Description of SQUASH, the SFU Question Answering Summary Handler for the DUC-2006 Summarization Task

Gabor Melli, Zhongmin Shi, Yang Wang, Yudong Liu,
Anoop Sarkar and Fred Popowich
School of Computing Science, Simon Fraser University
Burnaby, BC V5A 1S6, Canada
http://natlang.cs.sfu.ca/

Abstract

This paper describes the design of the SQUASH system, the SFU Question Answering Summary Handler, developed by members of the Natural Language Lab from the SFU School of Computing Science in order to participate in the 2006 Document Understanding Conference (DUC-2006) summarization task. The SQUASH system for DUC-2006 built upon the SQUASH system that was built for the last year's DUC-2005 competition (Melli et al., 2005). We also present and discuss the various evaluations performed on our system output, comparing our performance to the other systems that took part in DUC-2006.

1 Introduction

The SFU Question Answering Summary Handler (SQUASH) is a summarization system that incorporates semantic role labelling, semantic subgraph-based sentence selection and automatic post-editing to create a question-based 250 word summary from a set of documents, all of which are relevant to the question topic (Melli et al., 2005). The performance of this system has been improved through the introduction of:

1. More accurate semantic-role labeling,
2. Concept identifiers which were created using specific relations from Wordnet, and
3. Training of features by exploiting the DUC-2005 data, specifically by optimizing the ROUGE score of the system on the DUC-2005 data.

In this paper, we introduce the improved system, SQUASH-06, focusing on the differences between SQUASH-05 and SQUASH-06, and present the results obtained during the DUC-2006 competition.

The system produces summaries by first annotating the documents and the question text in a phase we will call the **Annotator** phase. These annotations are then fed to two summarization stages: The **Extractor** phase focuses on content selection and the ROUGE-2/ROUGE-SU4 scores, while the **Editor** phase focuses on linguistic readability and the human evaluation scores.

In each of the sections that follow, we will describe the different modules of the SQUASH-06 system, and detail how we used the the DUC-2005 human summaries to optimize the ROUGE-2 and ROUGE-SU4 scores for summaries produced by our system for DUC-2006.

2 The Annotator Module

The Annotator module in our submitted system provides both syntactic and semantic annotations of questions and documents. The syntactic annotations include the output of a statistical parser, a named-entity finder and a co-reference resolver. We use the same methods to produce the syntactic annotations as described in (Melli et al., 2005). The semantic annotations consist of the output of a semantic role labeler and conceptual information retrieved from WordNet (Fellbaum, 1998). The conceptual information provides ontological relations among words/phrases, thus linking questions and anwsers at the semantic level.

2.1 Semantic Role Labelling

A semantic role is the relationship that a syntactic constituent has with a predicate. Typical semantic roles include Agent, Patient, Instrument, etc. and also adjunctive arguments indicating Locative, Temporal, Manner, Cause, etc. aspects. The task of semantic role labelling is, for each verb in a sentence, to identify all constituents that fill a semantic role, and to determine their roles, if any (for more details see (Palmer et al., 2005)). For example:

> [A0 Late buying] [V gave] [A2 the Paris Bourse] [A1 a parachute] [AM-TMP after its free fall early in the day].

Here, the arguments for the predicate *gave* are defined in the PropBank annotation (Palmer et al., 2005) as: V: verb, A0: giver, A1: thing given, A2: beneficiary, AM-TMP: temporal.

In our SQUASH system, we use ASSERT(Pradhan et al., 2004)[1] to extract the shallow semantic relations, as defined in Propbank, from the syntactic constituents of full parse trees produced by a statistical parser. In SQUASH the extracted semantic knowledge is used in sentence selection and sentence compression.

In many cases, for a particular sentence, the propositions that we focus on sometimes take only a small part of the sentence's tokens. It is interesting to see that by simply checking these small portions of a sentence, some good sentence candidates can be selected out.

