Query Independent Sentence Scoring approach to multi-document Summarization

by

Jagadeesh J, Prasad Pingali, Vasudeva Varma

in


Report No: IIIT/TR/2008/47

Centre for Search and Information Extraction Lab
International Institute of Information Technology
Hyderabad - 500 032, INDIA
June 2008
Capturing Sentence Prior for Query-Based Multi-Document Summarization

Jagadeesh J ∗
Microsoft Research Lab, India
jags@microsoft.com

Prasad Pingali & Vasudeva Varma
International Institute of Information Technology, India
{pvvpr,vv}@iiit.ac.in

Abstract

In this paper, we have considered a real world information synthesis task, generation of a fixed length multi-document summary which satisfies a specific information need. This task was mapped to a topic-oriented, informative multi-document summarization. We also tried to estimate, given the human written reference summaries and the document set, the maximum performance (ROUGE1 scores) that can be achieved by an extraction-based summarization technique. Motivated by the observation that the current approaches are far behind the estimated maximum performance, we have looked at Information Retrieval techniques to improve the relevance scoring of sentences towards information need. Following information theoretic approach we have identified a measure to capture the notion of importance or prior of a sentence. Following a different decomposition of Probability Ranking Principle, the calculated importance/prior is incorporated into the final sentence scoring by weighted linear combination. In order to evaluate the performance, we have explored information sources like WWW and encyclopedia in computing the information measure in a set of different experiments. The t-test analysis of the improvement on DUC2 2005 data set is found to be significant (p ~ 0.05). The same system has outperformed rest of the systems at DUC 2006 challenge in terms of ROUGE scores with a significant margin over the next best system.

1 Introduction

In this era of information overload, it has become increasingly important to provide mechanisms to find and present the textual information effectively. While search engines were designed to deal with this abundant information, even they output a large number of documents for a user’s query. In this situation, systems that can automatically generate answers from one or more documents are becoming increasingly desirable.

In this paper we have considered modeling such a real world complex information access scenario. The goal is to, given user’s information need, expressed as multiple questions, produce a single text as a compressed version of a set of documents with all and only the relevant information. Here we assume that the set/cluster of documents will contain the answer. The system task is to create, from the document set, a response which answers the information need expressed. The information need or topic consists of mainly two components. First part is the title of the topic (e.g. Airport Security). Second is the actual information need, expressed as multiple questions

∗Work done as a student of International Institute of Information Technology
1Recall Oriented Understudy for Gisting Evaluation
2Document Understanding Conferences
System | ROUGE-2 | ROUGE-SU4
---|---|---
Maximum Performance | 0.09965 | 0.15407
HAL feature | 0.07618 | 0.13805

Table 1: Bounds on the ROUGE scores that can be obtained by sentence extraction, on DUC 2005 data (section 5)

(Example: “What threatens airport security? What are the measures to be taken to improve it?”). The information need is quite different from TREC factoid type or definitional based question [Dumais et al., 2002, Zheng, 2002], as it cannot be met by just stating a name, date or quantity etc. Moreover the answer may not be found in a single document, instead it is more likely that, retrieving multiple documents, locating portions of answers in them and combining them to form a single response will satisfy the information need. So this task is different from a factoid question answering task and, at a broader level, can be seen as a topic-oriented, informative multi-document summarization [Schlesinger and Baker, 2001, White et al., 2001], where the goal is to produce a focussed summary from a set of multiple documents. In this paper we have combined Information Retrieval techniques (section 6), to calculate the relevance of a sentence (section 4.1), with summarization techniques, in producing the extracts, to address this task.

A system that addresses such a complex task may involve the following stages: information need enrichment, content selection and summary generation. Building such systems will not only take considerable amount of resources but also significant time to produce the summary, as it involves deep analysis of large number of sentences, once the input and the data cluster is provided. But approaches like Language Modeling [Jagadeesh et al., 2005], Concept linkages and Bayesian framework [Daume and Marcu, 2005, Blair-Goldensohn, 2005, Ye et al., 2005, Li et al., 2005] provided a way to achieve good performance without involving such a deep processing of text. None of these approaches capture the sentence bias which is independent of the questions. Incorporating sentence importance which is independent of query into the final scoring is the main idea of this paper.

