Evaluating DUC 2004 Tasks with the QARLA Framework

Enrique Amigó, Julio Gonzalo, Anselmo Peñas, Felisa Verdejo
Departamento de Lenguajes y Sistemas Informáticos
Universidad Nacional de Educación a Distancia
c/Juan del Rosal, 16 - 28040 Madrid - Spain
{enrique,julio,anselmo,felisa}@lsi.uned.es

Abstract

This papers reports the application of the QARLA evaluation framework to the DUC 2004 testbed (tasks 2 and 5). Our experiment addresses two issues: how well QARLA evaluation measures correlate with human judgements, and what additional insights can be provided by the QARLA framework to the DUC evaluation exercises.

1 Introduction

QARLA (Amigó et al., 2005) is a framework that uses similarity to models as a building block for the evaluation of automatic summarisation systems. The input of QARLA is a summarisation task, a set of test cases, a set of similarity metrics, and sets of models and automatic summaries (peers) for each test case. With such a testbed, QARLA provides:

- A measure, QUEEN, which combines assorted similarity metrics to estimate the quality of automatic summarisers.
- A measure, KING, to select the best combination of similarity metrics.
- An estimation, JACK, of the reliability of the testbed for evaluation purposes.

The QARLA framework does not rely on human judges. It is interesting, however, to find out how well an evaluation using QARLA correlates with human judges, and whether QARLA can provide additional insights into an evaluation based on human assessments.

In this paper, we apply the QARLA framework (QUEEN, KING and JACK measures) to the output of two different evaluation exercises: DUC 2004 tasks 2 and 5 (Over and Yen, 2004). Task 2 requires short (one-hundred word) summaries for assorted document sets; Task 5 consists of generating a short summary in response to a "Who is" question.

In Section 2, we summarise the QARLA evaluation framework; in Section 3, we describe the similarity metrics used in the experiments. Section 4 discusses the results of the QARLA framework using such metrics on the DUC testbeds. Finally, Section 5 draws some conclusions.

2 The QARLA evaluation framework

QARLA uses similarity to models for the evaluation of automatic summarisation systems. Here we summarise its main features; the reader may refer to (Amigó et al., 2005) for details.

The input of the framework is:

- A summarisation task (e.g. topic oriented, informative multi-document summarisation on a given domain/corpus).
- A set T of test cases (e.g. topic/document set pairs for the example above)
- A set of summaries M produced by humans (*models*), and a set of automatic summaries A (*peers*), for every test case.
- A set X of similarity metrics to compare summaries.

With this input, QARLA provides three main measures that we describe below.

Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 49–56, Ann Arbor, June 2005. ©2005 Association for Computational Linguistics

2.1 $QUEEN$: Estimating the quality of an automatic summary

QUEEN operates under the assumption that a summary is better if it is closer to the model summaries according to all metrics; it is defined as the probability, measured on $M \times M \times M$, that for every metric in X the automatic summary a is closer to a model than two models to each other:

$$\text{QUEEN}_{X,M}(a) \equiv P(\forall x \in X.x(a,m) \geq x(m',m''))$$

where a is the automatic summary being evaluated, $\langle m, m', m'' \rangle$ are three models in M, and $x(a,m)$ stands for the similarity of m to a. QUEEN is stated as a probability, and therefore its range of values is $[0,1]$.

We can think of the QUEEN measure as using a set of tests (every similarity metric in X) to falsify the hypothesis that a given summary a is a model. Given $\langle a, m, m', m'' \rangle$, we test $x(a,m) \geq x(m',m'')$ for each metric x. a is accepted as a model only if it passes the test for every metric. QUEEN(a) is, then, the probability of acceptance for a in the sample space $M \times M \times M$.

This measure has some interesting properties: **(i)** it is able to combine different similarity metrics into a single evaluation measure; **(ii)** it is not affected by the scale properties of individual metrics, i.e. it does not require metric normalisation and it is not affected by metric weighting. **(iii)** Peers which are very far from the set of models all receive QUEEN=0. In other words, QUEEN does not distinguish between very poor summarisation strategies. **(iv)** The value of QUEEN is maximised for peers that "merge" with the models under all metrics in X. **(v)** The universal quantifier on the metric parameter x implies that adding redundant metrics do not bias the result of QUEEN.