For example, the following sentence is from Document 03 of question q0442g, DUC-2005. According to ASSERT output, it is labelled as:

> *The editor tried to console her by telling about* "[A0 *the guy*] *in the crowd*" [R-A0 *who*] "[V *saved*] [A1 *the President's life*]."

This sentence is obviously a good candidate for the given question "What are some outstanding instances of heroic acts when the hero was in danger of losing his/her life while saving others in imminent danger of losing their own lives?" (q0442g in DUC-2005).

2.2 Ontological Relations

The identification of relations between questions and sentences in the documents is fundamental to the question answering summarization task. Last year we applied a string-matching strategy to finding such relations at the lexical level, which however cannot capture ontological relations such as synonyms, *is-a* and *part-of*. For instance, the string-matching may not reward sentences containing "car" or "wagon" if the question asks about "automobile" (as is the case in topic q0608h).

For our purposes, we define **Super Concept**, a synonym, hypernym (*is-a*) or holonym (*part-of*) of a concept. Therefore, super concept is reflexive and transitive:

- Any concept is a super concept of itself.
- If concept A is a super concept of B and B is a super concept of C, then A is a super concept of C.

We say two concepts are related ontologically if one is the super concept of the other. In our submitted system this year the ontological relations are retrieved from WordNet in the following steps:

1 Assign each word/phrase in the questions and documents the corresponding Concept ID (CID) in WordNet as shown in Table 1. This step requires word sense disambiguation in the first place, which is accomplished by the WordNet::SenseRelate::AllWords module from CPAN[2].

[1] http://oak.colorado.edu/assert/

[2] http://senserelate.sourceforge.net/

Word	Base	Word Sense	CID
What	What	-	-
devices	device	1	03150574
and	and	-	-
procedures	procedure	3	06494814
have	have	2	02604841
been	be	1	02579744
implemented	implement	2	02535851
to	to	-	-
improve	improve	1	00202884
automobile	automobile	1	02929975
safety	safety	1	14345754
?	?	-	-

Table 1: An annotation example of a question with CIDs

Concept	Super concept
02660695	02929975
02787848	02929975
02890177	02929975
02929975	03749282
03749282	04122442

Table 2: An example of Super concepts. The table lists four concepts (including 02929975) whose super concept is 02929975 and three concepts that are super concepts of 02929975.

2 Identify ontological relations between two concepts by finding that one concept is a super concept of the other. All super concepts of words/phrases in the questions and documents are retrieved from Wordnet (as shown in Table 6) using the WordNet::querydata module from CPAN.

When associating questions and sentences in the documents, we expect to see a concept in sentence and its super concept in the question. The ontological relationships between, for instance, "car#n#1" (02929975) and "automobile#n#1" (02929975), "wagon#n#5" (02787848) and "automobile#n#1" can now be identified by looking up the Table 6.

The ontological relations among words/phrases provide essential semantic information to sentence clustering and rewarding sentences that address the question. Intuitively the integration of ontological relations into annotations would result in more relevant information in the summary. The intuition has been verified by experiments on the DUC-2005 data set, in that ontological relations improve the ROUGE-2 score from 0.07269 to 0.07375.

2.3 Baseline Model

We also built a knowledge poor baseline system (called GREEdy News Summarizer or GREENS) whose output was not submitted to the DUC-2006 evaluation. The baseline system does not emphasize readability but rather

System	ROUGE-2	ROUGE-SU4
SQUASH-05 (v22)	0.0632	0.1218
SQUASH-06 (v33.6)	0.0695	0.1235
Best peer	0.0725	0.1316
Worst model	0.0886	0.1484
GREENS (v1.5)	**0.0921**	**0.1700**
Best model	0.1180	0.1782

Table 3: ROUGE-2 and SU4 recalls of several systems on the DUC-2005 data set

focuses on content word selection using a simple n-gram model. During development, we evaluated our main system against the baseline using ROUGE scores on the DUC-2005 data set. GREENS performs significantly better than any submitted peers in DUC-2005 and even better than some human generated models, in terms of ROUGE-2/SU4 scores as shown in Table 3. GREENS may be ideal for summarization systems that do not require very high grammaticality and readability, such as summarization of document titles.

