Query Independent Sentence Scoring approach to DUC 2006

Jagadeesh Jagarlamudi j_jagdeesh@research.iiit.ac.in
Prasad Pingali pvvpr@iiit.ac.in
Vasudeva Varma vv@iiit.ac.in

Language Technologies Research Centre
International Institute of Information Technology
Hyderabad, India

Abstract

The task in Document Understanding Conferences (DUC[1]) 2006 is to generate a fixed length, user oriented, multi document summary, which remains same as that of DUC 2005. We have used two features to score the sentences based. The sentences are picked to form the summary based on the calculated score. The first feature is a query dependent scoring of a sentence which is an improvement over the HAL feature. The second feature is based on the observation that sentence importance, which is independent of the query, needs to be captured in the current approaches. We have explored the use of web in scoring the sentences in a query independent manner. Experiments show a performance gain of 6-7% over HAL feature by the inclusion of two new features. Our summarization system was ranked 1^{st} in all automatic evaluations with significant margin from second best system, 5^{th} in responsiveness and 9^{th} in linguistic quality evaluations in DUC 2006. Relatively lower performance in linguistic quality can be attributed to the stripped off sentences at the end of summary, when the summary length is exceeding 250 words.

[1]http://duc.nist.gov

1 Introduction

The task in DUC 2006 remains same as that of DUC 2005 with some small modifications. The task was to synthesize from a set of 25 documents a brief, well-organized, fluent answer to a need for information that cannot be met by just stating a name, date, quantity etc. That is, given a user's information need, as a DUC topic, and a cluster of documents relevant to the DUC topic, the system needs to create, from the document set, a summary which answers the information need expressed. A DUC topic is made up of two parts. First part is the title of the topic while second part is the actual information need expressed as a single question or multiple questions. One major difference in DUC 2006 compared to DUC 2005 task is that the granularity of the required summary which is a part topic was removed. The title of the topic is general enough to relate to all the documents in the cluster. The information need is both specific and complex, hence this task differs from the normal query based summarization where techniques like bag of words and lexical chains(Okumura et al., 1999) are used to represent the query and documents. This task is also different from normal factoid or definition based question answering tasks(Dumais et al., 2002; Hovy et al., 2000; Zheng, 2002), because the information need cannot be met by just stating a name, date or quantity etc. However, at a broader level it can be seen as topic-oriented, informative multi-document summarization(Berger and Mittal, 2000; Schlesinger and Baker, 2001; Radev et al., 2003), where the goal is to produce a summary, from a set of multiple docu-

ments, which is biased towards the topic.

A system that addresses such a complex task may involve the following stages; information need enrichment, content selection and summary generation. All the three stages need an effective combination of Natural Language Processing and Information Retrieval techniques. 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, Concept linkages, and Bayesian framework (Daume and Marcu, 2005; Blair-Goldensohn, 2005; Ye et al., 2005; Li et al., 2005; Schlesinger and Baker, 2001) provided a way to achieve good performance with out involving such a deep processing of text.

Most of these approaches involve scoring of a sentence based on its relevance towards the query/information need. And they don't capture the notion of sentence importance/prior which is independent of query during the content selection. A sentence which is not related to the query might still be important because of either user preferences or some external knowledge. PageRank is an example of query independent ranking of a document in case of web search engine. In this paper we tried to calculate the query independent importance for a sentence using the documents which are pseudo relevant to the topic. We have also modified HAL feature(Jagadeesh et al., 2005) to incorporate weighted query phrases into sentence scoring.

The rest of the paper is organized as follows. Section 2 describes the motivation for considering sentence importance which is independent of the query. Section 3 gives an overview of the system while sections 4 and 5 give a detailed description of the features used to score a sentence. Section 6 discusses the parameter selection based on DUC 2005 data set and report the official results indicating the performance of our summarization system relative to other systems.

