sentences	.60	.65

Table 1: Precision Scores

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.). MMR is also extremely useful in extraction of passages from multiple documents about the same topics. News stories contain much repetition of background information. Our preliminary results for multi-document summarization show that in the top 10 passages returned for news story collections in response to a query, there is significant repetition in content over the retrieved passages and the passages often contain duplicate or near-replication in the sentences. MMR reduces or eliminates such redundancy.

5 Concluding Remarks

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. We are currently performing studies on how this extends to several document collections as well as studies on the effectiveness of our system.

References

[1] Buckley C. Implementation of the smart information retrieval system. Technical Report TR 85-686, Cornell University.

[2] J.G.Carbonell, Y. Geng, and J. Goldstein. Automated query-relevant summarization and diversity-based reranking. In *15th International Joint Conference on Artificial Intelligence, Workshop: AI in Digital Libraries*, pages 9–14, Nagoya, Japan, August 1997.

[3] J.M. Kupiec, J. Pedersen, and F. Chen. A trainable document summarizer. In *Proceedings of the 18th Annual Int. ACM/SIGIR Conference on Research and Development in IR*, pages 68–73, Seattle, WA, July 1995.

[4] P.H. Luhn. Automatic creation of literature abstracts. *IBM Journal*, pages 159–165, 1958.

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

[6] In *TIPSTER Text Phase III 18-Month Workshop*, Fairfax, VA, May 1998.