3 The Extractor Module

The task of the Extractor module is to produce a ranked set of sentences that are relevant to the question. To accomplish this the Extractor module follows a procedure similar to the one employed in DUC-2005. First it identifies all of the concepts reported by the Annotator module. As in DUC-2005 the concepts were derived from NER and Noun part-of-speech annotation, this year however WordNet was also used to improve the precision and recall of the concepts identified. Next a text graph is created to facilitate subsequent processing. The nodes of the graph are the identified concepts and the edges represent the different linguistic associations required downstream. The graph structure, for example, facilitates the calculation of a weight for each concept. A change to this calculation step from DUC-2005 is that the distance (semantic relatedness) between a document's concept and each of the concepts in the question now influences the weight calculation. Next each of the SRL propostions within each sentence are ranked based on the concepts they contain and where they place in the concept vector space. This step identifies proposition that makes concise use of concepts that are relevant to the summary. Propositions are then iteratively selected based on their ranking. Each selection however also reduces the score of (penalizes) propositions that were similar to the one selected. Once the propositions are selected what is reported are the sentences from which the propositions originated. This is one of the main differences from other systems described in the literature. Extractor ranks not the full sentences but their SRL propositions. This section will emphasize the description given to those topics that were modified since DUC-2005.

3.1 Concept Identification and Weight Assignment

A simple mechanism is used to identify the concepts contained in all documents and to then assign them a measure of their significance. Concept identification is based on whether a token sequence was a WordNet concept, a named entity, or whether it was a noun. For example, the tokens *J. Smith* would likely be marked as a concept by the named-entity recognizer, the tokens *police dog* would be annotated as a concept found in WordNet, and the token *ex-combatant*, which is not in WordNet, would be identified as a noun if the part-of-speech annotation labeled it as a Noun. When an overlap existed between the three sources of concept identification the following priority is given: NER, WordNet and finally Noun. Concept identification also included a simple multi-document NER based on token overlap. If a named entity matched another on one the same type by one token then they were merged. For DUC-2006 multi-document NER was performed not only on PERSON but also on LOCATION and ORGANIZATION.

The weight assigned to each concept is based on a point system, where points are associated to five factors. Each of these factor is described below. The justification for the value given to each factor however is described in Section 5.

- *pointsForBeingAQuestionConcept*: If a concept is present in both a document and the question then add 1000 points to the concept's significance to the extraction metric. For example, assume that the word *drug* is present in both the question and in one or more documents, but the word *Colombia* is only present in the documents. The points assigned to the concept *drug* will be increased by 1000.

- *pointsForBeingRelatedToAQuestionConcept*: New to DUC-2006 is the use of WordNet concept distance into the calculation of the concept weight. Given a concept *X* from a document and a concept *Y* from the question, the weight of concept *X* will be increased if they are both WordNet concepts and *X* is a type-of *Y*. For DUC-2006 the points given to *X* where increased by one-half (0.5) of *pointsForBeingAQuestionConcept*. For example, if *X=German Shepard* and *Y=dog* then the concept *X* would be given 500 more points.

- *pointsForBeingNER*: If a concept is a named entity then subtract seven (7) points. For example, assume that the word *drug* is present in eight documents, while the word *Colombia* is present in only two documents. The points assigned to the concept *Colom-*

bia will be decreased by seven (7) points, the word *drug* by zero (0) points.

- *pointsForInADoc*: Add 12 points for each document that the concept is located in. For example, assume that the word *drug* is present in eight (8) documents, while the word *Colombia* is present in only two (2) documents. The points assigned to the concept *drug* will be increased by 80 points, the word *Colombia* by 20.