2 Motivation

Given the random variables Q, D and R denote query, document and the relevance, the Probability Ranking Principle (Robertson S.E., 1977) says 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 users query. Equivalently, we may use the following log-odds ratio to rank the documents;

$$\begin{aligned} log(rank(D)) &= log\frac{p(R|D,Q)}{p(\bar{R}|D,Q)} \\ &= log\frac{p(D,Q|R)\ p(R)}{p(D,Q|\bar{R})\ p(\bar{R})} \end{aligned} \quad (1)$$

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 $p(Q|D,R)\ p(D|R)$ (Laferty and Zhai, 2003). Before making any assumptions, the two types of models are equivalent in a probabilistic sense. Making the assumption (Laferty and Zhai, 2003) conditioned on the event of non-relevance, the document is independent of the query given non-relevance i.e. $p(D,Q|\bar{R}) = p(D|\bar{R})\ p(Q|\bar{R})$ equation 1 can be rewritten as;

$$log(rank(D)) = log\frac{p(D,Q|R)\ p(R)}{p(D|\bar{R})\ p(Q|\bar{R})\ p(\bar{R})}$$

Neglecting the probabilities which are independent of document, it can be rewritten as;

$$\begin{aligned} log(rank(D)) &\overset{rank}{=} log\frac{p(D,Q|R)}{p(D|\bar{R})} \\ &= log\frac{p(Q|D,R)\ p(D|R)}{p(D|\bar{R})} \\ &= log\ p(Q|D,R) + log\frac{p(D|R)}{p(D|\bar{R})} \end{aligned} \quad (2)$$

The first part of the equation 2, $p(Q|D,R)$, is responsible for query dependent ranking of a document and researchers have attempted to calculate it using language modeling (Ponte and Croft, 1998). The second part, $p(D|R)/p(D|\bar{R})$, essentially captures the explicit notion of importance or prior of a document. This allows other forms of evidence that are query independent to be incorporated into the ranking process.

Regarding query-based summarization, Relevance based language modeling framework along with the representation of words in higher dimensional spaces (Lund and Burgess, 1996) as discussed

in (Jagadeesh et al., 2005), can be seen as an instantiation of the first term to score a sentence. In this paper we discussed an improvement which considers the query as a set of weighted phrases instead of a bag of equally important key words. The second term in equation 2 has motivated us to look for sentence ranking mechanisms which are independent of query yet capturing its importance. One issue in computing $p(D|R)/p(D|\bar{R})$ is to identify the relevant (R) and non-relevant ($\bar{R}$) set of documents. In this paper we used web to obtain the relevant set of documents and discussed ways to use them in sentence scoring.

3 Architecture

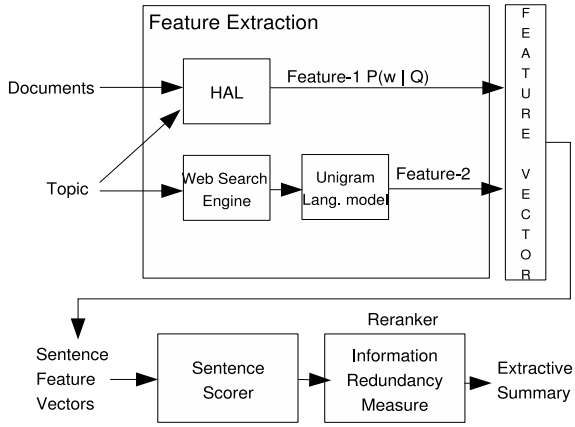

Figure 1: Architecture of our summarization system

In our system, the summaries are generated using a three step architecture (Figure 1) which is very similar to MEAD (Radev et al., 2003). In the first step, called feature extraction, the system extracts feature values for each sentence. These features may/may not be dependent on the query. 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. This normalization will facilitate an easier combination of different feature scores for a sentence. The second phase involves combining the information obtained in the form of features for each sentence. Currently the system supports only a weighted linear combination of the feature values with the weights being assigned by the user. The third phase, called re-ranker, takes the scored sentences and selects a subset of sentences which form the summary satisfying the required constraints. The re-ranker checks for the redundancy of information, cosine similarity measure, across the summary sentences and selects sentences towards final summary. To improve the conciseness of the summary, it also removes phrases like 'according to ⟨RegEx to match News agency⟩','on ⟨weekday⟩', 'for example', 'this year' and some stop words like 'and', 'but', 'also'.