2 Motivation

Similar task has been addressed in [Jagadeesh et al., 2005] and it has been shown that the HAL feature (section 4.1) has outperformed all the other approaches. In our initial experiments, we have compared the performance achieved using HAL feature with the maximum performance that can be achieved by an extraction based summarization system, on a given data set. As a step towards estimating the maximum performance, we have generated answer summaries by picking the sentences, from the document set, that have high cosine similarity with the model answers\(^3\). A comparison between the maximum performance and that of HAL feature (table 1), has revealed the fact that there is still a lot of scope for improvement in generating better answer extracts using

\(^3\)human written summaries are referred as model summaries, while system generated summaries are referred as peer summaries
sentence extraction framework itself, before employing any abstraction techniques. The result is valid even though the aim of summarization is not to maximize the ROUGE scores, as ROUGE is found to correlate well with the manual evaluations [Dang, 2005]. Naturally a summary with a better ROUGE score has more chances of being better. The possible improvement in the extraction framework, has motivated us to look for Information Retrieval techniques which can be used along with HAL feature, to move performance curve towards the attainable maximum.

3 Our Contributions

In general, in a document some sentences are more important than others, the importance could denote the centrality or any specific function of that sentence. In fact summarization systems try to explore this property of a sentence in producing generic summaries. This paper is an attempt to explore the effect of capturing sentence prior on the generated answer summaries. The contributions of this paper include:

- Based on information theoretic approach, we have defined a new measure called Information Measure (section 4.2.1) to capture a sentence prior. This measure can be shown as a special case of a more general Risk Minimization Framework [Lafferty and Zhai, 2001]. We have also discussed couple of information sources that can be used to compute Information Measure.

- Latter (section 6) we have justified, using Probability Ranking Principle [Robertson S.E., 1977], that such a measure calculated using the distribution of its constituent words among relevant documents will improve the relevance ranking of the candidates consistently and not by chance.

In evaluation section (section 5) we showed that the final system has generated best answers for a set of widely varying complex questions. Using t-test analysis we found the improvement to be statistically significant.

4 Features

As in any other extraction based summarization system, the sentences are scored using a set of features. For each feature, the score obtained by all sentences is normalized by the maximum score, so that the new maximum feature value corresponds to 1. The final score of a sentence is calculated using a weighted linear combination of the individual feature values. The top scoring sentences are selected greedily until the summary length reaches the desired limit. In the following section we will discuss the application of Relevance Based Language modeling to summarization [Jagadeesh et al., 2005] very briefly. In section 4.2.1 we introduce a new feature called Information Measure which measures the prior domain knowledge carried by the sentence.
4.1 Relevance-Based Language Model

Unlike Statistical Language Models applied to IR [Ponte and Croft, 1998], Relevance Based Language Model does not assume the query as a sample from any specific document model, instead it assumes both the query and the document as samples from an unknown relevance model $R$. It approximates $p(w|R)$, the probability of observing a word $w$ in the documents relevant to a particular information need, using the probability of observing the word in the context of the query $p(w|Q) = p(w, Q)/p(Q)$. Using Conditional Sampling [Lavrenko and Croft, 2001] assumption that the query words $q_i \in Q$ to be independent of each other while keeping their dependencies on $w$ intact, it computes the required joint probability $p(w, q_1, \cdots, q_k) = p(w) \prod_{i=1}^{k} p(q_i|w)$.

The required term dependencies, $p(q_i|w)$, are incorporated using the probabilistic interpretation of Hyperspace Analogue to Language (HAL) model. The HAL matrix is constructed by moving a window of length $K$ words across the entire corpus and weighting the co-occurring pair of words inversely proportional to the distance between them. Now the $pHAL$, probabilistic HAL, interpreted as given a word $w$ the probability of associating a word $w'$ with $w$ in a window of size $K$, can be expressed in terms of probability of observing $w'$ at a distance of $k < K$ from $w$;

$$pHAL(w'|w) = \sum_{k=0}^{K} p(k) p(w'|w, k)$$

(1)

where $p(w'|w, k) = \frac{n(w, k, w')}{\sum_{w'} n(w, k, w')}$. To ensure a valid probability distribution, the constraint

$$\sum_{w'} pHAL(w'|w) = 1$$

is imposed.