Now the question is: which similarity metrics are adequate to evaluate summaries? Imagine that we use a similarity metric based on sentence co-selection; it might happen that humans do not agree on which sentences to select, and therefore emulating their sentence selection behaviour is both easy (nobody agrees with each other) and useless. We need to take into account which are the features that human summaries do share, and evaluate according to them. This is provided by the KING measure.

2.2 $KING$: estimating the quality of similarity metrics

The measure KING$_{M,A}(X)$ estimates the quality of a set of similarity metrics X using a set of models M and a set of peers A. KING is defined as the probability that a model has higher QUEEN value than any peer in a test sample. Formally:

$$\text{KING}_{M,A}(X) \equiv$$

$$P(\forall a \in A, \text{QUEEN}_{M,X}(m) > \text{QUEEN}_{M,X}(a))$$

For example, an ideal metric -that puts all models together-would give QUEEN(m) $= 1$ for all models, and QUEEN(a) $= 0$ for all peers which are not put together with the models, obtaining KING $= 1$.

KING satisfies several interesting properties: (i) KING does not depend on the scale properties of the metric; (ii) Adding repeated or very similar peers do not alter the KING measure, which avoids one way of biasing the measure. (iii) the KING value of random and constant metrics is zero or close to zero.

2.3 JACK: reliability of the peer set

Once we detect a difference in quality between two summarisation systems, the question is now whether this result is reliable. Would we get the same results using a different test set (different examples, different human summarisers (models) or different baseline systems)?

The first step is obviously to apply statistical significance tests to the results. But even if they give a positive result, it might be insufficient. The problem is that the estimation of the probabilities in KING assumes that the sample sets M, A are not biased. If M, A are biased, the results can be statistically significant and yet unreliable. The set of examples and the behaviour of human summarisers (models) should be somehow controlled either for homogeneity (if the intended profile of examples and/or users is narrow) or representativity (if it is wide). But how to know whether the set of automatic summaries is representative and therefore is not penalising certain automatic summarisation strategies?

This is addressed by the JACK measure:

$$\text{JACK}(X, M, A) \equiv P(\exists a, a' \in A|$$
$$\forall x \in X.x(a, a') \leq x(a, m) \wedge x(a', a) \leq x(a', m) \wedge$$
$$\text{QUEEN}(a) > 0 \wedge \text{QUEEN}(a') > 0)$$

i.e. the probability over all model summaries m of finding a couple of automatic summaries a, a' which are closer to m than to each other according to all metrics. This measure satisfies three desirable properties: (i) it can be enlarged by increasing the similarity of the peers to the models (the $x(m, a)$ factor in the inequalities), i.e. enhancing the quality of the peer set; (ii) it can also be enlarged by decreasing the similarity between automatic summaries (the $x(a, a')$ factor in the inequality), i.e. augmenting the diversity of (independent) automatic summarisation strategies represented in the test bed; (iii) adding elements to A cannot diminish the JACK value, because of the existential quantifier on a, a'.

3 Selection of similarity metrics

Each different similarity metric characterises different features of a summary. Our first objective is to select the best set of metrics, that is, the metrics which best characterise the human summaries (models) as opposed to automatic summaries. The second objective is to obtain as much information as possible about the behaviour of automatic summaries.

In this Section, we begin by describing a set of 59 metrics used as a starting point. Some of them provide overlapping information; the second step is then to select a subset of metrics that minimises redundancy and, at the same time, maximises quality (KING values). Finally, we analyse the characteristics of the selected metrics.

