# Feature Selection for Summarising: The Sunderland DUC 2004 Experience

Shao-Fen Liang Siobhan Devlin John Tait
University of Sunderland School of Computing & Technology
St. Peters Campus Sunderland, Tyne & Wear SR6 0DD
+44(0)191 5153410
{shaofen.liang, siobhan.devlin, john.tait}@sunderland.ac.uk

## Abstract

*In this paper we describe our participation in task 1-very short single-document summaries in DUC 2004. The task chosen is related to our research project, which aims to produce abstracting summaries to improve search engine result summaries. DUC allowed us to produce summaries no longer than 75 characters, therefore we focused on feature selection to produce a set of key words as summaries instead of complete sentences. Three descriptions of our summarisers are given. Each of the summarisers performs very differently in the six ROUGE metrics. One of our summarisers which uses a simple algorithm to produce summaries without any supervised learning or complicated NLP technique performs surprisingly well among different ROUGE evaluations. Finally we give an analysis of ROUGE and participants' results. ROUGE is an automatic evaluation of summaries package, which uses n-gram matching to calculate the overlapping between machine and human summaries, and indeed saves time for human evaluation. However, the different ROUGE metrics give different results and it is hard to judge which is the best for automatic summaries evaluation. Also it does not include complete sentences evaluation. Therefore we suggest some work needs to be done on ROUGE in the future to make it really effective.*

## 1 Introduction

This is the first year we have participated in the Document Understanding Conference (DUC) [1]. We believe that DUC provides a good exercise environment to help us on our own research project. Our project is mainly concerned with improving search engine result summaries. Current search engines use sentence extraction techniques to produce snippet result summaries, which are less coherent and readable than the original documents. We believe users have to spend more time thinking about each summary and finding desired pages because the summary may not express the content of the page well. Our project aims to produce abstracting summaries which are coherent and easy to read thereby lessening users' time in judging the relevance of pages. However, automatic abstracting techniques have domain restrictions. For solving this problem we employ text classification techniques to classify web pages into different categories and produce very short summaries as search engine result summaries [5]. This is the reason that we decided in this competition to focus on only task 1- very short single-document summaries. However, the target length of the summaries in DUC 2004 was even shorter than our project's requirements. Therefore our system is focusing on feature selection [4] to present summaries as a set of the most important key words. The rest of the paper is organised as follows. In

Section 2 we give a description of our three systems, and in Section 3 we give the results of six ROUGE evaluations. A comparison among participants in ROUGE is given in Section 4. Finally we conclude in Section 5 and indicate our plans for future work.

# 2 System Description

As we are first year participants, our system was initially developed using data from DUC 2003, which was slightly different to the real data from DUC 2004. In DUC 2003 the data collection had 60 TDT English clusters and each cluster contained from 8 to 14 documents. The format in each document had a few lines of title tags, and the main body started from <TEXT> tag and ended with </TEXT> tag. Inside the main body, paragraphs were split by <P> and </P> tags. Therefore a normalisation process to remove noise from each document was required. But in DUC 2004 the data had 50 TDT English clusters and each cluster contains 10 documents as the input documents, so 500 summaries in total are required. (The documents come from the AP newswire and New York Times newswire). The format in DUC 2004 did not split documents into several paragraphs but only into one big paragraph, which was the only difference. For adapting to the new format, our system had a little modification of the normalisation process to extract text between <TEXT> and </TEXT> tags. Task 1 gave participants a limitation to produce each summary of no more than 75 characters including punctuation and spaces. The length of each summary is only about 2/3 of one line in a standard American letter size. It is almost impossible to produce a complete sentence to address the concepts of an original news document half a page to two pages in length. Therefore, we focused on feature selection to pick up people's names, groups, events, places and so on to produce headline-like summaries. The following sections describe our three entries.

## 2.1 System one

Sentence selection from a large text is a useful step for document summarisation. Related work includes Teufel [9], Goldstein [3] and McDonald [7]. The approach we used is also based on the sentence level. The input documents therefore need to be segmented into a set of sentences. In addition, we included the consideration of cue words, title words, key words and sentence location from Edmundson [2] and the term weighting from Salton [8].