Once the relevance of a word towards the information need is calculated, the relevance of a sentence can be computed by making the independence assumption between the words. This was referred as HAL feature. The reader is encouraged to refer to [Jagadeesh et al., 2005] for further clarification. We have retained the parameters of HAL to be same as those used in original paper, for reproducing the same performance as was reported hence a fair comparison of the features.

4.2 Sentence Prior

In a document some sentences are more important than others, the importance could denote the centrality or any specific function of that sentence. In fact systems try to explore this property in producing generic summaries. Most of the current query-based summarization systems concentrate only on features that measure the relevance of sentences towards the query. They do not explicitly attempt to capture centrality/prior knowledge carried by a sentence pertaining to a domain. In this section we defined a new measure which captures the sentence importance based on the distribution of its constituent words in the domain corpus. First we will describe the representation of this measure assuming the domain corpus is available. Latter part of the section will discuss some sources which can be used as necessary information sources for the domain corpus.
4.2.1 Information Measure

We have used entropy measure to compute the information content of a sentence based on a unigram model learned from document corpus. If \( p(x) \) is a probability density function (with respect to the counting measure) for a discrete random variable \( X \), Shannon [Shanon and Weaver, 1983] defined a measure of information content called self-information or surprisal of a message \( x \) given by:

\[
I(x) = - \log p(X = x),
\]

And the entropy, or uncertainty, of a message \( x_1, x_2, \cdots, x_k \) in that discrete message space is the expected self-information of the message:

\[
H(X) = \mathbb{E}\{I(x)\} = \sum_{x \in X} p(x) I(x) = - \sum_{x \in X} p(x) \log p(x)
\]  

(2)

Entropy can be seen as a measure of information content in a message. If a symbol has zero probability, which means it never occurred, it should not affect the entropy. So we let \( 0 \log 0 = 0 \).

Following the information theoretic approach, if \( p(w|r) \) denote the unigram language model conditioned on the relevant domain documents, which also denote the probability of a word being central to the domain and \( S = w_1, w_2, \cdots, w_k \) be a message in the words space then, by the definition of entropy, the amount of prior knowledge pertaining to that domain is the expected self-information (equation 2) of the message given by:

\[
H(S|r) = - \sum_{w \in S} p(w|r) \log p(w|r)
\]

We have defined this measure as Information Measure.

This model of ranking sentences can be seen as an instance of Risk Minimization Retrieval Framework [Lafferty and Zhai, 2001]. It views a query as being the output of some probabilistic process associated with user and similarly a document being as the output of some probabilistic process associated with an author or document source. If \( \theta_Q \) and \( \theta_D \) denote the parameters of query and document models respectively and \( L \), the loss associated with the action of returning a list of documents in response to a given query \( q \), is assumed to depend only on \( \theta_Q \) and \( \theta_D \), i.e. \( L(\theta_Q, \theta_D, R) = c \Delta(\theta_D, \theta_Q) \). Then the risk in returning these ranked set can be simplified as \( R(d; q) \propto \Delta(\hat{\theta}_d, \hat{\theta}_q) \). In the original paper, the authors defined the distance function to be relative entropy between the query and document probability distributions. Instead of relative entropy if we use information content in the document according to the query model as the distance function \( \Delta \), then the risk is proportional to the measure that we have defined. The reason for defining entropy rather than relative entropy can be attributed to the sparseness of the sentence probability models. In a way, this scoring mechanism is a special case of risk minimization framework.