3.1 Similarity metrics

For this work, we have considered the following similarity metrics:

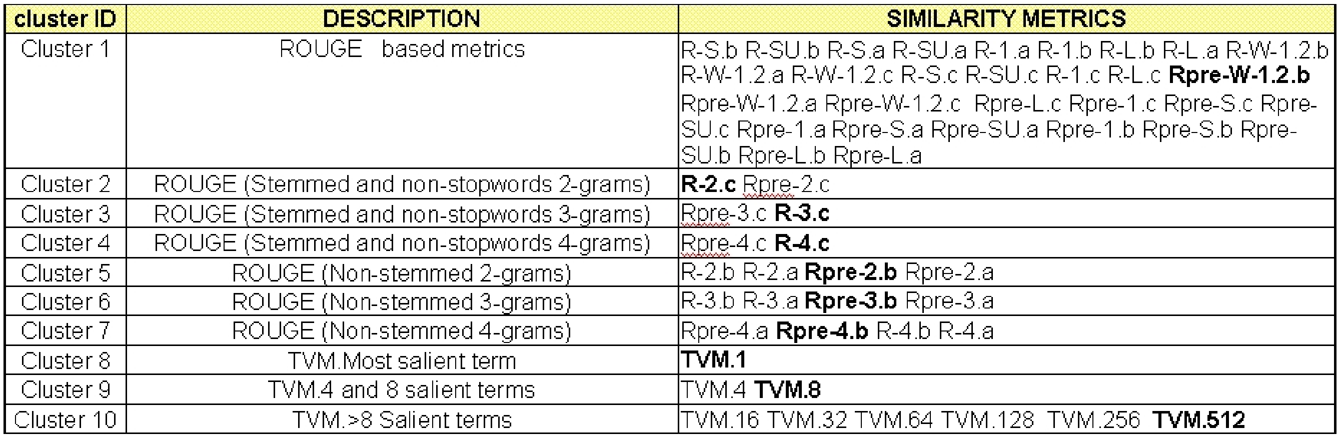
ROUGE based metrics (R): ROUGE (Lin and Hovy, 2003) estimates the quality of an automatic summary on the basis of the n-gram coverage related to a set of human summaries (models). Although ROUGE is an evaluation metric, we can adapt it to behave as a similarity metric between pairs of summaries if we consider only one model in the computation. There are different kinds of ROUGE metrics such as ROUGE-W, ROUGE-L, ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4, etc. (Lin, 2004b). Each of these metrics has been applied over summaries with three preprocessing options: with stemming and stopword removal (type c); only with stopwords removal (type b); or without any kind of preprocessing (type a). All these combinations give 24 similarity metrics based on ROUGE.

Inverted ROUGE based metrics (Rpre): ROUGE metrics are recall oriented. If we reverse the direction of the similarity computation, we obtain precision oriented metrics (i.e. Rpre(a, b) = R(b, a)). In this way, we generate another 24 metrics based on inverted ROUGE.

TruncatedVectModel (TVM$_n$): This family of metrics compares the distribution of the n most relevant terms from original documents in the summaries. The process is the following: (1) obtaining the n most frequent lemmas ignoring stopwords; (2) generating a vector with the relative frequency of each term in the summary; (3) calculating the similarity between two vectors as the inverse of the Euclidean distance. We have used 9 variants of this measure with $n = 1, 4, 8, 16, 32, 64, 128, 256, 512$.

AveragedSentencelengthSim (AVLS): This is a very simple metric that compares the average length of the sentences in two summaries. It can be useful to compare the degree of abstraction of the summaries.

GRAMSIM: This similarity metric compares the distribution of the part-of-speech tags in the two summaries. The processing is the following: (1) part-of-speech tagging of summaries using TreeTagger ; (2) generation of a vector with the tags frequency for each summary; (3) calculation of the similarity between two vectors as the inverse of the Euclidean distance. This similarity metric is not content oriented, but syntax-oriented.



