# Query Focus Guided Sentence Selection Strategy for DUC 2006

**Wenjie Li, Baoli Li, Mingli Wu**

Department of Computing
The Hong Kong Polytechnic University, Hong Kong
{cswjli, csblli, csmlwu}@polyu.edu.hk

## Abstract

This paper presents our new query-based multi-document summarization system for DUC 2006. It is an extended version of a generic multi-document summarization system developed previously (namely PoluS 1.0) which incorporates latent semantic analysis (LSA) technology. To make the generated summaries satisfying user's information need as possible as we can, we propose a query focus guided sentence selection strategy. The evaluation results show that our system ranks in the middle among 34 submitted systems. Although there is still room to improve the current version of PoluS, it provides a good framework for our future research on multi-document summarization.

## 1 Introduction

This is the second time that our group attended the DUC evaluations. In DUC2005, we built our system on MEAD[1] framework with four especially designed features. They are entity-based feature, pattern-based feature, term-based feature, and semantic-based feature (Li et al 2005). These new features help to enhance the system's judgment on whether and to what extent a sentence is relevant to a user's query. Our system ranks competitively in DUC 2005, especially in ROUGE evaluations.

In the last year's evaluation, we utilized several different components, such as GATE for POS tagging and named entities tagging, sentence segmentation tool for sentence splitting, WordNet-SenseRelate-AllWords package for word sense disambiguation, WordNet-Similarity package for calculating word sense based similarity, and so on. They are written in different languages and run on different platforms. It is inconvenient to do extensive experiments with many different components. Therefore, we decided to design and implement a brand new integrated multi-document summarization system, which does not process a cluster with separate components. Our new system named **PoluS** (HK **Pol**y**U** **S**ummarization system) takes a cluster as input, and outputs the final summary according to user's requirement. We hope that PoluS would be a good foundation for our future research on multi-document summarization.

As MEAD is successful in the past DUC evaluations (Erkan and Radev 2004a), we decided to incorporate good practices of MEAD in our PoluS system, such as centroid-based strategy (Radev et al 2004). In PoluS system, we also integrated the English analysis tool OAK (Sekine 2002) to do sentence splitting, tokenizing, stemming, POS tagging, named entities tagging, and chunking.

The task for DUC 2006 is the same as the last year's task, except that the documents come from the AQUAINT corpus rather than TDT corpus and each topic is not specified a desired granularity of a generated summary. It is essentially a query-based multi-document summarization task, but the query is expressed by a coherent paragraph or several related sentences. Although our system achieved much better performance in DUC 2005, we decided not to reuse the old system for this year's evaluation. We focus more on research rather than simply striving for the best scores. We expect that we could contribute more by exploring other strategies.

The rest of this paper is organized as follows. Section 2 gives the overview of our PoluS system for DUC 2006. Then we focus on our kernel strategy for dealing with user's information need. Section 4 presents the experiments and evaluation results. Finally, we conclude in section 5.

[1] http://www.summarization.com/mead/

## 2 Our System for DUC2006

Currently, our PoluS system is a sentence extraction summarization system, which does not resolve coreference, and does not try to compress or fuse sentences.

The basic PoluS system focuses on generic multi-document summarization, and like MEAD it takes the centroid based strategy to select sentences according to the similarity between a sentence and the centroid. Each sentence is represented as a term vector, and then the average of all the sentence vectors is regarded as their centroid. With the centroid-based values, we can obtain a ranked list of sentences. The final summary is generated with the Diversity-Based Reranking technology, i.e. MMR (Carbonell and Goldstein 1998).

The above description gives an overview of our basic PoluS system. Actually, we further enhance it with the following technology: named entities recognition and latent semantic analysis. We obtain named entities and 2-grams in noun phrases to enrich the document representation, i.e. including these derived terms besides words. Moreover, as term relatedness is completely ignored in calculating the similarity between two vectors, two highly similar sentences may be judged as irrelevant due to the terms' mismatch. Therefore, we try to use latent semantic analysis to implicitly discover the underlying term relations. We replace the original term/sentence matrix with the singular value decomposed one. The centroid and a sentence's centroid-based salience value are computed on this derived matrix.

To adapt the PoluS system to DUC 2006's task, we add two topic related salience values to the sentence importance: query term based value and query focus based value. They are linearly combined with the centroid based value to determine the importance of a sentence. In addition, we propose a query focus guided sentence selection strategy to direct the summary generation procedure. We hope that the generated summary could cover user's information requirement as many as possible.