The following sections describe different features that were used to score a sentence. The first feature is a query dependent scoring of a sentence which is an improvement over the HAL feature discussed in (Jagadeesh et al., 2005), while the second feature is independent of the query and attempts to capture importance of the sentence based on a set of pseudo relevant documents.

4 Query Dependent Sentence Score

In (Jagadeesh et al., 2005), it has been shown that relevance-based language modeling(Lavrenko and Croft, 2001) along with semantic representation of words in higher dimensions using HAL spaces(Lund and Burgess, 1996) can be extended to calculate the relevance score of a sentence towards the information need. But the authors treated the query as a bag of words and all of them are equally important. Here we have considered extending it to phrasal level and giving additional importance to a query word/phrase.

For a given DUC topic, its title and description are passed through a chunker to identify chunks or meaningful phrases. Now the probabilistic dependencies of an identified phrase on a word, as required by HAL spaces, are calculated using a normalized weighted linear combination of the constituent words of the phrase. The weights being proportional to the POS tag of each constituent word. Noun phrases are given more importance than verb phrases and adjectives, while no importance is given to conjunctions and determiners. If qp denotes a phrase, i.e. the new dimension which needs to be added, then the projections of this phrase on a word

w can be calculated as:

$$\begin{aligned} p(qp|w) &= c\,\frac{\text{co-occurrence strength of } qp \text{ with } w}{n(w)\times K} \\ &\approx \frac{c}{n(w)\times K}\sum_{w_j\in qp}\text{co-occurrence}(w_j,w) \\ &= c`\sum_{w_i\in qp} p(w_i|w) \end{aligned}$$

where $n(w)$ denote the unigram frequency of the word, and K denote the window size considered during the calculation of co-occurrence strength of word pairs. To ensure a valid probability distribution, the constraint $\sum p(\cdot|w) = 1$ is imposed. The reader is encouraged to refer to (Jagadeesh et al., 2005) for detailed description about application of HAL spaces to summarization.

We have also considered giving additional importance to word/phrase of the query. A query word or phrase is given a weight equal to its TFIDF score, the IDF of a phrase being the average IDF of all its constituent words. To incorporate a query word/phrase importance into sentence scoring, the joint probability $p(w, q_1\ ...\ q_k) = p(w)\prod_{q_j} p(q_j|w)$ is modified as

$$p(w, q_1\ ...\ q_k) = p(w)\prod_{q_j} p(q_j|w)^{\frac{1}{TFIDF(q_j)+1}}$$

Lower the value of TFIDF score for a word/phrase lower is the final score of a sentence.

5 Query Independent Sentence Score

The issue in calculating the prior/importance of a sentence lies in identifying the relevant document set. Here we will discuss how external resources like web can be used to compute the prior of a sentence. Web is a huge source of information and today search engines have achieved some level of accuracies in determining the relevant information to be presented to the users. For any given topic web can be used to get relevant document set. With the DUC topic title being the query, we have used Yahoo search engine (Yahoo,), to get a ranked set of retrieved documents from web. Limited by the capability of processing of different file types, we have restricted the search process only to html documents. Of the result set, at most top n documents are marked as relevant and retrieved from their corresponding source websites. Since there is a possibility for some of these documents to be non-relevant we refer to this set as pseudo relevant documents. These documents are parsed to extract text content. After performing the removal of stop words and stemming, a unigram language model, is learnt on the extracted text content, which can be interpreted as the probability of a word being related to the information need.

The following subsections describe two different scoring mechanisms that we have explored in scoring a sentence using the learnt unigram language model.