cluster ID	DESCRIPTION	SIMILARITY METRICS
Cluster 1	ROUGE based metrics	R-S.b R-SU.b R-S.a R-SU.a R-1.a R-1.b R-L.b R-L.a R-W-1.2.b R-W-1.2.a R-W-1.2.c R-S.c R-SU.c R-1.c R-L.c **Rpre-W-1.2.b** Rpre-W-1.2.a Rpre-W-1.2.c Rpre-L.c Rpre-1.c Rpre-S.c Rpre-SU.c Rpre-1.a Rpre-S.a Rpre-SU.a Rpre-1.b Rpre-S.b Rpre-SU.b Rpre-L.b Rpre-L.a
Cluster 2	ROUGE (Stemmed and non-stopwords 2-grams)	**R-2.c** Rpre-2.c
Cluster 3	ROUGE (Stemmed and non-stopwords 3-grams)	Rpre-3.c **R-3.c**
Cluster 4	ROUGE (Stemmed and non-stopwords 4-grams)	Rpre-4.c **R-4.c**
Cluster 5	ROUGE (Non-stemmed 2-grams)	R-2.b R-2.a **Rpre-2.b** Rpre-2.a
Cluster 6	ROUGE (Non-stemmed 3-grams)	R-3.b R-3.a **Rpre-3.b** Rpre-3.a
Cluster 7	ROUGE (Non-stemmed 4-grams)	Rpre-4.a **Rpre-4.b** R-4.b R-4.a
Cluster 8	TVM.Most salient term	**TVM.1**
Cluster 9	TVM.4 and 8 salient terms	TVM.4 **TVM.8**
Cluster 10	TVM.>8 Salient terms	TVM.16 TVM.32 TVM.64 TVM.128 TVM.256 **TVM.512**

Figure 1: Similarity Metric Clusters

3.2 Clustering similarity metrics

From the set of metrics described above we have 57 (24+24+9) content oriented metrics, plus two metrics based on stylistic features (AVLS and GRAMSIM). However, the 57 metrics characterising summary contents are highly redundant. Thus, clustering similar metrics seems desirable.

We perform an automatic clustering process using the following notion of proximity between two metric sets:

$$sim(X, X') \equiv Prob[H(X) \leftrightarrow H(X')]$$

$$\text{where} \quad H(X) \equiv \forall x \in X.x(a,m) \geq x(m',m'')$$

Two metrics sets are similar, according to the formula, if they behave similarly with respect to the $QUEEN$ condition (H predicate in the formula), i.e. the probability that the two sets of metrics discriminate the same automatic summaries when they are compared to the same pair of models.

Figure 1 shows the clustering of similarity metrics for the DUC 2004 Task 2. The number of clusters was fixed in 10. After the clustering process, the 48 ROUGE metrics are grouped in 7 sets, and the 9 TVM metrics are grouped in 3 sets. In each cluster, the metric with highest KING has been marked in boldface. Note that the ROUGE-c metrics (with stemming) with highest KING are those based on recall whereas the ROUGE-a/b metrics (without stemming) are those based on precision. Regarding TVM clusters, the metrics with highest KING in each cluster are those based on a higher number of terms.

Finally, we select the metric with highest KING in each group, obtaining the 10 most representative metrics.

3.3 Best evaluation metric: KING values

Figure 2 shows the KING values for the selected similarity metrics, which represent how every metric characterises model summaries as opposed to automatic summaries. These are the main results:

- The last column shows the best metric set, considering all possible metric combinations. In both DUC tasks, the best combination is $\{$Rpre-W-1.2.b, TVM.512. This metric set gets better KING values than any individual metric in isolation (17% better than the second best for task 2, and 23% better for task 5). This is an interesting result confirming that we can improve our ability to characterise human summaries just by combining standard similarity metrics in the QARLA framework. Note also that both metrics in the best set are content-oriented.

- Rpre-W.1.2.b (inverted ROUGE measure, using non-contiguous word sequences, removing stopwords, without stemming) obtains the highest individual KING for task 2, and is one of the best in task 5, confirming that ROUGE-based metrics are a robust way of evaluating summaries, and indicating that non-contiguous word sequences can be more useful for evaluation purposes than n-grams.

