NICTA's Update and Question-based Summarisation Systems at DUC 2007

Nicola Stokes, Jiawen Rong, Brianna Laugher, Yi Li and Lawrence Cavedon.
NICTA Victoria Lab, Department of Computer Science and Software Engineering,
The University of Melbourne,
Australia.
{nstokes, rongj, blaugher, yli8, lcavedon}@csse.unimelb.edu.au

Abstract

In this paper we describe an extractive summarisation approach that combines a variety of feature-based relevance measures to generate question-focussed and update-style summaries. Our results show that cosine, centroid and query expansion-based measures are the most effective similarity metrics for choosing appropriate summary sentences given a complex information need. Overall we achieved high to mid-table performance results for both of these tasks.

1 Introduction

In this paper, we describe NICTA's automatic summarisation performance at the DUC 2007 workshop. Two tasks were defined for the DUC challenge this year. The main task was Question-based Summarisation (QBS) - a complex question answering-type task that requires information synthesis from various text sources. This is the second year that this task has been running. A pilot study task referred to as Update Summarisation (UpS) was also investigated. UpS is similar to QBS in that the system is presented with a topic statement (consisting of one or more questions) and a cluster of on-topic documents; however, in this information searching scenario it is assumed that the user is already familiar with some aspects of the topic (represented by a set of earlier documents). Hence, the system is required to present the user with the information from a subsequent set of news stories that is both *novel* and *relevant* to their query.

This was our first year participating in the Question-based Summarisation (QBS) task. We modified our system so that we could also participate in the pilot study. Overall we achieved high to mid-table results for both of these tasks. However, a lack of training data on the part of the UpS task meant that we were unable to empirically determine similarity thresholds before submitting our results. Our post-submission runs show that with careful selection of our similarity-based relevance features and optimisation of our thresholds, we can achieve a 7.12% and 26.57% gain in our system performance for the QBS and UpS tasks respectively.

The rest of this paper is organised as follows. Our QBS and UpS systems are described in Section 2, which is followed by a detailed analysis of the performance of these systems on the DUC 2007 topics and document collection.

2 System Descriptions

Figure 1 shows the three processing steps of the NICTA Question-based Summarisation System: preprocessing, sentence relevance ranking and summary generation.

For the DUC QBS task the system is required to generate a 250-word summary given a user information need and a cluster of 25 relevant documents. We view this problem as an Information Re-

trieval/Synthesis task that first requires that information units (in this case sentences) are ranked with respect to their relevance to the query; and second, selected for inclusion in the summary because they do not discuss any information that is already covered in the summary, i.e. they are non-redundant. Sentences are added to the summary until the 250-word limit is *just* exceeded. Due to time constraints we were not able to implement the *summary editing* process of the generation step, which would have enforced the 250-word limit by removing dispensable information such as that found in relative clauses. Each of these summarisation steps will now be explained in more detail.

2.1 Preprocessing

The preprocessing step performs document cleaning, meta-information extraction, sentence detection, part-of-speech tagging, chunking, name entity recognition and detection, question classification and co-reference resolution.

First, HTML tags and tables (e.g. containing financial information) are removed from the original documents. Then some meta-information from the document, such as the document publication date, document identification number, the headline and the category are extracted. The publication date and document number are used for sentence indexing and sentence ordering. Sentence boundaries are then detected. The "cleaned" documents are fed into various NLP tools (part-of-speech tagging, chunking, name entity recognition and classification, and coreference resolution) which annotate the text with particular linguistic information (OpenNLP, 2006; LingPipe, 2006).