The value for these parameters are first multiplied by the number of instances and then summed. The values used for the DUC-2006 submission are presented in Section 5.

3.1.1 Example of Semantic Relatedness

As mentioned in the section on Ontological Relations, this year's version makes use of the semantic relatedness function to look for concepts in the documents whose *Super Concepts* include a concept in the question. This addition proved valuable to empirical performance, so an example of its functioning is provided. The example is drawn from topic 426 of DUC-2005. The question read as follows: *What [sorts of] [law-enforcement] [tasks] are [dogs] being used for worldwide? What [law enforcement] [agencies] are using [dogs]? What [breeds] of [dogs] are being used?* The words in brackets represent the concepts extracted from the question. The WordNet ConceptIds for these concepts are: dog[02064081], breed[01417230], sort of[00018345], task,[00784188], law enforcement agency,[08234204]. As seen in Figure 1 as the depth of the Super Concepts considered is increased several concepts from the documents that appear to be highly relevant to the question become more important.

WordNet ID	ConceptName	Rank			
		d=3	d=2	d=1	d=0
2064081	dog(*)	1	1	1	1
8096606	police	2	2	2	5
8234204	law enforcement agency(*)	3	3	3	2
784188	project	4	4	4	3
1417230	breed(*)	5	5	5	4
10292737	Police Officer	6	6		
2086825	Police dog	7	7		
2086633	German shephed	8			
...	...	...	...	...	...

Figure 1: As the depth of Super Concepts considered is increased the order of the most important concepts change for the better. For example German Shepard appears once a depth of four degrees of separation to the Super Concept are considered.

3.2 Semantic Text Graph Creation

As in (Mani and Bloedorn, 1997) a *semantic text graph* is used to facilitate the summarization of a document. Extractor's text graph however also support topic-based summarization of multiple documents. The details of the semantic text graph creation method however was largely unchanged from DUC-2005. The base nodes of the graph are the concepts identified by the procedure just presented. Other node types include the propositions, sentences and documents in the textset. The graph's edges on the other hand represent the different structural, syntactic, and semantic associations required downstream. Edges include, for example, what semantic arguments, propositions, sentences and documents each concept is found in.

3.3 Proposition Significance Assignment

Each proposition in every document is given a significance score that will be used to rank the relevance of each sentence to the summary. The significance score is based on the concepts contained in the proposition. The intuition here is that between two sentences the one with the proposition that contains the more signficant concepts will be given preference. The contribution based on concept weight is calculated as the summation of the individual weight of the concepts in the proposition. For example, if *Gen Noriega* is assigned a significance score of 0.8 and *cartel* the score 0.7 then the first score becomes 1.5.

Furthermore, a penalty was given to propositions with too many concepts. This penalty is based on a multiplier that decreases linearly from a minimum threshold to a maximum threshold of concepts in the proposition: *minOfTooManyConcepts* and *maxOfTooManyConcepts*. Propositions with more concepts than *maxOfTooManyConcepts* were excluded from the selection process by assigning them a score of zero. For DUC-2006 the settings used were a minimum threshold of one concept and a maximum threhold of ten concepts. Section 5 presents the values tested.

New to this year was the addition of a penalty for sentences that were far from the first sentence in the document. The intuition for this factor is that the first few sentences in a document are typically more general than the later sentences. The penalty function used was a linear decay: the propositions from a document's first sentence received no penaly, while the last sentence in the document was stripped of half of its points. This setting was select though the optimization reported in Section 5.

3.4 Proposition and Sentence Selection

The selection of propositions and sentences was largely unchanged from DUC-2005. Propositions were iteratively selected based on their ranking. After each selection the score of similar propositions was then reduced

in order to reward non-redundant information. The sentences are finally selected based on the propositions selected.