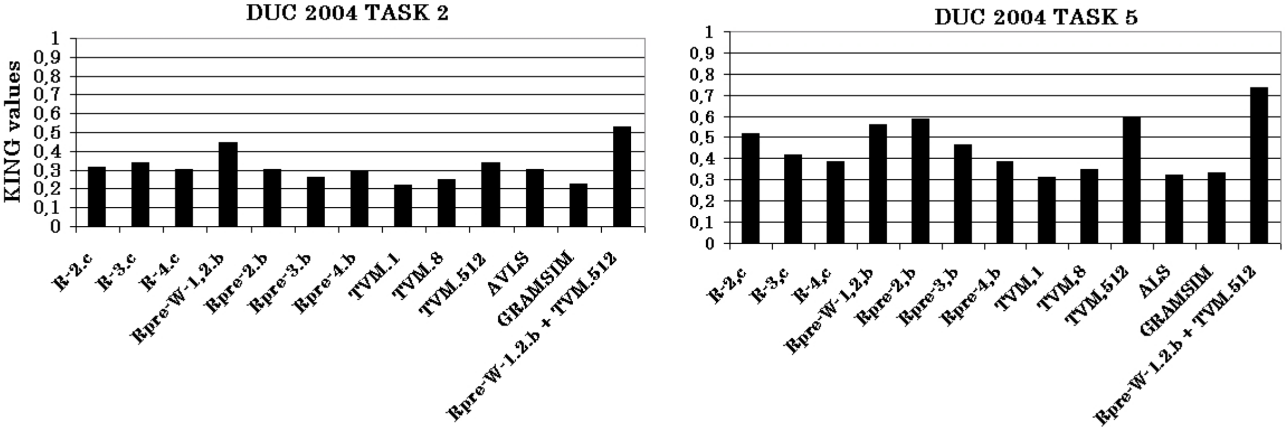


Figure 2: Similarity Metric quality

- TVM metrics get higher values when considering more terms (TVM.512), confirming that comparing with just a few terms (e.g. TVM.4) is not informative enough.

- Overall, KING values are higher for task 5, suggesting that there is more agreement between human summaries in topic-oriented tasks.

3.4 Reliability of the results

The JACK measure estimates the reliability of QARLA results, and is correlated with the diversity of automatic summarisation strategies included in the testbed. In principle, the larger the number of automatic summaries, the higher the JACK values we should obtain. The important point is to determine when JACK values tend to stabilise; at this point, it is not useful to add more automatic summaries without introducing new summarisation strategies.

Figure 3 shows how $JACK_{\text{Rpre-W,TVM.512}}$ values grow when adding automatic summaries. For more than 10 systems, JACK values grow slower in both tasks. Absolute JACK values are higher in Task 2 than in task 5, indicating that systems tend to produce more similar summaries in Task 5 (perhaps because it is a topic-oriented task). This result suggests that we should incorporate more diverse summarisation strategies in Task 5 to enhance the reliability of the testbed for evaluation purposes with QARLA.

4 Evaluation of automatic summarisers: QUEEN values

The QUEEN measure provides two kinds of information to compare automatic summarisation systems: which are the best systems -according to the best metric set-, and which are the individual features of every automatic summariser -according to individual similarity metrics-.

4.1 System ranking

The best metric combination for both tasks was $\{\text{Rpre-W}, \text{TVM.512}\}$; therefore, our global system evaluation uses this combination of content-oriented metrics. Figure 4 shows the $\text{QUEEN}_{\{\text{Rpre-W,TVM.512}\}}$ values for each participating system in DUC 2004, also including the model summaries. As expected, model summaries obtain the highest QUEEN values in both DUC tasks, with a significant distance with respect to the automatic summaries.

4.2 Correlation with human judgements

The manual ranking generated in DUC is based on a set of human-produced evaluation criteria, whereas the QARLA framework gives more weight to the aspects that characterise model summaries as opposed to automatic summaries. It is interesting, however, to find out whether both evaluation methodologies are correlated. Indeed, this is the case: the Pearson correlation between manual and QUEEN rankings is 0.92 for the Task 2 and 0.96 for the Task 5.