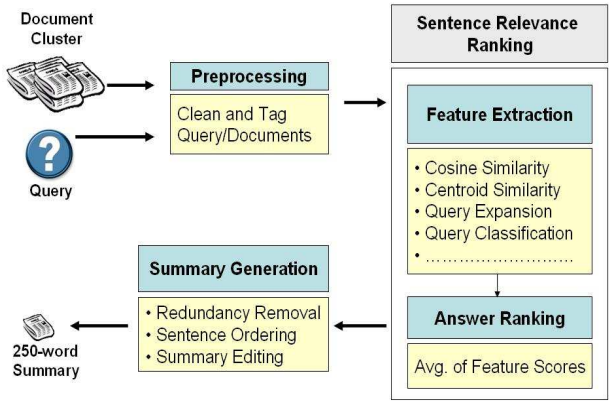

Figure 1: Question-based Summarisation System Architecture

At this stage the topic statement is assigned a question-classification type, and each sentence is assigned an answer type. We based our question-classification (QC) technique on (Li and Roth, 2002). QC can be described as a fine-grained version of named entity classification, where questions are assigned more specific labels such as *country* in the case of "What countries are having chronic potable water shortages and why?" rather than a high-level label such as "location". Although the majority of DUC topic statements are complex questions, QC can help in situation where factoid-type answers are required by increasing the relevance of sentences that contain the appropriate answer type. Hence, a sentence that lists countries as well as the topic terms "chronic potable water shortages" will be considered more relevant than a sentence that mentions topic terms only. Unfortunately, very few of the DUC 2007 topics required factoid-type answers, and there is little observed benefit from this analysis in our DUC performance scores.

2.2 Sentence Relevance Ranking

Once the DUC document collection and queries have been preprocessed, each sentence and its corresponding linguistic representations (e.g. part-of-speech tagged, chunk tags) are indexed using the Lucene IR engine (Lucene, 2006). Lucene defines a basic cosine term overlap metric for ranking documents, in our case sentences, where terms are weighted using a *tf.idf* weighting scheme. We extended Lucene by adding additional similarity functions to its API that take advantage of the different linguistic sentence representations provided by the preprocessing step. In this paper we refer to these similarity functions as *relevance features*. Each of these features are now explained in more detail:

1. Term-based Similarity (**Cos**): calculates the cosine similarity between content bearing terms (minus stopwords) in the query and each sentence in the topic cluster.

2. Named Entity-based Similarity (**NERC**): calculates the cosine similarity between named

entities (LOCATION, PERSON, ORGANISATION, DATE, MONEY, PERCENTAGE) in the sentence and the given query.

3. Centroid Similarity (**Cen**): calculates the cosine similarity between each sentence and the cluster centroid. This relevance feature is motivated by the fact that sentences that are both relevant to the query and central to the topic are considered stronger summary sentences.

4. Query Expansion (**QExp**): calculates the cosine similarity between each sentence and an expanded form of the query. This feature uses the *relevance feedback* technique commonly used in IR to expand the original query with additional relevant terms, with the hope that this will improve the recall of the retrieval system. First the original query is used to retrieve an initial set of relevant documents, and then the top 20 most frequent content words in this ranked list are extracted and used to expand the original query.

5. Density of Numeric References (**Num**): calculates the number of numeric references in the sentence divided by the number of terms in the sentence. This feature is based on the assumption that sentences that contain statistical information are more useful summary sentences.

6. Chunk-based Similarity (**Chunk**): calculates the cosine similarity between noun phrase chunks in the query and each sentence. This feature is like Named Entity-based cosine similarity except the system limits the overlap type to noun phrases.

7. Position of the Sentence (**Pos**): calculates sentence importance based on the inverse of its position in the document. For example, if the position of the sentence is 10 then its score is 0.1. Usually, important information is situated at the beginning of a new story.

8. Synonym Expansion (**Syn**): calculates the cosine similarity between the synonym expanded query and each sentence. Like the QExp feature, the original topic statement is expanded, but this time with extracted synonyms from WordNet. No disambiguation is performed prior to expansion, which implies that in many instances erroneous terms will be added. This may explain the subsequent poor performance of this feature.

9. Question Classification (**QC**): calculates the number of matched question and answer types between the query and each sentence divided by the number of identified question types in the query.

