Using Query Term Order for Result Summarisation

Shao Fen Liang, Siobhan Devlin and John Tait
The University of Sunderland
School of Computing and Technology
Sunderland SR6 0DD, UK
+44(0)191 515 3410
{ShaoFen.Liang, Siobhan.Devlin, John.Tait }@sunderland.ac.uk

ABSTRACT
We report on two experiments performed to test the importance of Term Order in automatic summarisation. Experiment one was undertaken as part of DUC 2004 to which three systems were submitted, each with a different summarisation approach. The system that used document Term Order outperformed those that did not use Term Order in the ROUGE evaluation. Experiment two made use of human evaluations of search engine results, comparing our Query Term Order summaries with a simulation of current Google search engine result summaries in terms of summary quality. Our QTO system's summaries aided users' relevance judgements to a significantly greater extent than Google's.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval - *Query formulation*.

General Terms
Algorithms, Performance, Experimentation.

Keywords
Query Term Order, Term Order, Summary Quality.

1. INTRODUCTION
Automatic summarisation approaches commonly use a bag of words model, title terms, term frequency and sentence order [1]. These approaches more or less ignore term order. We believe that term order in a document or query is an implicit indicator of a term's importance. Therefore, our hypothesis is that including term order in automatic summarisation produces better summaries than not including it. To prove our hypothesis we have done two experiments testing the importance of Term Order in documents summarisation and Query Term Order in search result summarisation.

The rest of the poster describes our weighting scheme, which takes term order into account. The first experiment was implemented while participating in task 1 of DUC 2004 and the second experiment used human evaluation to determine summaries' quality by comparing our QTO and a simulation of the then current Google system.

SIGIR'05, August 15–19, 2005, Salvador, Brazil.
ACM 1-59593-034-5/05/0008.

2. EXPERIMENT 1: TERM ORDER

2.1 System set up
The first experiment aimed to test the importance of term order in a document. The input data was 50 TDT English newswire clusters and each cluster contained 10 documents. Each of the 500 output summaries was required to be no more than 75 characters (a set of key words or phrases) including punctuation and spaces. We submitted three systems for the competition which were named *76*, *77* and *78* [2].

Run *76* extracted relevant sentences by adding Term Frequency (TF) and Sentence Order (SO) to produce a sentence weighting.

- TF: The frequency of each term in the document except stop words was calculated as a percentage of terms in the document (not collection).
- SO: Title tags and HTML tags were removed from the document. The text was broken into a set of sentences. The original order of these sentences was preserved.

Run *77* added Term Order (TO) to expand the weighting scheme of system 76, i.e. it used TF, SO and TO.

- TO: The first 50 words were extracted from the document. Stop words were removed if they appeared among the 50 words. This broke the sentence into a set of segments which were assigned scores in descending order of sentence weighting.

Run *78* adopted TF only to weight sentences.

2.2 Result

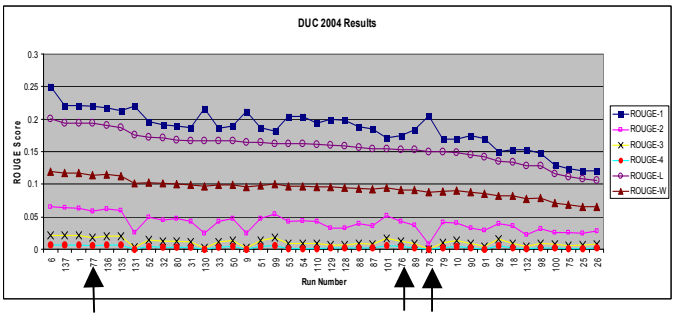

Figure 1. The DUC 2004 result

Figure 1 shows the DUC 2004 result. The result was evaluated by ROUGE and the figure is sorted by ROUGE-L scores. ROUGE-L scored each system with longest common subsequence mapping between human and system summaries. Our system *77* ranked 4th of the 40 participating automatic systems. In addition, *77* was no worse than 7th place using any ROUGE metrics. However, *76* and

78 scored far lower than *77*, ranking 26th and 28th. The result shows that combining TO, TF and SO can produce better summaries than TF alone or TF and SO.

3. EXPERIMENT 2:QUERY TERM ORDER

3.1 System set up

The second experiment aimed to test whether counting Query Term Order could produce better search result summaries to help search engine users in making relevance judgements. In this experiment, our system was designed slightly differently to the first experiment because we changed the sentence weighting scheme by combining Query Term Order (QTO), TF and SO to extract sentences as summaries so that we could use Google's summaries as the baseline comparison with our QTO system summaries. We selected 6 TREC9 queries and used each to retrieve 10 web pages (in English) from Google. The 60 summaries were output with exactly the same format and font in order not to be visibly distinct for the selected subjects during the evaluation process. Ten subjects were selected and split into two groups to evaluate summary quality. Each summary's quality was scored for the extent to which it accurately represented the original page's content (*representativeness*) and the extent to which it allowed the original page's relevance to the particular query to be judged (*meaningfulness*).