Of course, QUEEN values depend on the chosen metric set X; it is also interesting to check whether

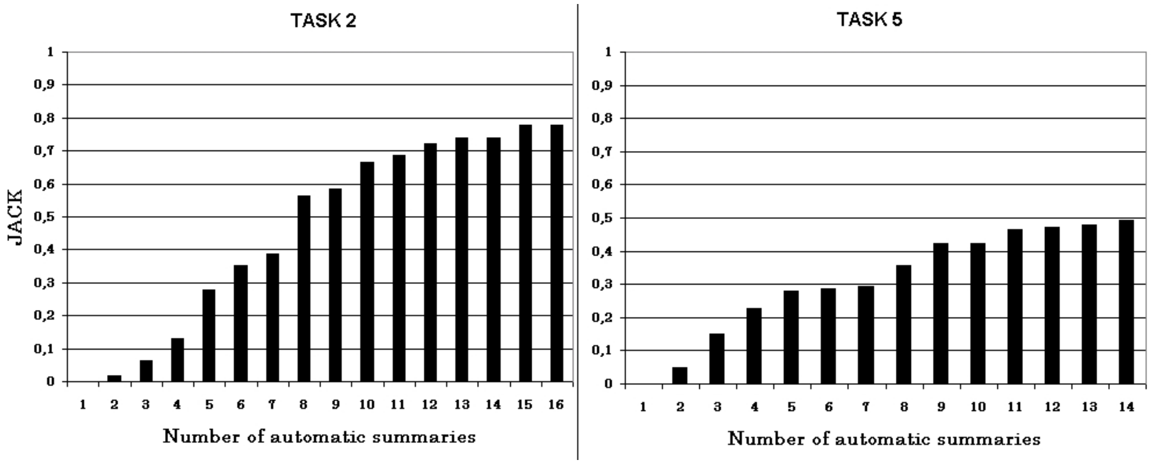


Figure 3: JACK vs. Number of Automatic Summaries

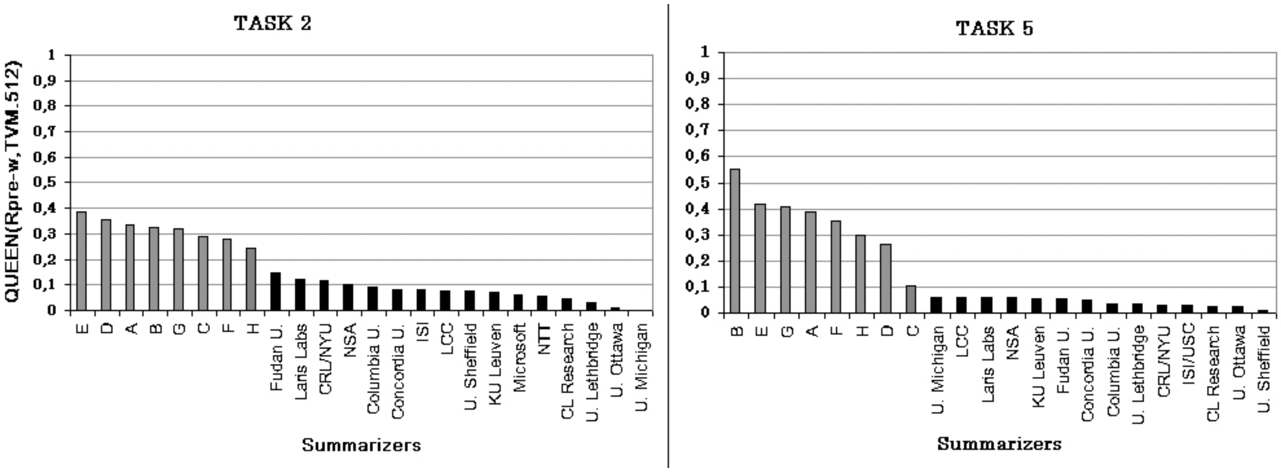


Figure 4: QUEEN system ranking for the best metric set (A-H are models)