Having discussed about the representation of the measure we will discuss few possible information sources which can be used to identify the relevant domain documents. Given a user’s information
need any of the external sources, like Wikipedia or World Wide Web, can be used to collect relevant domain corpus. Yahoo search engine [Yahoo] was used, to get a ranked set of retrieved html documents (limited by capability of processing other document types) from Wikipedia and WWW, with the topic title being the query. At most top \( n \) documents are assumed to be relevant and retrieved from their source website. These documents are parsed to extract text content and to remove any known unnecessary links when the structure of web page is known, e.g. all the web pages from Wikipedia follow the same html template. After performing the removal of stop words and stemming, required unigram language model, \( p(w|r) \), is learnt on the extracted text content, which can be interpreted as the probability of a word being central to the domain. This unigram model was then used in scoring the sentences as discussed above.

5 Evaluation

We have evaluated the effect of capturing sentence prior using DUC 2005 and 2006 data sets. The DUC task in 2005 and 2006 was to generate a brief (length \( \leq 250 \) words), well-organized, fluent answer to an information need given in the topic. In 2005, the data has 50 topics, and each topic is associated with a set of 25-50 relevant documents. Each topic has either 4 (for 30 topics) or 9 (for the rest of 20 topics) model summaries for the evaluation of peer summaries. 2005 topics also specify the granularity of the summary required, however we did not consider this while generating the summaries. 2006 data set has another 50 topics, but each with only 4 model summaries and the relevant documents for each topic were reduced to 25. We have used 2005 data set as a development data set to tune the parameters such as weights for each of the features and \( n \) (the number of results to be retrieved using search engine). DUC 2006 data set was used as test data to compare the effect of introducing the new feature.

In DUC, the evaluation of peer summaries was done both manually, for the Responsiveness, and by automatic evaluation techniques like ROUGE [Lin, 2004]. Responsiveness was primarily measured in terms of the amount of information in the summary that actually helps to satisfy the information need expressed in the topic. ROUGE is a recall oriented n-gram based similarity technique, which does a pair wise comparison between the peer summary and each reference summaries. Different ROUGE metrics [Lin, 2004] were proposed based on different similarity metrics. It was also shown in [Dang, 2005] that the automatic scores calculated using ROUGE-2 and ROUGE-SU4 correlated very well, Spearman coefficient of 0.95, 0.94 respectively, with the manual evaluations. So we have evaluated the performance of the new features using ROUGE system.

Table 2 shows the effect of introducing Information Measure during the sentence scoring. Baseline system returns the first 250 words of the most recent document for each topic as the summary. The system ID ‘Human Mean’ denotes the average ROUGE scores obtained by each of the model summaries when evaluated against rest of them. Table 2 suggests a significant performance improvement, compared to the score gap between the best performing peers (15 and 17), is possible when Information Measure was included in scoring the sentences. This improvement was achieved with a weight combination of 1 and 0.0004 for HAL and Information Measure respectively. The value of \( n \), the number of results to be retrieved from web was set to 10.
<table>
<thead>
<tr>
<th>System ID</th>
<th>ROUGE-2</th>
<th>ROUGE-SU4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human Mean</td>
<td>0.1025</td>
<td>0.1624</td>
</tr>
<tr>
<td>HAL + Information Measure</td>
<td>0.08591</td>
<td>0.14777</td>
</tr>
<tr>
<td>HAL</td>
<td>0.07618</td>
<td>0.13805</td>
</tr>
<tr>
<td>15</td>
<td>0.0725</td>
<td>0.1316</td>
</tr>
<tr>
<td>17</td>
<td>0.0717</td>
<td>0.1297</td>
</tr>
<tr>
<td>BaseLine</td>
<td>0.0402</td>
<td>0.0871</td>
</tr>
</tbody>
</table>

Table 2: Effect of including Information Measure on development data set

<table>
<thead>
<tr>
<th>Feature</th>
<th>ROUGE-2</th>
<th>ROUGE-SU4</th>
</tr>
</thead>
<tbody>
<tr>
<td>HAL + Info. Measure (web)</td>
<td>0.08591</td>
<td>0.14777</td>
</tr>
<tr>
<td>HAL + Info. Measure (wiki)</td>
<td>0.08247</td>
<td>0.14393</td>
</tr>
<tr>
<td>HAL</td>
<td>0.07618</td>
<td>0.13805</td>
</tr>
<tr>
<td>Information Measure (web)</td>
<td>0.07009</td>
<td>0.12803</td>
</tr>
<tr>
<td>Information Measure (wiki)</td>
<td>0.06061</td>
<td>0.11686</td>
</tr>
</tbody>
</table>