5.1 Equation based measure

Since we have identified only the relevant document set, we neglect the non-relevant document set while scoring sentences. After the unigram language model is learnt from the relevant documents, the second part of equation 2 can be calculated by making an independence assumption between the words. If S denotes a sentence from the cluster of input documents then it is assigned a score of

$$log\ p(S|R) = log\ \prod_{w\in S} p(w|R) = \sum_{w\in S} log\ p(w|R)$$

5.2 Information Measure in a Sentence

We have used entropy to compute the information content of a sentence based on the learnt unigram model. If $P(x)$ is a probability density function (with respect to the counting measure) for a discrete random variable X, Shanon(Shanon and Weaver, 1983) defined a measure of information content called self-information or surprisal of a message x;

$$I(x) = -\log p(x),$$

And the entropy, or uncertainty, of a discrete message space X is the expected self-information of a message x from that message space:

$$\begin{aligned} H(X) = E\{I(x)\} &= \sum_{x\in X} p(x)I(x) \\ &= -\sum_{x\in X} p(x)\log p(x) \end{aligned}$$

Entropy can be seen as a measure of information content in a message. If a symbol has zero proba-

bility, which means it never occurred, it should not affect the entropy. So we let $0log0 = 0$.

If $p(w|R)$ denote the unigram language model learnt from the identified relevant documents, which also denote the probability of a word being relevant to the information need, and $S = w_1, w_2, ..., w_k$ be a message in the words space then by the definition of entropy, the amount of relevant information contained in that sentence can be defined as;

$$H(S|R) = -\sum_{w \in S} p(w|R) \log p(w|R)$$

In both the cases, the final sentences are ranked based on weighted linear combination of modified HAL feature and one of the above defined measures. The final sentence score is

$$W_1 \cdot log\ p(Q|D,R) + W_2 \cdot \sum_{w \in S} p(w|R)$$

or

$$W_1 \cdot log\ p(Q|D,R) + W_2 \cdot H(S|R)$$

6 Evaluation

In this section we report the performance of individual features as well as combination of these features. For combining the features, the parameter values are calculated from DUC 2005 data set. In DUC 2005, the evaluation of peer summaries was done both manually, for responsiveness, and by automatic evaluation techniques like ROUGE(Lin and Hovy, 2003). 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 DUC topic. It was also shown in (Dang, 2005) that the automatic scores calculated using ROUGE-2 and ROUGE-SU4 correlated very well with the manual evaluations. So we have evaluated the performance of the new features using ROUGE system. Table 1 shows the ROUGE scores obtained by the summaries generated for DUC 2005 data set, when individual features are used to score the sentences. The modified HAL feature is able to generate more relevant summaries when compared to other features. From the results it is clear that considering the query as a set weighted phrases brings a considerable improvement.

The value of n, the number of results to be retrieved for pseudo relevant documents is set to 10.

Feature	ROUGE-2	ROUGE-SU4
Modified HAL	0.07926	0.14069
HAL	0.07618	0.13805
Information Measure	0.07083	0.12876
Equation based	0.02361	0.07291

Table 1: Performance of individual features on DUC 2005 data set

Table 2 shows the effect of top n documents retrieved on the unigram model learnt, hence on the Information Measure based sentence scoring. In all the experiments reported from here, it is assumed that the value of n is set to 10 in collecting the pseudo relevant documents, unless mentioned explicitly.

No. Results	ROUGE-2	ROUGE-SU4
n = 2	0.06930	0.12585
n = 5	0.06686	0.12427
n = 10	0.07083	0.12876

Table 2: Performance of Information Measure with respect to the number of results extracted from web

As in many of the extractive summarization systems, once sentences in the input document cluster are scored based on different features, the final score of sentence is computed as a weighted linear combination of the individual feature values. We finalized the weights based on a trial and error method on DUC 2005 data set. We picked the two features namely modified HAL and Information Measure and explored the weighted combination of both these features. 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 for other combination of weights which could give better results. Table 3 shows the ROUGE scores obtained during the search for appropriate weight combination. The first column of the table give the weights used to both modified HAL and Information measure features. We have started the search process with assigning equal weights to both features. The ranking of the sentences achieved with this combination is almost same as the ranking obtained when