In the following subsections, we will give some technical details of latent semantic analysis and query term based salience used in our system. We will deal with query focus based value and query focus guided sentence selection strategy in the next section.

### 2.1 Latent Semantic Analysis

Gong and Liu (2001) first introduce latent semantic analysis in generic single document summarization. Dou et al. (2004) also validate its effectiveness by an extrinsic evaluation. Its success in single document summarization attracts us to explore whether LSA is more applicable for multi-document summarization, as a larger sentence collection may determine a more accurate semantic space.

Unfortunately, there are few efforts in this direction. Yeh et al. propose a LSA-based T.R.M (text relationship map (Salton et al. 1997)) method to do both single and multiple document summarizations. Their experiments on a small corpus (5 clusters with totally 100 articles on politics) conclude that LSA can be employed to promote text summarization from keyword-level analysis to semantic-level analysis. Their method essentially combines LSA with (degree) centrality based summarization method (Erkan and Radev 2004b), where we will try to combine LSA with centroid based summarization method, because Erkan and Radev (2004b) show that centroid method is as good as the (degree) centrality based method.

In the past DUC evaluations, we find only one participant explicitly indicates that their method applies the LSA technology (Hachey et al., 2005). They utilized singular value decomposition via the Infomap tool to derive a semantic word space from a 100-million-word corpus that consists of Acquaint and DUC 2005 data. Each word is represented by a word vector. Then a given sentence can be represented as the average of its constituent word vectors. Their system ranked median in DUC 2005. We need point out there are two problems in deriving a word space from Acquaint and DUC 2005 data. Firstly, DUC 2005 data should not be deemed as an available resource when processing the data. Secondly, a general background corpus may bring bias on words, which will distort term relations in a specific domain built by a document cluster. Therefore, we think it is more reasonable to derive an explicit semantic space from a document cluster to be processed.

Deerwester et al. (1990) provide an excellent introduction of using LSA in information retrieval.

Here we give a brief description for applying LSA in centroid based document summarization.

Let's suppose a term/sentence matrix, *X*, obtained from a document cluster. With singular value decomposition, *X* can be decomposed into the product of three other matrices: $X = T_0 S_0 D_0^{'}$, where $T_0$ is the matrix of left singular vectors, $D_0$ the matrix of right singular vector, and $S_0$ the diagonal matrix of singular values. With a simple strategy of selecting the first *k* largest singular values in $S_0$, we can derive an optimal approximate fit of *X* using smaller matrices. After deleting the zero rows and columns of $S_0$ and the corresponding columns of $T_0$ and $D_0$, we get three new matrices: *S*, *T*, and *D* respectively. Then, $X \approx \hat{X} = TSD^{'}$. The rows of a *DS* matrix could be considered as coordinates for sentences. Thus, the similarity between two sentences will be dot product between their corresponding rows in the matrix *DS*. For calculating the centroid value of a sentence under the reduced semantic space, a conversion from $C_0$ to *C* should be done for the original centroid vector $C_0$ as follows: $C = C_0^{'} T S^{-1}$. The centroid based salience value for a sentence is then derived from the dot product of vector *C* and the sentence's corresponding row in the matrix *DS*.

### 2.2 Query Term Based Salience

To make PoluS system capable of doing query-based summarization, one strategy of ours is to derive a query term based salience value for each sentence. We first construct a query term vector from both the title and narrative parts of a topic. Stopwords and verb words appearing at the begging of a sentence such as "discuss/identify", are discarded. In our opinion, the frequencies of words in topic may not accurately reflect the user's information need. So we assign the term weights of the derived query term vector same as their values in the original centroid vector, because we assume the centroid vector of a cluster is a good indicator of user's requirement. The query term based salience value of each sentence is then calculated as we compute the centroid based value, where an original vector should be converted to a vector under a reduced semantic space.

## 3 Query Focus Guided Sentence Selection Strategy

We observe that the topics in DUC 2006 are much more general or abstract than those in DUC 2005. An example of a topic is given in figure 1. A general sentence can be entailed by many specific sentences. The mismatch between terms in topics and sentences poses a great difficulty for methods that rely on surface match. To alleviate this problem, we devise a query focus based salience value and propose a query focus guided sentence selection strategy for summary generation.