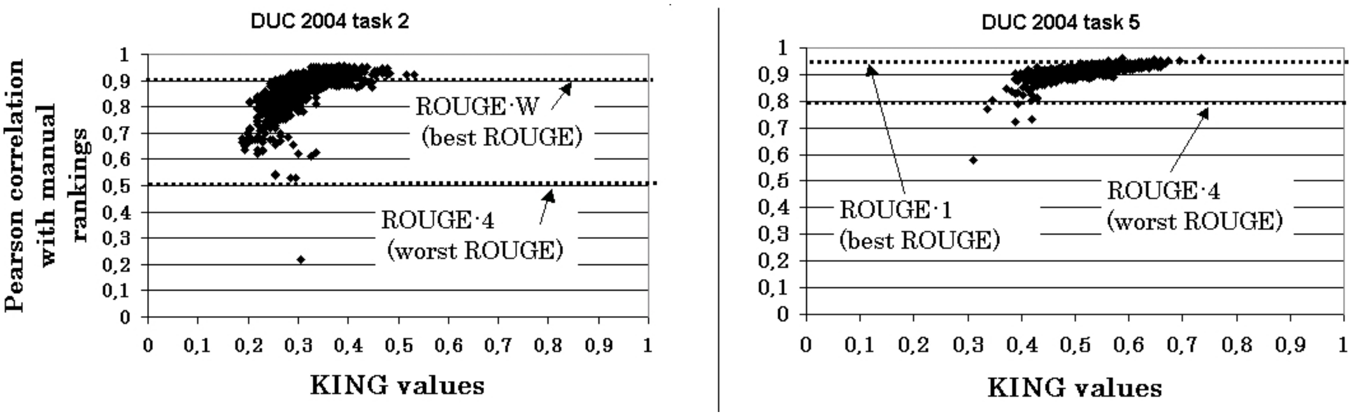


Figure 5: Correlation Between DUC and QARLA results

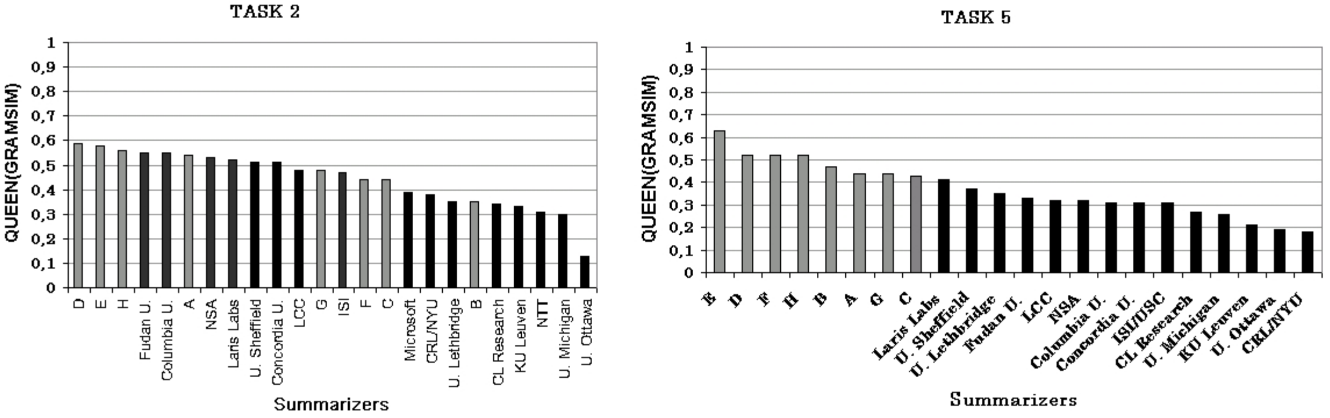


Figure 6: QUEEN values over GRAMSIM

metrics with higher KING values lead to QUEEN rankings more similar to human judgements. Figure 5 shows the Pearson correlation between manual and QUEEN rankings for 1024 metric combinations with different KING values. The figure confirms that higher KING values are associated with rankings closer to human judgements.

4.3 Stylistic features

The best metric combination leaves out similarity metrics based on stylistic features. It is interesting, however, to see how automatic summaries behave with respect to this kind of features. Perhaps the most remarkable fact about stylistic similarities is that, in the case of the GRAMSIM metric, task 2 and task 5 exhibit a rather different behaviour (see Figure 6). In task 2, systems merge with the models, while in task 5 the QUEEN values of the systems are inferior to the models. This suggests that there is some stylistic component in models that systems are not capturing in the topic-oriented task.