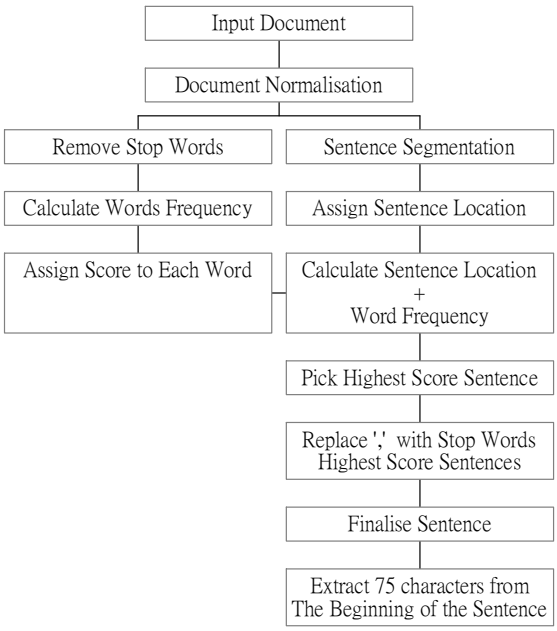


Diagram 1 The Approach of System One

System one (diagram 1) was our priority of the three systems, which received a run number 76 from the ROUGE evaluation. The algorithm first of all normalises each input document into desired text

then the process splits into two parallel ones. The left hand side process splits each document into a set of words and removes stop words. The remaining words then have their frequency calculated. Next they are put into an array in descending order of frequency as [$W_1$, $W_2$, ... $W_n$], where $W_1$ is the most frequent word and $W_n$ the least. These words are then given a score each as $W_1$=n, $W_2$=n-1, $W_3$=n-2, … $W_n$=1. On the right hand side process, firstly the original document (D) is split into a set of sentences which retain their original sequence as D = [$S_1$, $S_2$, $S_3$, …$S_n$]. Each sentence is assigned a location score as $S_{1L}$=n, $S_{2L}$=n-1, … $S_{nL}$=1. Therefore each sentence can be weighted as following:

$$S_W = S_L + Ws \qquad (1)$$

$S_W$ representing Sentence Weighting, $S_L$ representing Sentence Location and Ws representing Word Score. The Ws can be calculated as follows.

$$Ws = \sum_{i=1}^{n} Wi \qquad (2)$$

$\sum_{i=1}^{n} Wi$ representing total score of i words, which appear in the sentence. The highest scoring sentence is chosen as the most important sentence for further summarisation processing. The summarisation process firstly removes stop words from the chosen sentence and replaces them with a ‘,’. The idea for using ‘,’ instead of just removing stop words or replacing them with white spaces was to split the sentence into elements and observe more easily any error that might happen. Secondly, if two or more ‘,’s appear between two words, they would be reduced into one ‘,’. Thirdly, we finalise the sentence by removing redundant spaces appearing in front of or after a ‘,’. Finally, the sentence from the first character to the 75th character is selected as the summary. An example is given as follows.

Government and opposition parties have asked King Norodom Sihanouk to host a summit meeting after a series of post-election negotiations between the two opposition groups and Hun Sen*'s party* to form a new government failed. -----The highest score sentence from APW19981016.0240 in D30001t

Government , opposition parties , , King Norodom Sihanouk , host , summit meeting , , series , post-election negotiations , , two opposition groups , Hun Sen*'s party* , form , new government failed-----Removed stop words

Government,opposition parties,King Norodom Sihanouk,host,summit meeting,series,post-election negotiations,two opposition groups,Hun Sen*'s party*, form,new government failed ----- Finished sentence

Government,opposition parties,asked King Norodom Sihanouk,host,summit meeti ----- Final summary

## 2.2 System 2

System two was constructed and implemented in the very last stages of the competition as we didn’t wish to lose any opportunity in the three runs. Yet it proved the best of the three. It received a run number 77 from the ROUGE evaluation. We used a simple algorithm to pick up the first 300 characters from each document. The reason for picking up 300 characters was that we did not know how many

characters would remain after removing words in the sentence. Four things are removed and do not appear in our summaries, which are: 1. A list of stop words. 2. Words appearing between a pair of brackets like (…). 3. Reported speech such as ``No-one should internationalize Cambodian affairs. It is detrimental to the sovereignty of Cambodia,'' he said. 4. Date and Time such as January … December, Monday … Sunday, morning, afternoon, evening, night and so on. We decided to cut the document at 4 times the allotted character length of 75 i.e. 300 characters, to avoid the finished sentence being shorter than 75 after the removal. The cut sentence was put in the summarisation process as described in section 2.1 to produce summaries. The following is an example.

Cambodian leader Hun Sen on Friday rejected opposition parties' demands for talks outside the country, accusing them of trying to ``internationalize'' the political crisis. Government and opposition parties have asked King Norodom Sihanouk to host a summit meeting after a series of post-election neg ----- 300 characters were cut from the beginning of the document APW19981016.0240 in D30001t.