Figure 1. An example topic of DUC 2006 (D0601A).

```
<topic>
<num> D0601A </num>
<title> Native American Reservation System - pros and cons </title>
<narr>Discuss conditions on American Indian reservations or among Native American communities. Include the benefits and drawbacks of the reservation system. Include legal privileges and problems.</narr>
</topic>
```

### 3.1 Query Focus of a Topic

We define it as a profile of a topic, which indicate the factual and non-factual aspects of user's information need. Factual focuses include named entities, specific or non-specific, where non-factual focuses include positive/negative attitude, advantage/disadvantage, and cause/result. It is very difficult to automatically build a complete profile from a topic definition. Here, we have to reduce our expectation and try to derive a rough description of a topic as possible as we can. We use a focus frame to indicate whether a user requires that the generated summary include a special or a non-specific named entity, such as date/time, location, person, organization, quantity, frequency, and country, and whether the selected sentences describe a cause, a result, an advantage of something or simply positive attitude to it, and a disadvantage of something or negative attitude to it.

For example, we obtain the following focus frame for the topic D0601A in figure 1:

```
Focus_D0601A{
        Positive/advantage: yes;
        Negative/disadvantage: yes;
},
```

which indicates that user prefers some sentences that express positive or negative attitudes.

### 3.2 Detecting the User's Query Focus

At present, we take a very simple strategy. We build a focus indicator lexicon, and annotate for each item two focus attributes (query and answer), which indicate what focuses are related to the word when it appears in the topic description and in the documents to be summarized respectively. We first derive a wordlist from "General Inquirer"[2] dictionary for positive/advantage and negative/disadvantage focuses. Then, we manually add and annotate other words for other types of focuses. For example, we add a word "who" and indicate that if this word appears in the topic description, (at least) a non-specific named entity (person name) is required to be included in the final summary. The focus indicator lexicon consists of 3,706 words. For a topic, its query focus will be the union of the query focus attribute of each word in the topic description. A concrete named entity in topic will be a specific named entity focus.

### 3.3 Query Focus Based Salience Value

Following the method of obtaining the query focus of a topic, we can derive the answer focus of a sentence. Thus, we can compute whether a sentence provides some key points according to user's requirement and to what degree. The query focus based salience value is calculated as the overlapping percentage between the query focus of a topic and the answer focus of a sentence. The formal definition of this value is as follows:

$$SQ_{focus} = \frac{2*|M_{focus}|}{|Q_{focus}|+|A_{focus}|}.$$

In the above equation, | $Q_{focus}$ | indicates how many points a user requires, | $A_{focus}$ | how many points a sentence provides, and $M_{focus}$ is the overlapping set of $Q_{focus}$ and $A_{focus}$.

### 3.4 Sentence Selection Guided by Query Focus

We also use the derived query focus to guide the final sentence selection. Our goal is to satisfy user's information need in the generated summary as much as we can. We try to include in the final summary a sentence which could provide more uncovered points in the selected sentences. This strategy is executed at higher level, and whether a sentence should be included in the final summary is also determined by its salience value and the similarities between it and the already selected sentences.

## 4 Evaluations

DUC 2006 provides fifty document clusters for evaluation, each of which includes 25 documents from AQUAINT corpus. All submitted systems are manually or automatically evaluated according to the summary's linguistic quality, its responsiveness, Rouge measures (ROUGE-2 and ROUGE-SU4), BE score, and optionally Pyramid score.

In our official submitted system, we take the following parameters: 0.6 for MMR similarity threshold, 0.8 for parameter $\alpha$ in MMR, and 80% singular values. The centroid based value, query term based value, and query focus based value are combined equally to be the final salience value of a sentence.

Table 1 gives the official results of our system, which clearly show that our system ranks middle in the 34 systems.

Table 1. Official scores of DUC 2006.

| | **Ling. Quan.** | **R2-Score** | **SU4-Score** | **BE-Score** | **Pyd. Score** |
|---|---|---|---|---|---|
| **Our System** | 3.59 | 0.07479 | 0.13161 | 0.03735 | 0.17410 |
| **MAX** | 4.08 | 0.09558 | 0.15529 | 0.05107 | 0.25711 |
| **AVG.** | 3.35 | 0.07463 | 0.13021 | 0.03686 | 0.19185 |
| **MIN** | 2.32 | 0.02834 | 0.06394 | 0.00459 | 0.13042 |
| **Rank** | 15 | 20 | 19 | 19 | 16(21[3]) |

Table 2. Detailed Language Quality scores of DUC 2006.