4 The Editor Module

The task of the Editor module is to produce a summary with high linguistic quality. Specifically, the focus was on sentence compression and sentence ordering. The sentence compression component edits out hypothesized irrelevant content by heuristically dropping certain constituents of the sentence based on the semantic role label and manually designed rules. In this step, the dangling references and context cue information in candidate sentences are removed, which helps the second component to focus on reordering. The sentence reordering component selects and groups sentences based on the questions and lexical cohesion between sentences and attempts to find the most plausible sentence ordering in the summary text.

Unlike last year (Melli et al., 2005), the pronoun reference resolution component was removed. The reason is that only very few sentences got correct pronoun resolutions in DUC-2005 summaries based on our analysis. The current pronoun resolution across multi-document algorithm could not produce a good result for our task. To reduce the risk of introducing some unnecessary errors, we applied a simple strategy this year: removes all sentence output from Extractor starting with pronouns excluding "it".

4.1 Sentence Compression

Sentences from multi-documents often contain information that is not only irrelevant to the answer of questions but is also specific to the context of the original document. Most of the existing sentence compression algorithms require a training corpus (Knight and Marcu, 2000; Turner and Charniak, 2005). None of the existing compression methods are question oriented and they might remove the important constituents that contribute to the answer of the question. To preserve grammaticality with the minimum loss of content information, we only did compression on the surface sentence level and removed the context specific information in the original document.

Based on constraints similar to ones used in (Melli et al., 2005), we removed chronological phrases and transitional words in the sentence. However for DUC-2006, we applied Semantic Role Labelling (SRL) Information to remove these two types of constituents, since most of them can be captured by ARG-TMP (Temporal markers) and ARG-DIS (Discourse Markers). To summarize, main features used in the compression this year are:

- Temporal and discourse constituents labelled by SRL.
- The sub-header from the first sentence of the document.
- Person titles: Mr, Miss, Mrs, etc.
- Chunks ($\leq$ 5 words) if it contains original and inflective forms of say, report, argue, suggest etc.
- Words inside parentheses, dash lines, etc.
- Sentences starting with pronouns excluding "it".

SQUASH	**ROUGE-2**	**ROUGE-SU4**
W/O COMPRESSION	0.0649	0.1171
W/ COMPRESSION	0.0714	0.1284

Table 4: ROUGE-2 and SU4 recalls of SQUASH with and without compression on the DUC 2005 data set

Table 4 shows the ROUGE results on the DUC-2005 data.

4.2 Sentence Ordering

Instead of dynamically selecting sentences while doing the ordering in DUC-2005 (Melli et al., 2005), the sentence ordering component in DUC 2006 picked the right number of sentences that satisfied the 250 words length, and then reordered this fixed set of selected sentences.

4.2.1 Sentence selection and categorization

Given multiple questions to answer in summarization, the main goal in this step is to select and group every sentence by its contribution to the answer of each question. Extractor treated the questions as a bag of WordNet concepts instead of individual questions, it was possible that sentences that answered a specific question were ranked higher than all other sentences. To avoid such situation, Extractor provided Editor sentences that had twice the size of the final summary. Editor then categorized all the extracted sentences based on their word concept overlap and super concept relations with each question. Based on the sentence proportion that each question was answered in a larger pool of sentences from the Extractor, Editor estimated the number of sentences n_i to answer each question q_i within 250 words length limit.

This year, we first did automatic weight tuning as in the Extractor to optimize ROUGE scores on DUC-2005 data. But we didn't get a significant improvement in the result. This is reasonable since sentences selected from Extractor had already been optimized on ROUGE with the above features considered. To guarantee a minimum loss of content importance, we decided to select the top n_i sentences for each question q_i based only on their content importance score assigned by Extractor. In this way, every question is answered while preserving content information.

4.2.2 Sentence Reordering

Sentence ordering within each question cluster was performed based on the contextual information from original document and lexical cohesion between sentences (Halliday and Hasan, 1976). We approximated the lexical cohesion between sentences by their semantic relatedness among text entities and proposed a WordNet-based sentence similarity measure.