Wt. for modified Hal : Information measure	ROUGE-2	ROUGE-SU4
Only HAL	0.07618	0.13805
1.0 : 1.0	0.07083	0.12876
1.0 : 0.1	0.07167	0.12949
1.0 : 0.01	0.07298	0.13178
1.0 : 0.001	0.08393	0.14533
1.0 : 0.0001	0.08271	0.14421
1.0 : 0.0002	0.08389	0.14504
1.0 : 0.0003	0.08486	0.14588
1.0 : 0.0004	0.08494	0.14666
1.0 : 0.0005	**0.08640**	**0.14724**
1.0 : 0.0006	0.08596	0.14706
1.0 : 0.0007	0.08557	0.14698

Table 3: The ROUGE scores obtained for different weight combinations for DUC 2005 data

System ID	ROUGE-2	ROUGE-SU4	ROUGE-BE	Responsiveness
Human Mean	0.11	0.17	0.07	4.74
24	0.09558(1)	0.15529(1)	0.05107(1)	2.88(5)
15	0.09097(2)	0.14733(3)	0.04852(3)	2.48(23)
12	0.08987(3)	0.14755(2)	0.04710(7)	2.92(4)
8	0.08954(4)	0.14607(4)	0.04783(5)	2.58(15)
23	0.08792(5)	0.14486(6)	0.05049(2)	3 (2)
BaseLine	0.07	0.13	0.04	2.54

Table 4: Official scores of summarization systems at DUC-2006, sorted based on ROUGE-2 scores

only Information Measure is used, and hence the same ROUGE scores. This is because of the fact that the scores of modified HAL feature value for different sentences are very close, so the ranking of the sentences obtained by Information Measure dominated the effect of modified HAL feature. A weight combination of 1 for modified HAL and 0.0005 for Information Measure has generated best summaries, row shown in bold. Even though modified HAL alone does better than Information Measure (from table 1), better performance is achieved when both the features are combined. When Information Measure feature has been included, with the value of n being set to 10, the performance gain is 6.58% and 4.65% when compared to HAL and modified HAL respectively. Since the web is more dynamic, the pseudo relevant set may not be same each time you query and hence the performance improvement may not be exactly reproducible. However based on the consistent improvement across all topics and its wide coverage have made a good reason for Information Measure to be included in our DUC 2006 participation.

In the system that we have submitted, based on the experiments done on DUC 2005 data set, we used a weights combination of 1 for modified HAL feature and 0.0005 for Information Measure. The evaluation criteria in DUC 2006 remained same as DUC 2005, except the fact that all the topics have same number of model summaries in contrast with some

topics having 9 and rest having 4 model summaries in 2005. Table 4 shows the performance of our system when compared to the best 10 systems interms of ROUGE-2 scores. In each case the rank obtained by a DUC participant system, under the evaluation criteria mentioned as the column name, is shown in braces. Our summarization system, with system ID 24, was ranked 1^{st} in all automatic evaluations, 5^{th} in responsiveness and 9^{th} in linguistic quality evaluations respectively. It can also be observed from the ROUGE scores that there is a significant margin between our system and next best system. Of the 50 summaries generated, more than half of our summaries were exceeding the length limit of 250 words. The stripped off sentences, when the summary length is more than 250 words, might be a reason for its relatively less impressive performance in linguistic quality evaluations. Lower linguistic quality might also be the reason for its relatively lower performance in responsiveness evaluations.

7 Conclusion

Our system performed well with the respect to the rest of DUC 2006 participant systems. We used a sentence extraction based technique to generate a multi-document summary which answers the specific information need given as a query. We have explored different ways to include the sentence importance/prior in sentence scoring. Even though the web is more dynamic, it is contributing positively towards the final summaries. We have also suggested some improvements for the HAL feature to include weighted query phrases instead of considering it as a bag of equally important words.