Once all of these relevance features have been calculated, the Lucene engine returns the final ranked list of relevant sentences by averaging these feature scores.

2.3 Summary Generation

The summary generation step performs the redundancy removal and sentence ordering on the ranked list of sentences. The redundancy removal process iteratively adds top ranking sentences to the final summary if the sentence meets the following criteria: the term-based cosine similarity of the current sentence with each of the existing summary sentences must be less than or equal to the *redundancy removal* threshold. Short sentences (less than 7 terms) and some sentences with quotations are also filtered out at this stage. Once the maximal word-limit has been met or *just* exceeded, sentences are then re-ordered so that older sentences (with respect to publication date) are pushed to the front of the summary. If two sentences are published on the same date (that is, are from the same document), then we also take into consideration their sentence positions. Some simple editing operations are then performed - leading conjunctions such as "and" and "but" are removed, as well as leading sentence phrases that mention the author's name or the publishing organisation. As already stated our future intention is to extend this sentence editing phase to include more linguistically-motivated pruning operations such as those discussed in (Zajic et al., 2004).

2.4 Update Summarisation System Modifications

The NICTA Update Summarisation System is a modified version of our Question-based Summarisation System. Before this modification is explained

it is helpful to first define the update task in more detail.

Given a DUC topic statement and three on-topic document clusters (A, B and C), generate three distinct summaries that answer the user's information need as follows: first summarise the documents in cluster A; then assuming that the user has read all the documents in cluster A, produce an *update summary* of the documents in B; then make a similar assumption for the third and final summary about the user's knowledge of clusters A and B, and generate an update summary for cluster C. Each summary must be no longer than 100 words.

To address the assumption that the user has prior knowledge of the topic from the preceding cluster, we introduce an additional threshold in the summary generation step explained in the previous subsection. This threshold is referred to as a *novelty* threshold, and when combined with the redundancy removal constraint ensures that highly relevant sentences will only be added to the final summary if their contribution to the topic has (a) not been covered by an existing summary sentence (redundancy), and (b) has not been reported in any of the sentences in the previous cluster (novelty).

3 DUC 2007 Results

The DUC evaluation methodology uses human judges to evaluate peer summaries based on the following criteria: the *responsiveness* of the summary to the topic, that is how well the summary addresses the user's information need; and the *linguistic quality* of the summary, that is how well-written the summary is perceived to be. Linguistic quality is defined with respect to five distinct summary quality attributes: Grammaticality, Non-redundancy, Referential clarity, Focus, and Structure and Coherence [1]. An average of the scores for each of these attributes results in a single linguistic quality score.

In recent years, the summarisation community has also become interested in automatic metrics for determining summary *responsiveness* or quality of content. Three metrics are focussed on in the DUC 2007 evaluation: two based on the ROUGE evaluation package (ROUGE-2 and ROUGE-SU4) and one defined by the Basic Elements evaluation package (Lin, 2004; Hovy et al., 2005). All summaries are truncated to 250 words before being evaluated by either the automatic or manual evaluation procedures.

As already explained, this year's DUC workshop focussed on two tasks: Query-based Summarisation (QBS) and Update Summarisation (UpS). The latter was a pilot task initiated by Microsoft researchers. 45 topics were defined for the QBS task and 10 for the UpS task. The DUC collection consisted of 1125 news documents, where a subset of 250 documents was used for the UpS task.

3.1 Question-based Summarisation Results

Table 1 shows the metric scores and corresponding rank of our official submitted system (system id 8) with respect to: the highest score achieved for each metric (Best), and the baseline system scores for runs BL1 and BL2. BL1 returns all the leading sentences (up to 250 words) in the TEXT field of the most recent document in the topic cluster, while BL2 is the automatic multi-document summarisation system with the highest mean SEE coverage score on task 2 at DUC 2004 (Conroy et al., 2004). In the QBS task, BL2 ignores the topic narrative and retrieves relevant information based only on the TITLE field.