The QTO system's sentences weighting scheme was as follows:

- QTO: The first step was to use stop words to break the query into a set of weighted segments. These segments were stored in their original order. In the second step each segment was checked in order to break the segment into a set of single terms if it contained more than one term. Each single term generated from the second step was stored after those from the first stage, and their original order was retained. For example: the input query of TREC9 No. 522 "*how is water supplied to mojave desert region*" generate to a set of terms as "*water supplied*", "*mojave desert region*", "*water*", "*supplied*", "*mojave*", "*desert*", "*region*".

- TF: Only the top ten percent of frequent words in the page were selected because web pages often contain more terms than DUC's data.

- SO: The approach was the same as in experiment one but script languages and style sheets appearing in the page were also removed.

We used the following equations to determine each term's weighting.

$$QTO(t) = \frac{N - t + 1}{N} \; ; \; TF(t) = \frac{M - t + 1}{M} \; ; \; SO(t) = \frac{K - t + 1}{K}$$

Where *N, M* and *K* are the total number of terms in each category of QTO, TF and SO respectively. Each score in these three categories was normalised to between 0 and 1 and also occupied an equal ratio of 33% in the total score.

3.2 Result

The summary's quality is calculated according to A,B and C formulas. In formula A, $\overline{S}$ represents the mean value of summaries' representativeness score and is normalised to between 0 and 1, q represents the number of subjects from 1 to m, l represents the number of summaries from 1 to n, S_{ql} represents each summary's representativeness score. In formula B, M_{score} represents the mean value of summary's meaningfulness score of each query, T represents the total number of judgements and U represents the number of *unknown* judgements. In formula C, S_{score} represents the summary's quality.

$$\overline{S} = \frac{\sum_{q=1}^{m} \frac{\sum_{l=1}^{n} S_{ql}}{5n}}{m} : 0 \le \overline{S} \le 1 : S_{ql} \in \{1,2,3,4,5\} \quad \text{.....(A)}$$

$$M_{score} = \overline{\sum_{q=1}^{m} \frac{T_q - U_q}{T_q}} \quad \text{.....(B)} \; ; \; S_{score} = \overline{S} \times M_{score} \quad \text{.....(C)}$$

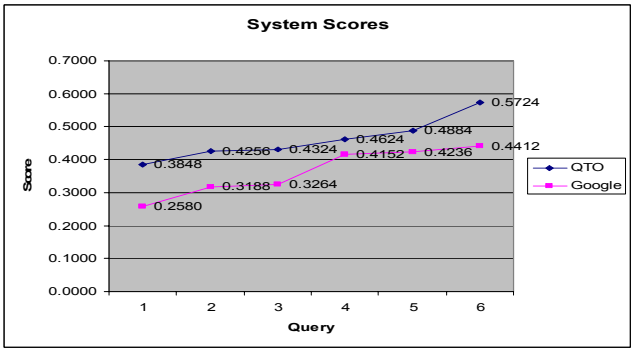

Figure 2. The human evaluation result

Figure 2 shows that QTO's summary quality scored higher than Google's among the six queries. The mean score of QTO's quality is 0.4610 and Google is 0.3639. A t-test indicates $r = 0.887$, $df = 5$, $t = 7.030$ and $P = 0.001$, which is significant.

4. CONCLUSION

We have reported two experiments that have shown that the use of Term Order in both documents and queries improves automatic summarisation. The first experiment formed our entry into DUC 2004. Three systems were submitted each with a different sentence weighting scheme. The system that combined TO, TF and SO performed much better than those without Term Order.

The second experiment used human judgement to evaluate QTO and Google's summary quality according to a summary's representativeness to its original page and it's meaningfulness in responding to a user's query. The result proves that Query Term Order is an important factor in producing better summary quality to help users' relevance judgements.

In the future we would like to expand the QTO algorithm into many steps instead of the current two, in order to achieve more detailed term weightings. Secondly, we will expand the second experiment with more users and more queries to test if the QTO algorithm also improves the speed of users' judgements.

5. REFERENCES

[1] Mani, I. *Automatic Summarisation*. John Benjamins, Amsterdam, 2001.

[2] Liang, S.F., Devlin, S. and Tait, J. (2004) Feature Selection for Summarising: The Sunderland DUC 2004 Experience. *In the proceedings of DUC 2004*, Boston USA.