The algorithm performs the following steps:

- Grouped sentences from the same original document together, and preserved their original presentation order. The intuition for this is that sentences from the same document often talk about the same events and they are more semantically related. We also observed that many sentences extracted by SQUASH were neighboring sentences in the original document, so we treated such sentence group as one sentence in the future processing.
- Picked the first sentence group based on the sentence original document position. The first sentence or sentence appear earlier in the original document is preferred.
- Greedily chose the next sentence group based on its semantic similarity to the current sentence within the same question cluster.

A series sentence similarity measures have been studied for the calculation of text cohesion. Taxonomy-based similarity measures are proved to be very effective (Lapata and Barzilay, 2005). We proposed a simple WordNet-based similarity measure using the WordNet concept identifier which captures entity repetition, synonyms and meronyms as discussed in Section2.2.

$$Sim(S_1, S_2) = \frac{2|concept(S_1) \cap concept(S_2)|}{|concept(S_1)| + |concept(S_2)|}$$

where concept S_i is the set of concepts in sentence i and $|concept(S_i)|$ is the number of concepts in sentence S_i

Sentence clusters were ordered after each cluster was internally ordered. The first sentence cluster was chosen if it contained sentence that was the first sentence in the original document. The next cluster were then decided by the semantic relatedness between its first sentence and the last sentence in the current cluster.

4.2.3 Evaluation

We evaluated whether the ordering module was doing a reasonable job by comparing the output summary with ordering algorithm to the summary without ordering, which was a ranked list of sentences by content importance from the compression module. We then gathered coherence judgments from 5 subjects. Each subject was presented with 10 pairs of summaries with their corresponding topic questions. Each summary pair contained summaries with and without ordering. Subjects were instructed to assign a judgment to each summary by a 2-point scale: "good" and "bad". Summaries from the same topic could be graded as the same scale, since multiple orderings for the same set of documents are allowed (Barzilay et al., 2002). The study results in the Table showed that across 50 topics, 20 summaries by Editor with ordering are rated better than those without ordering. 12 summaries with ordering are rated worse. The rest 14 summary pairs cannot be distinguished.

	with Ordering		
without Ordering		Good	Bad
	Good	8	12
	Bad	20	6

Table 5: Comparison of Summary coherence qualities by Editor with and without Ordering

5 Model Optimization

One of the main differences from DUC-2005 version were the values given to Extractor's five internal variables. For DUC-2005 these values were set through subjective means. This year we set the variables based on an optimization analysis of ROUGE-2 and SU-4 performance against the DUC-2005 data. The training process used was a greedy hill-climbing technique. Each variable was incrementally tested at a new setting (both higher and lower). Any change that resulted in an improved ROUGE score was selected and the process continued until no increase was attained. To test whether only a local maxima was attained the process was repeated several times with different starting points. All experiments resulted in the same optimima suggesting that the optimization surface may be convex. This section presents some of the more interesting changes from last year's submission.

5.1 Concept in Question

The internal variable that changed most significantly was *pointsForBeingAQuestionConcept*. For the DUC-2005 version a setting of 200 was used. The optimization experiments suggested that this value significantly underrepresented the importance of this association. For DUC-2006 a value 1000 was used.

5.2 Semantic Relatedness to Question Concept

A new variable to the DUC-2006 version was *pointsForBeingRelatedToAQuestionConcept*. For this variable the value of one half (0.5) was optimal. An interpretation of this setting is that concepts that are semantically related to a question concept will receive half of the points given to concepts that are perfectly synonymous to a question

Propositions	ROUGE-2
1	0.0724
2	0.0736
3	0.0738
4	0.0736

Table 6: The ROUGE-2 score as the maximum number of propositions considered in each sentence is increased from one through to four. The threshold of three was selected as optimal.

concept. To further validate the value of using this technique, when the variable was excluded from the system the ROUGE-2 score dropped from 0.0738 to 0.0727.