Table 3: Performance of individual features and a linear combination of features, on 2005 data

We have also tried to evaluate the contribution of individual features towards the final summary of our system. Table 3 show the performance of individual features Information Measure with Wikipedia and WWW, HAL and combination of Information Measure and HAL. It is clear that even though HAL is able to perform better than Information Measure individually, which is obvious as the latter does not take information need into consideration at all, a performance gain of 4-7% is observed when it is included along with HAL. The poor performance when Wikipedia is used as information source to collect the topic related documents is because of the fact that Wikipedia is sparse compared to WWW. For some DUC topics, e.g. d311i ‘VW/GM Industrial Espionage’, there were no related articles in Wikipedia. In such a case, we have directed the system to remove the first query word, and search again. This process is repeated as long as it finds some articles related to the query. During this process noise has been induced in to the model which resulted in poor performance. In the case of evidence collected from WWW, for most of the topics, we have found articles which have their heading same as the topic title, hence very likely to be a relevant document. So a better query model was built which resulted in better performance.

<table>
<thead>
<tr>
<th>No. Results</th>
<th>ROUGE-2</th>
<th>ROUGE-SU4</th>
</tr>
</thead>
<tbody>
<tr>
<td>n = 2</td>
<td>0.06930</td>
<td>0.12585</td>
</tr>
<tr>
<td>n = 5</td>
<td>0.06686</td>
<td>0.12427</td>
</tr>
<tr>
<td>n = 10</td>
<td>0.07009</td>
<td>0.12803</td>
</tr>
</tbody>
</table>

Table 4: Performance of Information Measure computed using Web with respect to the number of results extracted from WWW

Table 4, shows the effect of the top $n$ documents retrieved on the query model. In all the experiments reported from here, it is assumed that the value of $n$ is set to 10 in collecting the evidence,
unless mentioned explicitly. In the rest of the experiments we have reported the results only when WWW is used as source for Information Measure, as similar behavior was observed with Wikipedia also.

<table>
<thead>
<tr>
<th>Wt. for Hal; Info. Measure</th>
<th>ROUGE-2</th>
<th>ROUGE-SU4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Only HAL</td>
<td>0.07618</td>
<td>0.13805</td>
</tr>
<tr>
<td>1.0 : 1.0</td>
<td>0.07009</td>
<td>0.12804</td>
</tr>
<tr>
<td>1.0 : 0.1</td>
<td>0.07083</td>
<td>0.12858</td>
</tr>
<tr>
<td>1.0 : 0.01</td>
<td>0.07215</td>
<td>0.13089</td>
</tr>
<tr>
<td>1.0 : 0.001</td>
<td>0.08267</td>
<td>0.14351</td>
</tr>
<tr>
<td>1.0 : 0.0001</td>
<td>0.07893</td>
<td>0.14103</td>
</tr>
<tr>
<td>1.0 : 0.0003</td>
<td>0.08323</td>
<td>0.14516</td>
</tr>
<tr>
<td><strong>1.0 : 0.0004</strong></td>
<td><strong>0.08591</strong></td>
<td><strong>0.14777</strong></td>
</tr>
<tr>
<td>1.0 : 0.0005</td>
<td>0.08590</td>
<td>0.14748</td>
</tr>
<tr>
<td>1.0 : 0.0006</td>
<td>0.08623</td>
<td>0.14768</td>
</tr>
<tr>
<td>1.0 : 0.0007</td>
<td>0.08502</td>
<td>0.14648</td>
</tr>
</tbody>
</table>

Table 5: The ROUGE scores obtained for different weight combinations, on 2005 data