| | **LQ1** | **LQ2** | **LQ3** | **LQ4** | **LQ5** |
|---|---|---|---|---|---|
| **Our System** | 4.42 | 4.46 | 2.76 | 3.68 | 2.64 |
| **MAX** | 4.62 | 4.66 | 4.00 | 4.28 | 3.28 |
| **AVG.** | 3.57 | 4.22 | 3.07 | 3.57 | 2.34 |
| **MIN** | 1.38 | 3.76 | 1.90 | 2.50 | 1.16 |
| **Rank** | 4 | 8 | 27 | 15 | 7 |

[2] http://www.wjh.harvard.edu/~inquirer/

[3] 21 participants took part in the pyramid evaluation.

Table 2 details the language quality scores of our system. We achieve better scores on questions 1, 2 and 5. However, our system performs poorly on question 3, which results from unclear references in our generated summaries. Exploring how to improve our system in this aspect would be one important research direction for us in the near future.

After received DUC official results and model summaries, we further test how the previously discussed three features contribute to the overall performance in terms of automated evaluation metric ROUGE. The percentages of ROUGE-2 score changes with different weight settings over the submission are illustrated in Table 3. It tells very clearly that the current use of query focus is quite superficial. It is not well explored due to time limitation.

Table 3. ROUGE-2 score changes with different weight settings compared over the submission settings (1 1 1).

| **(F1, F2, F3)**[4] | **Percentage of Change** |
|---|---|
| (1 0 0) | +11.86% |
| (0 1 0) | -17.63% |
| (0 0 1) | -35.55% |
| (1 1 0) | +15.66% |
| (1 0 1) | -7.92% |
| (0 1 1) | -23.73% |
| (1 1 1) | 0 |

## 5 Conclusions

This paper introduces our system for DUC 2006. This system is extended from our generic multi-document summarization system PoluS, which incorporates latent semantic analysis technology. We derive a query focus from a topic description, and use it to guide the sentence selection procedure. Evaluation results show that our system ranks middle in the 34 participants of DUC 2006. It seems that our new multi-document summarization system PoluS places a good foundation for our future research.

## Acknowledgements

This work was supported by the Research Grants Council of Hong Kong (CERG reference number PolyU5181/03E). We thank Yunqing Xia for pointing out some useful material on sentiment analysis.

## References

Jaime Carbonell and Jade Goldstein. 1998. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. *In the proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR'98).

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407.

Gunes Erkan and Dragomir R. Radev. 2004a. The University of Michigan at DUC 2004. *In the proceedings of the Document Understanding Conference 2004* (DUC2004).

Gunes Erkan and Dragomir R. Radev. 2004b. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *Journal of Artificial Intelligence Research*, 22: 457-479.

Ben Hachey, Gabriel Murray, and David Reitter. 2005. The Embra System at DUC 2005: Query-oriented Multi-document Summarization with a Very Large Latent Semantic Space. *In the proceedings of the Document Understanding Conference 2005* (DUC2005).

Wenjie Li, Wei Li, Baoli Li, Qing Chen, and Mingli Wu. 2005. The Hong Kong Polytechnic University at DUC 2005. *In the proceedings of the Document Understanding Conference 2005* (DUC2005).

Yihong Gong and Xin Liu. 2001. Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis. *In the proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.*

Dragomir R. Radev, Hongyan Jing, Malgorzata Stys, and Daniel Tam. 2004. Centroid-based summariza-

[4] F1, F2 and F3 indicate centroid based, query term based and query focus based features, respectively.

tion of multiple documents. *Information Processing and Management*, 40:919-938.

G. Salton, A. Singhal, M. Mitra, and C. Buckley. 1997. Automatic Text Structuring and Summarization. *Information Processing & Management*, 33(2):193-207.

Satoshi Sekine. 2002. Manual of Oak System (version 0.1). Computer Science Department, New York University, http://nlp.cs.nyu.edu/oak.

Dou Shen, Zheng Chen, Qiang Yang, Hua-Jun Zeng, Benyu Zhang, Yuchang Lu, and Wei-Ying Ma. 2004. Web-page Classification through Summarization. *In the proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR'04).

Jen-Yuan Yeh, Hao-Ren Ke, Wei-Pang Yang, and I-Heng Meng. 2005. Text Summarization Using a Trainable Summarizer and Latent Semantic Analysis. *Information Processing and Management*, 41: 75-95.