5.3 Named Entities

The variable whose value changed in an unexpected direction was *pointsForBeingNER*. Last year this parameter was given a positive value of fifteen (15). The optimization experiments suggested a negative value of minus seven (-7) would result in better performance. An interpretation of this setting is that a summary should typically deemphasize the use of named entities.

5.4 Proposition Testing

For DUC-2005 Extractor only reference the first two propositions of each sentence. If a sentence had three propositions(i.e. predicates) only the first two would be tested. Experiments suggested that using the first two or three propositions is appropriate.

A test was also performed to quantify the value of using SRL propositions at all. The comparison was against a sentence-level approach in which the SRL proposition annotation was replaced with a single proposition that covered the whole sentence. After this change the ROUGE-2 score dropped from 0.0738 to 0.0702. This is a significant drop in performance relative to the removal of other types of information used by the Extraction component.

5.5 First Sentence Bias

The final variable tested was the other new internal variable introduced to the system which biasese for sentences near the beginning of a document. The optimal value for this variable was one half (0.5). An interpretation of this setting is that the last sentence of the document would be penalized by the removal of half of its accumulated points; the intermediary sentences would be apply a linear decay to the penalty; and the first sentence would receive no penalty. Note that when this variable was excluded from the system performance dropped from 0.0738 to 0.0714. This result reconfirms the value of this technique.

6 Results

SQUASH attained a more well-rounded set of results this year. Last year for example, SQUASH ranked lower than the median system in five metrics. This year on the other hand SQUASH ranked higher than the median of the 34 systems, on all but the Non-Redundancy metric. The other two systems that performed above the median on all metrics were systems 27 and 5. Figure 3 shows the Boxplot analysis of our system compared with other systems. The y-axis has been normalized by the min and max system performance as

$$y_i = \frac{x_i - \max(x_i)}{\max(x_i) - \min(x_i)}$$

where x_i is the original score for each system i, y_i is the The normalized score of best system got a score of 1 and worst system got 0 on the y-axis. To assist with the exploration of relative strenghts and weaknesses the x-axis is sorted by SQUASH's ranking.

Two of the metrics that SQUASH performed particularly well on were Responsiveness and Structure and Coherence linguistic metric. In terms of responsiveness, SQUASH ranked 10th in both the content and Overall versions of the metric. In terms of linguistic quality SQUASH improved significantly relative to its DUC05 performance. On the Structure and Coherence metric SQUASH ranked 6th (out of 34 systems) this year while last year the ranking was 22nd (out of 31 systems). This improvement is likely the result of two factors. The first is due to the improvement in content responsiveness; better content may lead to better coherence. A second reason is that we employed a new question-based lexical cohesion method for reordering.

Curiously, despite SQUASH's relatively good performance on Responsiveness, its ROUGE-based ranking dropped relative to other systems. This outcome was unexpected because, as described in Section 5, SQUASH was optimized on ROUGE. We notice however that ROUGE performance in general was significantly better relative to DUC-2005. We suspect that, like us, most teams this year attempted to optimize on ROUGE. However, the methods employed by the other teams to improve ROUGE performance appear to not have been generally as effective at optimizing actual Responsiveness.

7 Conclusion

Through SQUASH-06, we have illustrated how semantic role labelling, when combined with a concept-based relationship identification algorithm that associates document sentences and questions, can be used to obtain higher quality summaries. Effective use of Wordnet improved the concept identification mechanism, as well as the summary editing procedure. The use of