When more than one feature was available, a weighted linear combination of these individual feature values was used to compute the final sentence score. This section describes the search in the weights space that gave the best performance of the system. Note that we didn’t do an exhaustive search of the weights space, so there is a possibility of ending at local maximum. Table 5 show the ROUGE scores obtained during the search for appropriate weights combination. The first column of the table give the weights used to both HAL and Information Measure features. We have started the search process with assigning equal weights to both HAL and Information Measure, WWW is used as information source. The ranking of the sentences achieved with this combination was almost same as the ranking obtained when only Information Measure feature was used, and hence the same ROUGE scores. This is because of the fact that the relative order of both the feature values are not same, in one case it is product of probabilities and sum of log of probabilities in other case. For practical considerations we have raised the HAL scores by a huge factor (10000) because of which the rankings are biased. A weight combination of 1 for HAL and 0.0004 for Information Measure has generated best peer summaries. The winning combination of weights is also because of the difference in the relative order of the individual scores. Since the WWW is more dynamic, the performance improvement may not be exactly reproducible. However based on its behavior across the 2005 topics and the wide coverage of the topics we can see that performance improvement still lies in a range of 6-7%, when WWW is used as information source. The t-test analysis of the results show that the performance gain is statistically significant (p ∼ 0.05).

With these two features we have submitted our system to DUC 2006 challenge. Table 6 shows that our system (IIITH-Sum) with system ID 24 has outperformed the rest of the participant systems. The performance gap between our system and second best system is wider than the gap between second best and third best system. Table 7 shows the 95% confidence interval of each evaluation metric for our system in comparison to the confidence interval of next best system. It can be seen
Table 6: Official scores of summarization systems at DUC-2006, sorted based on ROUGE-2 scores

<table>
<thead>
<tr>
<th>System ID</th>
<th>ROUGE-2</th>
<th>ROUGE-SU4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human Mean</td>
<td>0.11</td>
<td>0.17</td>
</tr>
<tr>
<td>24</td>
<td>0.09558(1)</td>
<td>0.15529(1)</td>
</tr>
<tr>
<td>15</td>
<td>0.09097(2)</td>
<td>0.14733(3)</td>
</tr>
<tr>
<td>12</td>
<td>0.08987(3)</td>
<td>0.14755(2)</td>
</tr>
<tr>
<td>8</td>
<td>0.08954(4)</td>
<td>0.14607(4)</td>
</tr>
<tr>
<td>23</td>
<td>0.08792(5)</td>
<td>0.14486(6)</td>
</tr>
<tr>
<td>BaseLine</td>
<td>0.07</td>
<td>0.13</td>
</tr>
</tbody>
</table>

Table 7: 95% confidence intervals for the evaluation metric as mentioned by column name, official results of DUC 2006

<table>
<thead>
<tr>
<th>System</th>
<th>ROUGE-2</th>
<th>ROUGE-SU4</th>
</tr>
</thead>
<tbody>
<tr>
<td>IIITH-Sum</td>
<td>0.09144 - 0.09977</td>
<td>0.15126 - 0.15906</td>
</tr>
<tr>
<td>Next Best</td>
<td>0.08671 - 0.09478</td>
<td>0.14360 - 0.15142</td>
</tr>
</tbody>
</table>

6 Relevance Models

In this section we will argue, using Probability Ranking Principle [Robertson S.E., 1977], that a measure such as Information Measure which is independent of the actual users information need can be used to score the candidates based on their relevance towards the users query.

The fundamental question that applications such as Information Retrieval, Question Answering and Query-Based Summarization try to address is “How probable is that a candidate is relevant to a user’s information need?”. If \( Q, D \) and \( R \) denote query, document and the relevance, then Probability Ranking Principle [Robertson S.E., 1977] states that an optimal performance can be achieved by any Information Retrieval (IR) system, if the documents are ranked in the order of decreasing probability of relevance to the user’s query. If \( R = r \) denotes the document is relevant and \( R = \bar{r} \) denote it is non-relevant for the query, we may use the following log-odds ratio to rank the documents for optimal performance:

\[
\log(\text{rank}(D)) = \log \frac{p(R = r|D, Q)}{p(R = \bar{r}|D, Q)} = \log \frac{p(D, Q|r)}{p(D, Q|\bar{r})} \frac{p(r)}{p(\bar{r})}
\]