Cambodian leader Hun Sen on Friday rejected opposition parties' demands for talks outside the country, accusing them of trying to the political crisis. Government and opposition parties have asked King Norodom Sihanouk to host a summit meeting after a series of post-election neg ----- Words appearing in quotation marks were removed.

Cambodian leader Hun Sen,rejected opposition parties,demands,talks,country,accusing, trying,political crisis.Government,opposition parties, asked King Norodom Sihanouk,host,summit meeting, series,post-election neg ----- Finished sentence

Cambodian leader Hun Sen,rejected opposition parties,demands,talks,country, ----- Final summary

System 2 was modified from system one and the process only took a few minutes to make sure it ran successfully.

## 2.3 System 3

We started to design these three systems from late January 2004 so we didn't have sufficient time to implement a better summariser. Therefore we used one component of system 1 to be our system 3 and it received a run number 78 from the ROUGE evaluation. This system simply presented the most frequent words as the final summaries, extracting the first 75 characters from [W1, W2, ... Wn] list. An example result is shown below.

opposition Rainsy Sam two election country government Ranariddh form talks----- Most frequent words in the document APW19981016.0240 in D30001t.

# 3 The results from ROUGE evaluations

All participant results (Figure 4) were evaluated solely by ROUGE's n-gram matching [6]. ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation, which is an automatic evaluation

package. Our three runs performed very differently in the ROUGE evaluations. They are shown as numbers 76, 77, 78 in the figure below.

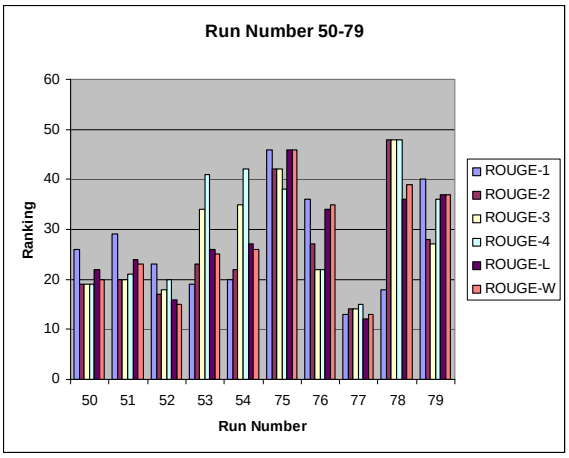


Figure 1 The ROUGE result of run number 50 to 79

The priority one run, number 76, didn't perform as well as we hoped: it was worse than 77 but better than 78 on average. It was ranked 36 in ROUGE-1, 27 in ROUGE-2, 22 in ROUGE-3, 22 in ROUGE-4, 34 in ROUGE-L and 35 in ROUGE-W. Although the algorithm combines weights for sentence location and word frequency scores in each sentence, the selected sentence didn't present a better feature selection approach if compared with human summaries (Figure 2). In addition, the sentence obviously did not include enough important key words, which did appear in the human summaries. The run number 77 was ranked 13 in ROUGE-1, 14 in ROUGE-2, 14 in ROUGE-3, 15 in ROUGE-4, 12 in ROUGE-L and 13 in ROUGE-W (Figure 3). The simple algorithm performed surprisingly well among different ROUGE evaluations. Excluding runs number A-H (human summaries), the run number 77 was ranked between 4$^{th}$ and 7$^{th}$.