These results show that the NICTA system outperforms both baseline systems on all evaluation metrics with the exception of BL1's Linguistic Quality score, which is exceptionally high since it only ever returns consecutive sentences from a single news document. The automatic metrics ROUGE-2, ROUGE-SU4 and BE place the NICTA system within the top 12 performing systems at this year's DUC, where 30 systems competed in total. However, the manual evaluation metrics tell a slightly different story. Average Content or Responsiveness puts us mid-table, while Linguistic Quality is relatively low. Looking at the breakdown of Linguistic Quality scores, we performed relatively well on Grammaticality (3.44), Non-redundancy (3.51), Referential clarity (3.16) and Focus (3.42); however, our Structure and Coherence score (2.40) was below either baseline. This can be explained by the fact that our peer summaries will have been automatically truncated in many cases since they exceeded the 250-word limit. We choose not to cut sentences

[1] See `http://www-nlpir.nist.gov/projects/duc/duc2007/quality-questions.txt` for definitions of the linguistic quality questions.

since ROUGE scores can be significantly reduced if summaries are less than 250 words, which can occur when the addition of a sentence is stopped because it puts the summary length in the *red-zone*. Of course, Structure and Coherence can also be hampered by an inadequate sentence ordering strategy. We will investigate this in due course to see if it was a contributing factor.

As already mentioned in Section 2, our system uses a selection of features to weight the relevance of the topic statement to each sentence in each of documents in the topic cluster. Post-submission, we ran a number of experiments to determine the "ranking value" of each of these features. A detailed explanation of each feature is provided in Section 2. Since the **Cos** feature was our strongest performer, we evaluated the performance of each of the other features by averaging their relevance scores with the **Cos** score. All features that produced a ROUGE-2 score gain over the Cos ROUGE-2 score are highlighted. All other features either had no effect or reduced the ROUGE score slightly. In our official NICTA submission, we combined the **Cos**, **Cen**, **Pos**, **Num**, **QC** and **QExp** features, which we empirically chose based on DUC 2006 ROUGE scores. However, our optimal list of features for the 2007 experiments is **Cos**, **Cen** and **QExp**, i.e. cosine similarity, centroid-based similarity, and query expansion respectively. This experimental run achieves the highest ROUGE-2 score (0.1115). Although this score shows a 7.12% gain over the submitted NICTA run (0.1041), it still falls short of the best performing system ROUGE-2 score (0.1245) for DUC 2007.

For our QBS system, we empirically determined the optimal *redundancy removal* threshold to be 0.3 from our experiments on the DUC 2006 data. To confirm the validity of this threshold on the 2007 data with our best performing features, we varied the threshold between 0.1 and 0.9. The results in Table 3 confirms the effectiveness of this threshold for the QBS task.

3.2 Update Summarisation Results

24 systems participated in the pilot UpS task. Our system results for this task are presented in Table 4. Overall our system (id 43) underperformed for this task: the BL2 baseline has a higher score across all metrics. Our official submission used the **Cos**, **Cen**, **QExp** and **Num** features; however, improved performance was achieved by dropping the **Num** feature. A thresholding variation experiment showed that we could have better optimised the *novelty* threshold if we had been provided with some additional training data. Our original novelty threshold

Metric	Rank	NICTA	Best	BL1	BL2
ROUGE-2	11	0.1041	0.1245	0.0604	0.0938
ROUGE-SU4	6	0.1640	0.1771	0.1051	0.1464
BE	12	0.0562	0.0663	0.0263	0.0462
Ling. Quality	18	3.19	3.40	4.30	3.56
Avg. Content	15	2.76	4.11	1.87	2.71

Table 1: Comparison of NICTA official Question-based Summarisation performance across metrics with respect to other automatic systems.