References

Adam L. Berger and Vibhu O. Mittal. 2000. Query-Relevant Summarization Using FAQs. In *Proceedings of Association for Computational Linguistics ACL 2000*, pages 294–301.

S. Blair-Goldensohn. 2005. Columbia university at duc 2005. In *Prceedings of Document Understanding Conferences*.

Hoa Trang Dang. 2005. Overview of duc 2005 (draft). In *Proceedings of Document Understanding Conferences*.

Hal Daume and Daniel Marcu. 2005. Bayesian multi-document summarization at mse. In *ACL 2005, Workshop on Multilingual Summarization Evaluation*.

Susan Dumais, Michele Banko, Eric Brill, Jimmy Lin, and Andrew Ng. 2002. Web question answering: is more always better? In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 291–298, New York, NY, USA. ACM Press.

E. Hovy, L. Gerber, U. Hermjakob, M. Junk, and C.-Y. Lin. 2000. Question answering in webclopedia. In *Proceedings of the TREC-9 Conference*, NIST, Gaithersburg, MD.

J Jagadeesh, Prasad Pingali, and Vasudeva Varma. 2005. A relevance-based language modeling approach to duc 2005. In *Proceedings of Document Understanding Conferences (along with HLT-EMNLP 2005), Vancouver, Canada*.

Karen Sparck Jones, Steve Walker, and Stephen E. Robertson. "2000". "a probabilistic model of information retrieval: development and comparative experiments - part 2". *"Information Processing and Management"*, "36"("6"):"809–840".

J. Laferty and C. Zhai. 2003. Probablistic relevance models based on document and query generation. In *Language Modeling for Information Retrieval*, pages 1–10. Kluwer Academic Publishers.

Victor Lavrenko and W. Bruce Croft. 2001. Relevance-based language models. In *International ACM SIGIR conference on Research and development in Information Retrieval*, pages 120–127.

W. Li, W. Li, B. Li, Q. Chen, and M. Wu. 2005. The hong kong polytechnic university at duc 2005. In *Proceedings of Document Understanding Conferences*.

Chin-Yew Lin and E.H. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of 2003 Language Technology Conference (HLT-NAACL 2003), Edmonton, Canada*.

K Lund and C Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. In *Behavior Research Methods, Instrumentation, and Computers*, pages 203–208.

Manabu Okumura, Hajime Mochizuki, and Hidetsugu Nanba. 1999. Query-Biased Summarization Based on Lexical Chaining. In *Proceedings of the Pacific Association for Computational Linguistics*, pages 324–334.

Jay M. Ponte and W. Bruce Croft. 1998. A language modeling approach to information retrieval. In *Research and Development in Information Retrieval*, pages 275–281.

Dragomir Radev, Timothy Allison, Sasha Blair-Goldensohn, John Blitzer, Arda Çelebi, Elliott Drabek, Wai Lam, Danyu Liu, Hong Qi, Horacio Saggion, Simone Teufel, Michael Topper, and Adam Winkel. 2003. The MEAD Multidocument Summarizer. http://www.summarization.com/mead/.

Robertson S.E. 1977. The probability ranking principle in ir. *Journal of Documentation*, 33:294–304.

Judith D. Schlesinger and Deborah J. Baker. 2001. Using Document Features and Statistical Modeling to Improve Query-based Summarization. In *Proceedings of Workshop on Document Understanding Conferences, DUC01*, New Orleans, LA.

C. E. Shanon and W. Weaver. 1983. *Mathematical Theory of Communication*. University of Illinois Press.

Yahoo. *http://search.yahoo.com/*.

S. Ye, L. Qiu, and T.-S. Chua. 2005. Nus at duc 2005: Understanding documents via concept links. In *Proceedings of Document Understanding Conferences*.

Z. Zheng. 2002. Answerbus question answering system. In *Human Language Technology Conference (HLT)*.