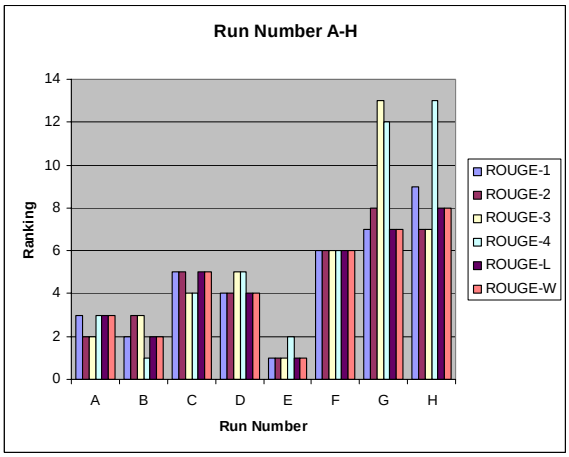


Figure 2 The ROUGE result of human summaries run number A to H

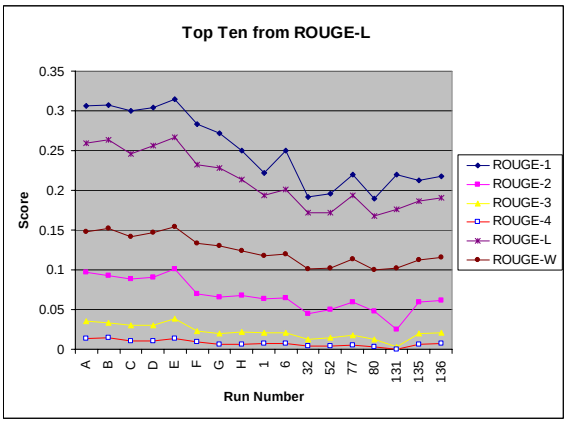


Figure 3 Top 10 run number based on ROUGE-L

The result indicates two points. Firstly, in the data collection in DUC 2004, the earlier the words appear in the original document, the more important they are. Secondly, human summaries tend to have similar word order to the original document. The result shows selecting features from the beginning of the input documents, to be a good algorithm for DUC 2004. The last run, number 78, performed slightly worse than run number 77 and ranked 18 in ROUGE-1 but it dropped to the last place in ROUGE-2, ROUGE-3 and ROUGE-4. In ROUGE-L

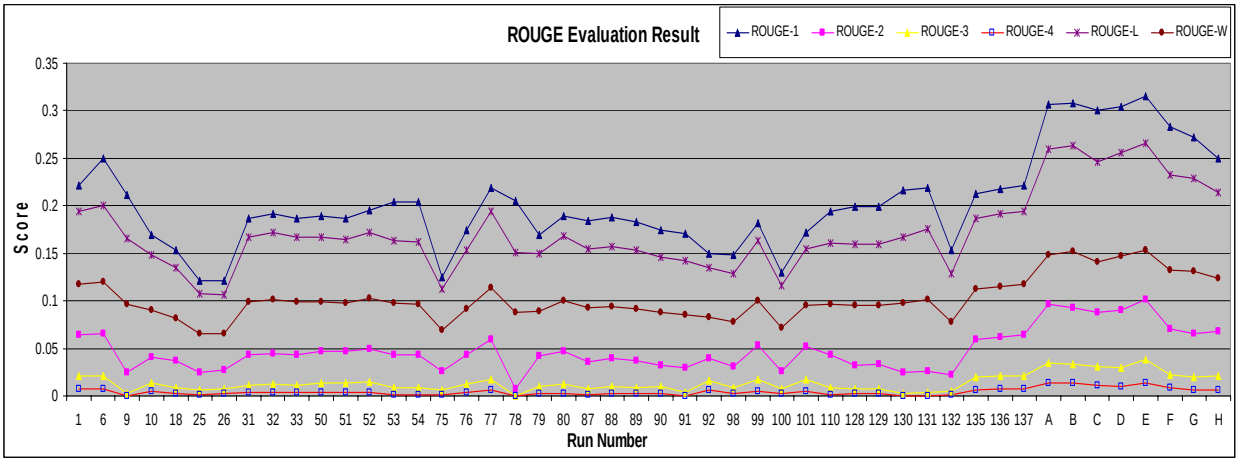


Figure 4 Results of all participants from six ROUGE evaluations

it improved its performance, to reach 36 and reached 39 in ROUGE-W. The result indicated that the most frequent words are likely to appear in human summaries. Therefore term frequency is still a good way to pick out key words from unknown documents. However number 78 dropped into the bottom in ROUGE-2, ROUGE-3 and ROUGE-4 evaluations because it is a unigram summariser, and these are n-gram metrics.

## 4 Comparison among all participants relating to ROUGE

Automatic evaluation is essential to conquer the time consuming task of human evaluation, especially for evaluating large numbers of results like the competition in DUC. ROUGE indeed saves time for humans and performs a good n-gram overlapping calculation algorithm but as can be seen in Figure 4 very few automatic systems perform well enough on ROUGE-3 (let alone ROUGE-4) for the results against these metrics to be meaningful. These two lines point out that although n-grams can be applied to any number of n, calculating the overlap between machine summaries and human summaries for n>2 cannot distinguish well between each machine system even between machine and human summaries, thus the need to use n-grams greater than 2 to calculate the overlapping between human summaries and machine summaries, which needs to be reconsidered in ROUGE's metric. Another phenomenon can be observed from Figure 5, which is that the selected run numbers show very different performances over the six ROUGEs. Each of them has a gap of over 20 between the lowest ranking and the highest ranking. They either score better in ROUGE-1 but worse in ROUGE-2, 3, 4, L and W (such as number 9, 53, 54, 78, 130, 131) or better in ROUGE-4 but worse in ROUGE-1 (like numbers 10 and 92). Especially in the case of run numbers 130 and 131, the summaries rank highly from ROUGE-1, ROUGE-L and ROUGE-W but poorly in ROUGE-2, 3 and 4. The ROUGE-L and ROUGE-W are designed for weighting and calculating the Longest Common Subsequence, when ROUGE-2, 3, 4 are worse ROUGE-L and W should also be relatively worse. However the results from 130 and 131 do