System	ROUGE-2
Cos	0.1074
Cos+Cen	0.1107
Cos+Chunk	0.1074
Cos+NERC	0.1074
Cos+Num	0.1071
Cos+Pos	0.1064
Cos+QC	0.1071
Cos+QExp	0.1099
Cos+Syn	0.1072
Cos+Cen+QExp	0.1115

Table 2: Comparision of feature performance with respect to ROUGE-2 metric on the Question-based Summarisation task.

Threshold	ROUGE-2
0.1	0.09416
0.2	0.10973
0.3	**0.11149**
0.4	0.10944
0.5	0.10852
0.6	0.10856

Table 3: The effect of varying the redundancy removal threshold on Question-based Summarisation ROUGE-2 scores for the Cos+Cen+QExp run.

Metric	Rank	NICTA	Best	BL1	BL2
ROUGE-2	16	0.0680	0.1119	0.0454	0.0850
ROUGE-SU4	15	0.1114	0.1431	0.0825	0.0122
BE	15	0.0348	0.0722	0.0178	0.0427
Avg. Content	17	2.1330	2.9670	1.6670	2.7000

Table 4: Comparison of NICTA official Update Summarisation performance across metrics with respect to other automatic systems.

was 0.2. Table 5, shows that low threshold values have a negative effect on ROUGE performance, and that a threshold of 0.9 achieves the highest ROUGE value of 0.08607 - a 26.57% gain over our submitted system result. This indicates that most information in the subsequent cluster (B or C) is novel. This run just outperforms the BL2 baseline but is still lagging behind the best system. The optimal *redundancy removal* threshold was keep static at 0.3 throughout this experiment.

Threshold	ROUGE-2
0.1	0.04363
0.2	0.06859
0.3	0.08266
0.4	0.08325
0.5	0.08560
0.6	0.08599
0.7	0.08599
0.8	0.08599
0.9	**0.08607**
1.0	0.08531

Table 5: The effect of varying the novelty threshold on Update Summarisation ROUGE-2 scores for the Cos+Cen+QExp run.

4 Conclusions

This was our first time participating in the Question-based Summarisation and Update Summarisation tasks at DUC. We built a system around the Lucene retrieval engine, by extending it with linguistically-motivated relevance features and a summary generation step. We modified this architecture slightly for the Update Summarisation task.

We performed best on the main question-focussed task, where we achieved high-table scores across automatic and manual metrics. Reducing our relevance feature set to the cosine, centroid-based, and query expansion similarity measures improved our performance with respect to the ROUGE-2 metric. However, on the Update task there is significant room for improvement, despite finding an optimal novelty threshold in our post-submission experiments. In future work we plan to conduct a detailed topic-level analysis of our results, with the aim of understanding why we underperformed on certain topics in both tasks.

Acknowledgments: National ICT Australia (NICTA) is funded by the Australian Government's Department of Communications, Information Technology, and the Arts, and the Australian Research Council through Backing Australia's Ability and the ICT Research Center of Excellence programs.

References

J. M. Conroy, J. D. Schlesinger, J.Goldstein, and D. P. O'Leary. 2004. Left-brain/right-brain multi-document summarization. In *Document Understanding Conference (DUC) Workshop at HLT/NAACL 2004*.

E. Hovy, C. Y. Lin, and L. Zhou. 2005. Evaluating DUC 2005 using basic elements. In *Document Understanding Conference (DUC) Workshop at HLT 2005*.

X. Li and D. Roth. 2002. Learning question classifiers. In *Proceedings of the 19th International Conference on Computational Linguistics*, pages 1–7.

C. Y. Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *the Workshop on Text Summarisation Branches out at ACL 2004*.

LingPipe. 2006. The LingPipe Natural Language Processing software: http://www.alias-i.com/lingpipe/.

Lucene. 2006. The Lucene search engine: http://lucene.apache.org/.

OpenNLP. 2006. The OpenNLP project: http://opennlp.sourceforge.net/.

D. Zajic, B. Dorr, and R. Schwartz. 2004. BBN/UMD at DUC-2004: Topiary. In *Document Understanding Conference (DUC) Workshop at HLT/NAACL 2004*.