In the Robertson-Sparck Jones approach [Jones et al., 2000], the probability \( p(D, Q|r) \) is factored as \( p(D, Q|r) = p(D|Q, r) p(Q|r) \). Instead, we can also decompose \( p(D, Q|r) \) as \( p(Q|D, r) p(D|r) \) [Lafferty and Zhai, 2003]. Intuitively the former decomposition corresponds to the probability that
an author will write a relevant document given query, while the latter associates the probability of generating a query from a document with it being relevant. These two decompositions lead to different statistical models: Robertson-Sparck Jones approach [Jones et al., 2000] and Language Modeling approach [Ponte and Croft, 1998]. Before making any assumptions to estimate the model parameters, the two types of models are equivalent in a probabilistic sense. Language model makes the assumption [Lafferty and Zhai, 2003] that conditioned on the event of non-relevance, the document is independent of the query given non-relevance i.e. $p(D, Q|\bar{r}) = p(D|\bar{r}) p(Q|\bar{r})$. Following this assumption, equation 3 can be rewritten as:

$$
\log(rank(D)) = \log \frac{p(D, Q|r) p(r)}{p(D|\bar{r}) p(Q|\bar{r})} \\
\overset{rank}{=} c \cdot \log \frac{p(D, Q|r)}{p(D|\bar{r})} \\
= c \cdot \log \frac{p(Q|D, r)}{p(D|\bar{r})} \\
= \alpha \log p(Q|D, r) + \beta \log \frac{p(D|r)}{p(D|\bar{r})}
$$

The first part of the equation 4, $p(Q|D, r)$, is responsible for query dependent ranking of a document/sentence and researchers have attempted [Ponte and Croft, 1998] to calculate it using language modeling. The second part, $p(D|r)/p(D|\bar{r})$, essentially captures the explicit notion of importance or prior of a document/sentence. This allows other forms of evidence that are query independent to be incorporated into the ranking process. This is calculated based on the distribution of its constituent words among relevant and non-relevant document sets. The Information Measure that we have defined will augment the functionality of the second term of the equation 4 for summarization application. This justifies the fact that capturing sentence prior using document corpus collected from external source will improve the performance consistently and not merely by chance. Apart from this fact, a feature such as this would capture the centrality of any word/sentence with respect to the domain corpus which is very essential for a summarization application as it brings some coherence into the generated summary.

7 Related Work

Li et al. (2005) explored the use of entity-based features, a blend of factoid question answering approaches involving identification of question type and answer type, in validating the appropriateness of a sentence to be included in the final summary. Ye et al. (2005) have discussed the concepts identified from the sentences and their definitions in Wordnet [Miller, 1990] to find the similarity of a sentence with the rest of the sentences. This approach is limited by the presence of a concept in the wordnet. Daume and Marcu (2005) assumed a graphical model in which words are drawn from a mixture of general english, document specific and query specific language models and used Bayesian framework to estimate the parameters. Even though these approaches perform competitively to HAL feature alone, a significant improvement is observed with the addition of Information Measure. SQUASH [Melli et al., 2005] explored the application of Natural Language Processing
techniques to identify and use the semantic dependencies across the document for summary generation. But their semantic role labeler perform with an F-measure of 42.2%, which will introduce lot of errors into the semantic graph.

8 Conclusion

In this paper we have presented a sentence extraction based technique to generate a fixed length summary from a cluster of documents, which answers the user’s information need. This task is different from normal factoid or definitional question answering task as the information need cannot be answered by simple facts. Based on the observation that a measure which captures the domain centrality of a sentence is needed for an application like query-based multi-document summarization, we have defined a generic measure called Information Measure which has roots in information theory and risk minimization framework. The experiments conducted on 2005 data set show a performance improvement of 4% and 7% when Wikipedia and WWW were used as information sources. The same system has outperformed the rest of the participant systems at DUC 2006 with a significant margin over the next best system.

References