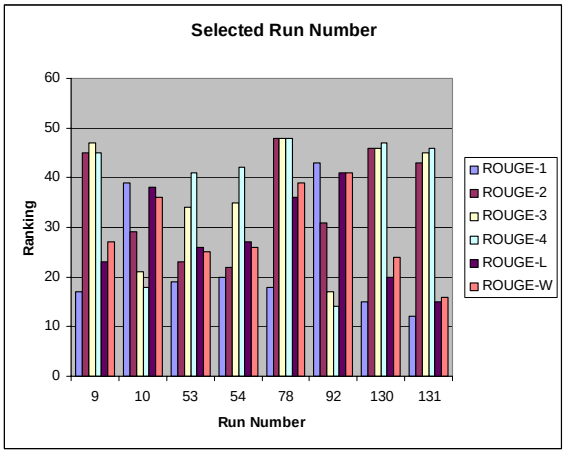


Figure 5 Gaps of over 20 between lowest and highest ranking

not follow this relation. On the other hand our 78 is reasonable following the ROUGE rule. Although ROUGE-1 is better, ROUGE-2, 3, 4 are worse and therefore ROUGE-L and W are also relatively worse. Another unusual situation happens in numbers 10 and 92. They both perform badly on ROUGE-1, ROUGE-L and W, which indicate their summaries do not contain enough overlapping with human summaries and also the Longest Common Subsequence mapping yet their results improve if the n-gram is bigger. This phenomenon is hard to explain. Theoretically, ROUGE-W should be the best evaluation model among ROUGEs but empirically from the task 1 result, we can see that there is conflict between different ROUGEs, which gives ROGUE more space to improve its evaluation algorithm. So for evaluating a complete sentence as a summary, we would suggest that current ROUGE needs further improvement.

## 5 Conclusion and future work

The primary conclusion from our work so far is that simple extraction of the first three hundred words works very well in DUC 2004 task 1: it performs better than combining weighting words and sentence location or unigram extraction. Similar conclusions have been drawn on previous work in the news domain. We have presented an analysis of participant results and ROUGE evaluations and find that although ROUGE has expanded from single word mapping to n-gram mapping and also to longest common sub-string mapping, we are still unsure which implementation of ROUGE is the best for evaluating summaries.

In our own work, in the future we will continue our aim to improve search engine result summaries by investigating other methods to improve our feature selection and also expand our summariser to produce complete sentence like summaries.

## References

[1] DUC. http://tides.nist.gov/

[2] Edmundson, H.P. 1969. New methods in automatic abstracting. Journal of the Association for Computing Machinery 16 (2): 264-285.

[3] Goldstein, J., Kantrowitz, M., Mittal, V. and Carbonell J. 1999 Summarizing text documents: sentence selection and evaluation metrics. In proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pages 121-128, August.

[4] Guyon, I. and Elisseeff, A. 2003. Special issue on special feature: An introduction to variable and feature selection. The Journal of

Machine Learning Research, Volume 3 pages 1157-1182.

[5] Liang, S.F. & Devlin, S. & Tait, J. 2004. Can automatic abstracting improve on current extracting techniques in aiding users to judge the relevance of pages in search engine results. In Proceeding of th 7th Computational Linguistics UK, pp. 154-159.

[6] Lin, C.Y. and Hovy, E. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In Proceedings of HLT-NAACL, Edmonton, Canada, May.

[7] McDonald, D., Chen, H. 2002. Summarization and question answering: Using sentence-selection heuristics to rank text segments in TXTRACTOR. In proceedings of the second ACM/IEEE-CS joint conference on Digital libraries, pages 28-35, July.

[8] Salton, G. and Buckley, C. 1988. *Term-weighting approaches in automatic text retrieval.* Information Processing and Management, Vol 24, 513-523.

[9] Teufel, S. and Moens, M. 1997. Sentence extraction as a classification task. In Proceedings of the ACL/EACL 1997 Workshop on Intelligent Scalable Text Summarisation, pages 58-65, Madrid, Spain, July.