5 Related work

The methodology which is closest to our framework is ORANGE (Lin, 2004a), which evaluates a similarity metric using the average ranks obtained by reference items within a baseline set. As in our framework, ORANGE performs an automatic meta-evaluation, there is no need for human assessments, and it does not depend on the scale properties of the metric being evaluated (because changes of scale preserve rankings). The ORANGE approach is, indeed, intimately related to the original QARLA measure introduced in (Amigo et al., 2004).

There are several approaches to the automatic evaluation of summarisation and Machine Translation systems (Culy and Riehemann, 2003; Coughlin, 2003). Probably the most significant improvement over ORANGE is the ability to combine automatically the information of different metrics. Our impression is that a comprehensive automatic evaluation of a summary must necessarily capture different aspects of the problem with different metrics, and that the results of every individual checking (metric) should not be combined in any prescribed algebraic way (such as a linear weighted combination). Our framework satisfies this condition.

ORANGE, however, has also an advantage over the QARLA framework, namely that it can be used for evaluation metrics which are not based on similarity between model/peer pairs. For instance, ROUGE can be applied directly in the ORANGE framework without any reformulation.

6 Conclusions

The application of the QARLA evaluation framework to the DUC testbed provides some useful insights into the problem of evaluating text summarisation systems:

- The results show that a combination of similarity metrics behaves better than any metric in isolation. The best metric set is $\{$Rpre-W, TVM.512$\}$, a combination of content-oriented metrics. Unsurprisingly, stylistic similarity is less useful for evaluation purposes.

- The evaluation provided by QARLA correlates well with the rankings provided by DUC human judges. For both tasks, metric sets with higher KING values slightly outperforms the best ROUGE evaluation measure.

- QARLA measures show that DUC tasks 2 and 5 are quite different in nature. In Task 5, human summaries are more similar, and the automatic summarisation strategies evaluated are less diverse.

Acknowledgements

We are indebted to Ed Hovy, Donna Harman, Paul Over, Hoa Dang and Chin-Yew Lin for their inspiring and generous feedback at different stages in the development of QARLA. We are also indebted to NIST for hosting Enrique Amigó as a visitor and for providing the DUC test beds. This work has been partially supported by the Spanish government, project R2D2 (TIC-2003-7180).

References

E. Amigó, J. Gonzalo, A. Peñas, and F. Verdejo. 2005. QARLA: a Framework for the Evaluation of Text Summarization Systems. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005).*

E. Amigo, V. Peinado, J. Gonzalo, A. Peñas, and F. Verdejo. 2004. An Empirical Study of Information Synthesis Tasks. In *Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics (ACL)*, Barcelona, July.

Deborah Coughlin. 2003. Correlating Automated and Human Assessments of Machine Translation Quality. In *In Proceedings of MT Summit IX*, New Orleans,LA.

Christopher Culy and Susanne Riehemann. 2003. The Limits of N-Gram Translation Evaluation Metrics. In *Proceedings of MT Summit IX*, New Orleans,LA.

C. Lin and E. H. Hovy. 2003. Automatic Evaluation of Summaries Using N-gram Co-ocurrence Statistics. In *Proceeding of 2003 Language Technology Conference (HLT-NAACL 2003).*

C. Lin. 2004a. Orange: a Method for Evaluating Automatic Metrics for Machine Translation. In *Proceedings of the 36th Annual Conference on Computational Linguisticsion for Computational Linguistics (Coling'04)*, Geneva, August.

Chin-Yew Lin. 2004b. Rouge: A Package for Automatic Evaluation of Summaries. In Marie-Francine Moens and Stan Szpakowicz, editors, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

P. Over and J. Yen. 2004. An introduction to DUC 2004 Intrinsic Evaluation of Generic New Text Summarization Systems. In *Proceedings of DUC 2004 Document Understanding Workshop, Boston*.