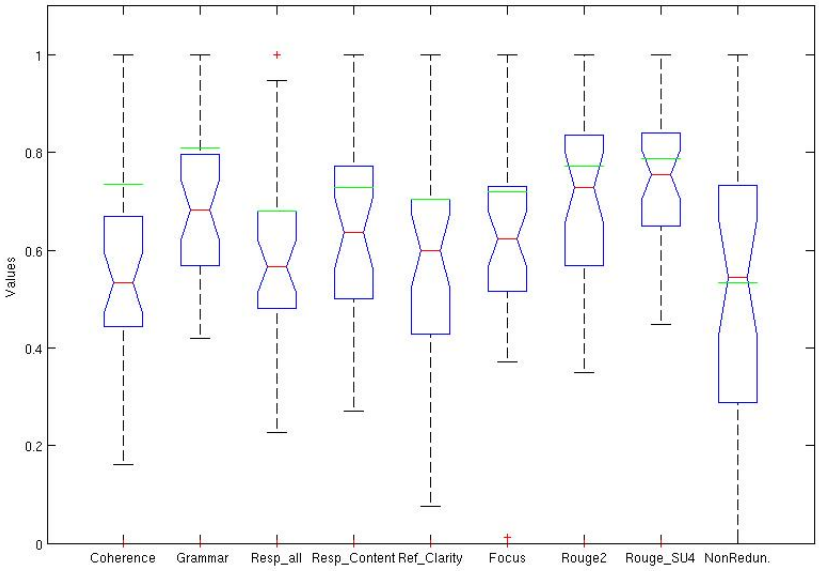

Figure 2: Statistical BoxPlot with the mean and both quartile values on each of the evaluation metrics, and an additional line to represent SQUASH's placement. The y-axis has been normalized by the Min and Max system performance. SQUASH ranked higher than the median of the 34 systems, on all but the Non-Redundancy metric. The other two systems that performed above the median on all metrics were systems 27 and 5.

a more sophisticated compression module allowed the elimination of irrelevant information, allowing more sentences to be included in the summary, thus improving the ROUGE score. Futhermore, the access to DUC-2005 data, also allowed for more accurate tuning of various system variables. A preliminary development version of the SQUASH system is available on the web at *http://natlang.cs.sfu.ca/qa*. The current web interface to SQUASH can only be used to summarize questions on the DUC 2005 document collection (selected using the topic identifiers), to avoid running the expensive annotation step on arbitrary user-specified document collections. However, the questions themselves can be arbitrary, and not just the ones in the DUC 2005 evaluation.

References

Regina Barzilay, Noemie Elhadad, and Kathleen McKeown. 2002. Inferring strategies for sentence ordering in multidocument news summarization. *Journal of Artificial Intelligence Research*, 17:35–55.

C. Fellbaum, editor. 1998. *Wordnet: An Electronic Lexical Database*. MIT Press.

M.A.K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman.

Kevin Knight and Daniel Marcu. 2000. Statistics-based summarization - step one: Sentence compression. In *AAAI/IAAI*, pages 703–710.

Mirella Lapata and Regina Barzilay. 2005. Automatic evaluation of text coherence: Models and representations. In *Proc. of ICAI-05*.

I. Mani and E. Bloedorn. 1997. Multi-document summarization by graph search and matching. In *Proceedings of the 14th National Conference on Artificial Intelligence*, pages 622–628, Providence, Rhode Island.

G. Melli, Y. Wang, Y. Liu, M. Kashani, Z. Shi, B. Gu, A. Sarkar, and F. Popowich. 2005. Description of squash, the sfu question answering summary handler for the duc-2005 summarization task. In *In Proceeding of Document Understanding Conference (DUC-2005)*.

M. Palmer, D. Gildea, and P. Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1).

Sameer Pradhan, Wayne Ward, Kadri Hacioglu, James Martin, and Dan Jurafsky. 2004. Shallow semantic parsing using support vector machines. In *Proceedings of the Human Language Technology Conference/North American chapter of the Association of Computational Linguistics (HLT/NAACL)*, Boston, MA.

Jenine Turner and Eugene Charniak. 2005. Supervised and unsupervised learning for sentence compression. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 290–297, Ann Arbor, Michigan, June. Association for Computational Linguistics.